Using Augmented Statistical Models and Score Spaces for Classification

Nathan Drysdale Smith

Christ's College

September 2003

Dissertation submitted for the

Degree of Doctor of Philosophy

Summary

Title: Using Augmented Statistical Models and Score Spaces for Classification

Name: Nathan Drysdale Smith

Many data sources in our world are stochastic in nature. They may be represented by statistical models which are applied to inference tasks such as classification. Unfortunately the precise nature of data sources is often unknown and sufficiently complicated that any statistical models proposed are much simpler and to some degree 'incorrect'. Model incorrectness harms the performance of classification algorithms. One technique for attaining better representation is to view statistical models as differentiable manifolds in the space of distributions. These manifolds may be augmented through application of the Taylor expansion to form much more flexible structures called fibre bundles. The definition of these structures can be extended to the space of scalar functions. This thesis develops the associated concept of score spaces. Score spaces may be used to facilitate the training of distributions in fibre bundles, or alternatively can simply be viewed as model-dependent feature spaces. Experiments were performed to classify fixed length and variable length patterns of speech data using score spaces, and promising performance was obtained. A useful characteristic of score spaces is that they permit the application of static classifiers such as SVMs to the classification of variable length patterns.

Acknowledgments

The research described in this thesis has been conducted under the supervision of Mark Gales who has made numerous suggestions and given much advise. Ideas and concepts for score spaces have been developed in discussion with Mark. Credit for originality should be attributed to Mark as well as to the author. Mark also proofread drafts of much of this thesis and made helpful comments and suggestions for its improvement. Martin Webber provided the initial derivation of the dependence of the Taylor expansion on the parameters of the statistical model as described in Appendix B.2. Both Martin and Tom Drummond gave help on understanding concepts, for example Martin with vector bundles, coordinate spaces and the path of the Taylor expansion along the manifold, and Tom with tensors and embedding spaces.

With respect to software, Thorsten Joachims provided SVM^{light} [54] [53] which was used to train and test SVM classfiers and provided the platform for the author's own code to map samples into score spaces. The tools and functions provided in HTK [114] [49], a HMM toolkit, were also used extensively to train and test HMM and GMM classifiers. For MMI training in the ISOLET experiments of Chapter 6, software embedded in HTK was provided by K.K.Chin to produce lattices and Dan Povey to train GMMs and HMMs. Phil Woodland provided helpful advice for applying MMI estimation to this task. Dan's software was also applied for MMI estimation in the experiments on Deterding vowel data in Chapter 5. Dan provided helpful advice for MMI training on both these tasks. Mark Gales, in effect, provided some software embedded in HTK to extract state posteriors during HMM training, these state posteriors being used in calculating the mapping of samples into score spaces. The versions of software used or modified are detailed where possible and relevant in this thesis. Patrick Gosling provided help and advice in terms of software, and in maintaining and supporting the computer system. Some experiments made use of equipment kindly provided by IBM under an SUR award.

This thesis has been prepared using LaTeX and associated software including BIBTeX, with some LaTeX code copied from the computer help webpages at Cambridge University Engineering Department and also from other sources, and with the Emacs [27] and Pico [74]

editors. Diagrams have been drawn with XFig [113] and MATLAB [68]. Reference material has been cited where relevant. For passages of text which describe a series of definitions, a single reference to the source has sometimes been made to avoid unneccessary repetition of the citation. Hopefully this is clear from context. References have been omitted when deemed common knowledge. An attempt has been made to cite all necessary references to indicate where originality should not be attributed to this thesis, though the success of this attempt cannot be guaranteed with certainty.

Some of the theoretical developments in this thesis and the ISOLET experimental results have been described in conference proceedings in [93], [94] and [91], and in the technical reports [92] and [95]. The M.Phil. thesis [90] contained the derivation for the mapping into score space as defined on Gaussian Mixture Models and a log likelihood scalar field, and contained some of the initial work on score spaces and associated concepts. The derivation for the mapping was later developed in the two technical reports with constraints on the mixture weights, for HMMs, and also for the log likelihood-ratio and log posterior scalar fields. The mappings appear without derivations in Appendix B.3.

The EPSRC kindly provided funding for three years as part of a CASE award cosponsored with IBM U.K. Laboratories. Both of these I would like to thank, in particular Eric Janke at IBM U.K. Laboratories. I am also grateful for an internship at the IBM T.J. Watson Research Center, New York State, and in particular to Ramesh Gopinath who helped to organise this and supervised me while there. I am also thankful to Christ's College for financial assistance towards conference and travel costs.

Finally, I would personally like to thank my supervisor for his patience and advice, my colleagues in the Machine Intelligence Laboratory for their help and friendship, and friends who have brought happiness to my time in Cambridge. My greatest thanks is to my family for their constant support and encouragement, especially to my brother Gavin without whom I would never have reached this stage, and to my parents whose gracious love is beyond the measure of words. To them I dedicate this thesis.

By the grace of God and in His strength alone.

Declaration

I declare that this thesis is substantially my own work, within the constraints and with the exceptions detailed in the Acknowledgments above, and within limits which seem reasonable to the author. In addition, ammendments suggested by the examiners have also been incorporated into this thesis.

To the best of my belief this thesis is less than 65'000 words, where any word count relating to a figure has been omitted even when text appears in the figure, the word count assigned to a table has been the number of words appearing in the table rather than the equivalent number of words estimated for the area occupied by the table, an estimate of 10 words has been used for each line of an equation, and maths embedded in the text has been counted with respect to intervening white spaces. This thesis also contains 29 figures, where subfigures have not been counted separately.

While care has been taken in the mathematical and theoretical derivations, experiments and analysis, the author accepts no responsibility for the accuracy of the contents of this thesis. Readers are advised that the inclusion or application of the contents of this thesis to their own work is entirely at their own risk.

Signature:

Date:

Contents

Chapter 1

Introduction

Processes exist in the world around us which can be characterised by measurable quantities, for example the passage of the sun and moon across the sky or the strength of the wind and currents in the sea. These processes may be viewed as data sources and are often highly complicated. Their intrinsic nature is the result of a myriad of interactions constrained by the laws of nature, and the study and understanding of such interactions is the goal of scientific research. If the outcome of a process is predictable then the process is deterministic. However many processes are stochastic in nature and the outcome of such a process cannot be predicted with certainty but nevertheless with a measure of probability. The spread of this measure of probability over the set of possible outcomes forms a probability mass function, or in the limit of a continuum of possible outcomes, a probability density function or *distribution*. Therefore since stochastic data sources have a mathematical description in terms of probability mass or density functions, it is possible to deduce these descriptions. In practice, a mathematical model called a *statistical model* is proposed and its parameters estimated.

Statistical models find widespread application in pattern recognition. For example, a quantity of data exists constituting a pattern. The pattern may belong to any one of a number of classes, where a class is either a single data source or a set of data sources with some common semantic meaning. The goal is to assign the pattern to a particular

class. However the stochastic nature of the sources implies the decision can only be made with a measure of confidence. The degree of confidence in the decision should be maximised. This is formalised in what is known as Bayes decision rule. Unfortunately the exact nature of each data source is often unknown, so an appropriate feature space for the statistical models for those sources and appropriate functional forms for the models are usually unknown. This is called model incorrectness. Counteracting model incorrectness is the motivation for the techniques described in this thesis. Two methods are proposed. First, given a sensible feature space and set of statistical models as a starting point, an alternative model-based feature space called a *score space* can be derived in which much simpler classifiers or distributions can be trained. Alternatively and from the same starting point, augmented forms of the original statistical models called *fibre bundles* can be defined and distributions trained within them. Under certain constraints on the fibre bundle and the score space classifier, these two approaches are identical and simply exemplify the well-known degeneracy between the feature extraction process and the forms of statistical models in the extracted feature space.

The thesis applies these techniques to the classification of patterns of fixed and variable length. The particular application is speech recognition since speech is a naturally occurring source of fixed and variable length patterns. The speech production process is also sufficiently complicated that any statistical model proposed is unlikely to replicate the underlying or correct data source, thereby introducing model incorrectness. The overall aim of the thesis is to furnish a better understanding of score spaces and fibre bundles in the context of classification, with respect to their limitations and advantages. The information geometric viewpoint of statistical models, developed by other researchers and applied in this thesis, should be of more general interest than the particular application to speech data. For those interested in speech recognition, the techniques permit the application of static classifiers such as SVMs to the classification of variable length patterns. This thesis does not address continuous speech recognition which is a much more difficult task, except in some comments for future work.

The thesis is organised as follows. This chapter describes the context and motivation for the thesis, and also summarises the main contributions in the thesis. Chapter 2 describes

classifiers for fixed and variable length patterns. Chapter 3 describes the theory of fibre bundles and score spaces. The development extends somewhat beyond that required for the experiments in the thesis but targets a more general understanding to stimulate further research. Chapter 4 describes score spaces in more detail in the context of classification. Chapter 5 then applies score spaces to the classification of fixed length patterns, and Chapter 6 to variable length patterns. Finally, Chapter 7 draws the thesis to a close with conclusions and ideas for future research.

It is helpful to summarise the central contributions in the thesis. The list is not self-explanatory since it is intended as a reference once the reader is familiar with the contents of the thesis. The theoretical developments are based on the fibre bundle of local exponential families from Section 4.8.1 of [3], developed there in the context of generalising asymptotic theory to families of distributions which were not embedded in exponential families. The author also draws heavily on this source for details of viewing statistical models from an information geometric perspective. The research on the Fisher kernel in [52] also provided a valuable starting point for applying score spaces to real tasks. The value of both of these sources cannot be understated. Many of the mathematical definitions originate from [50] and [108] and various sections of the thesis draw heavily on these sources. The contributions of this thesis are difficult to isolate from the ideas and suggestions of other researchers and supervisor, and are subject to the author's current knowledge. However with this caveat, contributions with respect to fibre bundles include,

- some further developments for the fibre bundle based on Taylor expansions defined and described in Section 4.8.1 of [3] (since the Taylor expansion is dependent on the choice of parameters for the statistical model, the bundle has limited general interest),

- introducing the concept of a space of scalar functions $L(\breve{p})$ to accomodate the evaluation of truncated Taylor expansions using fibre bundles,

- developing the relation between such fibre bundles, vector bundles and score spaces,

- developing an understanding of suitable metrics for score spaces,

- introducing within this context a manifold for multiple statistical models,
- developing the concept of score spaces defined on scalar fields other than the log likelihood, and the consequences for the semantic meaning of distributions,
- developing the concept of estimating distributions outside the statistical manifold but within the total space of the fibre bundle,
- detailing the concept and constraints under which training a linear discriminant in score space implicitly trains distributions within the total space of a fibre bundle,
- introducing 'fibre hopping', a technique which progressively maximises the log likelihood of a set of training samples by defining 'fibre bundles on fibres'.

With respect to score spaces and their application,

- introducing various 'appended posterior score spaces',
- comparing score spaces for the classification of fixed and variable length patterns,
- using score spaces to verify the relative importance of HMM parameters for discriminating letters in a simple isolated letter speech classification task.

Chapter 2

Classifiers

The techniques described and developed in this thesis are presented primarily in the context of classification. For context and motivation, this chapter first introduces the optimal Bayes decision rule and its implementation through statistical models in Section 2.1. Section 2.2 then reviews a variety of statistical models and estimation criteria. Since statistical models are often incorrect and the resulting classifier suboptimal, there are often advantages in training discriminants, both linear and nonlinear, via nonparametric techniques. Some common techniques are presented in Section 2.3. The complexity of classifiers can be reduced by increasing the complexity of the feature extraction process, and some relevant techniques are presented in Section 2.4. Finally, the classification of patterns of variable length is a demanding task. Section 2.5 reviews techniques for applying classifiers of fixed length patterns to this task, particularly in the context of speech recognition.

2.1 The optimal classifier

This section presents the concept of the optimal classifier from a decision theoretic perspective. The definitions and approach are taken from Chapter 2 of [25]. Parametric and nonparametric techniques are then introduced, and the concept of regularisation.

An open set of samples $L(\boldsymbol{O})$ exists where a sample is a pattern of data. If patterns are of fixed length, then they are often called *static data*, if of variable length then *dynamic data*. This space is assumed continuous and the samples are distributed according to the probability density function $p(\boldsymbol{O})$. Each sample $\boldsymbol{O} \in L(\boldsymbol{O})$ is assumed drawn from a class $\omega(\boldsymbol{O}) \in L(\omega)$ where $L(\omega) = \{\omega_1, \ldots \omega_Q\}$. There are therefore a finite number of Q possible classes. A decision rule or classifier D is required which assigns to the sample $\boldsymbol{O}$ the class $\hat{\omega}(\boldsymbol{O})$, so $D : \boldsymbol{O} \mapsto \hat{\omega}(\boldsymbol{O}), \forall \boldsymbol{O} \in L(\boldsymbol{O})$. From a decision theoretic approach, the optimal classifier minimises the *overall risk* R where,

$$R = \int R(\hat{\omega}(\boldsymbol{O})|\boldsymbol{O}) p(\boldsymbol{O}) \mathrm{d}\boldsymbol{O} \tag{2.1}$$

and where $R(\hat{\omega}(\boldsymbol{O})|\boldsymbol{O})$ is the *conditional risk*,

$$R(\hat{\omega}(\boldsymbol{O})|\boldsymbol{O}) = \sum_{q=1}^{Q} l(\hat{\omega}(\boldsymbol{O})|\omega_q) P(\omega_q|\boldsymbol{O}) \tag{2.2}$$

The term $P(\omega_q|\boldsymbol{O})$ is the posterior probability of class ω_q given sample $\boldsymbol{O}$ and $l(\omega_i|\omega_q)$ is the *loss* associated with selecting class ω_i for a sample which belongs truly to class ω_q. A popular loss function is one which does not penalise a correct classification but is equally injurious to incorrect classifications. So,

$$l(\omega_i|\omega_q) = \begin{cases} 0 & \text{if} \quad i = q \\ 1 & \text{if} \quad i \neq q \end{cases} \tag{2.3}$$

With this loss function, the conditional risk becomes,

$$R(\hat{\omega}(\boldsymbol{O})|\boldsymbol{O}) = \sum_{\substack{q=1 \\ \hat{\omega}_q \neq \omega(\boldsymbol{O})}}^{Q} P(\omega_q|\boldsymbol{O}) \tag{2.4}$$

The decision rule which minimises the conditional risk for each sample $\boldsymbol{O} \in L(\boldsymbol{O})$, and hence the overall risk, is that which dogmatically assigns each sample $\boldsymbol{O}$ to the class with maximum posterior probability for that sample. This decision rule is called the *optimal Bayes decision rule* and can be summarised by $D_{\text{opt}} : \boldsymbol{O} \mapsto \hat{\omega}(\boldsymbol{O}), \forall \boldsymbol{O} \in L(\boldsymbol{O})$ where,

$$\hat{\omega}(\boldsymbol{O}) = \underset{\omega_q \in L(\omega)}{\text{argmax}} \quad P(\omega_q|\boldsymbol{O}) \tag{2.5}$$

The optimal Bayes decision rule is sometimes called the *optimal Maximum A-Posteriori* (MAP) decision rule. The probability of error $\mathcal{E}_{\text{opt}}$ is then the overall risk R_{opt} so,

$$\mathcal{E}_{\text{opt}} \quad = \quad R_{\text{opt}} = \int \Big(1 - P(\hat{\omega}(\boldsymbol{O})|\boldsymbol{O})\Big) p(\boldsymbol{O}) \mathrm{d}\boldsymbol{O} \tag{2.6}$$

The optimal Bayes decision rule is strictly optimal in the sense of minimising the probability of error, and only optimal for minimising the overall risk subject to the 0-1 loss function. One of the most elegant and meaningful expressions of this decision rule is provided by Bayes Theorem where the posterior probability or *class posterior* for class ω_q is,

$$P(\omega_q|\boldsymbol{O}) \quad = \quad \frac{p(\boldsymbol{O}|\omega_q)P(\omega_q)}{p(\boldsymbol{O})} \tag{2.7}$$

The term $p(\boldsymbol{O}|\omega_q)$ is the *class likelihood* and is the output of a probability density function, the term $P(\omega_q)$ is the *class prior* and is the output of a probability mass function if Q is finite, and the term $p(\boldsymbol{O})$ is the *evidence* where,

$$p(\boldsymbol{O}) \quad = \quad \sum_{q=1}^{Q} p(\boldsymbol{O}|\omega_q)P(\omega_q) \tag{2.8}$$

Then, introducing the log term, the optimal Bayes decision rule selects the class,

$$\hat{\omega}(\boldsymbol{O}) \quad = \quad \underset{\omega_q \in L(\omega)}{\text{argmax}} \quad \Big(\ln p(\boldsymbol{O}|\omega_q) + \ln P(\omega_q)\Big) \tag{2.9}$$

Although the input space has been assumed continuous, the same theoretical development is possible for a discrete input space but with the relevant replacement of probability density functions and integrals by probability mass functions and summations. Bayes decision rule mimics the human reasoning process where decisions are typically based on previous knowledge and observing circumstantial information.

So far, the explanation has assumed that class posteriors, or class likelihoods and priors are known. Simply put, it assumes that correct models are known. In real-world applications, the only information pertaining to the correct models is usually available through a quantity of samples called *training data*. The goal of classification is to estimate the optimal Bayes decision rule through observing patterns in the training data. There are two common approaches.

- Parametric approaches use the training data to estimate models for class posteriors directly, or indirectly through estimating models for class likelihoods and priors. A decision rule in the form of Equation 2.5 or 2.9 is then applied to the estimates. However if the models are incorrect and do not perfectly capture the statistical relationships between samples, then the implementation of the decision rule is not necessarily optimal and the error rate $\mathcal{E}$ does not necessarily attain the lower bound, i.e. $\mathcal{E} \geq \mathcal{E}_{\text{opt}}$. This thesis applies the term 'Bayes decision rule' or 'MAP decision rule' to any rule of the form of Equations 2.5 or 2.9, but strictly reserves the term 'optimal decision rule' for the Bayes decision rule defined on correct models. An example of a class posterior model is a sigmoid trained about a linear discriminant. Examples of class likelihood and class prior models are prevalent in many applications of statistical pattern classification such as speech recognition.

- Nonparametric approaches learn decision rules directly from the data according to certain criteria but without inferring class models. These techniques make no assumptions about the forms of class models and may be more robust when there is little training data. Examples include Minimum Square Error classifiers and Support Vector Machines.

For clarity, this thesis reserves the term 'error rate' for the probability of error, and *training error rate* and *test error rate* for the empirically measured error rates on respectively training and test data, unless clear from context.

Regularisation is an important concept for estimating classifiers via parametric or nonparametric techniques. Taking the example of a parametric approach with statistical models where the correct forms of those models are not known, then increasing the complexity of the proposed models lowers the training error rate. However beyond a certain complexity the test error rate, and hence the error rate which it approximates, increases. The classifier becomes *overtrained* and its ability to *generalise* to unseen data is impaired. This is illustrated in Figure 2.1. The plots illustrate that there is often an optimal complexity for a given task, ideally yielding an error rate as close as possible to $\mathcal{E}_{\text{opt}}$. Overtraining is due to a mismatch between the relevant statistics for each class of training samples and

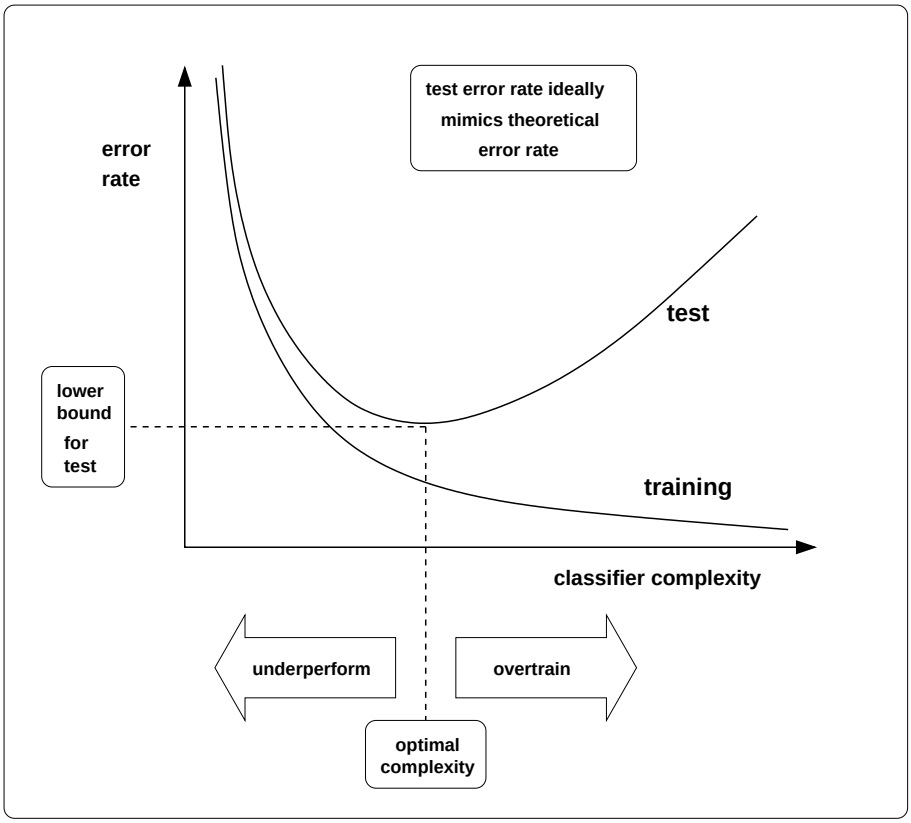


Figure 2.1: Illustrating the effect of overtraining on training and test error rates

the corresponding underlying distributions, where relevant is defined as that required to learn the classifier. The mismatch can be minimised by increasing the number of training samples drawn from the underlying distribution. The absolute size of the training set is less important than its size relative to the number of parameters to be estimated in the classifier. *Regularisation* is a general concept applicable to any form of classifier. It counteracts overtraining by favouring simpler classifiers where possible and so loosely implements Occam's Razor. In a wider sense regularisation is any method which constrains complexity with the intention of improving generalisation. This includes 'hard' regularisation, i.e. enforcing a-priori restrictions on the form of the classifier, or 'soft' regularisation, i.e. optimising an objective function to select the complexity which best trades-off reduced training error rate for better generalisation. Examples of 'soft' regularisation are weight decay in neural networks [6], optimising the curvature in polynomial regression, regularisation operators [6], priors in a Bayesian formulation [6], and the learning criterion in Support Vector Machines (SVMs) [19]. In the experiments in this thesis, linear classifiers

are estimated in linear spaces of high dimension relative to the number of training samples. Regularisation is then imperative for robust estimation.

This thesis focuses on estimating the optimal decision rule or classifier through modelling class likelihoods and class priors. The models for class likelihoods are called *statistical models*. They are typically incorrect. The main theme of this thesis is the development of statistical models, and their combination with nonparametric techniques, to attain better estimates of optimal classifiers.

2.2 Parametric techniques with statistical models

2.2.1 Statistical models

For a parametric approach with statistical models there are the following.

- The data source: this is the real-world process or underlying distribution which yields samples.
- The statistical model: this is proposed as a description of the source. Although it may be used in a generative fashion, its application is often constrained to analysis, typically to calculate likelihoods for samples. If the model is correct, then it perfectly captures the statistical relationships in the source.

Data sources may be static or dynamic in nature. The statistical properties of samples within sequences generated by these sources are respectively invariant and variant to the location of those samples in the sequence. A collection of samples of static, though not dynamic, data can be randomly permuted without affecting its integrity to the source (see [85]). If a static statistical model mimics a dynamic data source, the model can only capture the 'average' statistical properties of samples. A dynamic statistical model can mimic a static data source but with considerable redundancy. In this thesis a sample $\boldsymbol{O}$

is an ordered sequence of T observations,

$$\boldsymbol{O} \quad = \quad (\boldsymbol{o}_1, \ldots \boldsymbol{o}_t, \ldots \boldsymbol{o}_T) \tag{2.10}$$

Each observation is a $(d \times 1)$ column vector. If T is fixed, the sample is a fixed length pattern, else it is a variable length pattern. In this thesis, the type of patterns appropriate for a distribution $p(\boldsymbol{O}; \boldsymbol{\theta})$ should be clear from context.

A statistical model is summarised as $S(\boldsymbol{\theta})$ where $\boldsymbol{O} \in L(\boldsymbol{O})$ and $L(\boldsymbol{O})$ is an open set of samples,

$$S(\boldsymbol{\theta}) = \{p(\boldsymbol{O}; \boldsymbol{\theta}) \quad | \quad \boldsymbol{\theta} \in L(\boldsymbol{\theta}; S)\} \tag{2.11}$$

If each component of the parameter vector $\boldsymbol{\theta}$ is linearly independent, then $\text{size}(\boldsymbol{\theta}) = \dim(L(\boldsymbol{\theta}; S)) = n$, where $\text{size}(\cdot)$ and $\dim(\cdot)$ respectively denote the number of components and dimension of their arguments[1]. The constraints on the functional form of the model are implicit in the definition of $S(\boldsymbol{\theta})$. A parameterisation defines a probability density function or *distribution* [3][2], and $\boldsymbol{\theta} \in L(\boldsymbol{\theta}; S)$ ensures valid distributions.

In this thesis, statistical models are restricted to Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) with state-conditional likelihoods modelled by GMMs. A distribution within one of these models is respectively denoted by GMD or HMD. A Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ over the open set of observations $L(\boldsymbol{o})$ is defined as follows, where $\text{size}(\boldsymbol{o}) = d$,

$$\mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad = \quad \frac{1}{(2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\boldsymbol{o} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{o} - \boldsymbol{\mu})\} \tag{2.12}$$

A GMM with K mixture components is the static model $S(\boldsymbol{\theta})$,

$$S(\boldsymbol{\theta}) = \{p(\boldsymbol{o}; \boldsymbol{\theta}) \quad | \quad \boldsymbol{\theta} \in L(\boldsymbol{\theta}; S)\} \tag{2.13}$$

where,

$$p(\boldsymbol{o}; \boldsymbol{\theta}) \quad = \quad \sum_{k=1}^{K} w_k b_k(\boldsymbol{o}) \tag{2.14}$$

[1] This section always assumes an Identity metric tensor for parameter space so there is no need to distinguish covariant and contravariant components.

[2] The application of the term 'distribution' to descriptions which are not probability density functions should be clear from context, for example those prior 'distributions' or posterior 'distributions' which are probability mass functions.

and,

$$b_k(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2.15}$$

The set $L(\boldsymbol{\theta}; S)$ implies $\boldsymbol{\Sigma}_k$ is positive definite and symmetric, and w_k is such that,

$$0 \leq w_k \leq 1 \tag{2.16}$$

$$\sum_{k=1}^{K} w_k = 1 \tag{2.17}$$

The statistical model may be augmented with a first order Markov process to create a dynamic statistical model called an HMM,

$$S(\boldsymbol{\theta}) = \{p(\boldsymbol{O}; \boldsymbol{\theta}) \quad | \quad \boldsymbol{\theta} \in L(\boldsymbol{\theta}; S)\} \tag{2.18}$$

There are N emitting states[3] each modelled by a GMM. The transition probability from state i to state j is the transition probability $a(i, j)$. The models of primary interest in speech recognition are those which are left-to-right and, unless modelling silence, have no skips. So enforcing these constraints in this thesis, $a(i, j) = 0$ if $j < i$, and $a(i, i) = 1 - a(i, i + 1)$ with $a(i, j) = 0$ if $j > i + 1$. This thesis also defines $a(s(T), s(T + 1)) = 1$ where $s(t)$ is the state at sequence location, or *time*, t. The HMM has initial and final state distributions respectively π and ϖ where,

$$\pi = \{\pi_1, \ldots \pi_N\}, \quad \pi_j = P(s(1) = j) \tag{2.19}$$

$$\varpi = \{\varpi_1, \ldots \varpi_N\}, \quad \varpi_j = P(s(T) = j) \tag{2.20}$$

In this thesis $\pi_1 = 1$ and $\pi_j = 0$ for $j \neq 1$, and $\varpi_N = 1$ and $\varpi_j = 0$ for $j \neq N$. The HMM has an exponential state duration probability density function, where the rate of decay is determined by the self-transition probability of the state. It is possible to apply more realistic duration modelling [80]. Then letting ψ denote a state-level path through the HMM, and Ψ the entire set of such paths,

$$\begin{aligned} p(\boldsymbol{O}; \boldsymbol{\theta}) &= \sum_{\psi \in \Psi} p(\boldsymbol{O}, \psi; \boldsymbol{\theta}) \\ &= \sum_{\psi \in \Psi} p(\boldsymbol{O}; \psi, \boldsymbol{\theta}) P(\psi; \boldsymbol{\theta}) \end{aligned} \tag{2.21}$$

[3]There are no nonemitting states.

$$= \sum_{\psi \in \Psi} \Big(\prod_{t=1}^{T} b^{\psi}_{s(t)}(\boldsymbol{o}_t) \Big) \Big(\prod_{t=1}^{T} a^{\psi}(s(t), s(t+1)) \Big)$$

$$= \sum_{\psi \in \Psi} \prod_{t=1}^{T} b^{\psi}_{s(t)}(\boldsymbol{o}_t) a^{\psi}(s(t), s(t+1)) \tag{2.22}$$

where,

$$b_j(\boldsymbol{o}_t) = \sum_{k=1}^{K} w_{jk} b_{jk}(\boldsymbol{o}_t) \tag{2.23}$$

and $b_{jk}(\boldsymbol{o}_t)$ is defined as for $b_k(\boldsymbol{o})$ in Equation 2.15 except the index k is replaced by jk.

For a Q-class problem,

$$\mathcal{S}(\boldsymbol{\xi}) = \{ S(\boldsymbol{\theta}_1), \ldots S(\boldsymbol{\theta}_Q) \quad | \quad \boldsymbol{\xi} = (\boldsymbol{\theta}_1^{\top}, \ldots \boldsymbol{\theta}_Q^{\top})^{\top} \in L(\boldsymbol{\xi}; \mathcal{S}) \} \tag{2.24}$$

where,

$$L(\boldsymbol{\xi}; \mathcal{S}) = \oplus_{q=1}^{Q} L(\boldsymbol{\theta}_q; S) \tag{2.25}$$

and $\oplus$ is the module direct sum (see Section 275.F [50], [108]). The model $S(\boldsymbol{\theta}_q)$ represents class ω_q. When this framework is applied to HMMs, the index q distinguishes the parameters of one statistical model from another. For example, for state j of the model $S(\boldsymbol{\theta}_q)$, the state-conditional likelihood for $\boldsymbol{o}_t$ is $b_{qj}(\boldsymbol{o}_t)$.

First order Markov processes cannot model long term correlations in the underlying signal. Other statistical models may be of interest.

- Linear Gaussian Models (LGMs) [85] [84] are often restricted to those which are either static models or first order Markov processes. Besides GMMs and HMMs, other LGMs which may be viewed as statistical models include Factor Analysis, Special Principal Component Analysis, Independent Component Analysis and Kalman Filters [85]. Gaussians may also be replaced by any other forms of distribution.

- Markov Random Fields (MRFs) [100] are an extension of one-dimensional first order Markov dependencies between neighbouring states to two-dimensional dependencies which may be applied, for example, to pixels in images.

- Autoregressive (AR) models, under conditions, approximate probability density functions [41]. Then mixtures of AR models may be used to define state-conditional likelihoods for AR-HMMs [80] [66].
- Connectionist models include time-recurrent neural network [83] which can model long-term correlations in a sequence. Unfortunately connectionist models typically estimate class posterior probabilities. They define statistical models only in exceptional cases.

2.2.2 Training criteria

This section describes some popular estimation criteria for statistical models particularly relevant to speech recognition. For a Q-class problem, the unknown sources are described by the distributions $p''(\boldsymbol{O}|\omega_q), q = \{1, \ldots Q\}$, and the statistical models as in Equation 2.24. For supervised training, the training set of ordered pairs is,

$$\{(\boldsymbol{O}_1, y_1), \ldots (\boldsymbol{O}_\ell, y_\ell)\} \tag{2.26}$$

where each sample $\boldsymbol{O}_l \in L(\boldsymbol{O})$ has label $y_l \in \{1, \ldots Q\}$ and $y_l = q$ indicates the sample is drawn from class ω_q. Then,

$$\mathcal{O}_{\text{train}} = \{\boldsymbol{O}_1, \ldots \boldsymbol{O}_\ell\} \tag{2.27}$$

$$\Omega_{\text{train}} = \{\omega_{y_1}, \ldots \omega_{y_\ell}\} \tag{2.28}$$

There are ℓ_q samples with label $y_l = q$ and these form the subset $\mathcal{O}_{\text{train(q)}}$ and their class identities, all identical, the subset $\Omega_{\text{train(q)}}$. The samples in $\mathcal{O}_{\text{train}}$ are assumed independently and identically distributed (i.i.d.) and the class ω_{y_l} is only dependent on $\boldsymbol{O}_l$. In this section, the class priors are assumed known and independent of the parameters $\boldsymbol{\xi}$.

First, Maximum Likelihood (ML) estimation seeks the parameters $\boldsymbol{\xi}_{\text{ML}}$ to maximise the likelihood of $\mathcal{O}_{\text{train}}$,

$$\boldsymbol{\xi}_{\text{ML}} = \underset{\boldsymbol{\xi}}{\text{argmax}} \sum_{q=1}^{Q} \sum_{\substack{l=1 \\ y_l = q}}^{\ell} \ln p(\boldsymbol{O}_l|\omega_q) \tag{2.29}$$

Asymptotically in the limit of an infinite number of training samples,

$$\boldsymbol{\xi}_{\mathrm{ML}} = \underset{\boldsymbol{\xi}}{\operatorname{argmax}} \sum_{q=1}^{Q} \int \ln p(\boldsymbol{O}|\omega_q) p''(\boldsymbol{O}|\omega_q) \mathrm{d}\boldsymbol{O} \tag{2.30}$$

Only when $p(\boldsymbol{O}|\omega_q)$ and $p''(\boldsymbol{O}|\omega_q)$ coincide is the expected log likelihood for class ω_q maximised. The ML estimator is only consistent if the functional form of the statistical model is correct, there is an infinite number of training samples, the optimisation method guarantees global maximisation, and with certain constraints on the initial model parameters (see [28]). These conditions cannot usually be guaranteed. ML estimation is typically implemented by the EM algorithm for GMMs [21] and the Baum-Welch algorithm for HMMs [80].

In Maximum Mutual Information (MMI) estimation, all parameters are optimised concurrently. The mutual information between the random variables $\mathcal{O}$ and Ω is,

$$I(\mathcal{O};\Omega) = \sum_{\mathcal{O}\in L(\mathcal{O})} \sum_{\Omega\in L(\Omega)} P(\mathcal{O},\Omega) \ln \frac{P(\mathcal{O},\Omega)}{p(\mathcal{O})P(\Omega)} \tag{2.31}$$

where $L(\mathcal{O})$ and $L(\Omega)$ are respectively the open set of all datasets and the discrete open set of all possible class permutations. Mutual information may be approximated as below, where the approximation may be negative (see [28]) and hence not a valid distance metric. Using this,

$$\begin{aligned} \boldsymbol{\xi}_{\mathrm{MMI}} &= \underset{\boldsymbol{\xi}}{\operatorname{argmax}} \; I(\mathcal{O};\Omega)\Big|_{\substack{\mathcal{O}=\mathcal{O}_{\mathrm{train}} \\ \Omega=\Omega_{\mathrm{train}}}} \\ &= \underset{\boldsymbol{\xi}}{\operatorname{argmax}} \sum_{q=1}^{Q} \sum_{\substack{l=1 \\ y_l=q}}^{\ell} \ln P(\omega_q|\boldsymbol{O}_l) \end{aligned} \tag{2.32}$$

MMI estimation was first presented in the context of speech recognition in [4] using a gradient descent optimisation, though versions of the extended Baum-Welch (EBW) algorithm are now available [111].

Maximum A-Posteriori (MAP) estimation [39] searches for the model parameters $\boldsymbol{\xi}_{\mathrm{MAP}}$

which maximise the a-posteriori distribution over $\boldsymbol{\xi}$,

$$\boldsymbol{\xi}_{\text{MAP}} = \underset{\boldsymbol{\xi}}{\text{argmax}} \sum_{q=1}^{Q} \Big(\ln p(\boldsymbol{\theta}_q) + \sum_{\substack{l=1 \\ y_l = q}}^{\ell} \ln p(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \Big) \tag{2.33}$$

In the limit of an infinite number of training samples for each class, $\boldsymbol{\xi}_{\text{MAP}} \rightarrow \boldsymbol{\xi}_{\text{ML}}$. However the parameter priors typically introduce robustness when there are relatively few training samples.

Training criteria may be viewed as minimising differences between sources and proposed models [28]. A principled approach is introduced in the Kullback-Leibler (KL) information [28] where for two continuous distributions $p(\boldsymbol{O})$ and $q(\boldsymbol{O})$ over $L(\boldsymbol{O})$,

$$\text{KL}(q(\boldsymbol{O})||p(\boldsymbol{O})) = \int q(\boldsymbol{O}) \ln \frac{q(\boldsymbol{O})}{p(\boldsymbol{O})} \text{d}\boldsymbol{O} \tag{2.34}$$

Then $\text{KL}(q(\boldsymbol{O})||p(\boldsymbol{O})) = 0$ if $p(\boldsymbol{O}) = q(\boldsymbol{O}), \forall \boldsymbol{O} \in L(\boldsymbol{O})$. The KL information is not a valid distance metric since for example it is not symmetric. Nonzero values indicate dissimilarity. If both distributions are drawn from the same statistical model, then the KL information is the α-divergence where $\alpha = -1$ (see Section 3.2 [3]). KL information relationships for ML and MMI training are developed in [28] for a discrete space of samples of data sequences (there is no straightforward extension to continuous space through the application of dirac delta functions since their definition is inconsistent, see Section 21.9-2 of [59]). However the separation between adjacent samples may be made arbitrarily small. To describe these relationships, each continuous distribution is given a discrete analogue by replacing the lowercase letter by its uppercase, for example $p(\boldsymbol{O}|\omega_q)$ by $P(\boldsymbol{O}|\omega_q)$. It is also necessary to assume the source distribution $P''(\boldsymbol{O}|\omega_q)$ is modelled by a probability mass function $R(\boldsymbol{O}|\omega_q)$ called the *assumed source*. Where relevant $P''(\omega_q)$, the correct prior for class ω_q, is modelled by the *assumed prior* $R(\omega_q)$.

First for ML estimation,

$$\boldsymbol{\xi}_{\text{ML}} = \underset{\boldsymbol{\xi}}{\text{argmin}} \sum_{q=1}^{Q} \alpha_q \text{ KL}(R(\mathcal{O}|\Omega)||P(\mathcal{O}|\Omega))\Big|_{\mathcal{O}=\mathcal{O}_{\text{train(q)}}, \Omega=\Omega_{\text{train(q)}}} \tag{2.35}$$

where α_q is any class-specific multiplier independent of $\boldsymbol{\xi}$. ML estimation therefore selects the parameters $\boldsymbol{\xi}_{\text{ML}}$ which minimise the weighted average KL information between the

assumed source and model distributions. For MMI estimation,

$$\boldsymbol{\xi}_{\mathrm{MMI}} = \underset{\boldsymbol{\xi}}{\operatorname{argmin}} \sum_{q=1}^{Q} \mathrm{KL}(R(\Omega|\mathcal{O})||P(\Omega|\mathcal{O}))\Big|_{\mathcal{O}=\mathcal{O}_{\mathrm{train(q)}},\Omega=\Omega_{\mathrm{train(q)}}} \tag{2.36}$$

This estimate minimises the average KL information between the assumed source class posteriors and the model class posteriors. Strictly, Equation 2.35 is evaluated at $\Omega_{\mathrm{train(q)}}$ and all probability mass in $L(\mathcal{O})$ is then located at $\mathcal{O}_{\mathrm{train(q)}}$, whereas Equation 2.36 is evaluated at $\mathcal{O}_{\mathrm{train(q)}}$ and assumes all probability mass in $L(\Omega)$ is then located at $\Omega_{\mathrm{train(q)}}$.

So far the correct source $P''(\boldsymbol{O}|\omega_q)$ has been approximated by the assumed source $R(\boldsymbol{O}|\omega_q)$ defined as a collection of discrete dirac delta functions. However the intervening KL information tends to zero as $\ell_q \to \infty$ since,

$$\lim_{\ell_q\to\infty} R(\boldsymbol{O}|\omega_q) = \lim_{\ell_q\to\infty} \sum_{\substack{l=1 \\ y_l = q}}^{\ell} \frac{1}{\ell_q}\delta(\boldsymbol{O}-\boldsymbol{O}_l) = P''(\boldsymbol{O}|\omega_q) \tag{2.37}$$

With increasing numbers of samples, the assumed source reflects the true source. Consequently with few samples, the ML estimate may poorly approximate the true source.

It is useful to examine the relationship between probability mass functions and probability density functions, for example $P(\boldsymbol{O}|\omega_q)$ and $p(\boldsymbol{O}|\omega_q)$, when the mass function is assumed derived from the density function under a discrete sampling of continuous space $L(\boldsymbol{O})$. First, let the discrete intervals $\boldsymbol{\Delta}$ describe a 'grid' $L(\mathrm{gr},\boldsymbol{O})$ on $L(\boldsymbol{O})$. The probability mass function is defined at the discrete points of this grid. A continuous density 'block' $h(\cdot)$ is then fitted to each grid point and weighted by the value of the mass function at the grid point. The resulting distribution is $p_{\mathrm{disc}}(\boldsymbol{O}|\omega_q)$ where,

$$p_{\mathrm{disc}}(\boldsymbol{O}|\omega_q) = \sum_{\boldsymbol{O}_g \in L(\mathrm{gr},\boldsymbol{O})} P(\boldsymbol{O}_g|\omega_q)h(\boldsymbol{O}-\boldsymbol{O}_g) \tag{2.38}$$

and where,

$$h(\boldsymbol{O}) = \begin{cases} \frac{1}{\Delta} & -\frac{\boldsymbol{\Delta}}{2} < \boldsymbol{O} \leq \frac{\boldsymbol{\Delta}}{2} \\ 0 & \text{otherwise} \end{cases} \tag{2.39}$$

Then as $\boldsymbol{\Delta} \rightarrow \mathbf{0}$, $p_{\text{disc}}(\boldsymbol{O}|\omega_q) \rightarrow p(\boldsymbol{O}|\omega_q)$. These 'continuous space versions' of probability mass functions converge to the corresponding probability density functions as the interval of discretisation becomes infinitessimal. In this limit and assuming the probability mass function is formed by the discrete sampling process detailed above, it seems reasonable to transfer deductions on relations between probability mass functions to probability density functions.

2.2.3 Maximum Likelihood Estimation (MLE) discriminant

A popular decision rule for classes ω_a and ω_b may be derived from statistical models with respective distributions,

$$p(\boldsymbol{O};\boldsymbol{\theta}_a) = \mathcal{N}(\boldsymbol{O};\boldsymbol{\mu}_a,\boldsymbol{\Sigma}) \tag{2.40}$$

$$p(\boldsymbol{O};\boldsymbol{\theta}_b) = \mathcal{N}(\boldsymbol{O};\boldsymbol{\mu}_b,\boldsymbol{\Sigma}) \tag{2.41}$$

where $\boldsymbol{O} \in L(\boldsymbol{O})$ and $L(\boldsymbol{O})$ is restricted to patterns of fixed length. Assuming equal class priors, application of Bayes decision rule yields the linear discriminant [25],

$$\boldsymbol{w}^\top \boldsymbol{O} + b = 0 \tag{2.42}$$

where,

$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}(y_a\boldsymbol{\mu}_a + y_b\boldsymbol{\mu}_b) \tag{2.43}$$

$$b = -\frac{1}{2}(y_a\boldsymbol{\mu}_a^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_a + y_b\boldsymbol{\mu}_b^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_b) \tag{2.44}$$

and where without loss in generality $y_a = +1$ and $y_b = -1$. When the single Gaussians are fitted to individual classes by ML estimation, the discriminant is here called the Maximum Likelihood Estimation (MLE) discriminant. The tied covariance matrix may either model the within-class or global covariance.

2.3 Nonparametric techniques

2.3.1 Linear discriminants

An important subset of classifiers are piecewise linear, or for binary problems, linear discriminants. A linear discriminant can separate all possible dichotomies, or binary divisions, for at least one set of $(d+1)$ samples in a d-dimensional input space. The Vapnik-Chervonenkis (VC) dimension of the linear discriminant is $h = (d+1)$ [9]. The VC dimension is a measure of capacity [104]. According to [25], when the number of samples approaches $2h$ linear separation becomes more difficult for a random dichotomy, and the discriminant becomes overdetermined. This implies some robustness if the learning algorithm can accomodate errors. Since the capacity of linear discriminants is relatively low compared to other discriminants, they generally yield more robust classifiers when training data is sparse as in the experiments in this thesis. Robustness can be further improved by additional capacity control.

A linear discriminant is characterised by a weight $\boldsymbol{w}$ and bias b. In application, the discriminant induces a *functional distance* $d(\boldsymbol{O}_l)$ for the sample $\boldsymbol{O}_l \in L(\boldsymbol{O})$ where $L(\boldsymbol{O})$ is constrained to samples of fixed length and[4],

$$d(\boldsymbol{O}_l) = (\boldsymbol{w}, \boldsymbol{O}_l) + b \tag{2.45}$$

and $(\cdot,\cdot)$ is the scalar product between members of the same space. Normalising the weight vector by its norm $||\boldsymbol{w}||$ yields the *geometric distance* $\gamma(\boldsymbol{O}_l)$,

$$\begin{aligned} \gamma(\boldsymbol{O}_l) &= \frac{1}{||\boldsymbol{w}||}((\boldsymbol{w}, \boldsymbol{O}_l) + b) \\ &= (\boldsymbol{w}_{\text{unit}}, \boldsymbol{O}_l) + b_{\text{unit}} \end{aligned} \tag{2.46}$$

Therefore,

$$\gamma(\boldsymbol{O}_l) = \frac{d(\boldsymbol{O}_l)}{||\boldsymbol{w}||} \tag{2.47}$$

[4]It is also possible to define a vector $\tilde{\boldsymbol{w}} = (\boldsymbol{w}^\top, b)^\top$ and view the learning algorithms and classification rules relative to $\tilde{\boldsymbol{w}}$.

The linear discriminant defines two half-spaces, one for each class. Assuming samples from class ω_a and ω_b are respectively labelled with $y_a = 1$ and $y_b = -1$, then a decision rule consistent with this labelling is as follows, where the assignment at equality is otherwise arbitrary,

$$\hat{\omega}(\boldsymbol{O}_l) \quad = \quad \begin{cases} \omega_a & \text{if} \quad d(\boldsymbol{O}_l) \geq 0 \\ \omega_b & \text{if} \quad d(\boldsymbol{O}_l) < 0 \end{cases} \tag{2.48}$$

2.3.1.1 Minimum Square Error (MSE) learning machine

The Minimum Square Error (MSE) learning machine [25] minimises the following criterion, where all samples contribute errors through a quadratic loss function,

$$\mathcal{F}_{\text{MSE}}(\boldsymbol{w}, b) \quad = \quad \sum_{\boldsymbol{O}_l \in \mathcal{O}_{\text{train}}} \Big(-y_l((\boldsymbol{w}, \boldsymbol{O}_l) + b) + d_l \Big)^2 \tag{2.49}$$

where $d_l \geq 0$ is a predefined target or functional margin for the sample $\boldsymbol{O}_l$, and $\mathcal{O}_{\text{train}}$ is the training data. MSE learning is a regression technique and functional margins are assigned a-priori. All samples from a given class are usually given the same margin, so requiring $d_a > 0$ and $d_b > 0$,

$$d_l \quad = \quad \begin{cases} +d_a & \text{if} \quad y_l = 1 \\ -d_b & \text{if} \quad y_l = -1 \end{cases} \tag{2.50}$$

The criterion $\mathcal{F}_{\text{MSE}}(\boldsymbol{w}, b)$ can be minimised through calculating the pseudoinverse solution of a linear algebraic equation [25]. Providing the solution exists MSE learning returns a unique solution. However when the data is linearly separable, it does not guarantee a discriminant which perfectly separates the two classes. MSE learning finds the linear discriminant which forces all samples to lie as close as possible to the $+\gamma_a$ and $-\gamma_b$ geometric hyperplanes, where γ_i and d_i, $i = \{a, b\}$ are related as in Equation 2.47. Each sample has equal influence in training the discriminant. If the pseudoinverse solution does not exist, a unique solution can be obtained through ridge regression [25] which introduces some regularisation.

The MSE solution has an important relationship with the MLE linear discriminant de-

scribed in Section 2.2.3 when,

$$d_a = \frac{\ell}{\ell_a} \tag{2.51}$$

$$d_b = \frac{\ell}{\ell_b} \tag{2.52}$$

and ℓ_a and ℓ_b are respectively the numbers of samples in classes ω_a and ω_b, and $\ell = \ell_a + \ell_b$. Referring to Section 2.2.3 for notation, then if the tied covariance matrix is the weighted within-class covariance matrix, i.e. $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\text{wtd}}$, then [25] shows that the weight vector of the MSE discriminant coincides with that of the MLE discriminant up to a scaling constant. When normalised the two weight vectors $\boldsymbol{w}_{\text{unit}}(MSE)$ and $\boldsymbol{w}_{\text{unit}}(MLE)$ are identical (and also identical to the normalised weight vector yielded by Fisher Discriminant Analysis), but the corresponding biases $b_{\text{unit}}(MSE)$ and $b_{\text{unit}}(MLE)$ differ,

$$b_{\text{unit}}(MSE) = -\frac{1}{||\boldsymbol{w}(MSE)||}(y_a \boldsymbol{\mu}_a^\top \boldsymbol{\Sigma}_{\text{wtd}}^{-1} \boldsymbol{\mu}_{\text{glob}} + y_b \boldsymbol{\mu}_b^\top \boldsymbol{\Sigma}_{\text{wtd}}^{-1} \boldsymbol{\mu}_{\text{glob}}) \tag{2.53}$$

$$b_{\text{unit}}(MLE) = -\frac{1}{2||\boldsymbol{w}(MLE)||}(y_a \boldsymbol{\mu}_a^\top \boldsymbol{\Sigma}_{\text{wtd}}^{-1} \boldsymbol{\mu}_a + y_b \boldsymbol{\mu}_b^\top \boldsymbol{\Sigma}_{\text{wtd}}^{-1} \boldsymbol{\mu}_b) \tag{2.54}$$

where,

$$\boldsymbol{\mu}_{\text{glob}} = \frac{1}{\ell}(\ell_a \boldsymbol{\mu}_a + \ell_b \boldsymbol{\mu}_b) \tag{2.55}$$

However if $\ell_a = \ell_b$, the two biases and linear discrimants coincide.

2.3.1.2 Support Vector Machine (SVM)

MSE learning penalises samples which lie on the correct side of the hyperplanes geometrically defined at $+\gamma_a$ and $-\gamma_b$. It is sensible to limit penalisation to the set of samples $\mathcal{O}_{\text{ms}}$ which lie only on the incorrect side of these hyperplanes, i.e. those which fail the margin constraint. This forms the basis of the Support Vector Machine (SVM) [19]. The SVM discriminant combines characteristics of the MSE formulation and perceptron learning with a margin constraint [25]. The SVM minimises, for $\beta = \{1, 2\}$,

$$\mathcal{F}_{\text{SVM}}(\boldsymbol{w}, b) = C \sum_{\boldsymbol{O}_l \in \mathcal{O}_{\text{ms}}} \Big(-y_l((\boldsymbol{w}, \boldsymbol{O}_l) + b + d_l)\Big)^\beta + \frac{1}{2^{(2-\beta)}}(\boldsymbol{w}, \boldsymbol{w}) \tag{2.56}$$

Setting[5] $\beta = 2$ yields the '2-norm soft margin' SVM and setting $\beta = 1$ yields the '1-norm soft margin' SVM [19]. The loss functions are respectively quadratic hinge and linear

[5]The multiplier $1/(2^{(2-\beta)})$ may be subsumed within C.

hinge loss functions. The most common SVM formulation is the 1-norm soft margin SVM and is henceforth implied by the term 'SVM'. Next canonical hyperplanes either side of the linear discriminant are defined by $d_a = d_b = 1$ and $\gamma_a = \gamma_b = \gamma_{\text{can}}$. Then,

$$\mathcal{F}_{\text{SVM}}(\boldsymbol{w}, b) \;=\; C \sum_{\boldsymbol{O}_l \in \mathcal{O}_{\text{ms}}} \Big(-y_l((\boldsymbol{w}, \boldsymbol{O}_l) + b) + 1 \Big) + \frac{1}{2}(\boldsymbol{w}, \boldsymbol{w}) \tag{2.57}$$

The choice $d_a = d_b = 1$ is inconsequential since $\boldsymbol{w}$ is permitted to have non-unit length and hence the geometric margin $\gamma_{\text{can}} = 1/||\boldsymbol{w}||$ is free to vary. The SVM finds the solution which maximises γ_{can} while simultaneously minimising the sum of errors. The parameter C controls the trade-off. If the two classes are linearly separable, then minimisation does not guarantee a separating linear discriminant. Capacity control ensures a unique solution and positions the discriminant 'midway between the two classes'. The discriminant is influenced by the samples near to the decision boundary rather than by the data density as for MSE learning. In the extreme case the solution is minimally defined by two samples and so is sensitive to class outliers.

Quadratic programming techniques are applied. It is common practice to map the constrained optimisation into its Lagrangian dual [19]. Possible optimisation techniques include interior-point methods [112], chunking methods [8] [54], and the Sequential Minimal Optimisation (SMO) algorithm [76] [57] (see also comments on [75]). The solution for $\boldsymbol{w}$ is,

$$\boldsymbol{w} \;=\; \sum_{\boldsymbol{O}_l \in \mathcal{O}_{\text{train}}} \alpha_l y_l \boldsymbol{O}_l \tag{2.58}$$

where $0 \leq \alpha_l \leq C$ [19]. Any sample for which $\alpha_l > 0$ is called a *support vector* since it defines the weight $\boldsymbol{w}$. Support vectors necessarily fail the margin constraint or lie on the canonical hyperplanes. Ideally the solution should be sparse with few support vectors. The bias b is available from application of the Karush-Kuhn-Tucker complementarity conditions.

The MSE and SVM learning machines are compared in Figure 2.2, where the MSE discriminant is equivalent to an MLE discriminant where the tied covariance matrix models within-class covariance. Briefly, MSE learning trains $\boldsymbol{w}$ to select appropriate geometrical margins γ_a and γ_b to minimise errors, while the SVM maximises its geometrical margin γ_{can} while simultaneously minimising errors.

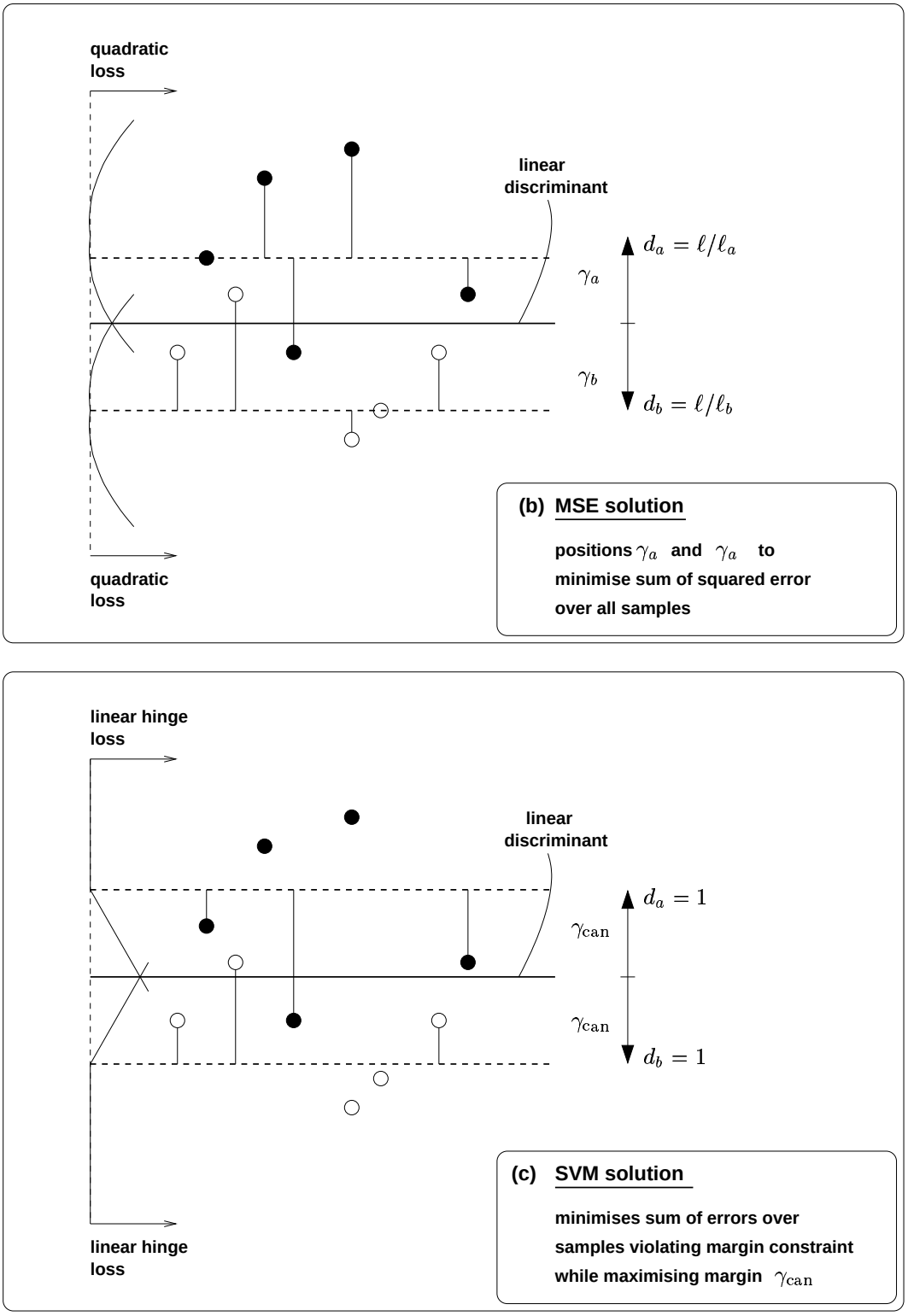


Figure 2.2: Pictorial comparison of MSE and SVM linear discriminants

2.3.2 Kernelisation and nonlinear discriminants

Linear discriminants are poor approximations to optimal classifiers when there is a clear nonlinear separation between the classes. Nonlinear discriminants perform better in these cases. However it is often difficult to implement capacity control to ensure good generalisation. Kernelisation is a technique which extends linear learning machines to yield nonlinear discriminants in input space while retaining some of the regularisation properties inherent in linear discriminants. This section briefly describes kernelisation and references other nonlinear discriminants.

A prerequisite for kernelisation is that the learning and classification algorithms must only access the samples through scalar products. Any learning machine which fails in this condition cannot be kernelised. The learning phase may be a closed form solution as for the MLE linear discriminants based on Identity class-conditional covariance matrices, or may require iterative optimisation such as for the linear SVM. The scalar product may be defined in any feature space. Of importance are feature spaces which are transformations of the input space $L(\boldsymbol{O})$. Defining a linear or nonlinear mapping φ from input space into the new feature space or *image space* $\varphi(\boldsymbol{O})$, then $\varphi : \boldsymbol{O}_l \mapsto \boldsymbol{\varphi}(\boldsymbol{O}_l)$ where $\boldsymbol{O}_l \in L(\boldsymbol{O})$ and $\boldsymbol{\varphi}(\boldsymbol{O}_l) \in \varphi(\boldsymbol{O})$. The image space $\varphi(\boldsymbol{O})$ is often a nonlinear subspace within a linear space. For $\boldsymbol{O}_i, \boldsymbol{O}_j \in L(\boldsymbol{O})$, then the scalar product in the learning algorithm in input space may be replaced by,

$$(\boldsymbol{O}_i, \boldsymbol{O}_j) \quad \mapsto \quad (\boldsymbol{\varphi}(\boldsymbol{O}_i), \boldsymbol{\varphi}(\boldsymbol{O}_j)) \tag{2.59}$$

Application of the learning algorithm learns discriminants in the image space rather than in the input space. If the mapping φ is nonlinear, a linear discriminant learnt in image space maps back to a nonlinear discriminant in input space. Any capacity control in the image space provides regularisation within the constraints of the mapping φ. The scalar product in image space can also be written as a function $k(\cdot, \cdot)$ called a *kernel* with arguments in the input space,

$$k(\boldsymbol{O}_i, \boldsymbol{O}_j) \quad \stackrel{\text{def}}{=} \quad (\boldsymbol{\varphi}(\boldsymbol{O}_i), \boldsymbol{\varphi}(\boldsymbol{O}_j)) \tag{2.60}$$

The kernel embodies both the mapping φ and the metric tensor in image space. Conversely, rather than define a mapping φ and derive the functional form of the kernel, it is also

possible to propose an arbitrary function and accept the mapping φ and metric tensor implicit in the proposal. No computation is then required in the image space or even its explicit definition. However the function must be a *Mercer kernel* [19]. Common kernels include the linear kernel, the Gaussian Radial Basis Function (GRBF) kernel with width w and the polynomial kernel with degree d and offset c. These are respectively,

$$k_{\text{lin}}(\boldsymbol{O}_i, \boldsymbol{O}_j) = \boldsymbol{O}_i^\top \boldsymbol{O}_j \tag{2.61}$$

$$k_{\text{GRBF}}(\boldsymbol{O}_i, \boldsymbol{O}_j) = \exp\{-\frac{1}{2w^2}(\boldsymbol{O}_i - \boldsymbol{O}_j)^\top(\boldsymbol{O}_i - \boldsymbol{O}_j)\} \tag{2.62}$$

$$k_{\text{poly}}(\boldsymbol{O}_i, \boldsymbol{O}_j) = (c + \boldsymbol{O}_i^\top \boldsymbol{O}_j)^d \tag{2.63}$$

If $c = 0$, then the polynomial is homogeneous, else inhomogeneous. Later kernels are described which are defined on statistical models. Examples of learning machines which can be kernelised are the SVM, perceptron and MSE learning machines. Learning machines which cannot be kernelised include those which rely on collecting second order moments in the image space, for example for the estimation of a covariance matrix required by a MLE learning machine operating in image space.

Kernelisation is only one method for learning nonlinear discriminants. An advantage is a principled method to incorporate capacity control through a linear discriminant in image space. Other techniques for nonlinear discriminants in input space include the training of polynomial discriminant functions [25] or splines, k-Nearest Neighbour classifiers [25], Condensed Nearest Neighbour classifiers [46], Parzen windows [25] [34], and appropriate neural networks [6]. Alternatively, more complicated GMM classifiers can be implemented. For example, if the class covariances defining the MLE discriminant are not tied, hyper-quadric discriminants are yielded [25]. In addition, the number of mixture components in the mixture models for each class can be increased yielding nonlinear discriminants. This thesis is restricted to the comparison and application of SVM and GMM classifiers since the former includes a principled approach to capacity control, and the latter permits flexibility in varying complexity.

2.4 Feature selection and extraction

Each linear space is described by a number of components called *features*. Ideally the set of features should perfectly capture characteristics of the underlying signal for the intended task such as representation or classification [100]. For representation, relevant features are those which preserve as much of the semantically meaningful structure in the original signal. For classification, relevant features are those which enable distinctions between classes. Some of the complexity of the classifier can be transferred into the feature extraction process and vice versa, illustrating degeneracy.

Feature space mappings may be proposed either to extract the most relevant features from a given feature space, or simply to reduce the number of features, thereby limiting the effects of the 'curse of dimensionality' for subsequent classifiers. The mappings fall into two general categories. *Feature selection* selects those features which contain the most information according to the task-dependent criterion. *Feature extraction* is identical but with the extra degree of freedom that new features may be formed through linear or nonlinear combination of the original features. Both techniques assume that a low dimensional representation of the underlying signal exists. Feature selection assumes it is linear in the frame of the original feature space, feature extraction assumes it is linear or nonlinear depending on the technique. Otherwise, the techniques are simply principled approaches to reducing the size of feature space. Unfortunately feature selection may select features which are highly correlated but which possess little mutually exclusive information. Feature selection and extraction can be subdivided [58] into *filter techniques* which are preprocessing methods independent of performance in the selected space, and *wrapper techniques* which embed the feature selection or extraction process in a feedback loop monitoring performance. Wrapper techniques are more computationally expensive, are more prone to overfitting without proper regularisation, but often perform better. For computational reasons, this thesis is restricted to filter methods.

Methods of feature selection and extraction which preserve representation include linear techniques such as Principal Component Analysis (PCA) and Common Factor Analysis (CFA) [78]. Nonlinear techniques include the Sammon mapping [100] and the isometric

feature mapping (Isomap) [98] which are based on preserving as closely as possible the geodesics of low dimensional structures, and locally linear embedding [86]. In kernel PCA [71], it is not immediately clear how to ensure the nonlinear mapping inherent in the kernel extracts representative information.

Techniques which emphasise separability and are therefore appropriate as preprocessing for classifiers include feature selection based on divergence-like 'distances' or the relative entropy [100] or Fisher ratios. These are described below. An alternative is selection by correlation coefficients [44]. Common examples of linear feature extraction techniques are Fisher Discriminant Analysis (FDA) [100] and its Q-class extension to Multiple Discriminant Analysis (MDA) [25], and Linear Discriminant Analysis [34] and its heteroscedastic extensions [61] [87] [36]. Examples of nonlinear feature selection and extraction techniques include kernel FDA [70] and kernel LDA [42]. Likelihood scaling [111] is a specialised technique applied to a Q-dimensional space of linear class posteriors (later called an appended zeroth order linear posterior score space). It is also applied in this thesis to increase confusion for MMI training.

2.4.1 Feature selection using Fisher ratios

The experiments in this thesis use Fisher ratios for feature selection. It is instructive to view the relationship of the Fisher ratio to other measures. Assume there are two classes ω_a and ω_b with class-conditional distributions over $L(\boldsymbol{O})$ defined by $p(\boldsymbol{O}|\omega_a)$ and $p(\boldsymbol{O}|\omega_b)$. The separation between distributions can be measured by the KL information [3],

$$\mathrm{KL}(p(\boldsymbol{O}|\omega_a)||p(\boldsymbol{O}|\omega_b)) = \int p(\boldsymbol{O}|\omega_a) \ln \frac{p(\boldsymbol{O}|\omega_a)}{p(\boldsymbol{O}|\omega_b)} \mathrm{d}\boldsymbol{O} \tag{2.64}$$

the measure called the 'divergence' [100] [20],

$$D(p(\boldsymbol{O}|\omega_a)||p(\boldsymbol{O}|\omega_b)) = \mathrm{KL}(p(\boldsymbol{O}|\omega_a)||p(\boldsymbol{O}|\omega_b)) + \mathrm{KL}(p(\boldsymbol{O}|\omega_b)||p(\boldsymbol{O}|\omega_a)) \tag{2.65}$$

or the Bhattacharyya 'distance' [100],

$$B(p(\boldsymbol{O}|\omega_a)||p(\boldsymbol{O}|\omega_b)) = -\ln \int \sqrt{p(\boldsymbol{O}|\omega_a)p(\boldsymbol{O}|\omega_b)} \mathrm{d}\boldsymbol{O} \tag{2.66}$$

The KL information is not symmetric, and none of these measures obey the triangle inequality [100]. Hence none are valid distance measures. When the class conditional distributions are single Gaussians with tied covariance matrices,

$$p_a = p(\boldsymbol{O}|\omega_a) = \mathcal{N}(\boldsymbol{O}; \boldsymbol{\mu}_a, \boldsymbol{\Sigma}) \tag{2.67}$$

$$p_b = p(\boldsymbol{O}|\omega_b) = \mathcal{N}(\boldsymbol{O}; \boldsymbol{\mu}_b, \boldsymbol{\Sigma}) \tag{2.68}$$

then [100] shows,

$$D(p_a||p_b) = 8B(p_a||p_b) = (\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_a - \boldsymbol{\mu}_b) \tag{2.69}$$

and also,

$$D(p_a||p_b) = 2\mathrm{KL}(p_a||p_b) = 2\mathrm{KL}(p_b||p_a) \tag{2.70}$$

The similarity measures only differ in a scaling constant. Assuming the covariance matrix $\boldsymbol{\Sigma}$ is diagonal, they return identical rankings for individual features. The ranking for the ith component of $L(\boldsymbol{O})$ is determined by $F(i)$ where,

$$F(i) = \frac{(\boldsymbol{\mu}_a(i) - \boldsymbol{\mu}_b(i))^2}{\boldsymbol{v}(i)} \tag{2.71}$$

and $\boldsymbol{\mu}_q(i)$ is the ith component for a vector $\boldsymbol{\mu}_q$, and $\boldsymbol{v}(i) = \boldsymbol{\Sigma}(i,i)$. The selection process retains features with highest values of $F(i)$ since these separate the two classes most distinctly. The measure $F(i)$ is sometimes called the Fisher ratio [65]. In this thesis, this feature selection technique is applied to large linear spaces called score spaces since it is a robust method to reduce their size. It is possible to define Bhattacharyya and divergence-based measures on more complicated distributions.

2.5 Applying static classifiers to dynamic data

The nonparametric linear discriminants detailed in Section 2.3.1 classify patterns of fixed length. Extensions to classify patterns of variable length would be useful. Since this is one of the main applications of this thesis, it is worthwhile to highlight some techniques to map variable length patterns to fixed length patterns. The review is biased towards preparing

input for SVM classifiers since the generalisation properties of SVMs have encouraged consideration of these mappings. The review is also biased towards speech applications since it represents a demanding dynamic classification task and forms the basis of the experiments in this thesis. The mappings may often be viewed as front end processing techniques and can sometimes be embedded in the subsequent classifier, for example through kernelisation. There is also discussion of more general techniques to incorporate static classifiers into dynamic classifiers. In this section the sample $\boldsymbol{O} \in L(\boldsymbol{O})$ is a sequence of observations and is sometimes called the input data sequence.

2.5.1 Front end processing techniques

Front end processing techniques map variable length data sequences into those of fixed length. In speech applications, a sequence of observations is often split into contiguous segments, where each segment is an acoustic realisation of a speech unit such as a monophone. The segments are usually of variable length.

- In [64], manually marked monophone units of variable duration were forced into patterns of fixed length by two methods: (a) dividing each segment into a fixed number of contiguous subsegments of equal duration and then averaging the obervations within each subsegment, or (b) linear compaction or elongation of segments by omitting observations from segments which are too long, or repeating some observations for segments which are too short. In their approach the duration of the segment was retained as an extra feature for a subsequent SVM classifier.

- A similar approach was taken in [38] [45] [23]. Each segment of variable length was divided into three contiguous subsegments according to a 3-4-3 ratio. The observations within each subsegment were mean averaged.

- In [5], observations were concatenated to form a variable length input vector. Regarding this sequence of observations as a trajectory, fixed length was enforced by discarding the necessary number of observations at the flattest points on the trajectory.

Averaging, omitting or repeating observations all corrupt the original signal and its information content. Sometimes this is useful, for example averaging removes noise, but at other times useful information may be lost. Front end processing techniques should therefore be selected considering the subsequent application.

- For example, if a data sequence is intended as input for an SVM classifier, then the similarity between this sequence and alternative sequences should provide the basis for SVM training. With this viewpoint, [89] proposed a dynamic time-alignment kernel which uses a dynamic time warping (DTW) algorithm to compare the similarity between two data sequences. The DTW path calculated is that which maximises the similarity measured by a kernel function. Alternatively, [89] noted linear time warping may be applied.

2.5.2 Model-based front end processing techniques

Alternative techniques use statistical models of data sequences to map the sequences into fixed length patterns, where the length is dependent on the number of parameters in the models. These techniques typically use all the data from the sequences.

- One of the most popular schemes for SVMs is the Fisher kernel [52]. In [52], the Fisher kernel was introduced as a similarity measure between two data sequences and was derived from a statistical model for the sequences. For two sequences $\boldsymbol{O}_a, \boldsymbol{O}_b \in L(\boldsymbol{O})$, the Fisher kernel returns the scalar $k(\boldsymbol{O}_a, \boldsymbol{O}_b)$ where,

$$k(\boldsymbol{O}_a, \boldsymbol{O}_b) = \left(\nabla_{\boldsymbol{\xi}} \ln p(\boldsymbol{O}_a; \boldsymbol{\xi})\right)^{\top} \bar{\boldsymbol{F}}^{-1} \nabla_{\boldsymbol{\xi}} \ln p(\boldsymbol{O}_b; \boldsymbol{\xi}) \tag{2.72}$$

  and $p(\boldsymbol{O}; \boldsymbol{\xi})$ is a probability distribution over $L(\boldsymbol{O})$ with parameter vector $\boldsymbol{\xi}$. The covariant derivatives with respect to model parameters are implicitly defined at a fixed model parameterisation, and for shorthand $p(\boldsymbol{O}_l; \boldsymbol{\xi})$ denotes $p(\boldsymbol{O}; \boldsymbol{\xi})|_{\boldsymbol{O}=\boldsymbol{O}_l}$. The matrix $\bar{\boldsymbol{F}}$ is the Fisher information matrix, as defined in Appendix D.1.3. The mapping implicit in the Fisher kernel is extended and investigated in this thesis. In [73], the Fisher kernel was extended to a larger family of natural kernels and the

regularisation properties of such kernels discussed. The natural kernel is defined by simply replacing the Fisher Information matrix with any positive definite matrix $\bar{\boldsymbol{A}}$. The 'plain kernel' is yielded if $\bar{\boldsymbol{A}}$ is the Identity matrix.

The 'tangent vector of posterior log-odds' (TOP) kernel was introduced in [102] using a Taylor expansion. A similar variant of the Fisher kernel was developed simultaneously in [91]. Both highlight the generic nature of the mapping from input space. In [102], the mapping is described as a family of "model-dependent feature extractors". In [51], an algorithm was described which has similarities to the SVM and incorporates the Fisher kernel. However the discriminant objective function differs from the SVM and may be viewed as an attempt to directly model the log class posterior-ratio between competing classes.

Applications of the Fisher kernel include speaker verification or identification [31] [32] [105] [106], speech recognition [91] [94], and multiaspect target recognition [60]. The researches in [52], [102] and [51] were applied in computational biology, for example for protein classification. Typical statistical models for this application are HMMs with discrete output distributions. However speech applications typically require HMMs with continuous density output distributions.

- An alternative model-based front end processing scheme was proposed in [10] in the context of SVMs and speaker recognition. Speaker recognition requires the definition of a scoring function and this was provided through sequence kernels based on general linear discriminants. In [11], the endpoints of sequences were provided by an HMM Viterbi segmentation. Also of relevance to score spaces, [10] noted that all support vectors can be collapsed down to a single vector when the feature space is explicitly calculated and linear kernels employed in feature space.

- String kernels are applied, for example, in protein classification and text categorisation. In [88], string kernels were viewed as Fisher kernels and extended to finite state automata thereby permitting similarity measures between variable length subsequences of a document. In [17], string kernels were viewed as a special case of 'rational kernels', and rational kernels were applied to measure topic similarity between word lattices.

2.5.3 Embedding static classifiers in standard dynamic classifiers

A different approach embeds a static classifier, such as an SVM, within a dynamic classifier.

- A popular approach uses the static classifier to process the input data sequence on an observation-by-observation basis, moderates its output, and relegates the modelling of temporal variation to the dynamic classifier. The outputs of SVMs are typically moderated by converting the functional margins for each observation to class posterior estimates, for example through a sigmoid fitted to the appropriate decision boundary [62] [77]. Moderated outputs fit naturally into the Bayesian framework and permit the application of Bayesian inference for the selection and interpretation of SVM parameters [63] [97] [96]. However since SVMs are trained as classifiers, there is some inconsistency in using them to return class posteriors.

  Moderated outputs were applied in the hybrid SVM/HMM architecture in [38]. A 1-v-rest SVM was trained for each HMM state and a sigmoid then discriminatively trained at the decision boundary thereby defining a state posterior distribution. The system was modified in [45] by replacing SVMs with Relevance Vector Machines (RVMs). Since the RVM directly models the state posterior, its output is probabilistic and more naturally fits into the hybrid framework. Another hybrid system, able to incorporate 1-v-1 classifiers, was applied to speaker verification in [24]. The posterior for a state was the output of a Gaussian distribution fitted to vectors of 1-v-1 class posteriors. Each element in the vector of 1-v-1 class posteriors was obtained by the familiar sigmoidal fit to 1-v-1 decision boundaries. Essentially SVMs were used in feature extraction.

- An alternative approach in [12] is the Forward-Decoding Kernel Machine applied to phone recognition. In this hybrid HMM/SVM system, state posteriors were obtained by a technique called *Gini*SVM which provided a sparse approximation to kernel logistic regression.

- A selection of techniques have been developed combining GMMs and SVMs. For example [31] applied SVMs, the Fisher kernel, the sigmoidal fit, and Error Correcting

Output Coding to a speaker identification task. The system was extended in [33] for speech recognition by using a conventional HMM network to yield an N-best list for each state, and SVMs to select the best candidate from each list. An SVM 'adviser' to a standard GMM classifier was described for speaker identification and verification in [32].

2.6 Summary

This chapter has introduced the optimal Bayes decision rule in the context of statistical models. Unfortunately there is an increase in error rate consequential on incorrect estimates of statistical models. For this reason, discriminants trained by nonparametric techniques are sometimes preferred since they circumvent any explicit representation of class models. Unfortunately, these discriminants often lack the rich probabilistic interpretation available to statistical models. This suggests there is promise in combining statistical models and nonparametric discriminants, for example through kernels defined on statistical models. These kernels, and the mappings implicit in them, are described in the remainder of this thesis. The degeneracy between the feature extraction process and the classifier is a recurring theme.

Chapter 3

Augmenting statistical models

This chapter introduces the concept of viewing a statistical model as a differentiable manifold in the space of probability distributions, or more generally in the space of scalar functions. Estimating distributions in the statistical model is analogous to estimating points on the manifold. First differentiable manifolds are introduced in Section 3.1 and applied to the description of statistical models in Section 3.2. Section 3.3 describes how a scalar field may vary over this statistical manifold and a Taylor expansion used to recover the value of the scalar field at distant points on the manifold. Section 3.4 then describes a structural augmentation as the statistical manifold forms the base of a fibre bundle. Applications are described in Section 3.5. Fibre bundles permit a formalisation of approximations to the Taylor expansion. They also permit distributions to be estimated outside the base manifold which are not in the original statistical model. Various techniques to estimate distributions in fibre bundles are presented. An important property of the base manifold is the metric defined in its tangent space, with repercussions for both calculating the Taylor expansion and estimating distributions. Suitable metrics are discussed. Section 3.6 details how some constraints are relaxed for certain experiments later in the thesis. This chapter assumes some familiarity with tensor algebra (e.g., see Chapter 19 of [81]). The simplicity of the task permits all summations to be written explicitly rather than implied in Einstein's convention.

3.1 Differentiable manifolds

This section introduces some concepts describing differentiable manifolds, namely their nature, the means of describing points upon them, the variation of scalar fields over them, and their tangent spaces. The concepts are later applied to the description of statistical models.

A *differentiable manifold* S is here viewed as a n-dimensional topological manifold which is a Hausdorff space (see Section 108 of [50]). Describing points on the manifold usually requires an atlas of coordinate neighbourhoods which provide an open covering of the manifold. However, in this thesis the manifold S is sufficiently described by a single coordinate neighbourhood (S, ψ) where,

$$\psi : S \quad \rightarrow \quad L(\boldsymbol{\theta}; S) \tag{3.1}$$

The mapping ψ is a homeomorphism called the *coordinate chart* [108], and $L(\boldsymbol{\theta}; S)$ is an open set of $\mathbb{R}^n$ isomorphic to points on S and sometimes called the 'coordinate space'. There is a set of real-valued continuous functions $(\theta^1, \ldots \theta^n)$ which operate on points on S and which are abbreviated to $[\theta^i]$ and called the *global coordinate system* in (S, ψ). The coordinates of a point $\psi(p) \in L(\boldsymbol{\theta}; S)$ are often assembled in column format and called the *coordinate vector* $\boldsymbol{\theta}$ where,

$$\boldsymbol{\theta} = (\theta^1(p), \ldots \theta^n(p))^\top \tag{3.2}$$

and where the dependence of $\boldsymbol{\theta}$ on p is implied. For convenience, this thesis often refers to $\boldsymbol{\theta}$ as the 'coordinate vector of a point p on the manifold S'. The manifolds in this thesis are assumed C^∞ (see Section 1.1 of [3] and Section 108.H of [50]).

Next, a scalar field f varies over the manifold S such that,

$$f : S \rightarrow \mathbb{R} \tag{3.3}$$

It is difficult to analyse a function which operates on points on a manifold but easier to analyse a function which operates on points in $L(\boldsymbol{\theta}; S)$. It is possible to define a scalar field $\bar{f}$,

$$\bar{f} \quad = \quad f \circ \psi^{-1} \tag{3.4}$$

where for a point $p \in S$ with coordinate vector $\boldsymbol{\theta} \in L(\boldsymbol{\theta}; S)$,

$$f(p) \quad = \quad \bar{f}(\theta^1(p), \ldots \theta^n(p)) = \bar{f}(\boldsymbol{\theta}) \tag{3.5}$$

and,

$$f : p \mapsto f(p) \tag{3.6}$$

$$\bar{f} : \boldsymbol{\theta} \mapsto \bar{f}(\boldsymbol{\theta}) \tag{3.7}$$

Then Section 1.1 of [3] defines,

$$\frac{\partial f}{\partial \theta^i} \quad \stackrel{\text{def}}{=} \quad \frac{\partial \bar{f}}{\partial \theta^i} \circ \psi : S \to \mathbb{R} \tag{3.8}$$

so for $\psi(p) = \boldsymbol{\theta}$,

$$\frac{\partial f}{\partial \theta^i}\Big|_p \quad \stackrel{\text{def}}{=} \quad \Big(\frac{\partial \bar{f}}{\partial \theta^i} \circ \psi\Big)\Big|_p = \frac{\partial \bar{f}}{\partial \theta^i}\Big|_{\boldsymbol{\theta}} \tag{3.9}$$

Similar relations follow for higher order partial derivatives and covariant derivatives. For example,

$$f_{;j_1;\ldots;j_r}\Big|_p \quad \stackrel{\text{def}}{=} \quad (\bar{f}_{;j_1;\ldots;j_r} \circ \psi)\Big|_p = \bar{f}_{;j_1;\ldots j_r}\Big|_{\boldsymbol{\theta}} \tag{3.10}$$

Indeed if $\bar{f}$ is C^∞, then f is called a C^∞ function on S. In this thesis, a scalar field f over S is strictly defined as a scalar function whose sole input argument is the location $p \in S$, hence $f : p \mapsto f(p)$. However there is an important set of scalar functions which are dependent both on the location $p \in S$ and on another random variable. Of relevance are scalar functions of the form $\varsigma(\boldsymbol{O}; p)$ where $\boldsymbol{O} \in L(\boldsymbol{O}), p \in S$ and $L(\boldsymbol{O})$ is an open set of input samples as defined later in Equation 3.21. If the random variable $\boldsymbol{O}$ is then fixed at $\boldsymbol{O}_l \in L(\boldsymbol{O})$, then a scalar field ς_l is defined,

$$\varsigma_l \quad : \quad S \to \mathbb{R} \tag{3.11}$$

$$\varsigma_l \quad : \quad p \mapsto \varsigma(\boldsymbol{O}_l; p) \tag{3.12}$$

and for $\boldsymbol{\theta} \in L(\boldsymbol{\theta}; S)$ where $\psi(p) = \boldsymbol{\theta}$,

$$\bar{\varsigma}_l \quad : \quad L(\boldsymbol{\theta}; S) \to \mathbb{R} \tag{3.13}$$

$$\bar{\varsigma}_l \quad : \quad \boldsymbol{\theta} \mapsto \bar{\varsigma}(\boldsymbol{O}_l; \boldsymbol{\theta}) \tag{3.14}$$

The linear space tangential to a differentiable manifold S at a point $p \in S$ is called the *tangent space* $T_p(S)$ and has dimension n. A suitable *basis* for $T_p(S)$ can be formed by assigning the ith element of the basis to a tangent vector parallel to the ith coordinate curve through point p, where a coordinate curve is the curve traced on the manifold by keeping all coordinates but θ^i constant. In particular, a *natural basis* for $T_p(S)$ for the coordinate system $[\theta^i]$ is $e_i, i = \{1, \dots n\}$ where for $\psi : p \mapsto \boldsymbol{\theta}$,

$$e_i \;=\; \frac{\partial}{\partial \theta^i}\Big|_p \tag{3.15}$$

Any scalar field over S can be mapped to a point in tangent space. Of particular interest is the class of all real-valued C^∞ functions on S. As an example, a function in this class ς_l is mapped to a member $\nabla\varsigma_l$ of $T_p(S)$ where,

$$\nabla\varsigma_l\Big|_p \;=\; \sum_{i=1}^{n}\Big(\nabla_{\frac{\partial}{\partial\theta^i}}\varsigma_l\Big)\Big|_p = \sum_{i=1}^{n}(\varsigma_l)_{;i}\Big|_p \tag{3.16}$$

the term $(\nabla_{\frac{\partial}{\partial\theta^i}}\varsigma_l)|_p$ is the *covariant derivative* of ς_l with respect to $\partial/\partial\theta^i$ evaluated at point $p \in S$, abbreviated to $(\varsigma_l)_{;i}|_p$. Numerically, for $\boldsymbol{\theta} = \psi(p)$,

$$(\varsigma_l)_{;i}\Big|_p \;=\; \Big(\frac{\partial\varsigma_l}{\partial\theta^i}\Big)\Big|_p = \Big(\frac{\partial\bar{\varsigma}_l}{\partial\theta^i}\circ\psi\Big)\Big|_p = \frac{\partial\bar{\varsigma}_l}{\partial\theta^i}\Big|_{\boldsymbol{\theta}} \tag{3.17}$$

For convenience the following abbreviation is applied in this thesis for $\boldsymbol{\theta}_0 \in L(\boldsymbol{\theta}; S)$,

$$\bar{\varsigma}(\boldsymbol{O}_l; \boldsymbol{\theta}_0) \;\stackrel{\text{def}}{=}\; \bar{\varsigma}(\boldsymbol{O}_l; \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \bar{\varsigma}_l\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \tag{3.18}$$

3.2 Statistical manifolds

Having introduced differentiable manifolds, this section applies these concepts to a statistical model thereby defining a structure called a statistical manifold. This structure exists in the space of probability distributions or more generally the space of scalar functions. The section proceeds to make various assumptions concerning the statistical manifold to aid tractability in the analysis. Many of the definitions and descriptions in Section 3.2 are summarised without further reference from Chapters 1 and 2 of [3], and the only references explicitly stated are those not from [3].

A statistical model or family of distributions may be written as $S(\boldsymbol{\theta})$ with parameter vector $\boldsymbol{\theta}$,

$$S(\boldsymbol{\theta}) \quad = \quad \{p = p(\boldsymbol{O};\boldsymbol{\theta}) \quad | \quad \boldsymbol{\theta} \in L(\boldsymbol{\theta};S)\} \tag{3.19}$$

The scalar function $p(\boldsymbol{O};\boldsymbol{\theta})$ has unit integral over $L(\boldsymbol{O})$ and is a probability density function, simply called a distribution. Hence,

$$\int p(\boldsymbol{O};\boldsymbol{\theta})\mathrm{d}\boldsymbol{O} \quad = \quad 1 \tag{3.20}$$

The term $L(\boldsymbol{\theta};S)$ describes an open set in $\mathbb{R}^n$ defining valid distributions. The n parameters of the statistical model are assumed linearly independent so the dimension of the coordinate space and size of the parameter vector[1] are both n, i.e. $\dim(L(\boldsymbol{\theta};S)) = \mathrm{size}(\boldsymbol{\theta}) = n$. The variable $\boldsymbol{O}$ is called a *sample* and $\boldsymbol{O} \in L(\boldsymbol{O})$ where,

$$L(\boldsymbol{O}) \quad = \quad \mathrm{supp}(p) \stackrel{\mathrm{def}}{=} \{\boldsymbol{O} \quad | \quad p(\boldsymbol{O}) > 0\} \tag{3.21}$$

and $\mathrm{supp}(p)$ is assumed invariant to the parameterisation of $p \in S(\boldsymbol{\theta})$. This thesis also requires invariance of the support across different statistical models. Hence $L(\boldsymbol{O})$ may be viewed as defined by a typically unknown source.

It is possible to view the statistical model $S(\boldsymbol{\theta})$ as a manifold in the space of probability distributions, or more widely scalar functions. The manifold has a global coordinate system $[\theta^i]$ and coordinate vector[2] $\boldsymbol{\theta}$. If $S(\boldsymbol{\theta})$ is a C^∞ differentiable manifold, then it is called a *statistical manifold*. All parameterisations which are C^∞ diffeomorphic to $[\theta^i]$ are regarded as 'equivalent'.

The most useful properties of a differentiable manifold are those which are invariant to a change in coordinate system (see Section 1.1 of [3]). However this chapter references a manifold through its coordinate system, for example the manifold S expressed in the coordinate system $[\theta^i]$ is written as $S(\boldsymbol{\theta})$ where the implication is that $\boldsymbol{\theta} \in L(\boldsymbol{\theta};S)$. This

[1] This thesis assumes that the parameters of $S(\boldsymbol{\theta})$ are components of a tensor of type $(1,0)$. This does not seriously restrict the analysis since if the parameters of the statistical model include tensors of higher rank, then suitable isomorphisms may be defined mapping them to a type $(1,0)$ tensor.

[2] The term parameter vector may be used when emphasising a distribution in the model and the term coordinate vector when emphasising a point on the manifold.

is simply for clarity and any analysis transfers from one coordinate system to another unless explicitly stated. An example of a property which is not invariant to a change in coordinate system is the Taylor expansion along the manifold, as detailed in Appendix B.2.

Various structural assumptions for $S(\boldsymbol{\theta})$ ease analysis. In most physical processes, the structure of a differentiable manifold is defined relative to an embedding space. However the frame of reference for $S(\boldsymbol{\theta})$ in the space of probability distibutions or scalar functions is not defined. Hence, it is reasonable to endow the manifold with properties and thereby implicitly define such a frame of reference. An analogy in our 3-dimensional world is contorting a flexible 2-dimensional surface till it acquires a shape with desirable properties.

The first structural assumption or augmentation concerns defining the tangent space of the statistical manifold as a metric space, or more usefully as a Hilbert space (see Appendix C). An affine space is then defined on this Hilbert space with an affine frame conistent with the natural basis in tangent space. The metric can then be fully described by a *metric tensor* g with fully covariant components $g_{ij}, i = \{1, \ldots n\}, j = \{1, \ldots n\}$ determined by the relation of the tangent space to the embedding space. For example, assuming the embedding space is Euclidean with its own Identity metric tensor, then the scalar product between the ith and jth elements of the basis for $T_p(S)$, as calculated in the embedding space is,

$$(e_i, e_j) \quad = \quad \left(\frac{\partial}{\partial \theta^i}, \frac{\partial}{\partial \theta^j}\right) = g_{ij} \tag{3.22}$$

The manifold becomes a *Riemannian manifold* $(S(\boldsymbol{\theta}), g)$, and g is known as the *Riemannian metric* or *fundamental tensor* of $S(\boldsymbol{\theta})$. There are an infinite number of possible metrics and hence an infinite number of possible Riemannian manifolds for $S(\boldsymbol{\theta})$.

As the location of the point $p \in S(\boldsymbol{\theta})$ varies, the orientation and magnitude of each element of the natural basis varies. For tractability, the manifold is assumed equipped with an *affine connection* described by n^3 real numbers called *connection coefficients* $\Gamma_{ij}^k, k = \{1, \ldots n\}, i = \{1, \ldots n\}, j = \{1, \ldots n\}$. The coefficients generally vary with location on the manifold and are defined by,

$$\left(\nabla_{e_i} e_j\right)\Big|_p \quad = \quad \sum_{k=1}^{n} (\Gamma_{ij}^k)|_p (e_k)|_p, \quad \forall i, \forall j \tag{3.23}$$

where $\nabla_{e_i} e_j$ is the covariant derivative of e_j with respect to e_i and is evaluated at point p, and $e_i = \partial/\partial\theta^i$. The only constraint on the functions yielding the connection coefficents is that they are C^∞. Hence the affine connection has this degree of freedom. For convenience, an affine connection is sometimes simply called a *connection* and abbreviated to ∇. If the basis is invariant to location on the manifold, then the manifold is *flat* with respect to ∇ and the connection coefficients $\Gamma_{ij}^k = 0, \forall i, j, k$. Then it is always possible to characterise the tangent space with an Identity metric tensor $g_{ij} = \delta_{ij}$, where δ_{ij} is the Kronecker delta. The corresponding coordinate system for the manifold is a *Euclidean coordinate system* with respect to g.

The manifold is also characterised by a *curvature tensor* and a *torsion tensor*. If the curvature tensor is zero, then the manifold is flat with respect to the connection ∇. This is very restrictive. However the torsion tensor is often assumed zero and consequently the connection ∇ is symmetric so $\Gamma_{ij}^k = \Gamma_{ji}^k, \forall i, j, k$. Next, the affine connection ∇ is assumed compatible with the metric g so that it becomes a *metric connection* with respect to g. A connection which is both symmetric and metric is a *Riemannian connection*, also known as a *Levi-Civita connection* with respect to g. For a Riemannian manifold, a unique Riemannian connection exists with connection coefficients,

$$\Gamma_{ij}^k = \frac{1}{2} \sum_{m=1}^{n} g^{km} \Big(\frac{\partial g_{jm}}{\partial \theta^i} + \frac{\partial g_{mi}}{\partial \theta^j} - \frac{\partial g_{ij}}{\partial \theta^m} \Big) \tag{3.24}$$

The numerical values of the components of g and the connection coefficients generally vary across the manifold.

To summarise, a statistical model, under various conditions, may be viewed as a C^∞ differentiable manifold and called a statistical manifold. This manifold is assumed Riemannian with an affine connection. The connection is assumed torsion-free and hence symmetric. It is also assumed compatible with the metric, so the connection becomes a Riemannian connection with respect to this metric. The curvature tensor is only zero as a special case. If the statistical manifold is single-dimensional, then the torsion and curvature tensors are zero and the manifold is flat with respect to the connection.

3.3 Taylor expansions along the manifold

Given a statistical manifold and a scalar field which varies over the manifold, a sensible analysis is the Taylor expansion. This permits the value of the scalar field at a point on the manifold to be recovered from analysing the variations of the scalar field at a possibly distant point on the manifold. The familiar form of the Taylor expansion (e.g. Section 4.7 of [81]) assumes the manifold is flat with respect to its connection and the coordinate system is Euclidean. This section presents the more general expression for a manifold with nonzero curvature. It also shows how, by employing suitable isomorphisms, the Taylor expansion can be reduced to a single bilinear form between two members of a unit rank tensor space. These tensor spaces are later called score spaces.

In terms of the coordinate space,

$$\bar{\varsigma}_l : \boldsymbol{\theta} \mapsto \bar{\varsigma}(\boldsymbol{O}_l; \boldsymbol{\theta}) \tag{3.25}$$

Recovering the value of the scalar field ς_l at point $p' \in S(\boldsymbol{\theta})$ is identical to recovering the value of the scalar field $\bar{\varsigma}_l$ at point $\boldsymbol{\theta'}$ where $\boldsymbol{\theta'} = \psi(p')$, but according to the structure of the manifold. However the only information available is at point $\boldsymbol{\theta}_0$ where in general $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta'}$. The following power series is proposed,

$$\bar{z}(\bar{\varsigma}_l, \boldsymbol{\theta'}, \boldsymbol{\theta}_0) = \bar{\varsigma}_l + \sum_{j_1=1}^{n} (\bar{\varsigma}_l)_{;j_1} (\theta'^{j_1} - \theta_0^{j_1}) + \frac{1}{2} \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} (\bar{\varsigma}_l)_{;j_1;j_2} (\theta'^{j_1} - \theta_0^{j_1})(\theta'^{j_2} - \theta_0^{j_2}) + \ldots \tag{3.26}$$

where the scalar field $\bar{\varsigma}_l$ and all covariant derivatives are evaluated at point $\boldsymbol{\theta}_0$ and are expressed in the coordinate space. More precisely,

$$\bar{z}(\bar{\varsigma}_l, \boldsymbol{\theta'}, \boldsymbol{\theta}_0) = \sum_{r=0}^{\infty} \sum_{j_1 \ldots j_r} T_{j_1 \ldots j_r}(\boldsymbol{O}_l) \alpha^{j_1 \ldots j_r} \tag{3.27}$$

where the summation over $j_1 \ldots j_r$ implies all possible permutations for the r indeces where $j_i = \{1, \ldots n\}, \forall i$, the scalar $T(\boldsymbol{O}_l) = \bar{\varsigma}_l|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ and,

$$T_{j_1 \ldots j_r}(\boldsymbol{O}_l) = \frac{1}{r!} (\bar{\varsigma}_l)_{;j_1;\ldots;j_r} \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \tag{3.28}$$

$$\alpha^{j_1 \ldots j_r} = \prod_{i=1}^{r} (\theta'^{j_i} - \theta_0^{j_i}) \tag{3.29}$$

Expressions for the terms of this series are given in Appendix B.1. When this power series $\bar{z}(\bar{\varsigma}_l, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ coincides with the value of the scalar field $\bar{\varsigma}_l$ evaluated at $\boldsymbol{\theta}'$, then the power series is a *Taylor expansion* of $\bar{\varsigma}_l$ about $\boldsymbol{\theta}_0$ along the manifold. However this requires certain conditions (see Section 25.B of [50]).

- First the scalar field $\bar{\varsigma}_l$ must be holomorphic about $\boldsymbol{\theta}_0$, i.e. the field must be continuous and first order partial derivatives must exist at $\boldsymbol{\theta}_0$.
- The scalar field $\bar{\varsigma}_l$ must be analytic at $\boldsymbol{\theta}_0$, i.e. a convergent power series called the Taylor expansion does exist centred at $\boldsymbol{\theta}_0$ so that the value of the power series and scalar field coincide in a neighbourhood of $\boldsymbol{\theta}_0$.
- The coordinate vector $\boldsymbol{\theta}'$ must lie within the *convergence domain* of $\bar{\varsigma}_l$ about $\boldsymbol{\theta}_0$ (see Section 25.B of [50]). A special case is a scalar field of a single variable. Then there is a single radius of convergence which is uniquely defined through the Cauchy-Hadamard formula (see Section 336.A of [50]), and the convergence domain is known as the *circle of convergence*. There must be at least one singularity at this radius, and the power series is guaranteed to converge in the open set inside this circle and diverge outside it.

The remainder of this section assumes that the power series $\bar{z}(\bar{\varsigma}_l, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ is a Taylor expansion of $\bar{\varsigma}_l$ at $\boldsymbol{\theta}'$ about $\boldsymbol{\theta}_0$. Since an infinite order Taylor expansion is difficult to calculate, the series of terms in the expansion can be truncated to yield an approximation with an error (see Appendix B.1). Denoting the ϱth order approximation by a $(\varrho+1)$ term series $\bar{\varsigma}(\boldsymbol{O}_l; \varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$,

$$\bar{\varsigma}(\boldsymbol{O}_l; \varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0) \;=\; \sum_{r=0}^{\varrho} \bar{\varsigma}^r(\boldsymbol{O}_l; \boldsymbol{\theta}', \boldsymbol{\theta}_0) \tag{3.30}$$

where,

$$\bar{\varsigma}^r(\boldsymbol{O}_l; \boldsymbol{\theta}', \boldsymbol{\theta}_0) \;=\; \sum_{j_1 \ldots j_r} T_{j_1 \ldots j_r}(\boldsymbol{O}_l) \alpha^{j_1 \ldots j_r} \tag{3.31}$$

and $T_{j_1 \ldots j_r}(\boldsymbol{O}_l)$, $\alpha^{j_1 \ldots j_r}$ and the summation notation are as detailed above. As $\varrho \to \infty$, so $\bar{\varsigma}(\boldsymbol{O}_l; \varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0) \to \bar{\varsigma}(\boldsymbol{O}_l; \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}$.

The Taylor expansion may also be written as a sum of bilinear forms where,

$$\bar{\varsigma}^r(\boldsymbol{O}_l;\boldsymbol{\theta}',\boldsymbol{\theta}_0) \;=\; <\alpha^{r(r,0)}(\boldsymbol{\theta}',\boldsymbol{\theta}_0),\varphi^{r(0,r)}(\boldsymbol{O}_l;\bar{\varsigma},\boldsymbol{\theta}_0)> \tag{3.32}$$

and,

$$\alpha^{r(r,0)}(\boldsymbol{\theta}',\boldsymbol{\theta}_0) \;=\; \sum_{j_1\ldots j_r}\alpha^{j_1\ldots j_r}\otimes_{i=1}^{r}e_{j_i} \tag{3.33}$$

$$\varphi^{r(0,r)}(\boldsymbol{O}_l;\bar{\varsigma},\boldsymbol{\theta}_0) \;=\; \sum_{j_1\ldots j_r}T_{j_1\ldots j_r}(\boldsymbol{O}_l)\otimes_{i=1}^{r}e^{j_i} \tag{3.34}$$

Again the summation over $j_1\ldots j_r$ implies all possible permutations. The $(r+1)$th term of the Taylor expansion is therefore a bilinear form between two tensors which are members of tensor spaces $\breve{L}^{r(r,0)}(\boldsymbol{\alpha};\bar{\varsigma},\boldsymbol{\theta}_0)$ and $\breve{L}^{r(0,r)}(\boldsymbol{\alpha};\bar{\varsigma},\boldsymbol{\theta}_0)$ of respective types $(r,0)$ and $(0,r)$ so,

$$\alpha^{r(r,0)}(\boldsymbol{\theta}',\boldsymbol{\theta}_0)\in\breve{L}^{r(r,0)}(\boldsymbol{\alpha};\bar{\varsigma},\boldsymbol{\theta}_0) \tag{3.35}$$

$$\varphi^{r(0,r)}(\boldsymbol{O}_l;\bar{\varsigma},\boldsymbol{\theta}_0)\in\breve{L}^{r(0,r)}(\boldsymbol{\alpha};\bar{\varsigma},\boldsymbol{\theta}_0) \tag{3.36}$$

where,

$$\breve{L}^{r(r,0)}(\boldsymbol{\alpha};\bar{\varsigma},\boldsymbol{\theta}_0) \;=\; \text{span}\{e_{j_1\ldots j_r}=\otimes_{i=1}^{r}e_{j_i},\forall j_1\ldots j_r\} \tag{3.37}$$

$$\breve{L}^{r(0,r)}(\boldsymbol{\alpha};\bar{\varsigma},\boldsymbol{\theta}_0) \;=\; \text{span}\{e^{j_1\ldots j_r}=\otimes_{i=1}^{r}e^{j_i},\forall j_1\ldots j_r\} \tag{3.38}$$

and $j_i=\{1,\ldots n\},\forall i$. The reason for this notation becomes apparent in later sections of this chapter. For the present it is sufficient to know that the breve ($\breve{}$) indicates an unbounded linear space. The tensor spaces are assumed affine spaces defined on Hilbert spaces, so scalar products and norms are permitted (see Appendix C). The two tensor spaces are dual spaces. Both take $\boldsymbol{\theta}_0$ as an argument because their bases are the natural bases defined at point $p_0\in S$ with coordinate vector $\boldsymbol{\theta}_0$, and points in both spaces are given an interpretation in terms of the scalar function $\bar{\varsigma}$. The two spaces have identical dimension. In practice the tensors $\alpha^{r(r,0)}(\boldsymbol{\theta}',\boldsymbol{\theta}_0)$ and $\varphi^{r(0,r)}(\boldsymbol{O}_l;\bar{\varsigma},\boldsymbol{\theta}_0)$ may be bounded to subspaces within these linear spaces.

Next, the Taylor expansion may be written as a single bilinear form by defining isomorphisms from the direct sum of tensor spaces to single unit rank tensor spaces. These unit rank tensor spaces are also assumed Hilbert spaces so scalar products and norms are

defined. The isomorphisms are,

$$\oplus_{r=0}^{\varrho} \breve{L}^{r(r,0)}(\boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0) \cong \widehat{\breve{L}}^{1(1,0)}(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) \tag{3.39}$$

$$\oplus_{r=0}^{\varrho} \breve{L}^{r(0,r)}(\boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0) \cong \widehat{\breve{L}}^{1(0,1)}(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) \tag{3.40}$$

The unit rank spaces are of dimension δ where $\delta = \sum_{r=0}^{\varrho} n^r$. The isomorphisms are defined by the mappings,

$$f_1 : \alpha^{r(r,0)}(\boldsymbol{\theta}', \boldsymbol{\theta}_0), r = \{0, \dots \varrho\} \mapsto \boldsymbol{\alpha}(\varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0) \tag{3.41}$$

$$f_2 : \varphi^{r(0,r)}(\boldsymbol{O}_l; \bar{\varsigma}, \boldsymbol{\theta}_0), r = \{0, \dots \varrho\} \mapsto \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) \tag{3.42}$$

Bold font without and with a bar respectively denote linear algebraic vectors formed from contravariant and covariant components (see Appendix C.2). Summarising,

$$\boldsymbol{\alpha}(\varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0) = (\alpha, \alpha^1, \dots \alpha^n, \alpha^{11}, \dots \alpha^{\overbrace{n \dots n}^{\varrho \text{ repetitions}}})^\top \tag{3.43}$$

$$\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) = (T, T_1, \dots T_n, T_{11}, \dots T_{\underbrace{n \dots n}_{\varrho \text{ repetitions}}})^\top \tag{3.44}$$

where for brevity $T_{j_1 \dots j_r}$ abbreviates $T_{j_1 \dots j_r}(\boldsymbol{O}_l)$. Using this column notation, the ϱth order approximation to the Taylor expansion may be written as,

$$\bar{\varsigma}(\boldsymbol{O}_l; \varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0) = < \boldsymbol{\alpha}(\varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0), \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) > \tag{3.45}$$

The approximation to the Taylor expansion is a bilinear form or scalar product between two tensors in unit rank tensor spaces. As $\varrho \to \infty$, the unit rank tensor spaces have infinite dimension. It is more conventional to express scalar products in terms of two members of the same tensor space. Therefore, letting $\bar{\boldsymbol{A}}$ denote the metric matrix for the $(r, 0)$ tensor space, and $\boldsymbol{A}$ the metric matrix for the $(0, r)$ tensor space, and remarking that $\boldsymbol{A}^{-1} = \bar{\boldsymbol{A}}$,

$$\begin{aligned} \bar{\varsigma}(\boldsymbol{O}_l; \varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0) &= < \boldsymbol{\alpha}(\varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0), \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) > \\ &= \boldsymbol{\alpha}(\varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0)^\top \boldsymbol{A}^{-1} \boldsymbol{\varphi}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) \end{aligned} \tag{3.46}$$

or,

$$\bar{\varsigma}(\boldsymbol{O}_l; \varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0) = \bar{\boldsymbol{\alpha}}(\varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0)^\top \bar{\boldsymbol{A}}^{-1} \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) \tag{3.47}$$

The structures of $\boldsymbol{A}$ and $\bar{\boldsymbol{A}}$ must support the orthogonality of tensor spaces $\breve{L}^{r(0,r)}(\boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$ or $\breve{L}^{r(r,0)}(\boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$ of different degree r. They must also be consistent with the metric matrices $\boldsymbol{G}$ and $\bar{\boldsymbol{G}}$ for the tangent space and its dual respectively (see Section 3.5.4.2 for more

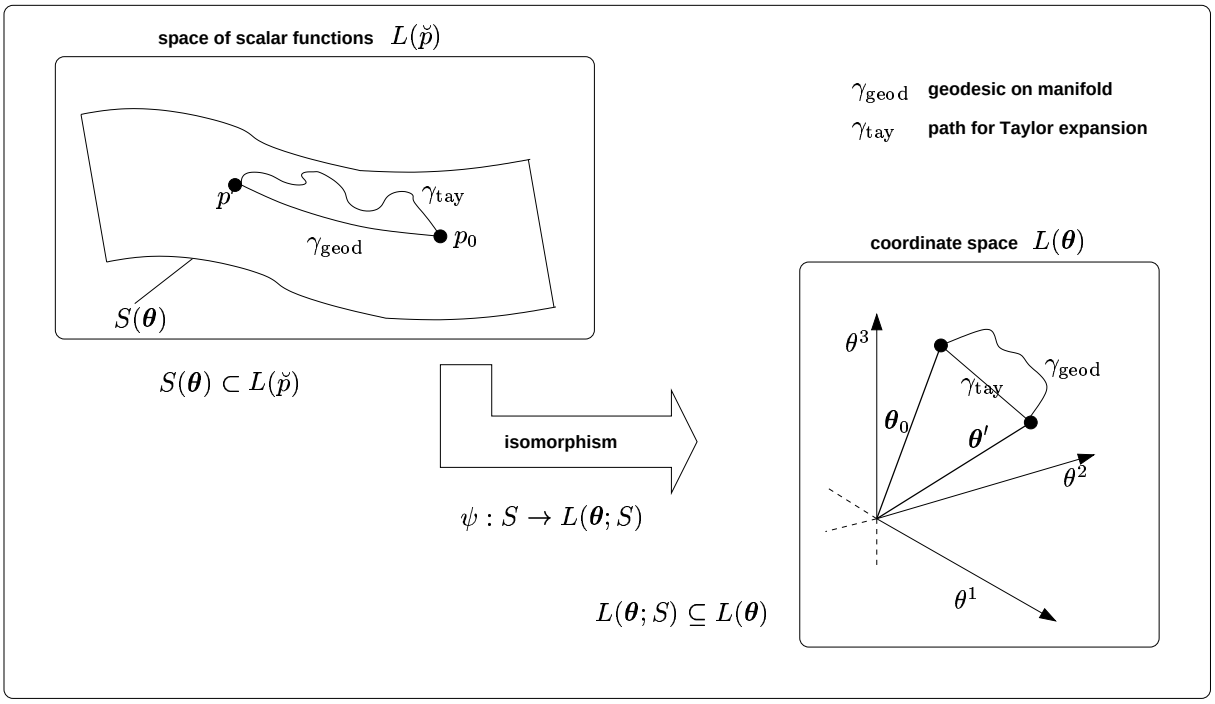


Figure 3.1: Relationship between the coordinate space and space of scalar functions for the manifold $S(\boldsymbol{\theta})$ and coordinate chart ψ

details). The Taylor expansion makes no claims about convergence of $||\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)||$ in the infinite dimensional Hilbert space $\widehat{\breve{L}}^{1(0,1)}(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)$ as $\varrho \to \infty$.

The Taylor expansion is not a property of the manifold alone, but of the manifold and its coordinate system. Generally, the ϱth order approximation to the Taylor expansion, where $\varrho > 1$, differs under different coordinate systems (see Appendix B.2). It is also useful to consider the path on the manifold along which the Taylor expansion is calculated [107]. Figure 3.1 illustrates two points p_0 and p' on the manifold $S(\boldsymbol{\theta})$, with respective coordinate vectors $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}'$ under the coordinate chart ψ. A geodesic between the two points along the manifold follows the path[3] γ_{geod} (for a Riemannian connection, the geodesic is locally the shortest path between two neighbouring points, see Section 360.C of [50]). However the geodesic does not necessarily trace the shortest path in the coordinate space. This shortest path in the coordinate space is that along which the Taylor expansion is calculated. This in turn does not necessarily map onto a geodesic of the manifold.

[3]The path of a geodesic, viewed in the coordinate space, is the solution of a second order ordinary differential equation, see Section 1.8 of [3].

3.4 Fibre bundles

Having proposed a statistical model $S(\boldsymbol{\theta})$, it is standard practice to identify a distribution $p_0 \in S(\boldsymbol{\theta})$ which is a representation of the data source being modelled. If the model is complicated, it is possible to propose a much simpler statistical model which is an exponential family and which locally approximates the original model $S(\boldsymbol{\theta})$ at the distribution p_0. This new statistical model is part of a fibre anchored at p_0. If fibres are defined at all points in $S(\boldsymbol{\theta})$, then the entire structure is called a fibre bundle (see [108] and Section 155 of [50]). A fibre bundle permits a principled approach to approximating the Taylor expansion and can furnish better estimates of the data source.

3.4.1 Describing fibre bundles in the space of scalar functions

First, it is necessary to formalise the concept of a statistical model in the space of distributions, and extend this to the space of scalar functions. First, using the definition in Section 2.1 of [3],

$$L(p) \stackrel{\text{def}}{=} \{p : L(\boldsymbol{O}) \rightarrow \mathbb{R} \quad | \quad p(\boldsymbol{O}) > 0 \quad \forall \boldsymbol{O} \in L(\boldsymbol{O}), \int p(\boldsymbol{O}) \mathrm{d}\boldsymbol{O} = 1\} \tag{3.48}$$

where $L(\boldsymbol{O})$ is as defined in Section 3.2 and $\mathbb{R}$ is the field of real numbers. Extending this definition to the space of scaled distributions $L(\tilde{p})$,

$$L(\tilde{p}) = \{\tilde{p} : L(\boldsymbol{O}) \rightarrow \mathbb{R} \quad | \quad p(\boldsymbol{O}) > 0 \quad \forall \boldsymbol{O} \in L(\boldsymbol{O}), \int \tilde{p}(\boldsymbol{O}) \mathrm{d}\boldsymbol{O} > 0\} \tag{3.49}$$

Finally, the space of all possible scalar functions $L(\breve{p})$ with possibly negative or undefined integral,

$$L(\breve{p}) = \{\breve{p} : L(\boldsymbol{O}) \rightarrow \mathbb{R}\} \tag{3.50}$$

These spaces form a nested structure so $L(p) \subset L(\tilde{p}) \subset L(\breve{p})$.

Next, let a n-dimensional C^∞ differentiable manifold in $L(p)$ represent the statistical model $S(\boldsymbol{\theta})$ as described in Section 3.2. The statistical manifold is,

$$S(\boldsymbol{\theta}) = \{p = p(\boldsymbol{O}; \boldsymbol{\theta}) \quad | \quad \boldsymbol{\theta} \in L(\boldsymbol{\theta}; S)\} \tag{3.51}$$

where the open subset $L(\boldsymbol{\theta}; S)$ ensures p is a valid distribution. The statistical manifold $S(\boldsymbol{\theta})$ exists within $L(p)$, $L(\tilde{p})$ and $L(\breve{p})$. The denormalisation of $S(\boldsymbol{\theta})$ yields the C^∞ differentiable manifold $\tilde{S}(\tau, \boldsymbol{\theta}) \subset L(\tilde{p})$ where,

$$\tilde{S}(\tau, \boldsymbol{\theta}) \quad = \quad \{\tilde{p} = \tilde{p}(\boldsymbol{O}; \tau, \boldsymbol{\theta}) = \tau p(\boldsymbol{O}; \boldsymbol{\theta}) \quad | \quad \tau \in L(\tau; \tilde{S}), \boldsymbol{\theta} \in L(\boldsymbol{\theta}; \tilde{S})\} \tag{3.52}$$

where $L(\tau; \tilde{S})$ is the open set such that $\tau > 0$ and $L(\boldsymbol{\theta}; \tilde{S}) = L(\boldsymbol{\theta}; S)$. The manifold has dimension $(n+1)$ since τ is linearly independent of the coordinate system $[\theta^i]$. Of course, $S(\boldsymbol{\theta})$ may be viewed as a 'slice' of the denormalisation of constant $\tau = 1$, i.e. $S(\boldsymbol{\theta}) = \tilde{S}(\tau, \boldsymbol{\theta})|_{\tau=1} = \tilde{S}(\boldsymbol{\theta}; \tau)|_{\tau=1}$. This 'slice' is a submanifold, where a submanifold is a smooth embedding in a manifold [3]. The denormalisation is then extended to include all realisations of $(\tau, \boldsymbol{\theta})$ including those which do not yield valid scaled distributions. Hence,

$$\breve{S}(\tau, \boldsymbol{\theta}) \quad = \quad \{\breve{p} = \breve{p}(\boldsymbol{O}; \tau, \boldsymbol{\theta}) \quad | \quad \tau \in L(\tau; \breve{S}), \boldsymbol{\theta} \in L(\boldsymbol{\theta}; \breve{S})\} \tag{3.53}$$

where $L(\tau; \breve{S}) = L(\tau) = \mathbb{R}$ and $L(\boldsymbol{\theta}; \breve{S}) = L(\boldsymbol{\theta}) = \mathbb{R}^n$. This is not a differentiable manifold though certain submanifolds may be. The analysis in this chapter only requires the introduction of $S(\boldsymbol{\theta})$ for the original statistical model, though the concept of the denormalisation and its extension are relevant for fibres. Although this explanation assumes the statistical manifold $S(\boldsymbol{\theta})$ is C^∞, it need only be differentiable up to the required order.

Next it is necessary to extend the definition of the scalar field. For $\boldsymbol{O}_l \in L(\boldsymbol{O})$, there is a scalar field ς_l,

$$\varsigma_l : p \quad \mapsto \quad \varsigma(\boldsymbol{O}_l; p) \tag{3.54}$$

which in terms of the coordinate space where $\psi(p) = \boldsymbol{\theta}$ yields,

$$\bar{\varsigma}_l : \boldsymbol{\theta} \quad \mapsto \quad \bar{\varsigma}(\boldsymbol{O}_l; \boldsymbol{\theta}) \tag{3.55}$$

At least in an abstract sense, the definition of the scalar field ς_l may be extended to all points within $L(p)$, and by introducing a dependency on τ, to all points within $L(\breve{p})$.

The fibre bundles of relevance to this thesis are described in Section 4.8 of [3] and called fibre bundles of local exponential families. As a preliminary description, let a statistical model $S(\boldsymbol{\theta})$ exist and let a point $p_0 \in S(\boldsymbol{\theta})$ be selected with coordinate vector $\boldsymbol{\theta}_0$. A

scalar function ς varies over samples and distributions in the statistical model. If the statistical model is very complicated, particularly if it is not an exponential family, its properties at point p_0 may be difficult to analyse. For this reason, Section 4.8 of [3] proposes that another statistical model, defined as a constrained exponential family and which replicates some of the properties of the original model at p_0, may be defined at p_0. In geometric terms, the statistical manifold $S(\boldsymbol{\theta})$ is locally approximated at p_0 by another manifold $S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ with coordinate vector $\boldsymbol{\alpha}$. The interpretation of the argument list is described later, but for the present it is important to note that the manifold is dependent on the scalar function $\bar{\varsigma}$. The term ϱ indicates that the new manifold *osculates* [3] $S(\boldsymbol{\theta})$ at p_0 to the ϱth order, indicating that both manifolds contain p_0, and their tangent spaces and 'higher-order tangent spaces' coincide up to and including the ϱth order. The manifold $S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ is actually a submanifold within a larger structure $\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$ called a *fibre*. Hence there is a fibre anchored at point $p_0 \in S(\boldsymbol{\theta})$. The manifold $S(\boldsymbol{\theta})$ is called the *base manifold* and each point in the base manifold has a distinct fibre. The collection of fibres and base manifold is called a *fibre bundle*. The fibre bundle which exists in the space of scalar functions $L(\breve{p})$ is abbreviated to $\eta_{\breve{p}}$ when its definition is clear from context (fibre bundles are described more generally in Appendix F). The application of different scalar functions yields different fibre bundles. Whether the semantics of points in the fibres and points in the base manifold are comparable depends on the choice of scalar function. By restricting the definition of the fibre bundle to the space of distributions gives what is called in this thesis an *augmented statistical model*, where the augmentation (see Section 4.8.1 of [3]) is relative to the model described by the base manifold.

It is necessary to describe a fibre and its submanifolds in more detail (a summary of the notation used to describe different submanifolds is given for reference in Appendix F.2). The base manifold is $S(\boldsymbol{\theta})$ and a point $p_0 \in S(\boldsymbol{\theta})$ is selected. The fibre at this point is fully described by the extended denormalisation $\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$,

$$\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0) \quad = \quad \{\breve{p} = \breve{p}(\boldsymbol{O}; \tau, \boldsymbol{\alpha}) \quad | \quad \tau \in L(\tau; \breve{S}), \boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; \breve{S})\} \tag{3.56}$$

where $L(\tau; \breve{S}) = L(\tau) = \mathbb{R}$ and $L(\boldsymbol{\alpha}; \breve{S}) = L(\boldsymbol{\alpha}) = \mathbb{R}^{\infty}$. Since each component of $\boldsymbol{\alpha}$ is assumed linearly independent then $\text{size}(\boldsymbol{\alpha}) = \dim(L(\boldsymbol{\alpha}))$. A submanifold of this fibre is

the denormalisation (see Section 2.6 of [3] for the concept of denormalisation),

$$\tilde{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0) \quad = \quad \{\tilde{p} = \tilde{p}(\boldsymbol{O}; \tau, \boldsymbol{\alpha}) = \tau p(\boldsymbol{O}; \boldsymbol{\alpha}) \quad | \quad \tau \in L(\tau; \tilde{S}), \boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; \tilde{S})\} \tag{3.57}$$

where $L(\tau; \tilde{S}) \in L(\tau)$ is the open set fulfilled by $\tau > 0$ and $L(\boldsymbol{\alpha}; \tilde{S}) = L(\boldsymbol{\alpha}; S)$ where $L(\boldsymbol{\alpha}; S)$ ensures valid distributions. A submanifold of the denormalisation is the statistical model defined by constraining $\tau = 1$,

$$S(\boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0) \quad = \quad \{p = p(\boldsymbol{O}; \boldsymbol{\alpha}) \quad | \quad \boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; S)\} \tag{3.58}$$

So,

$$S(\boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0) \quad = \quad \tilde{S}(\boldsymbol{\alpha}; \tau, \bar{\varsigma}, \boldsymbol{\theta}_0)\Big|_{\tau=1} = \tilde{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)\Big|_{\tau=1} \tag{3.59}$$

A nested structure is implied by the submanifolds so that $S(\boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0) \subset \tilde{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0) \subset \breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$. The first is the 'fibre in $L(p)$', the next the 'fibre in $L(\tilde{p})$', and the next the 'fibre in $L(\breve{p})$' or simply the 'fibre'. The fibres have infinite dimension.

These manifolds and submanifolds are constrained to exponential families since the properties of these families are well-known, and sufficient statistics of a fixed size are available facilitating the estimation of parameters of members of these families. For this reason, scalar functions which are log terms are preferred over scalar functions which are linear terms, for example the log likelihood over the linear likelihood. The constraints defining the exponential families have not been explicitly stated and are assumed implicit in the definition of the fibres relevant to this thesis. For example the form of a function $\breve{p}$ within the exponential family defining the fibre is,

$$\breve{p} = \breve{p}(\boldsymbol{O}; \boldsymbol{\alpha}) \quad = \quad \exp\{C(\boldsymbol{O}) + \sum_{r=1}^{\infty} \sum_{j_1 \ldots j_r} F_{j_1 \ldots j_r}(\boldsymbol{O}) \alpha^{j_1 \ldots j_r} - D(\boldsymbol{\alpha}) + \ln \tau\} \tag{3.60}$$

where the terms are as described for distributions in Appendix A, and where the summation over $j_1 \ldots j_r$ is over all possible permutations where $j_i = \{1, \ldots n\}, \forall i$. The following constraints $\mathcal{C}_{\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)}$ are implied, and applied to all submanifolds within the fibre,

$$\mathcal{C}_{\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)} \quad = \quad \left\{ \begin{array}{l} C(\boldsymbol{O}) = \ln p(\boldsymbol{O}; \boldsymbol{\theta}_0) \\ F_{j_1 \ldots j_r}(\boldsymbol{O}) = T_{j_1 \ldots j_r}(\boldsymbol{O}) \quad (\forall j_1 \ldots j_r) \\ D(\boldsymbol{\alpha}) = \ln \int \exp\{C(\boldsymbol{O}) + \sum_{r=1}^{\infty} \sum_{j_1 \ldots j_r} \alpha^{j_1 \ldots j_r} F_{j_1 \ldots j_r}(\boldsymbol{O})\} \mathrm{d}\boldsymbol{O} \end{array} \right\} \tag{3.61}$$

The term $T_{j_1 \cdots j_r}(\boldsymbol{O})$ is as detailed in Equation 3.28. Submanifolds such as the denormalisation and statistical model then simply differ in the constraints on the parameters $\boldsymbol{\alpha}$ and τ. Only if the scalar function is the log likelihood of a sample, i.e. $\varsigma = \ln p(\boldsymbol{O}; p)$ where $\boldsymbol{O} \in L(\boldsymbol{O})$, does a distribution in the fibre submanifold $S(\boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$ have the same semantic meaning as distributions in the base manifold. The fibre has a number of important submanifolds besides the denormalisation and the statistical model.

- A submanifold $\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ exists, all points on which replicate some properties of the base manifold at point p_0 where,

$$\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0) \quad = \quad \{\breve{p} = \breve{p}(\boldsymbol{O}; \tau, \boldsymbol{\alpha}) \quad | \quad \tau \in L(\tau), \boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; \boldsymbol{\theta}', \boldsymbol{\theta}_0)\} \tag{3.62}$$

  where the constraint $\boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; \boldsymbol{\theta}', \boldsymbol{\theta}_0) \subset L(\boldsymbol{\alpha})$ forces each component to be of form,

$$\alpha^{j_1 \cdots j_r} \quad = \quad \prod_{i=1}^{r} (\theta'^{j_i} - \theta_0^{j_i}) \tag{3.63}$$

  and where $\boldsymbol{\theta}', \boldsymbol{\theta}_0 \in L(\boldsymbol{\theta})$. The point p_0 is given when $\tau = 1$, the scalar $\alpha = 1$ and $\alpha^{j_1 \cdots j_r} = 0, r > 0$. The effective dimension of $\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ is n since each point within the submanifold is fully specified by the n components of the cordinate vector $\boldsymbol{\theta}'$, the coordinate vector $\boldsymbol{\theta}_0$ being known.

- Another important submanifold is $\breve{S}(\tau, \boldsymbol{\alpha}; \mathrm{cd}, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0) \subseteq \breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ where $\boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; \mathrm{cd}, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0) \subseteq L(\boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ and where $L(\boldsymbol{\alpha}; \mathrm{cd}, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ is the convergence domain of the function $\bar{\varsigma}$ about point $\boldsymbol{\theta}_0$. When limited to the space of distributions, then $S(\boldsymbol{\alpha}; \mathrm{cd}, \varrho, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ as $\varrho \to \infty$ becomes coincident with $S(\boldsymbol{\theta})$, but not necessarily over the whole surface of $S(\boldsymbol{\theta})$.

- Since the fibre is a family of scalar functions with an infinite number of parameters, this is impractical for analysis. Hence a submanifold is introduced $\breve{S}(\tau, \boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) \subset \breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$ where, for $0 \leq \varrho < \infty$,

$$\breve{S}(\tau, \boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) \quad = \quad \{\breve{p} = \breve{p}(\boldsymbol{O}; \tau, \boldsymbol{\alpha}) \quad | \quad \tau \in L(\tau), \boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; \varrho)\} \tag{3.64}$$

  where $\boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; \varrho)$ implies,

$$\alpha^{j_1 \cdots j_r} = 0, \quad \text{if} \quad r > \varrho \tag{3.65}$$

The dimension of $L(\boldsymbol{\alpha};\varrho)$ and effective size of $\boldsymbol{\alpha}$ are then reduced to δ where,

$$\delta = \sum_{r=0}^{\varrho} n^r \tag{3.66}$$

If $\boldsymbol{\alpha} \in L(\boldsymbol{\alpha};\varrho,\boldsymbol{\theta}',\boldsymbol{\theta}_0)$ where the argument ϱ implies the constraint in Equation 3.65 and the argument $\boldsymbol{\theta}'$ the constraint in Equation 3.63, then the submanifold is denoted by $\breve{S}(\tau,\boldsymbol{\alpha};\varrho,\bar{\varsigma},\boldsymbol{\theta}',\boldsymbol{\theta}_0) \subset \breve{S}(\tau,\boldsymbol{\alpha};\bar{\varsigma},\boldsymbol{\theta}',\boldsymbol{\theta}_0)$ for finite ϱ. The effective size of $\boldsymbol{\alpha}$ is still δ but the effective dimension is the lower of δ or n. Applied to the space of distributions, then all points within $S(\boldsymbol{\alpha};\varrho,\bar{\varsigma},\boldsymbol{\theta}',\boldsymbol{\theta}_0)$ osculate with the base manifold at p_0 up to and including the order ϱ.

Although the description of submanifolds has focussed on those for the fibre in $L(\breve{p})$, submanifolds in $L(\tilde{p})$ and $L(p)$ may be similarly defined with similar notation. Although τ is a variable, only two values are important in this thesis.

- $\tau = 1$: in the case of the denormalisation, a 'slice' of constant τ is $\tilde{S}(\tau,\boldsymbol{\alpha};\varrho,\bar{\varsigma},\boldsymbol{\theta}_0)|_{\tau=1}$ which is abbreviated to $S(\boldsymbol{\alpha};\varrho,\bar{\varsigma},\boldsymbol{\theta}_0)$. All points on this submanifold are distributions.

- $\tau = \tau^{\text{tay}} = \exp\{D(\boldsymbol{\alpha})\}$: the corresponding 'slice' of the extended denormalisation is $\breve{S}(\tau,\boldsymbol{\alpha};\varrho,\bar{\varsigma},\boldsymbol{\theta}_0)|_{\tau=\tau^{\text{tay}}}$. Points on this submanifold do not have unit integral over $L(\boldsymbol{O})$ and there is no need to evaluate the integral. As described later, these points yield ϱth order Taylor expansion approximations to $\bar{\varsigma}$ providing $\boldsymbol{\alpha}$ is additionally within the appropriate convergence domain $L(\boldsymbol{\alpha};\text{cd},\bar{\varsigma},\boldsymbol{\theta}',\boldsymbol{\theta}_0)$.

Since submanifolds are topological spaces, isomorphisms between them are homeomorphisms.

In summary, a base manifold $S(\boldsymbol{\theta})$ has been introduced and, at each point $p_0 \in S(\boldsymbol{\theta})$, a fibre $\breve{S}(\tau,\boldsymbol{\alpha};\bar{\varsigma},\boldsymbol{\theta}_0)$ defined. Each fibre contains a number of important submanifolds. Some of the relations between the submanifolds are illustrated in Figure 3.2. The whole structure is the fibre bundle $\eta_{\breve{p}}$.

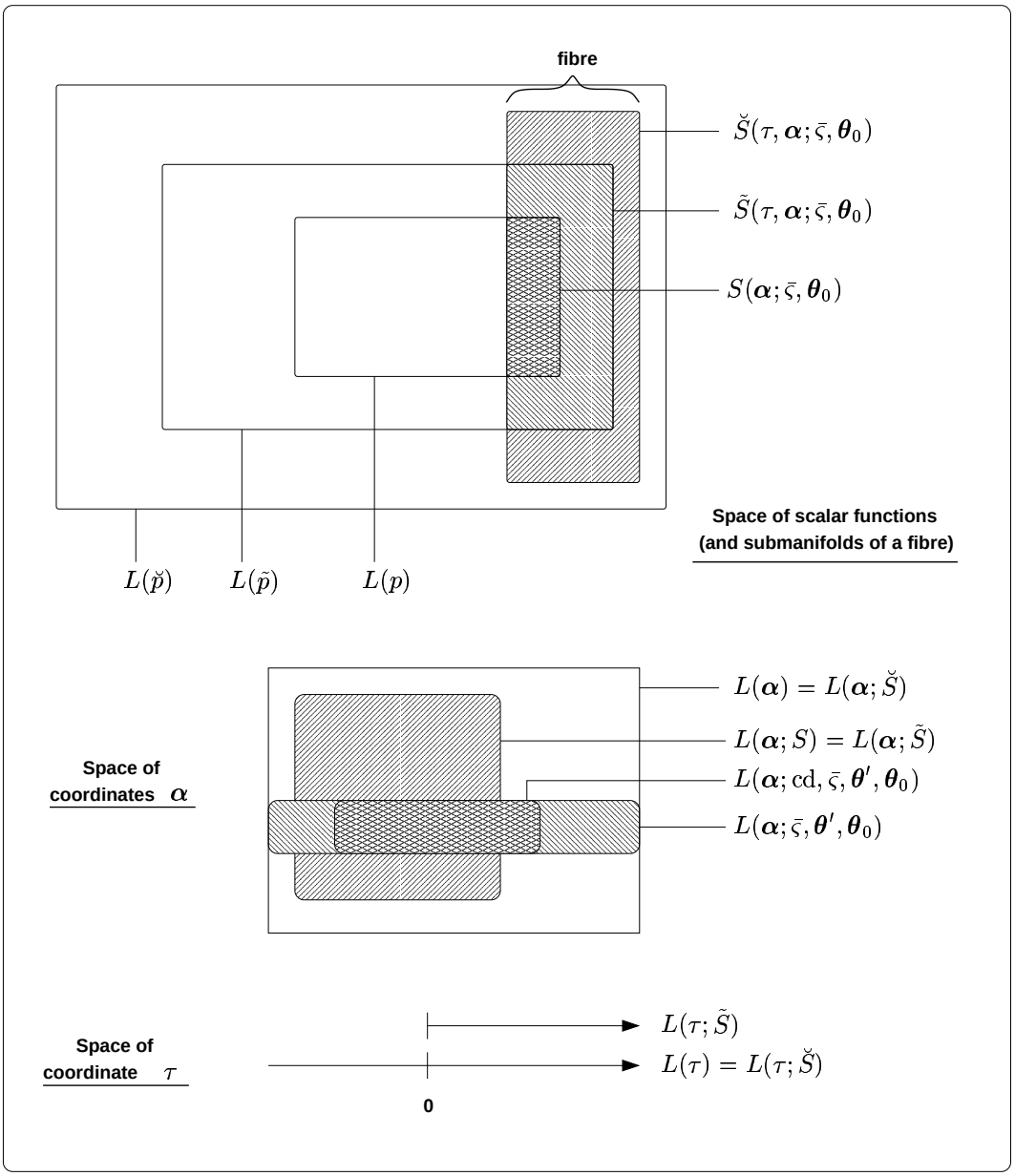


Figure 3.2: Illustrating some of the subspaces and submanifolds relevant to the fibre bundle $\eta_{\breve{p}}$

3.4.2 Introducing vector bundles

The fibre bundle described above is defined as a structure in the space of scalar functions $L(\breve{p})$. It is useful to estimate points in such a fibre bundle and this often requires the application of 'distance-based' learning algorithms in linear spaces. However a fibre within the space of scalar functions is not a linear space. Although metrics such as the Hellinger distance [3] apply along statistical manifolds, it is more convenient and versatile to work with linear spaces. For this reason an isomorphism is defined to map each point within a fibre to a distinct point within a linear space. It is then relatively straightforward to define sensible metrics with invariant properties within these linear spaces. The fibre bundles yielded by the fibre isomorphisms are examples of *vector bundles* (see [108] and Section 155 of [50]) as detailed in Appendix F.1.

Before describing a vector bundle, it is first necessary to define the isomorphism from $\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$ and its submanifolds to the linear space and its subspaces. From Equation 3.60, the parameter or coordinate vector $\boldsymbol{\alpha}$ is a set of fully contravariant tensors of ever-increasing degree. Denoting this set by $\{\alpha^{j_1 \cdots j_r}\}$ where,

$$\{\alpha^{j_1 \cdots j_r}\} = \{\alpha, \alpha^1, \ldots \alpha^n, \alpha^{11} \ldots \alpha^{\overbrace{n \ldots n}^{\infty \text{ repetitions}}}\} \tag{3.67}$$

and the corresponding linear space as $\breve{L}(\{\alpha^{j_1 \cdots j_r}\})$, then it is possible to define an isomorphism,

$$\breve{L}(\{\alpha^{j_1 \cdots j_r}\}) \cong \overset{\wedge}{\breve{L}}{}^{1(1,0)}(\boldsymbol{\alpha}) \tag{3.68}$$

This isomorphisms permits the convenient use of the term $\boldsymbol{\alpha}$ to define the parameters of a scalar function even when the parameters include tensors of degree greater than unity. It also permits the fibre isomorphism between points on $\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$ and the linear space $\overset{\wedge}{\breve{L}}{}^{1(1,0)}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$. Hence,

$$\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0) \cong \overset{\wedge}{\breve{L}}{}^{1(1,0)}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0) \tag{3.69}$$

where,

$$\overset{\wedge}{\breve{L}}{}^{1(1,0)}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0) = \overset{\wedge}{\breve{L}}{}^{1(1,0)}(\tau) \oplus \overset{\wedge}{\breve{L}}{}^{1(1,0)}(\boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0) \tag{3.70}$$

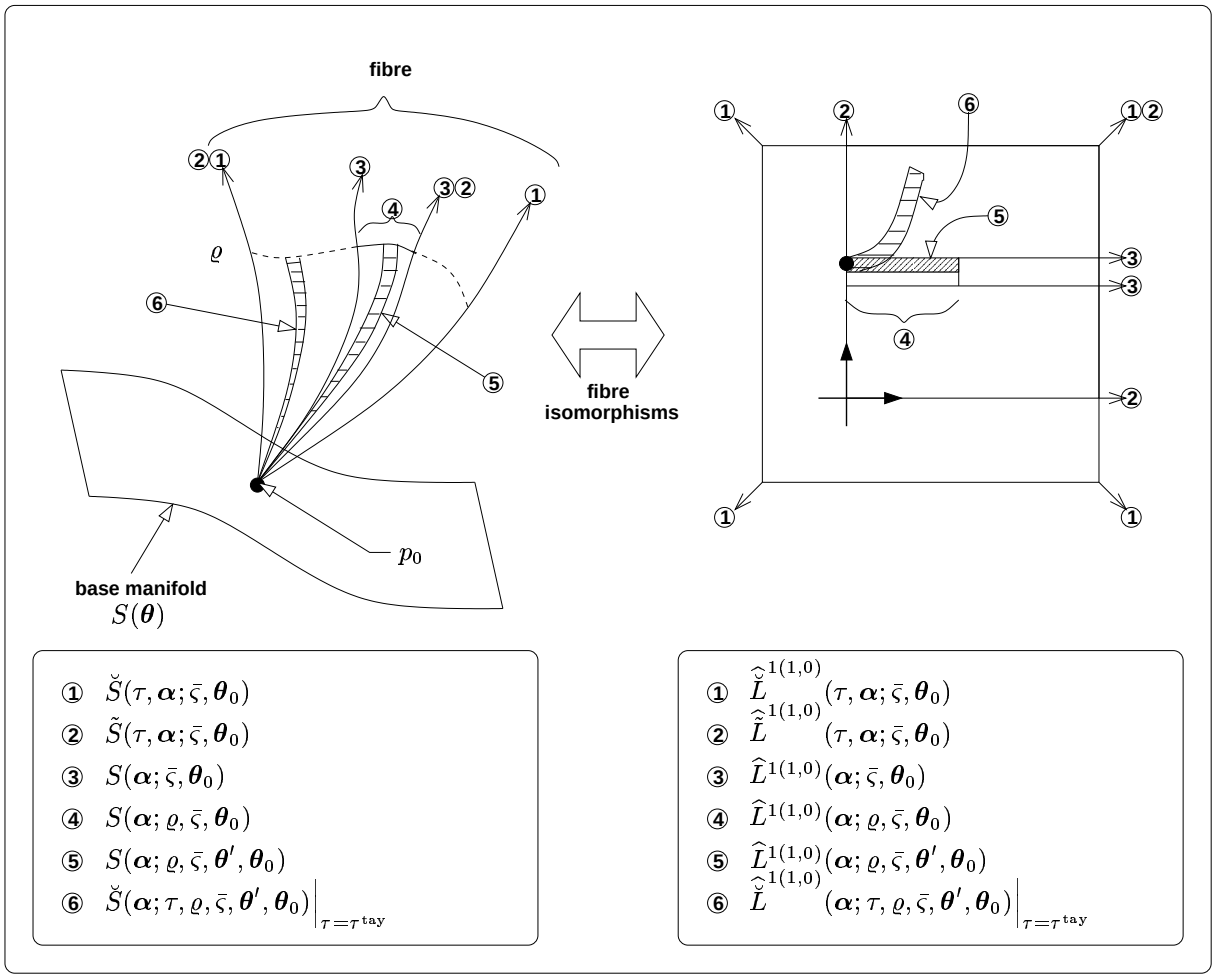


Figure 3.3: Illustrating the isomorphic relations between fibres of $\eta_{\check{p}}$ and η_{vec}

and where $\hat{\check{L}}^{1(1,0)}(\tau)$ is the tensor space of type $(1,0)$ which is isomorphic to $L(\tau)$ and defined by simply regarding the scalar τ as a contravariant component. Of course $\hat{\check{L}}^{1(1,0)}(\tau)$ and $\hat{\check{L}}^{1(1,0)}(\boldsymbol{\alpha};\bar{\varsigma},\boldsymbol{\theta}_0)$ are assumed linearly independent. The structure with base manifold $S(\boldsymbol{\theta})$ and fibres of form $L^{1(1,0)}(\tau,\boldsymbol{\alpha};\bar{\varsigma},\boldsymbol{\theta}_0)$ anchored at each point $p_0 \in S(\boldsymbol{\theta})$ is the vector bundle η_{vec}. The linear spaces necessarily include all possible realisations of τ and $\boldsymbol{\alpha}$ so the vector bundle η_{vec} is isomorphic[4] to $\eta_{\check{p}}$.

Given a single base manifold $S(\boldsymbol{\theta})$, the isomorphic mapping between the fibres of $\eta_{\check{\text{p}}}$ and fibres of η_{vec} permits the adoption of analogous notation for subspaces within $\hat{\check{L}}^{1(1,0)}(\tau,\boldsymbol{\alpha};\bar{\varsigma},\boldsymbol{\theta}_0)$ as for submanifolds within $\check{S}(\tau,\boldsymbol{\alpha};\bar{\varsigma},\boldsymbol{\theta}_0)$. This is briefly described below with illustrations

[4]It is also possible to define another vector bundle isomorphic to η_{vec} and $\eta_{\check{p}}$ which is the Whitney sum (see Section 155.F of [50]) of fibre bundles whose fibres are tensor spaces of ever increasing rank.

in Figure 3.3.

- $\widehat{\breve{L}}^{1(1,0)}(\tau, \boldsymbol{\alpha})$: this is the linear space which forms the fibre.
- $\widehat{\breve{L}}^{1(1,0)}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$: this is the linear space appropriated to the point $p_0 \in S(\boldsymbol{\theta})$, thereby defining the basis for tangent space at p_0 and hence for this linear space. Each point in the linear space is given an interpretation in terms of the scalar function $\bar{\varsigma}$. It is isomorphic to $\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$.
- $\widehat{\tilde{L}}^{1(1,0)}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$: the coordinates are constrained so $\tau \in L(\tau; \tilde{S})$ which is the open set where $\tau > 0$, and $\boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; \tilde{S}) = L(\boldsymbol{\alpha}; S)$, where $L(\boldsymbol{\alpha}; S)$ ensures a valid distribution. This subspace is isomorphic to $\tilde{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$.
- $\widehat{L}^{1(1,0)}(\boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$: the coordinates are constrained so $\boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; S)$ where $L(\boldsymbol{\alpha}; S)$ is as above. This subspace is isomorphic to $S(\boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$.

Elsewhere, the notation follows a similar pattern to that for submanifolds in Section 3.4.1.

3.4.3 Score spaces

This thesis is primarily concerned with estimating coordinates or parameters in the fibres of η_{vec}. An important linear space is the 'slice' of the fibre defined by $\tau = \tau^{\text{tay}}$ and some $\varrho \geq 0$, denoted by $\widehat{\breve{L}}^{1(1,0)}(\boldsymbol{\alpha}; \tau, \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)|_{\tau=\tau^{\text{tay}}}$. Its dual is the linear space $\widehat{\breve{L}}^{1(0,1)}(\boldsymbol{\alpha}; \tau, \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)|_{\tau=\tau^{\text{tay}}}$ which is called a *score space*. Score spaces are a convenient descriptive tool. A mapping exists from input space to score space,

$$\varphi : L(\boldsymbol{O}) \;\; \rightarrow \;\; \widehat{\breve{L}}^{1(0,1)}(\boldsymbol{\alpha}; \tau, \varrho, \boldsymbol{O}, \bar{\varsigma}, \boldsymbol{\theta}_0)\Big|_{\tau=\tau^{\text{tay}}} \subseteq \widehat{\breve{L}}^{1(0,1)}(\boldsymbol{\alpha}; \tau, \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)\Big|_{\tau=\tau^{\text{tay}}} \tag{3.71}$$

where the presence of the argument $\boldsymbol{O}$ simply identifies the subspace of score space which is 'reachable' from the input space. Then for $\boldsymbol{O}_l \in L(\boldsymbol{O})$,

$$\varphi : \boldsymbol{O}_l \;\; \mapsto \;\; \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) \tag{3.72}$$

where $\bar{\varphi}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)$ is a member of score space where in terms of previous notation,

$$\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) \;\; = \;\; (T, T_{11}, \ldots T_{nn}, \ldots T_{\underbrace{n \ldots n}_{\varrho \text{ repetitions}}})^\top \tag{3.73}$$

where for brevity $T_{j_1 \ldots j_r} = T_{j_1 \ldots j_r}(\boldsymbol{O}_l)$. The following terms are used in this thesis,

- *score mapping*: the mapping φ,
- *score space*: the linear space $\widehat{\breve{L}}^{1(0,1)}(\boldsymbol{\alpha}; \tau, \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)|_{\tau=\tau^{\text{tay}}}$,
- *score* for sample $\boldsymbol{O}_l$: the member of score space denoted by $\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)$.

Later in this thesis the score space and score are respectively abbreviated to $\varphi^{\text{sup}}(\varrho, \boldsymbol{\theta}_0)$ and $\bar{\boldsymbol{\varphi}}^{\text{sup}}(\boldsymbol{O}_l; \varrho, \boldsymbol{\theta}_0)$, where 'sup' identifies[5] the scalar function $\bar{\varsigma}$.

3.4.4 Introducing manifolds for multiple statistical models

Before proceeding onto applying fibre bundles to approximate Taylor expansions or estimate distributions, it is necessary to extend the concepts developed above to multiple statistical models, for example those for different classes.

First a space is defined,

$$L(\breve{\mathcal{P}}) = \oplus_{q=1}^{Q} L(\breve{p}_q) \tag{3.74}$$

where $L(\breve{p}_q)$ is the space of scalar functions for class ω_q. Next, a statistical model for class ω_q is defined as $S(\boldsymbol{\theta}_q) \subset L(\breve{p}_q)$ and a point on this manifold is $p_q \in S(\boldsymbol{\theta}_q)$ with coordinate vector $\boldsymbol{\theta}_q$. The entire set of Q statistical models is $\mathcal{S}(\boldsymbol{\xi}) \in L(\breve{\mathcal{P}})$ and a point within this set is $\mathcal{P} \in \mathcal{S}(\boldsymbol{\xi})$ where,

$$\mathcal{S}(\boldsymbol{\xi}) = \oplus_{q=1}^{Q} S(\boldsymbol{\theta}_q) \tag{3.75}$$

$$\mathcal{P} = (p_1, \ldots p_Q) \tag{3.76}$$

$$\boldsymbol{\xi} = (\boldsymbol{\theta}_1^\top, \ldots \boldsymbol{\theta}_Q^\top)^\top \tag{3.77}$$

If $\dim(S(\boldsymbol{\theta}_q)) = n_q$ and $\dim(\mathcal{S}(\boldsymbol{\xi})) = n$ then if the parameters of each statistical model

[5]The superscript 'sup' may be omitted when clear from context or score spaces are referenced in a generic sense.

are linearly independent,

$$n = \sum_{q=1}^{Q} n_q \tag{3.78}$$

This analysis assumes no parameter tying between or within statistical models so all parameters are linearly independent. Then for example,

$$L(\boldsymbol{\xi}; \mathcal{S}) = \oplus_{q=1}^{Q} L(\boldsymbol{\theta}_q; S) \tag{3.79}$$

and similarly for other subspaces. If there is parameter tying, then Equations 3.78 and 3.79 do not hold. Since $\mathcal{S}(\boldsymbol{\xi})$ is a manifold, it is possible to define a tangent space to this manifold as the direct sum of the tangent spaces of the individual statistical manifolds. The dimension of this tangent space is also n. Let a scalar function ς over the space of input samples $L(\boldsymbol{O})$ and the manifold $\mathcal{S}(\boldsymbol{\xi})$ be decomposed into a linear sum of class-specific scalar functions $\varsigma(q)$ over $L(\boldsymbol{O})$ and the individual manifolds $S(\boldsymbol{\theta}_q)$ such that,

$$\varsigma = \sum_{q=1}^{Q} c(q)\varsigma(q) \quad , \quad c(q) \neq \text{fn}(\boldsymbol{O}) \tag{3.80}$$

where $\text{fn}(\cdot)$ is a generic function which varies with its argument. Fixing a sample $\boldsymbol{O}_l \in L(\boldsymbol{O})$ then a scalar field is defined over $S(\boldsymbol{\xi})$ where,

$$\varsigma_l = \sum_{q=1}^{Q} c(q)(\varsigma_l)(q) \tag{3.81}$$

In terms of scalar fields over the coordinate space,

$$\bar{\varsigma}_l = \sum_{q=1}^{Q} c(q)(\bar{\varsigma}_l)(q) \tag{3.82}$$

where $\bar{\varsigma}_l = \bar{\varsigma}(\boldsymbol{O}_l; \boldsymbol{\xi})$, $\bar{\varsigma}(q)_l = (\bar{\varsigma}(q))(\boldsymbol{O}_l; \boldsymbol{\theta}_q)$ and $c(q) \in \mathbb{R}$. It is possible to define a fibre in $L(\breve{\mathcal{P}})$ anchored at point $\mathcal{P}_0 \in \mathcal{S}(\boldsymbol{\xi})$ with coordinate vector $\boldsymbol{\xi}_0$ as,

$$\breve{\mathcal{S}}(\mathcal{T}, \boldsymbol{\beta}; \bar{\varsigma}, \boldsymbol{\xi}_0) = \oplus_{q=1}^{Q} \breve{S}(\tau_q, \boldsymbol{\alpha}_q; c(q)\bar{\varsigma}(q), (\boldsymbol{\theta}_q)_0) \tag{3.83}$$

where,

$$\mathcal{T} = (\tau_1, \ldots \tau_Q) \tag{3.84}$$

$$\boldsymbol{\beta} = (\boldsymbol{\alpha}_1^\top, \ldots \boldsymbol{\alpha}_Q^\top)^\top \tag{3.85}$$

$$\boldsymbol{\xi}_0 = ((\boldsymbol{\theta}_1)_0^\top, \ldots (\boldsymbol{\theta}_Q)_0^\top)^\top \tag{3.86}$$

It is then straightforward to extend the analysis for a single statistical model to that for multiple statistical models. For the special case that each class-conditional statistical model $S(\boldsymbol{\alpha}_q; c(q)\bar{\varsigma}(q), (\boldsymbol{\theta}_q)_0)$ forms a C^∞ differentiable manifold, then the set of statistical models $\mathcal{S}(\boldsymbol{\beta}; \bar{\varsigma}, \boldsymbol{\xi}_0)$ is also a C^∞ differentiable manifold. However if the scalar function $\bar{\varsigma}$ cannot be decomposed in this manner, then the fibre at point $\mathcal{P}_0 \in \mathcal{S}(\boldsymbol{\xi})$ cannot be decomposed. The fibre is then $\check{\mathcal{S}}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\xi}_0)$ and has as coordinates a single value of τ and a single vector $\boldsymbol{\alpha}$. This fibre describes a single family of scalar functions. Since there is no guarantee that the distributions within this family can be related to a set of distributions for each class, this fibre has little obvious application. The analysis would however proceed in a similar manner to that for a statistical model for a single class.

3.5 Applications of fibre bundles

3.5.1 Estimating points on the base manifold

Given a statistical manifold and an unknown data source, it is instructive to consider how different learning criteria use the training samples from the source to estimate different points on the base manifold. These estimates may be subsequently used to define points from which to 'extend fibres'. Adopting the source/model approach and training criteria of Section 2.2.2, each class ω_q has correct and assumed source probability mass functions $P''(\boldsymbol{O}|\omega_q)$ and $R(\boldsymbol{O}|\omega_q)$ respectively. As explained in Section 2.2.2, assuming a probability mass function is viewed as a sampled version of a probability density function, then the mass function may be cast into continuous form as a 'block-like' probability density function. This 'block-like' function converges to the density function being sampled as the level of discretisation becomes infinitessimal. The convergence is also in terms of the KL information. Hence under these assumptions and in this limit of infinitessimal discretisation, the deductions concerning probability mass functions in Section 2.2.2 may be applied to probability density functions. These assumptions and this limit are applied in this section.

Transferring the analysis to the space of distributions, the correct source is represented by the point p''_q and the assumed source by the point r_q, where $p''_q, r_q \in L(p)$. Next $S(\boldsymbol{\theta}_q)$ is proposed as a statistical model of the source. In the special case that the model is of correct functional form so $p''_q \in S(\boldsymbol{\theta}_q)$, then ML estimation does not select the distribution $p'_q \in S(\boldsymbol{\theta}_q)$ which has minimal KL information with the correct source p''_q but the assumed source r_q. However as $\ell_q \to \infty$ where ℓ_q is the number of training samples in class ω_q, Section 2.2.2 shows that, assuming a consistent estimator, $r_q \to p''_q$ so $p'_q \to p''_q$. This justifies the popularity of the ML estimate. However if p''_q lies outside $S(\boldsymbol{\theta}_q)$, i.e. the proposed model is incorrect in functional form, then even with infinite training data, the estimate $p'_q \in S(\boldsymbol{\theta}_q)$ is only at best the 'minimum error' estimate. In this respect the 'error' is measured in terms of the KL information. In practice, finite training data or nonglobal maximisation of the training criterion implies that p'_q is probably only an approximate to the 'minimum error' estimate.

For the multiple class problem, the manifold $\mathcal{S}(\boldsymbol{\xi})$ defines a set of Q class-conditional models. A realisation of this set is $\mathcal{P}' = (p'_1, \ldots p'_Q) \in \mathcal{S}(\boldsymbol{\xi})$. The set of assumed source distributions is $\mathcal{R} = (r_1, \ldots r_Q)$ which in general does not coincide with the set of correct source distributions $\mathcal{P}'' = (p''_1, \ldots p''_Q)$. To compare the ML, MMI and MAP estimation criteria, a pictorial representation of their estimates is given in Figure 3.4 for $Q = 2$. The manifold $\mathcal{S}(\boldsymbol{\xi}) = S(\boldsymbol{\theta}_a) \oplus S(\boldsymbol{\theta}_b)$ is drawn as a plane. Each point on the plane lies on $\mathcal{S}(\boldsymbol{\xi})$, but each point above or below the plane cannot be described by distributions in $\mathcal{S}(\boldsymbol{\xi})$. In the diagram, the KL information is the norm defined on an Identity metric tensor (this assumption is purely for illustration since the KL information is not a valid distance metric). The assumed set of source models $\mathcal{R}$ lies above the plane $\mathcal{S}(\boldsymbol{\xi})$. The ML criterion estimates the point $\mathcal{P}'_{\mathrm{ML}} \in \mathcal{S}(\boldsymbol{\xi})$ which is the perpendicular projection of $\mathcal{R}$ onto $\mathcal{S}(\boldsymbol{\xi})$. In practice, a Bayes decision rule based on $\mathcal{P}'_{\mathrm{ML}}$ may not yield the lowest possible error rate on unseen data. The MMI criterion instead selects the distributions $\mathcal{P}'_{\mathrm{MMI}} \in \mathcal{S}(\boldsymbol{\xi})$ which minimise the sum of KL informations over all classes between the assumed source class posteriors and model class posteriors. The point $\mathcal{P}'_{\mathrm{MMI}}$ is the shortest 'distance' from $\mathcal{S}(\boldsymbol{\xi})$ to $\mathcal{R}$ along a path which denotes the sum of these KL informations. The MAP solution $\mathcal{P}'_{\mathrm{MAP}} \in \mathcal{S}(\boldsymbol{\xi})$ migrates to the ML solution $\mathcal{P}'_{\mathrm{ML}}$ as $\ell_a \to \infty$, $\ell_b \to \infty$. The migration is 'smooth' if the parameter priors provide good regularisation. Of course in practice the

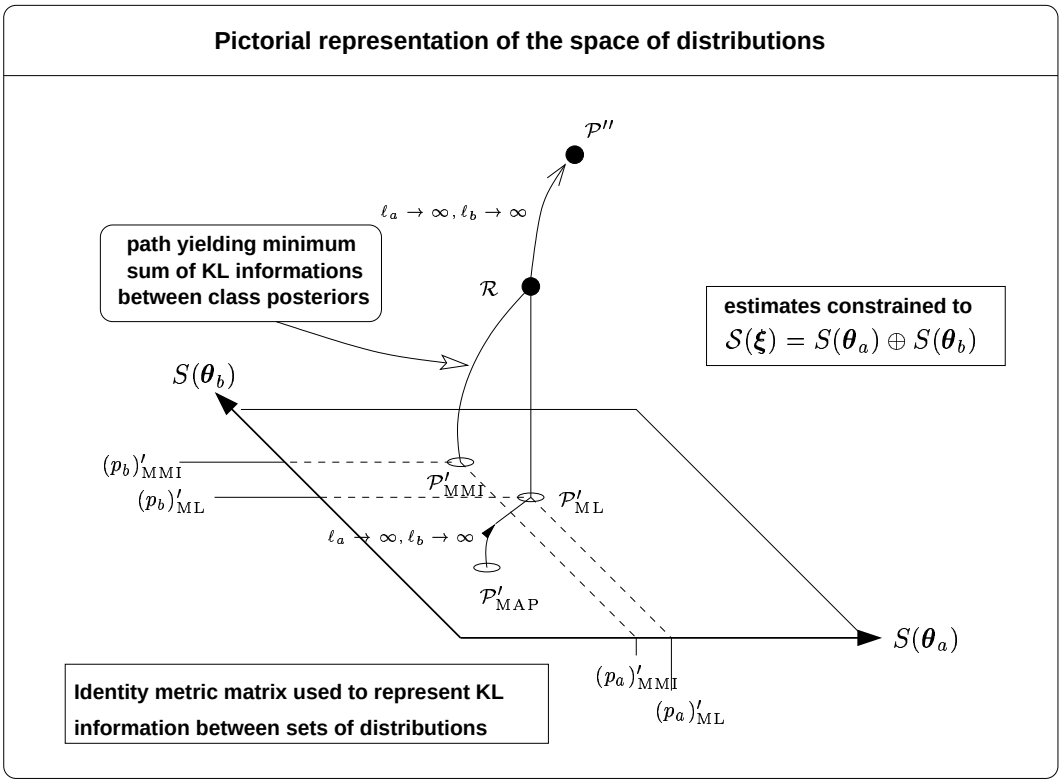


Figure 3.4: Comparing the different estimates on $\mathcal{S}(\boldsymbol{\xi})$ and how they relate to the correct and assumed distributions $\mathcal{P}''$ and $\mathcal{R}$ respectively

assumed sources $\mathcal{R}$ only approximate the correct sources $\mathcal{P}''$. However as $\ell_a \to \infty$ and $\ell_b \to \infty$, the assumed sources $\mathcal{R}$ converge to the correct sources $\mathcal{P}''$ in terms of the KL information. Consequently, the estimates $\mathcal{P}'_{\mathrm{ML}}$ and $\mathcal{P}'_{\mathrm{MMI}}$ both track the 'shadow' of $\mathcal{R}$ across the manifold $\mathcal{S}(\boldsymbol{\xi})$ according to their respective criteria. The MAP estimate $\mathcal{P}'_{\mathrm{MAP}}$ plays 'catch-up' to the ML estimate $\mathcal{P}'_{\mathrm{ML}}$ along the manifold. Hence different estimation criteria may be used to estimate the point on the base manifold $\mathcal{P}_0$ from which to 'extend a fibre'. In the experiments in this thesis, ML and MMI estimation are used.

3.5.2 Approximating Taylor expansions

Having defined fibre bundles $\eta_{\breve{p}}$ and η_{vec}, it is possible to formalise Taylor expansion approximations as the evaluation of scalar fields at points within the total space of the fibre bundle. To recap for the single model case, let ς_l denote a scalar field which varies

over the statistical model $S(\boldsymbol{\theta})$ such that $\varsigma_l : p \mapsto \varsigma(\boldsymbol{O}_l; p)$ where $p \in S(\boldsymbol{\theta})$. The value of the scalar field ς_l at point $p' \in S(\boldsymbol{\theta})$ is required by taking measurements at point $p_0 \in S(\boldsymbol{\theta})$ where $p' \neq p_0$. For the coordinate chart $\psi : p \mapsto \boldsymbol{\theta}$, then from Section 3.3, the ϱth order Taylor expansion approximation is,

$$\bar{\varsigma}(\boldsymbol{O}_l; \varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0) \quad = \quad \sum_{r=0}^{\varrho} \sum_{j_1 \ldots j_r} T_{j_1 \ldots j_r}(\boldsymbol{O}_l) \alpha^{j_1 \ldots j_r} \tag{3.87}$$

where $T_{j_1 \ldots j_r}(\boldsymbol{O}_l)$ is defined in Equation 3.28 and $\alpha^{j_1 \ldots j_r}$ is as defined in Equation 3.29. The value $\bar{\varsigma}(\boldsymbol{O}_l; \varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ is the value of the scalar field ς_l at point $p^\dagger \in \breve{S}(\boldsymbol{\alpha}; \tau, \varrho, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)|_{\tau=\tau^{\text{tay}}}$ but as a function of coordinates $\boldsymbol{\alpha}$. Since the Taylor expansion makes no claims as to whether $\boldsymbol{\alpha}$ permits a distribution, the Taylor expansion is the main motivation for defining fibres within $L(\breve{p})$ rather than within $L(\tilde{p})$. The evaluation of the scalar field at point $p^\dagger$ requires the subtle redefinition of the scalar field over the total space of the fibre bundle rather than simply the base manifold. As ϱ increases, the scalar field is evaluated at different points within the submanifold $\breve{S}(\boldsymbol{\alpha}; \tau, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)|_{\tau=\tau^{\text{tay}}}$. In the limit, and providing $\boldsymbol{\theta}'$ is within the domain of convergence of $\bar{\varsigma}_l$ about $\boldsymbol{\theta}_0$,

$$\bar{\varsigma}(\boldsymbol{O}_l; \varrho, \boldsymbol{\theta}', \boldsymbol{\theta}_0) \quad \stackrel{\varrho \to \infty}{\longrightarrow} \quad \bar{\varsigma}(\boldsymbol{O}_l; \boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} \tag{3.88}$$

If this is true of all scalar fields defined on the same scalar function, i.e. for $\bar{\varsigma}_l, \forall \boldsymbol{O}_l \in L(\boldsymbol{O})$, then,

$$p^\dagger \quad \stackrel{\varrho \to \infty}{\longrightarrow} \quad p' \tag{3.89}$$

In this case, it is also interesting to note that letting $p^\ddagger \in \breve{S}(\boldsymbol{\alpha}; \tau, \varrho, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)|_{\tau=1} = S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ be a point which has identical coordinates $\boldsymbol{\alpha}$ as $p^\dagger$ but a different value of τ, then since p' is a distribution,

$$\tau^{\text{tay}} \quad \stackrel{\varrho \to \infty}{\longrightarrow} \quad 1 \tag{3.90}$$

$$p^\ddagger \quad \stackrel{\varrho \to \infty}{\longrightarrow} \quad p' \tag{3.91}$$

Therefore, as $\varrho \to \infty$, the submanifolds $\breve{S}(\boldsymbol{\alpha}; \tau, \varrho, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)|_{\tau=\tau^{\text{tay}}}$ and $S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ coincide with $S(\boldsymbol{\theta})$ at least within the domain of convergence of $\bar{\varsigma}_l$ about $\boldsymbol{\theta}_0$. It is possible that different scalar fields have different domains of convergence. The description has focussed on submanifolds within the fibre bundle $\eta_{\breve{p}}$. This is because the fibres in η_{vec} do

not possess the same semantic meaning in terms of distributions or points in $L(\breve{p})$ except through an isomorphism.

Hence the truncated Taylor expansion of a scalar field ς_l at p' about the point p_0 where $p', p_0 \in S(\boldsymbol{\theta})$, may be regarded as evaluating the scalar field at a point $p^\dagger$, where $p^\dagger$ does not lie on the manifold $S(\boldsymbol{\theta})$ but within a fibre anchored at point p_0. The point $p^\dagger$ lies within a submanifold of the fibre characterised by the value $\tau = \tau^{\text{tay}}$. As $\varrho \to \infty$, and under the conditions described above, so $p^\dagger \to p'$. A similar analysis is possible for a manifold defined for multiple statistical models.

3.5.3 Estimating a point within the total space of a fibre bundle

Although the fibre bundle $\eta_{\breve{p}}$ provides semantics relating points to scalar functions or distributions, the vector bundle η_{vec} provides a convenient framework for 'distance-based' learning algorithms. The algorithms effectively operate in $\eta_{\breve{p}}$ due to the isomorphism between fibres in $\eta_{\breve{p}}$ and η_{vec}. An example is when a vector bundle η_{vec} is used to estimate distributions outside the base manifold but within the total space of $\eta_{\breve{p}}$. This section considers maximum likelihood and discriminative methods for estimating such distributions. It is important to note that for a fibre bundle defined on $S(\boldsymbol{\theta}_q)$, distributions outside the base manifold only have the same semantic meaning as those on the base manifold if the bundle is defined on the log likelihood scalar function for a sample.

3.5.3.1 Maximum likelihood estimation

A point in the total space of a fibre bundle for a single statistical model may be estimated to better represent a set of samples $\mathcal{O}$ where,

$$\mathcal{O} = \{\boldsymbol{O}_1, \boldsymbol{O}_2, \ldots \boldsymbol{O}_\ell\} \tag{3.92}$$

A suitable scalar field for achieving this is the log likelihood of the samples so $\varsigma_{\mathcal{O}} : p \mapsto \ln p(\mathcal{O}; p)$. A scalar field ς_l may also be defined on a single sample $\varsigma_l : p \mapsto \ln p(\boldsymbol{O}_l; p)$. Both these fields vary over points in the base manifold, and by extension to points in the

total space. Since the present task is restricted to the estimation of distributions, it is only necessary to consider the fibres in $L(p)$. Under the coordinate chart $\psi : p \mapsto \boldsymbol{\theta}$, the scalar fields are $\bar{\varsigma}_{\mathcal{O}} : \boldsymbol{\theta} \mapsto \ln p(\mathcal{O}; \boldsymbol{\theta})$ and $\bar{\varsigma}_l : \boldsymbol{\theta} \mapsto \ln p(\boldsymbol{O}_l; \boldsymbol{\theta})$, where for convenience the bar in the notation for the log likelihood is omitted. First, a point $p_0 \in S(\boldsymbol{\theta})$ with coordinate vector $\boldsymbol{\theta}_0$ is estimated to maximise likelihood,

$$\begin{aligned} \boldsymbol{\theta}_0 &= \underset{\boldsymbol{\theta} \in L(\boldsymbol{\theta}; S)}{\operatorname{argmax}} \quad \ln p(\mathcal{O}; \boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\theta} \in L(\boldsymbol{\theta}; S)}{\operatorname{argmax}} \quad \sum_{l=1}^{\ell} \ln p(\boldsymbol{O}_l; \boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\theta} \in L(\boldsymbol{\theta}; S)}{\operatorname{argmax}} \quad \sum_{l=1}^{\ell} \bar{\varsigma}_l \end{aligned} \tag{3.93}$$

where the samples are assumed i.i.d.. Next it is possible to define a fibre $\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$ at p_0 with submanifold $S(\boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$. The submanifold contains a nested structure of submanifolds $S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)$ indexed by $\varrho \geq 0$. After selecting an appropriate order of expansion ϱ for a submanifold $S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)$, a point $p'(\varrho)$ is estimated with coordinate vector $\boldsymbol{\alpha}'(\varrho)$ where,

$$\boldsymbol{\alpha}'(\varrho) = \underset{\boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; S)}{\operatorname{argmax}} \quad \ln p(\mathcal{O}; \varrho, \boldsymbol{\alpha}, \boldsymbol{\theta}_0) \tag{3.94}$$

The original point p_0 is contained within $S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)$ and corresponds to the coordinate vector $\boldsymbol{\alpha}_0$ where the scalar $\alpha = 1$ and $\alpha^{j_1 \ldots j_r} = 0, r > 0$, i.e. $\boldsymbol{\alpha}_0 = (1, 0, 0, \ldots 0)^\top$. So,

$$\ln p(\mathcal{O}; \varrho, \boldsymbol{\alpha}, \boldsymbol{\theta}_0)\Big|_{\boldsymbol{\alpha} = \boldsymbol{\alpha}'(\varrho)} \geq \ln p(\mathcal{O}; \varrho, \boldsymbol{\alpha}, \boldsymbol{\theta}_0)\Big|_{\boldsymbol{\alpha} = \boldsymbol{\alpha}_0} = \ln p(\mathcal{O}; \boldsymbol{\theta})\Big|_{\boldsymbol{\theta} = \boldsymbol{\theta}_0} \tag{3.95}$$

Hence the estimate $p'(\varrho) \in S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)$ is guaranteed to yield a likelihood for the samples $\mathcal{O}$ not less than the estimate $p_0 \in S(\boldsymbol{\theta})$. By a similar argument, the estimate $p'(\varrho)$ is guaranteed to return nondecreasing likelihoods as ϱ increases. Hence,

$$\ln p(\mathcal{O}; \varrho, \boldsymbol{\alpha}, \boldsymbol{\theta}_0)\Big|_{\substack{\varrho = s \\ \boldsymbol{\alpha} = \boldsymbol{\alpha}'(s)}} \geq \ln p(\mathcal{O}; \varrho, \boldsymbol{\alpha}, \boldsymbol{\theta}_0)\Big|_{\substack{\varrho = t \\ \boldsymbol{\alpha} = \boldsymbol{\alpha}'(t)}}, \quad s > t, \quad t \geq 0 \tag{3.96}$$

So far, the estimation has only been confined to submanifolds of the form $S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)$ rather than the more restrictive form $S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ which enforces osculation constraints including in the zeroth order. While this endows the estimate $p'(\varrho)$ with considerably more degrees of freedom, the osculation constraints are useful and act as a form of regularisation. The constraints need not be enforced if there is a lack of confidence in the estimate p_0. It

is also important to note that if $p'(\varrho) \in S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ then there are only n parameters that require estimation for $\varrho \geq 1$. However as ϱ increases and higher order covariant derivatives are included in the definition of the curved exponential family, the estimates for those n parameters may vary. The points $p'(\varrho) \in S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ necessarily vary with ϱ.

So far, the optimisation required to estimate points $p'(\varrho) \in S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)$ has been ignored and assumed tractable. Using the bilinear form from Section 3.3, optimisation requires,

$$\begin{aligned} \boldsymbol{\alpha}'(\varrho) &= \underset{\boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; S)}{\operatorname{argmax}} \quad \ln p(\mathcal{O}; \varrho, \boldsymbol{\alpha}, \boldsymbol{\theta}_0) \\ &= \underset{\boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; S)}{\operatorname{argmax}} \quad < \boldsymbol{\alpha}, \bar{\boldsymbol{\varphi}}(\mathcal{O}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) > - D(\boldsymbol{\alpha}) \\ &= \underset{\boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; S)}{\operatorname{argmax}} \quad \boldsymbol{\alpha}^\top \bar{\boldsymbol{\varphi}}(\mathcal{O}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) - D(\boldsymbol{\alpha}) \end{aligned} \tag{3.97}$$

where,

$$\bar{\boldsymbol{\varphi}}(\mathcal{O}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) = \sum_{l=1}^{\ell} \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) \tag{3.98}$$

$$D(\boldsymbol{\alpha}) = \ln \int \exp\{\boldsymbol{\alpha}^\top \bar{\boldsymbol{\varphi}}(\mathcal{O}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)\} \mathrm{d}\mathcal{O} \tag{3.99}$$

and $\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)$ is as in Equation 3.73. The maximisation is a constrained optimisation requiring an explicit expression for the normalisation term $D(\boldsymbol{\alpha})$. This increases the analytical complexity of the ML estimate $\boldsymbol{\alpha}'(\varrho)$. However since $\boldsymbol{\alpha}$ and $\bar{\boldsymbol{\varphi}}(\mathcal{O}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)$ are conveniently obtained in fully contravariant and fully covariant form respectively, there is no need to define a metric matrix for ML estimation.

It is possible to extend this estimation technique in an iterative manner as illustrated in Figure 3.5. Once a point $p'(\varrho)$ has been estimated on the submanifold $S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)$ or $S(\boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$, then the submanifold may itself form the base for a new fibre bundle. A fibre can be defined at $p'(\varrho)$ and this new fibre searched to yield a new estimate. This iterative procedure can be continued indefinitely, although with considerable computational expense. The procedure guarantees distributions which are nondecreasing in terms of the likelihood of the training samples. This procedure is called 'fibre hopping'.

- Fibre hopping may be compared to a decision tree search, where each level of the tree is a new submanifold and the number of branch points at each level corresponds

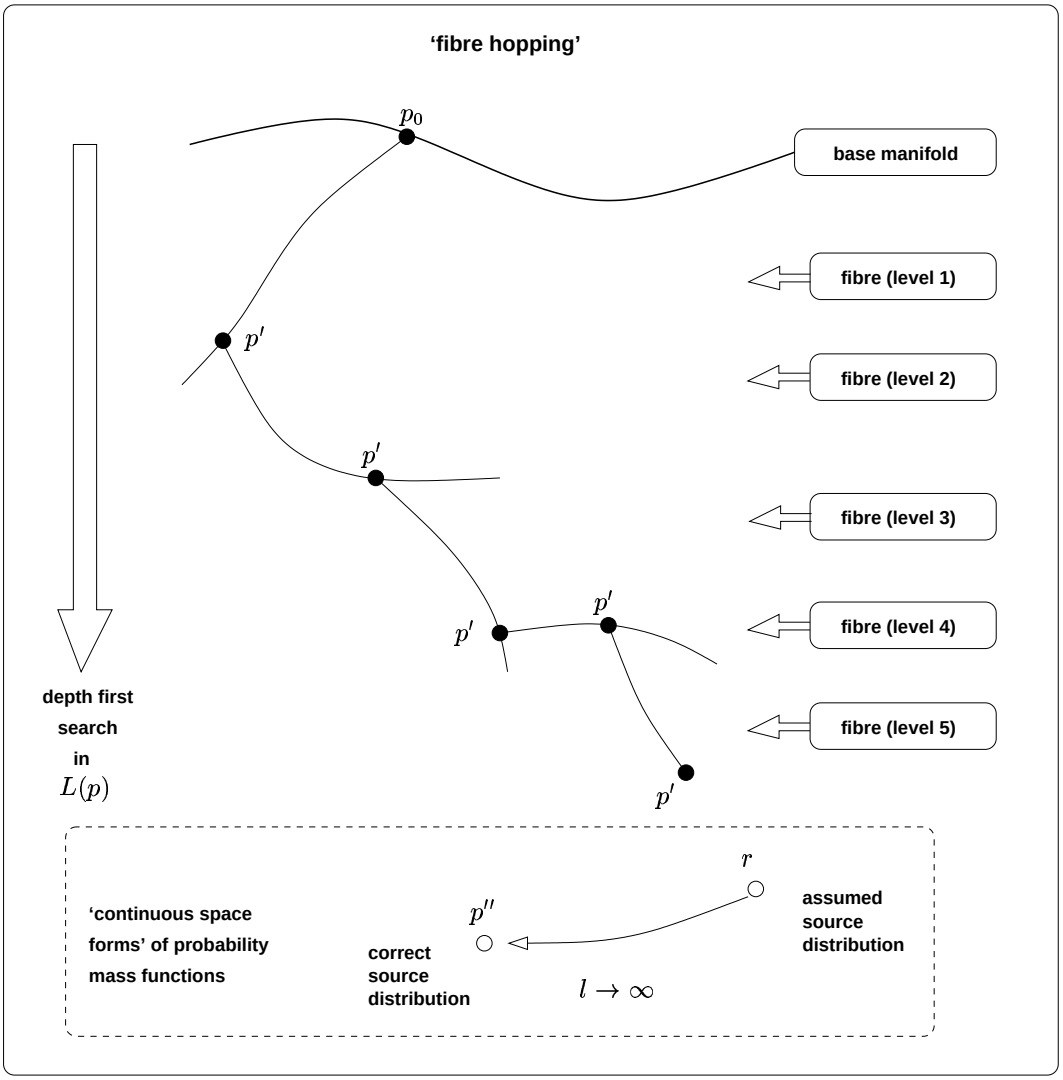


Figure 3.5: Estimating distributions p' by 'fibre hopping'

to the infinite number of points in the submanifold. Fibre hopping is then a depth-first search which makes a decision at each level of the tree to locally maximise the likelihood of the training samples. Alternative training criteria may be introduced at selected levels of the tree. While this destroys the trend of nondecreasing likelihoods, it may be useful to force the estimates away from local likelihood maxima.

- The training samples are regarded as produced by an unknown source p''. As detailed in Section 2.2.2, the only information concerning this source is that available from the training samples furnishing an assumed source r, which, as described in Section 3.5.1, is a 'continuous space' form of the probability mass function R. As illustrated in Figure 3.5, as the number of samples $\ell \to \infty$, so $r \to p''$. Hence for finite ℓ, the assumed source r may be distant from the correct source p''. With each iteration of fibre hopping, i.e. with each level of the decision tree, the ML estimates p' converge to the assumed source r in the sense of KL information. The estimates become overtrained to the assumed source r rather than to the correct source p''. For this reason, the depth of the decision tree should be in proportion to the number of training samples available.

3.5.3.2 Discriminative estimation by maximising mutual information

Discriminative estimation techniques can also be employed to estimate points within the total space of a fibre bundle. Discriminative techniques require the definition of more than one class. For MMI estimation, the fibre bundle has base manifold $\mathcal{S}(\boldsymbol{\xi})$ as defined in Section 3.4.4. The scalar field is $\bar{\varsigma}_{\mathcal{O}} = \ln P(\omega(\mathcal{O})|\mathcal{O};\boldsymbol{\xi}_0)$, where the training data is,

$$\mathcal{O} = (\boldsymbol{O}_1, \boldsymbol{O}_2, \ldots \boldsymbol{O}_\ell) \tag{3.100}$$

$$\omega(\mathcal{O}) = (\omega(\boldsymbol{O}_1), \omega(\boldsymbol{O}_2), \ldots \omega(\boldsymbol{O}_\ell)) \tag{3.101}$$

and where $\omega(\boldsymbol{O}) \in L(\omega) = \{\omega_1, \ldots \omega_Q\}$ is the correct class for the sample $\boldsymbol{O} \in L(\boldsymbol{O})$. The sample/label pairs are assumed i.i.d.. The definition of the scalar field is extended to the total space of the fibre bundle. A fibre is defined at a point $\mathcal{P}_0$ on the base manifold with coordinate vector $\boldsymbol{\xi}_0$. A point in this fibre may be estimated using MMI estimation. However since the scalar field $\bar{\varsigma}_{\mathcal{O}}$ cannot be decomposed into a linear sum of

class-conditional scalar fields as in Equation 3.81, the analysis must resort to a submanifold of form $\check{\mathcal{S}}(\tau, \boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0)$. Furthermore, since the class posterior is not a probability density function, there is no motivation to normalise the scalar function to yield unit integral over $L(\boldsymbol{O})$. A point $\mathcal{P}'$ may be estimated on the submanifold $\check{\mathcal{S}}(\boldsymbol{\alpha}; \tau, \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0)|_{\tau=\tau^{\text{tay}}}$ with coordinate vector $\boldsymbol{\alpha}'(\varrho)$ where,

$$\boldsymbol{\alpha}'(\varrho) \quad = \quad \underset{\boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; \check{\mathcal{S}})}{\operatorname{argmax}} \sum_{q=1}^{Q} \sum_{\substack{l=1 \\ \omega(\boldsymbol{O}_l) = \omega_q}}^{\ell} \ln P(\omega_q | \boldsymbol{O}_l; \varrho, \boldsymbol{\alpha}, \boldsymbol{\xi}_0) \tag{3.102}$$

and similarly to ML estimation in Section 3.5.3.1,

$$\ln P(\omega_q | \boldsymbol{O}_l; \varrho, \boldsymbol{\alpha}, \boldsymbol{\xi}_0) \quad = \quad < \boldsymbol{\alpha}, \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0) >= \boldsymbol{\alpha}^\top \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0) \tag{3.103}$$

where $\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0)$ is defined similarly. Hence $\boldsymbol{\alpha}'(\varrho)$ corresponds to a point which maximises the sum of correct log class posteriors for each sample, though without the constraint that the class posteriors are bounded by zero and unity. It is not straightforward to ensure the estimates correspond to a valid set of distributions.

3.5.3.3 Estimation by training a linear discriminant

Next, a discriminative estimation technique is presented based on calculating a linear discriminant in a fibre of η_{vec}. This permits the inclusion of learning algorithms such as the SVM into the training process. The training data is $\mathcal{O}$ and $\omega(\mathcal{O})$ as defined in Section 3.5.3.2. It is straightforward to relate the estimation technique to training distributions if the scalar function ς is a weighted linear sum of class-conditional functions $\varsigma(q)$ as in Equation 3.80, and the scalar field ς_l for a sample $\boldsymbol{O}_l \in L(\boldsymbol{O})$ is a weighted linear sum of functions $(\varsigma_l)(q)$ according to Equation 3.81. Otherwise, as for MMI estimation, it is difficult to relate the linear discriminant to a set of distributions with the same semantic meaning as those in the base manifold. Expressing the functions in terms of the coordinate space, the definitions of $\bar{\varsigma}$ and $\bar{\varsigma}(q)$ are extended to the total space of the bundle $\eta_{\check{\mathcal{P}}}$ with base manifold $\mathcal{S}(\boldsymbol{\xi})$.

The fibre bundle η_{vec} exists with a fibre anchored at $\mathcal{P}_0 \in \mathcal{S}(\boldsymbol{\xi})$. Following the notation in

Section 3.4.4, an important subspace of this fibre, for $\varrho \geq 0$, is,

$$\widehat{\breve{L}}^{1(1,0)}(\boldsymbol{\beta}; \mathcal{T}, \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0)\Big|_{\{\tau_q=\tau_q^{\text{tay}}, \forall q\}} = \oplus_{q=1}^{Q} \widehat{\breve{L}}^{1(1,0)}(\boldsymbol{\alpha}_q; \tau_q, \varrho, c(q)\bar{\varsigma}(q), (\boldsymbol{\theta}_q)_0)\Big|_{\tau_q=\tau_q^{\text{tay}}} \tag{3.104}$$

It is possible to map each training sample $\boldsymbol{O}_l \in \mathcal{O}$ into the dual to this fibre subspace. Hence the mapping φ is defined where,

$$\varphi : \boldsymbol{O}_l \mapsto \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0) \in \widehat{\breve{L}}^{1(0,1)}(\boldsymbol{\beta}; \mathcal{T}, \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0)\Big|_{\{\tau_q=\tau_q^{\text{tay}}, \forall q\}} \tag{3.105}$$

As described in Section 3.4.3, this dual subspace is called the score space. Since the samples are drawn from two classes, it is sensible to calculate a separating linear discriminant $(\boldsymbol{w}, b)$ in the score space. To relate the linear discriminant $(\boldsymbol{w}, b)$ to distributions for the Q classes, it is necessary to employ an isomorphism from the fibre subspace in η_{vec} to the corresponding fibre submanifold in $\eta_{\breve{\mathcal{P}}}$ for $\tau_q = \tau_q^{\text{tay}}, \forall q$, and thence to employ an isomorphism from this submanifold to the submanifold defined by $\tau_q = 1, \forall q$.

For the first isomorphism from the fibre subspace in η_{vec} to the fibre submanifold in $\eta_{\breve{\mathcal{P}}}$, it is necessary to consider the mapping from $\boldsymbol{w}$ to corresponding tensors,

$$\boldsymbol{w} \mapsto (w, w^1, \ldots w^n, w^{11}, \ldots w^{\overbrace{n \ldots n}^{\varrho \text{ repetitions}}}) \tag{3.106}$$

and for an unlabelled sample $\boldsymbol{O}_l \in L(\boldsymbol{O})$,

$$\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0) \mapsto (T, T_1, \ldots T_n, T_{11}, \ldots T_{\underbrace{n \ldots n}_{\varrho \text{ repetitions}}}) \tag{3.107}$$

where for brevity $T_{j_1 \ldots j_r} = T_{j_1 \ldots j_r}(\boldsymbol{O}_l)$ as defined in Equation 3.28 and the scalar field $\bar{\varsigma}_l$ is evaluated at point $\mathcal{P}_0 \in \mathcal{S}(\boldsymbol{\xi})$. Analogous to the decomposition of the scalar function, each tensor can be decomposed as follows,

$$T_{j_1 \ldots j_r}(\boldsymbol{O}_l) = \sum_{q=1}^{Q} c(q)(T_q)_{j_1 \ldots j_r}(\boldsymbol{O}_l) \tag{3.108}$$

where $(T_q)_{j_1 \ldots j_r}(\boldsymbol{O}_l)$ is the scaled covariant derivative defined on the scalar field $\bar{\varsigma}(q)_l$ at point $(p_q)_0 \in S(\boldsymbol{\theta}_q)$.

It is then possible to view the linear discriminant in terms of a ϱth order Taylor expansion approximation,

$$b = wT(\boldsymbol{O}_l) + \sum_{j_1=1}^{n} w^{j_1} T_{j_1}(\boldsymbol{O}_l) + \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} w^{j_1 j_2} T_{j_1 j_2}(\boldsymbol{O}_l) + \ldots$$

$$= \sum_{r=0}^{\varrho} \sum_{j_1 \ldots j_r} w^{j_1 \ldots j_r} T_{j_1 \ldots j_r}(\boldsymbol{O}_l) \tag{3.109}$$

where the summation over $j_1 \ldots j_r$ implies all possible permutations for $j_i = \{1, \ldots n\}, \forall i$. Division by the scalar w followed by decomposition yields,

$$\begin{aligned} \frac{b}{w} &= T(\boldsymbol{O}_l) + \sum_{r=1}^{\varrho} \sum_{j_1 \ldots j_r} \frac{w^{j_1 \ldots j_r}}{w} T_{j_1 \ldots j_r}(\boldsymbol{O}_l) \\ &= \sum_{q=1}^{Q} c(q) \sum_{r=0}^{\varrho} \sum_{j_1 \ldots j_r} \frac{w^{j_1 \ldots j_r}}{w} (T_q)_{j_1 \ldots j_r}(\boldsymbol{O}_l) \end{aligned} \tag{3.110}$$

Many of the tensors $(T_q)_{j_1 \ldots j_r}(\boldsymbol{O}_l)$ are zero due to the independence assumption between the parameters of statistical models. Next isomorphisms between $\boldsymbol{w}$ and the set of parameters for each statistical model is required. For each $q \in \{1, \ldots Q\}$,

$$(w, w^l, \ldots w^m, w^{ll}, \ldots w^{mm}, \ldots w^{\overbrace{m \ldots m}^{\varrho \text{ repetitions}}}) = (\alpha_q, \alpha_q^1, \ldots \alpha_q^{n_q}, \alpha_q^{11}, \ldots \alpha_q^{n_q n_q}, \ldots \alpha_q^{\overbrace{n_q \ldots n_q}^{\varrho \text{ repetitions}}}) \tag{3.111}$$

where,

$$l = \sum_{i=1}^{q-1} n_i + 1 \tag{3.112}$$

$$m = \sum_{i=1}^{q} n_i \tag{3.113}$$

Hence all 'cross terms' $w^{j_1 \ldots j_r}$, where all contravariant indeces do not refer to the same statistical model, can be ignored and should be forced to zero in training the linear discriminant $(\boldsymbol{w}, b)$. Then the linear discriminant may be written as follows, where the summation implies all permutations of $j_1 \ldots j_r$ but for $j_i = \{1, \ldots n_q\}, \forall i$,

$$\begin{aligned} \frac{b}{w} &= \sum_{q=1}^{Q} c(q) \sum_{r=0}^{\varrho} \sum_{j_1 \ldots j_r} \frac{(\alpha_q)^{j_1 \ldots j_r}}{\alpha_q} (T_q)_{j_1 \ldots j_r}(\boldsymbol{O}_l) \\ &= \sum_{q=1}^{Q} c(q)(\bar{\varsigma}(q))(\boldsymbol{O}_l; \tau_q, \varrho, (\boldsymbol{\theta}_q)_0)\Big|_{\tau_q = \tau_q^{\text{tay}}} \end{aligned} \tag{3.114}$$

However the points identified are not distributions. To remedy this the normalisation factors must be substituted into the expression so that,

$$\frac{b}{w} = \sum_{q=1}^{Q} \Big\{ c(q)(\bar{\varsigma}(q))(\boldsymbol{O}_l; \tau_q, \varrho, (\boldsymbol{\theta}_q)_0)|_{\tau_q = 1} + D(\boldsymbol{\alpha}_q) \Big\}$$

$$\begin{aligned}\frac{b}{w} - \sum_{q=1}^{Q} c(q)D(\boldsymbol{\alpha}_q) &= \sum_{q=1}^{Q} c(q)(\bar{\varsigma}(q))(\boldsymbol{O}_l; \tau_q, \varrho, (\boldsymbol{\theta}_q)_0)\Big|_{\tau_q=1} \\ &= \bar{\varsigma}(\boldsymbol{O}_l; \varrho, \boldsymbol{\xi}_0) \end{aligned} \tag{3.115}$$

The right hand side of the equation is the value of the scalar field $\bar{\varsigma}_l$ evaluated at a point $\mathcal{P}' \in \mathcal{S}(\boldsymbol{\beta}; \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0)$ which potentially lies outside the manifold $\mathcal{S}(\boldsymbol{\xi})$. This corresponds to a set of Q class-conditional distributions which potentially lies outside the corresponding statistical manifolds $S(\boldsymbol{\theta}_q), \forall q$. The left hand side of the equation is a new threshold b' where,

$$b' = \frac{b}{w} - \sum_{q=1}^{Q} c(q)D(\boldsymbol{\alpha}_q) \tag{3.116}$$

The linear discriminant therefore reduces to a threshold decision on the output of the scalar field $\bar{\varsigma}_l$ evaluated at $\mathcal{P}'$. Training the linear discriminant effectively trains the point $\mathcal{P}'$ and the threshold b'. Of course this assumes that each parameter $\boldsymbol{\alpha}_q$ gives rise to a valid distribution, i.e. $\boldsymbol{\alpha}_q \in L(\boldsymbol{\alpha}_q; S), \forall q$. The distributions are discriminatively trained. Learning machines which yield different solutions in $(\boldsymbol{w}, b)$ imply different estimates for the distributions $\mathcal{P}'$ and threshold b'.

It is important to understand the constraints and limitations of learning distributions through training linear discriminants.

- A mechanism is required to constrain the optimisation of $\boldsymbol{w}$ so that each $\boldsymbol{\alpha}_q \in L(\boldsymbol{\alpha}_q; S)$. Without this mechanism there is no guarantee that a solution $\boldsymbol{w}$ corresponds to a set of valid class distributions.

- It may be helpful to restrict each distribution p'_q to the corresponding submanifold $S(\boldsymbol{\alpha}_q; \varrho, c(q)\bar{\varsigma}(q), (\boldsymbol{\theta}_q)', (\boldsymbol{\theta}_q)_0)$ and hence enforce osculation to the ϱth order. At present, only osculation at the zeroth order is enforced since $\alpha_q/\alpha_q = 1, \forall q$.

- The optimisation is within the linear space $\overset{\smile}{\widehat{L}}^{1(0,1)}(\boldsymbol{\beta}; \mathcal{T}, \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0)|_{\{\tau_q=\tau_q^{\text{tay}}, \forall q\}}$. This is isomorphic, but via a nonlinear mapping, to the linear space $\widehat{L}^{1(0,1)}(\boldsymbol{\beta}; \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0) = \overset{\smile}{\widehat{L}}^{1(0,1)}(\boldsymbol{\beta}; \mathcal{T}, \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0)|_{\{\tau_q=1, \forall q\}}$. The estimation technique is more easily implemented to minimise errors in the first rather than the second linear space. Since the second

space is related to submanifolds of distributions, it is intuitively more desirable to optimise in this linear space. However this requires the calculation of normalisation terms either explicitly or within the optimisation process and this adds extra complexity.

- If the scalar function ς does not decompose in the manner described in Equation 3.80 with $\varsigma(q) = \ln p(\boldsymbol{O}; p)$, or if any of the multipliers $c(q)$ are functions of $\boldsymbol{O}$, then it is difficult to relate a linear discriminant to a set of class-conditional distributions. An example of such a scalar function is the log posterior for the correct class. In such cases, it is better to view $\mathcal{S}(\boldsymbol{\xi})$ as a single manifold rather than a direct sum of class-conditional manifolds. Even then if $\bar{\varsigma}_q$ is not the log likelihood function, then the set of distributions does not have the same semantic meaning as a point in the base manifold $\mathcal{S}(\boldsymbol{\xi})$.

- The linear discriminant effectively trains a single threshold classifier in the output of a scalar field. It is a binary decision. More complicated decision rules require, for example, methods to map binary decisions to multicategory decisions.

An example of a learning machine for calculating linear discriminants is the SVM with linear kernel. An SVM trains the discriminant through calculating scalar products between members of its input space. In the current context, scalar products are required in $\widehat{\widetilde{L}}^{1(0,1)}(\boldsymbol{\beta}; \mathcal{T}, \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0)|_{\{\tau_q = \tau_q^{\text{tay}}, \forall q\}}$. For two members of this space $\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0), l = i, j$, their scalar product is,

$$(\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_i; \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0), \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_j; \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0)) \quad = \quad \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_i; \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0)^{\top} \boldsymbol{A} \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_j; \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0) \qquad (3.117)$$

The matrix $\boldsymbol{A}$ is formed from fully contravariant components and is the metric matrix for score space. The solution $(\boldsymbol{w}, b)$ returned by the SVM ordinarily varies with the choice of the metric matrix, as does any learning algorithm which derives its solution through calculating distances between members of score space. The choice of metric is discussed next.

3.5.4 The choice of metric

The statistical models in this chapter are assumed Riemannian manifolds with a metric in tangent space. Tangent spaces are also assumed Hilbert spaces. An affine space is defined in tangent space with an affine frame defined on the natural basis. The elements of the natural basis are not in general orthonormal or even orthogonal, and their orientation and magnitude relative to one another generally vary across the manifold. The metric permits the calculation of distances in tangent or higher rank tensor spaces defined on the tangent space and its dual. The metric is fully specified by a metric tensor g. An example of the application of the metric tensor is the calculation of scalar products in tangent space. Assuming the statistical manifold has a Riemannian connection, the connection coefficients are uniquely determined by the components of the metric tensor and, as detailed in Appendix B.1, are required to evaluate the Taylor expansion along the manifold. Metric tensors are discussed in more detail in Appendix D. This section proposes some metric tensors suitable for tangent spaces. For brevity, this section only considers the tangent space to statistical manifolds. A similar analysis is possible for its denormalisation and those submanifolds in its extended denormalisation which are differentiable.

3.5.4.1 Metric tensors for tangent space

Log likelihood scalar field over $S(\boldsymbol{\theta}_q)$

For $\boldsymbol{O}_l \in L(\boldsymbol{O})$, a scalar field is defined as $\varsigma(q)_l : p_q \mapsto \ln p(\boldsymbol{O}_l; p_q)$ where $p_q \in S(\boldsymbol{\theta}_q)$. This corresponds through the coordinate chart $\psi_q : p_q \mapsto \boldsymbol{\theta}_q$ to the scalar field $\bar{\varsigma}(q)_l : \boldsymbol{\theta}_q \mapsto \ln p(\boldsymbol{O}_l; \boldsymbol{\theta}_q)$. The definition of $\varsigma(q)_l$ is extended to the total space of the fibre bundle with base manifold $S(\boldsymbol{\theta}_q)$, and $\bar{\varsigma}(q)_l$ similarly. When mapped into the tangent space, the scalar field over the coordinate space yields the gradient $\nabla\bar{\varsigma}(q)_l$ where,

$$\nabla\bar{\varsigma}(q)_l \;=\; \sum_{i=1}^{n_q} \frac{\partial}{\partial(\theta_q)^i} \ln p(\boldsymbol{O}_l; \boldsymbol{\theta}_q)\Big|_{\boldsymbol{\theta}_q=(\boldsymbol{\theta}_q)_0} \tag{3.118}$$

A suitable metric for this tangent space is,

$$g_{ij}(\boldsymbol{\theta}_q) \;=\; \int \frac{\partial}{\partial(\theta_q)^i} \ln p(\boldsymbol{O}; \boldsymbol{\theta}_q) \frac{\partial}{\partial(\theta_q)^j} \ln p(\boldsymbol{O}; \boldsymbol{\theta}_q) p(\boldsymbol{O}) \mathrm{d}\boldsymbol{O} \tag{3.119}$$

where the covariant derivatives are evaluated at point $(\boldsymbol{\theta}_q)_0$. This metric tensor is a second order moment in tangent space. If $p(\boldsymbol{O}) = p(\boldsymbol{O};\boldsymbol{\theta}_q)|_{\boldsymbol{\theta}_q=(\boldsymbol{\theta}_q)_0}$, then the first order moment is additionally zero and the metric tensor is also a second order central moment. It is then called the Fisher metric (see Section 2.2 of [3]). The Fisher metric is maximally noncommittal with respect to the relative importance of each component of tangent space to the squared norm of $\nabla\bar{\varsigma}(q)_l$ as $\boldsymbol{O}_l$ varies, but assuming a diagonal Fisher metric (see Appendix D.1 for more details). Of course the maximally noncommittal notion is only in terms of the distribution of samples $p(\boldsymbol{O};\boldsymbol{\theta}_q)$. The Fisher metric is also invariant to expressing distributions over their sufficient statistics (see Appendix D.1.3). Unfortunately, when $p(\boldsymbol{O}) \neq p(\boldsymbol{O};\boldsymbol{\theta}_q)|_{\boldsymbol{\theta}_q=(\boldsymbol{\theta}_q)_0}$ and the first order moment is nonzero, any distances calculated with this metric tensor are more sensitive to those components with large nonzero first order moments. An alternative metric tensor which counteracts this bias and is maximally noncommittal but with respect to the zero-mean gradient in score space $(\nabla\bar{\varsigma}(q)_l - \bar{\boldsymbol{\mu}})$ even when $p(\boldsymbol{O}) \neq p(\boldsymbol{O};\boldsymbol{\theta}_q)|_{\boldsymbol{\theta}_q=(\boldsymbol{\theta}_q)_0}$ is,

$$g_{ij}(\boldsymbol{\theta}_q) = \int \Big(\frac{\partial}{\partial(\theta_q)^i}\ln p(\boldsymbol{O};\boldsymbol{\theta}_q) - \mu_i\Big)\Big(\frac{\partial}{\partial(\theta_q)^j}\ln p(\boldsymbol{O};\boldsymbol{\theta}_q) - \mu_j\Big)p(\boldsymbol{O})\mathrm{d}\boldsymbol{O} \quad (3.120)$$

where,

$$\mu_i = \int \frac{\partial}{\partial(\theta_q)^i}\ln p(\boldsymbol{O};\boldsymbol{\theta}_q)p(\boldsymbol{O})\mathrm{d}\boldsymbol{O} \quad (3.121)$$

all covariant derivatives are evaluated at $(\boldsymbol{\theta}_q)_0$, and $\bar{\boldsymbol{\mu}} = (\mu_1, \ldots \mu_{n_q})^\top$. This metric is however only invariant to sufficient statistics for various distributions and sufficient statistics and no longer in the general case. The fully covariant components $g_{ij}(\boldsymbol{\theta}_q)$ may be assembled into a matrix written as $\bar{\boldsymbol{G}}(\boldsymbol{\theta}_q)$. The fully contravariant components $g^{ij}(\boldsymbol{\theta}_q)$ are defined such that when assembled into a matrix $\boldsymbol{G}(\boldsymbol{\theta}_q)$, then $\bar{\boldsymbol{G}}(\boldsymbol{\theta}_q) = [\boldsymbol{G}(\boldsymbol{\theta}_q)]^{-1}$. Where there is no ambiguity as to the manifold, then the metric matrices may be written as $\bar{\boldsymbol{G}}$ and $\boldsymbol{G}$.

Decomposable scalar field over $\mathcal{S}(\boldsymbol{\xi})$

A decomposable scalar function exists,

$$\bar{\varsigma}(\boldsymbol{O};\boldsymbol{\xi}) = \sum_{q=1}^{Q} c(q)\ln p(\boldsymbol{O};\boldsymbol{\theta}_q) \quad , c(q) \neq \mathrm{fn}(\boldsymbol{O}) \quad (3.122)$$

where fn$(\cdot)$ is a generic function which varies with its argument. An example of such a decomposable scalar function is the log likelihood ratio between two competing classes. The function is a field which varies over $L(\boldsymbol{O}) \times L(\boldsymbol{\xi}; \mathcal{S})$. For a fixed $\boldsymbol{O}_l \in L(\boldsymbol{O})$, a scalar field $\bar{\varsigma}_l : \boldsymbol{\xi} \mapsto \bar{\varsigma}(\boldsymbol{O}_l; \boldsymbol{\xi})$ varies over $L(\boldsymbol{\xi}; \mathcal{S})$. The definitions of both fields are extended to coordinate space corresponding to the total space of the fibre bundle. The tangent space at a point $\mathcal{P}_0 \in \mathcal{S}(\boldsymbol{\xi})$ with coordinate vector $\boldsymbol{\xi}_0$ is,

$$T_{\mathcal{P}_0}(\mathcal{S}) \quad = \quad \text{span}\Big\{ \frac{\partial}{\xi^1}, \ldots \frac{\partial}{\partial \xi^n} \Big\}\Big|_{\boldsymbol{\xi}=\boldsymbol{\xi}_0} \tag{3.123}$$

where $n = \sum_{q=1}^{Q} n_q$ is the dimension of $\mathcal{S}(\boldsymbol{\xi})$ and hence of the tangent space. The scalar field $\bar{\varsigma}_l$ can be decomposed into individual fields over individual statistical models. The tangent space is therefore the direct sum of Q class-conditional tangent spaces of form $T_{(p_q)_0}(S)$ which are linearly independent and where $\mathcal{P}_0 = ((p_1)_0, \ldots (p_Q)_0)$. The metric tensor for the tangent space to $\mathcal{S}(\boldsymbol{\xi})$ can therefore be defined from the metric tensors of the tangent spaces of each individual statistical model. In component form,

$$g_{ml}(\boldsymbol{\xi}) \quad = \quad \begin{cases} c(q)^2 g_{ij}(\boldsymbol{\theta}_q), & \left\{ \begin{array}{l} i = m - \sum_{k=1}^{q-1} n_k \\ j = l - \sum_{k=1}^{q-1} n_k \\ \sum_{k=1}^{q-1} n_k < m \leq \sum_{k=1}^{q} n_k \\ \sum_{k=1}^{q-1} n_k < l \leq \sum_{k=1}^{q} n_k \end{array} \right\}, \quad q = \{1, \ldots Q\} \\ 0, \quad \text{otherwise} \end{cases} \tag{3.124}$$

The metric tensor is a block-diagonal structure and implies the absense of 'cross terms' in the Taylor expansion. This is reasonable. For example if the metric tensor in tangent space is the global covariance of members of tangent space, then it is reasonable to assume that the covariant derivatives relative to the parameters of different statistical models are decorrelated, thereby implying the absense of 'cross terms'. This decorrelation assumption is not implied by the independence assumption between the parameters of different statistical models.

The fully covariant components of this metric tensor may be assembled into a matrix

$\bar{\boldsymbol{G}}(\boldsymbol{\xi})$,

$$\bar{\boldsymbol{G}}(\boldsymbol{\xi}) = \begin{bmatrix} c(1)^2\bar{\boldsymbol{G}}(\boldsymbol{\theta}_1) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & c(2)^2\bar{\boldsymbol{G}}(\boldsymbol{\theta}_2) & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & c(Q)^2\bar{\boldsymbol{G}}(\boldsymbol{\theta}_Q) \end{bmatrix} \tag{3.125}$$

By definition, the inverse relationship yields $\boldsymbol{G}(\boldsymbol{\xi}) = [\bar{\boldsymbol{G}}(\boldsymbol{\xi})]^{-1}$ and the components of $\boldsymbol{G}(\boldsymbol{\xi})$ are the fully contravariant components $g^{ml}(\boldsymbol{\xi})$ of the metric tensor.

Nondecomposable scalar field over $\mathcal{S}(\boldsymbol{\xi})$

If the scalar function $\bar{\varsigma}$ cannot be decomposed in the manner of Equation 3.122, then it is difficult to express a fibre over the base manifold $\mathcal{S}(\boldsymbol{\xi})$ as a direct sum of fibres over individual base manifolds $S(\boldsymbol{\theta}_q)$. In this case, a metric tensor is calculated directly in the tangent space for $\mathcal{S}(\boldsymbol{\xi})$ described in Equation 3.123. A sensible metric tensor is the covariance in tangent space which is maximally noncommittal in the sense described above. The covariance in tangent space for a general scalar function $\bar{\varsigma} : (\boldsymbol{O}, \boldsymbol{\xi}) \mapsto \bar{\varsigma}(\boldsymbol{O}; \boldsymbol{\xi})$ is then,

$$g_{ij}(\boldsymbol{\xi}) = \int \Big(\frac{\partial}{\partial \xi^i}\bar{\varsigma}(\boldsymbol{O}; \boldsymbol{\xi}) - \mu_i\Big)\Big(\frac{\partial}{\partial \xi^j}\bar{\varsigma}(\boldsymbol{O}; \boldsymbol{\xi}) - \mu_j\Big)p(\boldsymbol{O})\mathrm{d}\boldsymbol{O} \tag{3.126}$$

where,

$$\mu_i = \int \frac{\partial}{\partial \xi^i}\bar{\varsigma}(\boldsymbol{O}; \boldsymbol{\xi})p(\boldsymbol{O})\mathrm{d}\boldsymbol{O} \tag{3.127}$$

and where covariant derivatives are evaluated at $\boldsymbol{\xi}_0$. This metric tensor may have nonzero entries for all its components. However a block diagonal structure can be enforced. The metric tensor has a nonlinear relationship to the metric tensors for the tangent spaces of individual class manifolds, when those metric tensors are calculated as covariances on the log likelihood scalar function. The fully covariant components can be arranged into the metric matrix $\bar{\boldsymbol{G}}(\boldsymbol{\xi})$. The fully contravariant components are available from $\boldsymbol{G}(\boldsymbol{\xi}) = [\bar{\boldsymbol{G}}(\boldsymbol{\xi})]^{-1}$.

Calculating scalar products in tangent space

For $\boldsymbol{O}_l, \boldsymbol{O}_m \in L(\boldsymbol{O})$, two scalar fields are defined as $\bar{\varsigma}_l : \boldsymbol{\xi} \mapsto \bar{\varsigma}(\boldsymbol{O}_l; \boldsymbol{\xi})$ and $\bar{\varsigma}_m : \boldsymbol{\xi} \mapsto \bar{\varsigma}(\boldsymbol{O}_m; \boldsymbol{\xi})$ where $\boldsymbol{\xi} \in L(\boldsymbol{\xi}; \mathcal{S})$. Each of these scalar fields can be mapped to gradients in tangent

space at a point $\mathcal{P}_0 \in \mathcal{S}(\boldsymbol{\xi})$ with coordinate vector $\boldsymbol{\xi}_0$. The scalar product between the two gradients is,

$$(\nabla\bar{\varsigma_l}, \nabla\bar{\varsigma_m}) = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial}{\partial\xi^i}\bar{\varsigma}(\boldsymbol{O}_l;\boldsymbol{\xi})g(\boldsymbol{\xi})^{ij}\frac{\partial}{\partial\xi^j}\bar{\varsigma}(\boldsymbol{O}_m;\boldsymbol{\xi}) \tag{3.128}$$

where covariant derivatives are implicitly defined at point $\boldsymbol{\xi}_0$. In linear algebraic form,

$$\begin{aligned}(\nabla\bar{\varsigma_l}, \nabla\bar{\varsigma_m}) &= \nabla\bar{\varsigma_l} \quad \boldsymbol{G}(\boldsymbol{\xi}) \quad \nabla\bar{\varsigma_m} \\ &= \nabla\bar{\varsigma_l} \quad [\bar{\boldsymbol{G}}(\boldsymbol{\xi})]^{-1} \quad \nabla\bar{\varsigma_m}\end{aligned} \tag{3.129}$$

where $\bar{\boldsymbol{G}}(\boldsymbol{\xi})$ and $\boldsymbol{G}(\boldsymbol{\xi})$ are described above. Consequential to this is the definition of the 'natural gradient' [2] defined as $\tilde{\nabla}\bar{\varsigma_l} = [\bar{\boldsymbol{G}}(\boldsymbol{\xi})]^{-1}\nabla\bar{\varsigma_l}$ and which is described in [52] as the direction of steepest ascent for the scalar field along the manifold. The natural gradient is simply the conventional gradient in contravariant component form.

Having considered metric tensors for individual samples $\boldsymbol{O}$, it is sometimes useful to propose a metric tensor for a sequence $\mathcal{O}$ of ℓ i.i.d. samples. There is in general no simple relationship between the metric tensor $g(\boldsymbol{\xi};\boldsymbol{O})$ for a single sample, as denoted above by $g(\boldsymbol{\xi})$, and the metric tensor $g(\boldsymbol{\xi};\mathcal{O})$ for a set of samples. An exception is the Fisher metric where, for the statistical model $S(\boldsymbol{\theta})$,

$$g_{ij}(\boldsymbol{\theta};\mathcal{O}) = \ell g_{ij}(\boldsymbol{\theta};\boldsymbol{O}) \tag{3.130}$$

Ideally metric tensors should be invariant to sufficient statistics. However only the Fisher metric fulfills this invariance requirement for all distributions and sufficient statistics. It is defined on the log likelihood scalar function. However noninvariance to sufficient statistics is tolerated since it permits the application of a wider variety of scalar functions. However this ties the metrics to a particular set of sufficient statistics for the statistical models. This is analogous, though perhaps less severe, to a 'distance function' which is not a tensor and is tied to a particular coordinate system for the base manifold.

3.5.4.2 Metric tensors for score spaces

Having defined different metrics for the tangent space of a manifold $S(\boldsymbol{\theta}_q)$ or $\mathcal{S}(\boldsymbol{\xi})$, it is sensible to define metrics for the linear spaces defined on tangent spaces and their duals.

For example for linear spaces of form $\widehat{\breve{L}}^{1(0,1)}(\boldsymbol{\alpha}_q; \tau_q, \varrho, c(q)\bar{\varsigma}(q), (\boldsymbol{\theta}_q)_0)$ which, for $\tau_q = \tau_q^{\text{tay}}$ are elsewhere called score spaces, the metric tensor must be defined in a manner consistent with that for the tangent space. Once the metric tensor for the linear space is defined, it is straightforward to calculate the metric tensor for the dual space by the inverse relationship. From above, $\boldsymbol{G}(\boldsymbol{\theta}_q)$ is the metric matrix for tangent space and $\bar{\boldsymbol{G}}(\boldsymbol{\theta}_q)$ for its dual.

For brevity, the metric matrices for the dual of score space are analysed since it is conventional to consider the fully covariant components of metric tensors. Furthermore, the metric tensor is obtained in this form if calculated as the covariance of members in score space. For $\widehat{\breve{L}}^{1(1,0)}(\boldsymbol{\alpha}_q; \tau_q, \varrho, c(q)\bar{\varsigma}(q), (\boldsymbol{\theta}_q)_0)|_{\tau_q=\tau_q^{\text{tay}}}$, the dual of score space, the metric-like quantity is the $(\delta_q \times \delta_q)$ metric matrix $\bar{\boldsymbol{A}}(\varrho, \boldsymbol{\theta}_q)$ where,

$$\delta_q \;=\; \sum_{r=0}^{\varrho} (n_q)^r \tag{3.131}$$

and using $\otimes$ to denote the Kronecker product of matrices and $\mathbf{0}$ to represent a matrix of zeros of the appropriate size,

$$\bar{\boldsymbol{A}}(\varrho, \boldsymbol{\theta}_q) \;=\; \begin{bmatrix} 1 & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \bar{\boldsymbol{G}}(\boldsymbol{\theta}_q) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \otimes_{r=0}^{2} \bar{\boldsymbol{G}}(\boldsymbol{\theta}_q) & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \otimes_{r=0}^{\varrho} \bar{\boldsymbol{G}}(\boldsymbol{\theta}_q) \end{bmatrix} \tag{3.132}$$

The use of $\bar{\boldsymbol{A}}(\varrho, \boldsymbol{\theta}_q)$ ensures the isomorphism between the relevant tensor spaces and the dual of score space is isometric. Substituting $\boldsymbol{G}(\boldsymbol{\theta}_q)$ for $\bar{\boldsymbol{G}}(\boldsymbol{\theta}_q)$ in the right hand side of Equation 3.132 yields $\boldsymbol{A}(\varrho, \boldsymbol{\theta}_q)$ which is the metric matrix for score space. The two matrices are related by an inverse relationship so $\boldsymbol{A}(\varrho, \boldsymbol{\theta}_q) = [\bar{\boldsymbol{A}}(\varrho, \boldsymbol{\theta}_q)]^{-1}$.

Whether the scalar function is decomposable according to Equation 3.122 or not, the metric matrix for the dual of score space $\widehat{\breve{L}}^{1(1,0)}(\boldsymbol{\beta}; \mathcal{T}, \varrho, \bar{\varsigma}, \boldsymbol{\xi}_0)$ is of size $(\delta \times \delta)$ where,

$$\delta \;=\; \sum_{r=0}^{\varrho} \Big(\sum_{q=1}^{Q} n_q \Big)^r \tag{3.133}$$

Then if $\bar{\boldsymbol{G}}(\boldsymbol{\xi})$ is the metric matrix for the dual of tangent space formed from fully covariant

components, then $\bar{\boldsymbol{A}}(\varrho, \boldsymbol{\xi})$ is,

$$\bar{\boldsymbol{A}}(\varrho, \boldsymbol{\xi}) = \begin{bmatrix} 1 & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \bar{\boldsymbol{G}}(\boldsymbol{\xi}) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \otimes_{r=0}^{2} \bar{\boldsymbol{G}}(\boldsymbol{\xi}) & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots & \otimes_{r=0}^{\varrho} \bar{\boldsymbol{G}}(\boldsymbol{\xi}) \end{bmatrix} \tag{3.134}$$

Again replacing $\bar{\boldsymbol{G}}(\boldsymbol{\xi})$ by $\boldsymbol{G}(\boldsymbol{\xi})$ in the right hand side of the equation yields $\boldsymbol{A}(\varrho, \boldsymbol{\xi})$ which is the metric matrix for score space. Again $\boldsymbol{A}(\varrho, \boldsymbol{\xi}) = [\bar{\boldsymbol{A}}(\varrho, \boldsymbol{\xi})]^{-1}$. If the scalar function is decomposable, then $\bar{\boldsymbol{G}}(\boldsymbol{\xi})$ is as given in Equation 3.125.

3.6 Application to experiments

This section gives details concerning the relaxation of constraints permitted when training linear discriminants in score space. First, when scalar products are used to train such a linear discriminant, it is acceptable to relax the constraints on the metric tensor providing the weight vector of the discriminant fulfills the constraints required for the relevant submanifold. For example, the score space may simply be viewed as a type $(0, 1)$ tensor space and a metric defined in this space. Such an approach was adopted for the experiments in this thesis with $\varrho = 1$. In these experiments, the metric matrix for the dual of score space defined on $\mathcal{S}(\boldsymbol{\xi})$ was[6],

$$\bar{\boldsymbol{A}}(1, \boldsymbol{\xi}) = \begin{bmatrix} v & \mathbf{0} \\ \mathbf{0} & \bar{\boldsymbol{\Sigma}} \end{bmatrix} \tag{3.135}$$

where $\bar{\boldsymbol{\Sigma}}$ was the diagonal covariance matrix in tangent space and v was not necessarily unity but the variance in the zeroth degree subspace. The metric matrix for score space was then $\boldsymbol{A}(1, \boldsymbol{\xi}) = [\bar{\boldsymbol{A}}(1, \boldsymbol{\xi})]^{-1}$. By this means the contribution of the zeroth degree covariant derivative to the square of the norm of a member of score space was sensibly scaled by its inverse variance $1/v$.

[6] A similar approach was also applied to score spaces defined on single models such as $S(\boldsymbol{\theta}_q)$ or $S(\boldsymbol{\theta})$.

Furthermore, in the experiments no attempt was made to ascertain whether the relevant scalar fields were analytic about the points of expansion or what were their domains of convergence. This was of no consequence for the validity of the solutions obtained. However no attempt was made to verify whether it was possible to relate the weight vector of each linear discriminant trained in score space to valid distributions in the total space of the corresponding fibre bundle. This was simply for convenience, and constrained optimisation would otherwise be required to guarantee the relationship.

3.7 Summary

This chapter has introduced the concept of viewing a statistical model as a differentiable manifold in the space of scalar functions. Scalar fields vary over the statistical manifold and the Taylor expansion for recovering values of this scalar field at distant points on the manifold was presented. Rather than assume the coordinate system is Euclidean, expressions were presented permitting nonzero curvature in the statistical manifold. Next a fibre bundle was introduced as a structure also existing in the space of scalar functions and with the statistical manifold as its base. This permitted a principled approach to approximating the Taylor expansion by evaluating the scalar field at a point outside the base manifold but within a fibre. By considering submanifolds of fibres in the space of distributions, a fibre bundle could be defined as an augmented form of statistical model, where the augmentation was relative to its base model. It is then possible to estimate distributions within the total space of the bundle but not necessarily within the original statistical model. Maximum likelihood and discriminative estimation techniques were presented in the context of a vector bundle, and particularly in the context of fibre subspaces called score spaces. The vector bundle is closely related to the aforementioned fibre bundle, but more readily permits the application of 'distance-based' learning algorithms. Some suitable metrics were then proposed with a view to calculating distances within score spaces or defining connection coefficients on the base manifold.

Chapter 4

Score spaces for classification

The previous chapter introduced the concept of score spaces. Score spaces may be viewed as 'tools' to train scalar functions in the total space of fibre bundles. More generally, they may be viewed as model-based feature spaces in which statistical models or classifiers can be trained. This chapter adopts this general view. Section 4.1 introduces some simple score spaces relevant for the thesis. Sections 4.2 and 4.3 discuss the nature of score spaces and factors affecting classification performance in score space. Section 4.4 describes some common training criteria for statistical models from the perspective of a simple score space. Section 4.5 discusses why a distinct division between the score mapping and score space classifier may be advantageous. Section 4.6 then describes a normalisation technique applied for variable length patterns.

4.1 Description of different score spaces

In Chapter 3, statistical models were proposed to model samples in input space. Given a statistical model $S(\boldsymbol{\theta})$, a Taylor expansion can be taken about a point or distribution $p_0 \in S(\boldsymbol{\theta})$ up to the order ϱ. The covariant derivatives in the Taylor expansion can be assembled into a linear space called a score space as defined in Section 3.4.3. The technique is a model-based mapping from input space to a new 'feature space'. The term 'score map'

introduced in [73] is extended to describe all such model-based mappings.

In the experiments in this thesis, the set of statistical models $\mathcal{S}(\boldsymbol{\xi})$ is either a set of class-conditional GMMs or HMMs. For clarity, a point $\mathcal{P}_0 \in \mathcal{S}(\boldsymbol{\xi})$ corresponds to a particular set of Gaussian Mixture Distributions (GMDs) or Hidden Markov Distributions (HMDs). As detailed in Section 3.4.3, the notation for score spaces can be considerably simplified. Hence for the model for class ω_q, $\boldsymbol{\alpha}_q \in L(\boldsymbol{\alpha}_q)$ and $\tau_q = \tau_q^{\text{tay}}$ are assumed, and the linear space,

$$\widehat{\tilde{L}}^{1(0,1)}(\boldsymbol{\alpha}_q; \tau_q, \varrho, c(q)\bar{\varsigma}(q), (\boldsymbol{\theta}_q)_0)\Big|_{\tau_q=\tau_q^{\text{tay}}} \equiv \varphi^{\text{sup}}(\varrho, (\boldsymbol{\theta}_q)_0)$$

where the superscript 'sup' contains information relative to the scalar field, and ϱ is simply replaced by the order of exponentiation where $\varrho \in \{0, 1\}$. Though simple, these score spaces illustrate important principles without heavy analytical or computational cost. In this chapter, the unit degree covariant derivatives are often restricted to those with respect to Gaussian means. This is simply for convenience but also because these covariant derivatives are often the most discriminative (e.g., see Section 6.3.4). These simple score spaces may also be spliced together to form *appended score spaces*.

The notation for a set of Q class-conditional statistical models is as described in Section 2.2.1. Scores are detailed for a sample $\boldsymbol{O}_l \in L(\boldsymbol{O})$. The simplest of the score spaces are those where $\varrho = 0$.

- *Likelihood score space*, $\varphi^{\text{lk(q)}}(0, (\boldsymbol{\theta}_q)_0)$: for class ω_q, this is a 1-component space consisting of a log likelihood scalar field evaluated at point $(\boldsymbol{\theta}_q)_0$. The score for $\boldsymbol{O}_l$ is unbounded and,

$$\bar{\boldsymbol{\varphi}}^{\text{lk(q)}}(\boldsymbol{O}_l; 0, (\boldsymbol{\theta}_q)_0) = \left[\ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_q)_0) \right] \tag{4.1}$$

- *Likelihood-ratio score space* $\varphi^{\text{lr(a,b)}}(0, \boldsymbol{\xi}_0)$: the 1-component log likelihood-ratio score space between classes ω_a and ω_b evaluated at point $\boldsymbol{\xi}_0 = ((\boldsymbol{\theta}_a)_0^\top, (\boldsymbol{\theta}_b)_0^\top)^\top$. The score for $\boldsymbol{O}_l$ is unbounded and,

$$\bar{\boldsymbol{\varphi}}^{\text{lr(a,b)}}(\boldsymbol{O}_l; 0, \boldsymbol{\xi}_0) = \left[\ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_a)_0) - \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_b)_0) \right] \tag{4.2}$$

- *Posterior score space*, $\varphi^{\mathrm{ps(q)}}(0, \boldsymbol{\xi}_0)$: this 1-component score space for class ω_q is the output of a log class posterior scalar field. The score for $\boldsymbol{O}_l$ is unbounded and,

$$\bar{\boldsymbol{\varphi}}^{\mathrm{ps(q)}}(\boldsymbol{O}_l; 0, \boldsymbol{\xi}_0) \;=\; \left[\; \ln P(\omega_q|\boldsymbol{O}_l) \;\right] \tag{4.3}$$

- *Linear likelihood score space*, $\varphi^{\mathrm{lkl(q)}}(0, (\boldsymbol{\theta}_q)_0)$: the linear form of the above likelihood score space for class ω_q. Lower bounded by zero, the score for $\boldsymbol{O}_l$ is,

$$\bar{\boldsymbol{\varphi}}^{\mathrm{lkl(q)}}(\boldsymbol{O}_l; 0, (\boldsymbol{\theta}_q)_0) \;=\; \left[\; p(\boldsymbol{O}_l; (\boldsymbol{\theta}_q)_0) \;\right] \tag{4.4}$$

- *Linear posterior score space*, $\varphi^{\mathrm{psl(q)}}(0, \boldsymbol{\xi}_0)$: the linear form of the above posterior score space for class ω_q. The score for $\boldsymbol{O}_l$ is bounded by zero and unity and,

$$\bar{\boldsymbol{\varphi}}^{\mathrm{psl(q)}}(\boldsymbol{O}_l; 0, \boldsymbol{\xi}_0) \;=\; \left[\; P(\omega_q|\boldsymbol{O}_l) \;\right] \tag{4.5}$$

- *Appended likelihood score space*, $\varphi^{\mathrm{lk(all)}}(0, \boldsymbol{\xi}_0)$: formed by appending the individual likelihood score spaces for different classes. For Q classes, this yields a Q-component score space. The score for $\boldsymbol{O}_l$ is unbounded and,

$$\bar{\boldsymbol{\varphi}}^{\mathrm{lk(all)}}(\boldsymbol{O}_l; 0, \boldsymbol{\xi}_0) \;=\; \begin{bmatrix} \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_1)_0) \\ \vdots \\ \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_q)_0) \\ \vdots \\ \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_Q)_0) \end{bmatrix} \tag{4.6}$$

- *Appended posterior score space*, $\varphi^{\mathrm{ps(all)}}(0, \boldsymbol{\xi}_0)$: identical to the appended likelihood score space except that log class likelihoods are replaced by log class posteriors. The class posteriors provide all the information necessary to implement a MAP decision rule. For $\boldsymbol{O}_l$, the score is bounded by sum-to-unity constraints on the linear class posteriors and,

$$\bar{\boldsymbol{\varphi}}^{\mathrm{ps(all)}}(\boldsymbol{O}_l; 0, \boldsymbol{\xi}_0) \;=\; \begin{bmatrix} \ln P(\omega_1|\boldsymbol{O}_l) \\ \vdots \\ \ln P(\omega_q|\boldsymbol{O}_l) \\ \vdots \\ \ln P(\omega_Q|\boldsymbol{O}_l) \end{bmatrix} \tag{4.7}$$

Score spaces may also be defined with $\varrho = 1$, i.e. on zeroth and first degree covariant derivatives. Definitions follow an identical pattern to the corresponding zeroth order score spaces. The number of parameters and dimension of model $S(\boldsymbol{\theta}_q)$ is n_q. Covariant derivatives are implicitly taken at $(\boldsymbol{\theta}_q)_0$ or $\boldsymbol{\xi}_0$ as appropriate[1].

- *Likelihood score space*, $\varphi^{\mathrm{lk(q)}}(1, (\boldsymbol{\theta}_q)_0)$: this has $(n_q + 1)$ components and for $\boldsymbol{O}_l$, the score is unbounded and,

$$\bar{\boldsymbol{\varphi}}^{\mathrm{lk(q)}}(\boldsymbol{O}_l; 1, (\boldsymbol{\theta}_q)_0) = \begin{bmatrix} \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_q)_0) \\ \nabla_{\boldsymbol{\theta}_q} \ln p(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \end{bmatrix} \tag{4.8}$$

- *Likelihood-ratio score space*, $\varphi^{\mathrm{lr(a,b)}}(1, \boldsymbol{\xi}_0)$: this has $(n_a + n_b + 1)$ components with unbounded score for $\boldsymbol{O}_l$,

$$\bar{\boldsymbol{\varphi}}^{\mathrm{lr(a,b)}}(\boldsymbol{O}_l; 1, \boldsymbol{\xi}_0) = \begin{bmatrix} \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_a)_0) - \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_b)_0) \\ \nabla_{\boldsymbol{\theta}_a} \ln p(\boldsymbol{O}_l; \boldsymbol{\theta}_a) \\ -\nabla_{\boldsymbol{\theta}_b} \ln p(\boldsymbol{O}_l; \boldsymbol{\theta}_b) \end{bmatrix} \tag{4.9}$$

- *Posterior score space*, $\varphi^{\mathrm{ps(q)}}(1, \boldsymbol{\xi}_0)$: this has $(1 + \sum_{q=1}^{Q} n_q)$ components with score for $\boldsymbol{O}_l$,

$$\bar{\boldsymbol{\varphi}}^{\mathrm{ps(q)}}(\boldsymbol{O}_l; 1, \boldsymbol{\xi}_0) = \begin{bmatrix} \ln P(\omega_q | \boldsymbol{O}_l) \\ \nabla_{\boldsymbol{\xi}} \ln P(\omega_q | \boldsymbol{O}_l) \end{bmatrix} \tag{4.10}$$

The appended likelihood score space $\varphi^{\mathrm{lk(all)}}(1, \boldsymbol{\xi}_0)$ and appended posterior score space $\varphi^{\mathrm{ps(all)}}(1, \boldsymbol{\xi}_0)$ are defined similarly to the case where $\varrho = 0$. However there are many repeated components in the score space $\varphi^{\mathrm{ps(all)}}(1, \boldsymbol{\xi}_0)$. This linear space can be reduced in size without sacrificing any information to give the *reduced appended posterior score space* $\varphi^{\mathrm{psr(all)}}(1, \boldsymbol{\xi}_0)$ with $\sum_{q=1}^{Q}(2n_q + 1)$ components. A slightly more generalised form of this appended linear space is also proposed which has no straightforward relation to fibre bundles and is simply a discriminative feature space. The score space is called the *generalised appended posterior score space* denoted by $\varphi^{\mathrm{psg(all)}}(1, \boldsymbol{\xi}_0)$ and also has $\sum_{q=1}^{Q}(2n_q + 1)$

[1]For convenience there is a slight abuse of notation for covariant derivatives as explained in Appendix B.3.

components. These two score spaces are detailed in Appendix B.4. Also proposed and detailed in Appendix B.4 is a hybrid score space which combines the zeroth order linear posterior score space and the unit degree covariant derivatives of the appended likelihood score space. This is denoted by $\varphi^{\text{psh(all)}}(1, \boldsymbol{\xi}_0)$ and has fewer components at $\sum_{q=1}^{Q}(n_q + 1)$.

If $S(\boldsymbol{\theta})$ is a single statistical model for Q classes and has dimension and number of parameters n, then for $\varrho \in \{0, 1\}$,

- *Likelihood (Q-class) score space*, $\varphi^{\text{lk(1,...Q)}}(\varrho, \boldsymbol{\theta}_0)$: this is identical to the likelihood score space $\varphi^{\text{lk(q)}}(\varrho, (\boldsymbol{\theta}_q)_0)$ except the defining distribution is marginalised over the class labels.

When $Q = 2$, another score space implicitly defined by the Fisher kernel in [52] is,

- *Fisher score space*, $\varphi^{\text{fs(a,b)}}(\boldsymbol{\theta}_0)$: this is identical to the likelihood (2-class) score space $\varphi^{\text{lk(a,b)}}(1, \boldsymbol{\theta}_0)$ except for the omission of the zeroth order subspace. The score space has n components.

Whether a statistical model refers to a single class or is marginalised over more than one class should be clear from context.

The unit degree covariant derivatives for the log likelihood, log likelihood-ratio and log class posterior scalar fields are detailed in Appendix B.3 for statistical models which are HMMs (GMMs may be viewed as single state HMMs). All score spaces calculated in this thesis are defined on these statistical models.

4.2 The nature of the score mapping

An understanding of the nature of score spaces and score mappings is useful (see set theory in Section 376 of [50]). To recap, $L(\boldsymbol{O})$ is defined as the open set of all possible samples

determined by the source. For a statistical model $S(\boldsymbol{\theta})$ and a point $\boldsymbol{\theta}_0 \in L(\boldsymbol{\theta}; S)$, a score mapping φ may be defined[2],

$$\varphi : L(\boldsymbol{O}) \quad \rightarrow \quad \varphi(\boldsymbol{O}, \boldsymbol{\theta}_0) \tag{4.11}$$

where $\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$ is the image space and is a possibly open, closed or boundary set within the score space $\varphi(\boldsymbol{\theta}_0)$. The score space is assumed isomorphic to $\mathbb{R}^\delta$ where δ is the size of the score space. Hence for $\boldsymbol{O}_l \in L(\boldsymbol{O})$ and $\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \boldsymbol{\theta}_0) \in \varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$,

$$\varphi : \boldsymbol{O}_l \quad \mapsto \quad \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_l; \boldsymbol{\theta}_0) \tag{4.12}$$

The score mapping is a surjection with respect to the subspace $\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$, and certain regions of score space may be 'unreachable' from $L(\boldsymbol{O})$. In addition, the score mapping may either be injective or noninjective. Mappings which are surjective and injective are bijective and imply an inverse mapping exists. These principles are illustrated in Figure 4.1(a) for a noninjective mapping φ.

Two important subspaces of score space are defined by the data sequences or samples mapped into score space. If there are ℓ training samples summarised by the sample set $\mathcal{O}$, then their ℓ training scores define the vertices of a hyperpolyhedron $\varphi(\mathcal{O}, \boldsymbol{\theta}_0)$ in score space. The span of the hyperpolyhedron is a linear subspace $\varphi(\mathrm{sp}, \mathcal{O}, \boldsymbol{\theta}_0)$. In Figure 4.1(b), a magnified version of Figure 4.1(a), the hyperpolyhedron $\varphi(\mathcal{O}, \boldsymbol{\theta}_0)$ is represented by the shaded region and $\varphi(\mathrm{sp}, \mathcal{O}, \boldsymbol{\theta}_0)$ is labelled along its limiting boundaries. The subspace $\varphi(\mathcal{O}, \boldsymbol{\theta}_0)$ is strictly enclosed within the image space $\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$ and may be nonlinear due to restrictions induced by the distribution $p_0 \in S(\boldsymbol{\theta})$ in input space. The subspace $\varphi(\mathrm{sp}, \mathcal{O}, \boldsymbol{\theta}_0)$ is strictly enclosed within $\varphi(\mathrm{sp}, \boldsymbol{O}, \boldsymbol{\theta}_0)$ which is the span of the image space $\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$. Hence,

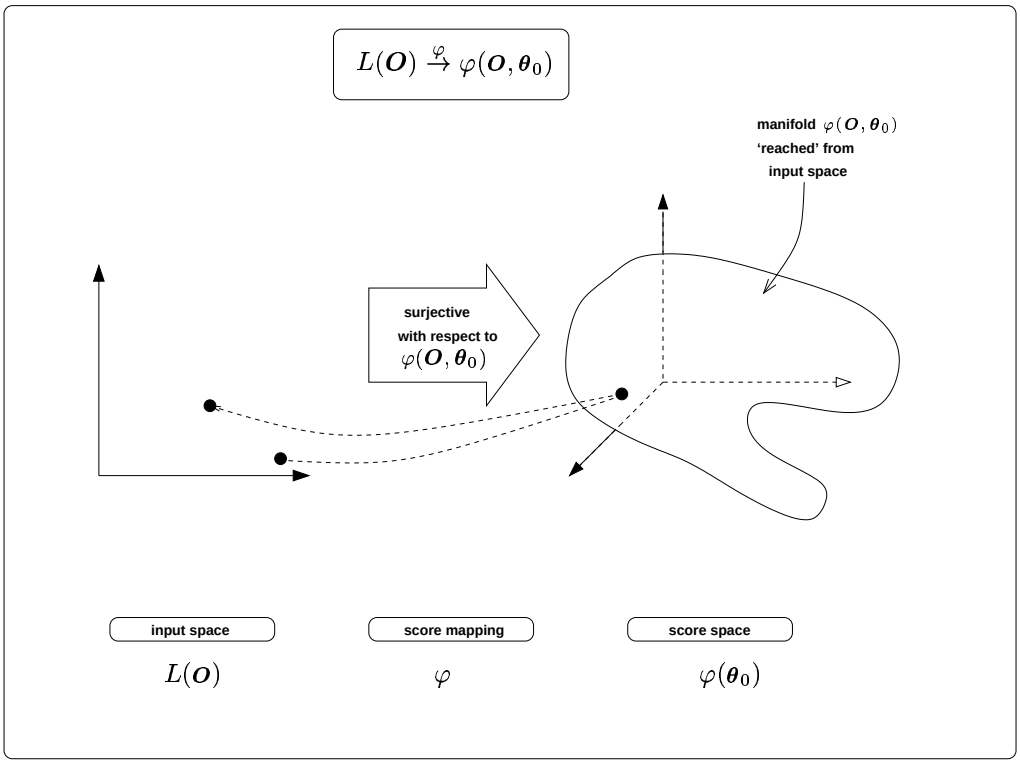
$$\varphi(\mathrm{sp}, \mathcal{O}, \boldsymbol{\theta}_0) \quad \subseteq \quad \varphi(\mathrm{sp}, \boldsymbol{O}, \boldsymbol{\theta}_0) \subseteq \varphi(\boldsymbol{\theta}_0) \tag{4.13}$$

$$\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0) \quad \subseteq \quad \varphi(\mathrm{sp}, \boldsymbol{O}, \boldsymbol{\theta}_0) \tag{4.14}$$

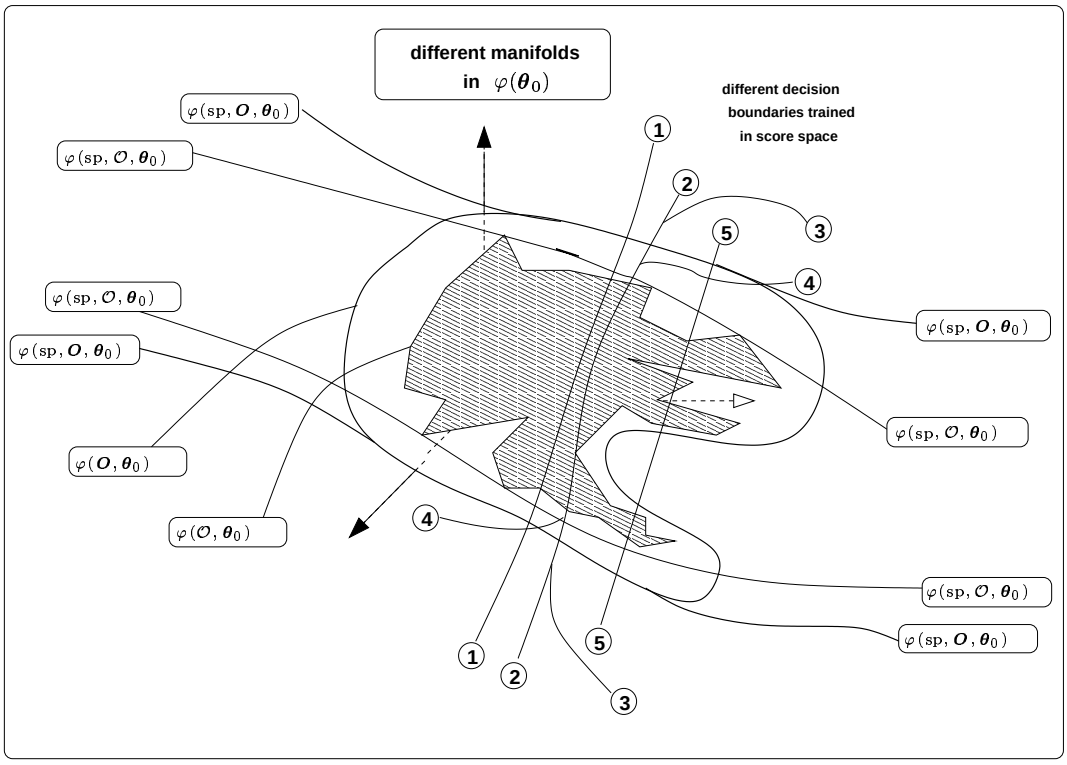
$$\varphi(\mathcal{O}, \boldsymbol{\theta}_0) \quad \subseteq \quad \varphi(\mathrm{sp}, \mathcal{O}, \boldsymbol{\theta}_0) \tag{4.15}$$

Also, $\dim(\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)) \leq \dim(L(\boldsymbol{O}))$ where $\dim(\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0))$ is sometimes called the *intrinsic dimensionality* of the structure in score space. A simple illustration is given in Figure 4.2

[2]The scalar field $\bar{\varsigma}$ and order of expansion ϱ are assumed known and are not explicitly detailed.



(a) The score mapping is noninjective



(b) Different subspaces within score space

Figure 4.1: Illustrating a score mapping and corresponding score space

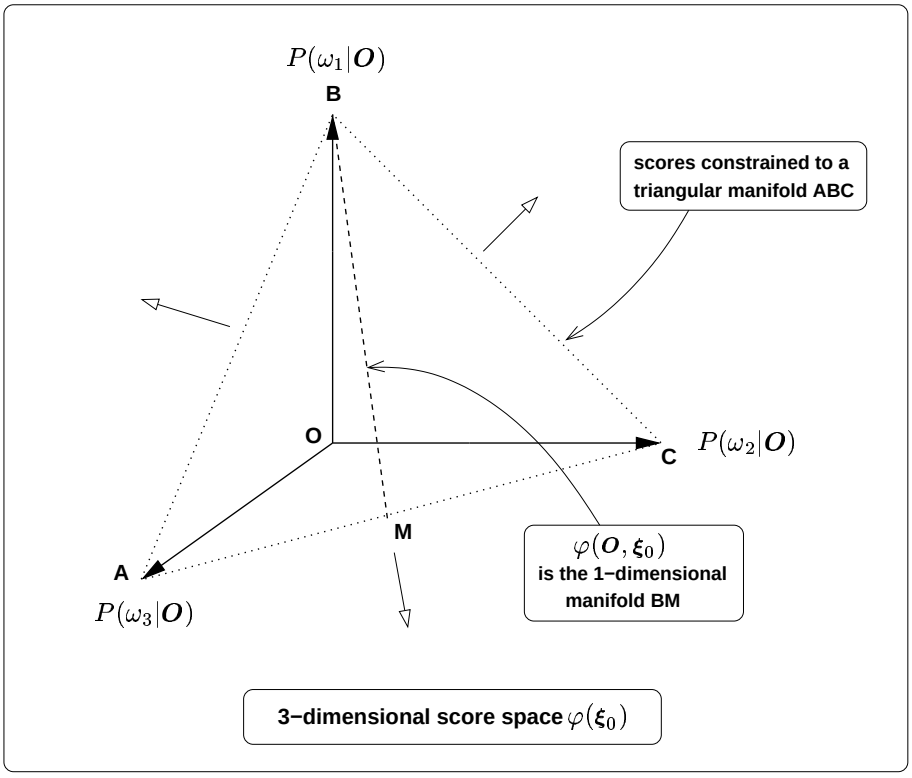


Figure 4.2: Example of an image space $\varphi(\boldsymbol{O},\boldsymbol{\xi}_0)$ in score space

for the appended linear posterior score space $\varphi^{\text{psl(all)}}(0,\boldsymbol{\xi}_0)$ and its associated score mapping. A 3-class task is presented. The entire space of input samples $L(\boldsymbol{O})$ is mapped onto a finite 2-dimensional boundary set ABC defined by the zero-unity bounds on class posteriors and their sum-to-unity constraint. Any point outside the set ABC cannot possibly be reached by the score mapping. However, the structure of the distributions $\mathcal{P}_0$ in input space may additionally constrain the permissible combination of class posteriors. As a result the image space $\varphi^{\text{psl(all)}}(\boldsymbol{O},\boldsymbol{\xi}_0)$ may be a subset within ABC. A trivial example is presented where the class posteriors $P(\omega_2|\boldsymbol{O})$ and $P(\omega_3|\boldsymbol{O})$ are always identical, e.g. when class priors are identical and class-conditional distributions coincide. This restricts the scores to a 1-dimensional line on ABC which is $\varphi^{\text{psl(all)}}(\boldsymbol{O},\boldsymbol{\xi}_0)$. Hence both the score mapping and distributions $\mathcal{P}_0$ impose constraints on the image space.

The nature of the mapping from input space to score space is also described by the metric it induces in input space. Both [1] and Remark 3.16 in [19] view the input space $L(\boldsymbol{O})$ as a Riemannian manifold and the image space here abbreviated to $\varphi(\boldsymbol{O})$ as a linear space.

A metric $g(\boldsymbol{O}_0)$ is defined on the tangent space to this manifold at a point with coordinate vector $\boldsymbol{O}_0 \in L(\boldsymbol{O})$ (this is in a similar manner to the metric $g(\boldsymbol{\theta}_0)$ defined on a Riemannian statistical manifold at a point with coordinate vector $\boldsymbol{\theta}_0$). A mapping is assumed,

$$\varphi : L(\boldsymbol{O}) \quad \rightarrow \quad \varphi(\boldsymbol{O}) \tag{4.16}$$

Following [1], the metric is described by a tensor with component $g_{ij}(\boldsymbol{O}_0)$ set to the scalar product of the gradients of $\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_a)$ and $\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_b)$, where $\boldsymbol{O}_a, \boldsymbol{O}_b \in L(\boldsymbol{O})$ and $\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_a), \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_b) \in \varphi(\boldsymbol{O})$. Hence,

$$\begin{aligned} g_{ij}(\boldsymbol{O}_0) &= g_{ij}(\boldsymbol{O})\Big|_{\boldsymbol{O}=\boldsymbol{O}_0} = \left(\frac{\partial}{\partial(O_a)^i}\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_a)\frac{\partial}{\partial(O_b)^j}\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_b)\right)\Big|_{\boldsymbol{O}_a=\boldsymbol{O}_b=\boldsymbol{O}_0} \\ &= \left(\frac{\partial^2}{\partial(O_a)^i\partial(O_b)^j}k(\boldsymbol{O}_a,\boldsymbol{O}_b)\right)\Big|_{\boldsymbol{O}_a=\boldsymbol{O}_b=\boldsymbol{O}_0} \end{aligned} \tag{4.17}$$

If both $\boldsymbol{O}_a$ and $\boldsymbol{O}_b$ are single observations of equal length with d components, then the components $(O_a)^i, i = \{1 \ldots d\}$ are linearly independent and $(O_b)^j, j = \{1 \ldots d\}$ are also linearly independent. The metric tensor $g(\boldsymbol{O}_0)$ has $(d \times d)$ components. If both $\boldsymbol{O}_a$ and $\boldsymbol{O}_b$ are sequences of observations and individual observations are not i.i.d., then the dimension of the input manifold is difficult to determine.

The metric $g(\boldsymbol{O}_0)$ describes the infinitessimal change in the image vector $\bar{\boldsymbol{\varphi}}(\boldsymbol{O})$ for infinitessimal changes in the components of the input space at $\boldsymbol{O}_0$. For a d-component input space,

$$||\mathrm{d}\bar{\boldsymbol{\varphi}}(\boldsymbol{O})||^2\Big|_{\boldsymbol{O}=\boldsymbol{O}_0} \quad = \quad \sum_{i=1}^{d}\sum_{j=1}^{d} g_{ij}(\boldsymbol{O}_0)\mathrm{d}(O^i)\mathrm{d}(O^j) \tag{4.18}$$

where $||\cdot||$ denotes the norm. This metric is induced by the mapping φ or by a kernel function which implicity defines the mapping. Alternatively any positive definite metric may be substituted for the metric in Equation 4.17, tantamount to assuming a different mapping or kernel. A different expression for this metric tensor is calculated in [19] and is derived from the second order term of a Taylor expansion but the derivation in [1] is preferred since it is clearer. Following this, [1] derive a magnification factor $M(\boldsymbol{O}_0)$ which defines the ratio increase for an infinitessimal volume about $\boldsymbol{O}_0$ as it is mapped into the image space. Letting $\det(\cdot)$ and $|\cdot|$ respectively denote the determinant and absolute value of their arguments,

$$M(\boldsymbol{O}_0) \quad = \quad \sqrt{|\det(\bar{\boldsymbol{G}}(\boldsymbol{O}_0))|} \tag{4.19}$$

4.3 Factors affecting classification performance

There are a number of factors which affect classification performance in score space. These include,

- the definition of the score space,
- the noninjective nature of the score mapping,
- the nature of the classifier in score space,
- the number of training samples,
- the magnification induced by the mapping near to the decision boundary.

These factors are described below and are related to the experiments later in the thesis. The factors have a complicated interaction and the effect of one cannot be viewed in isolation.

4.3.1 Definition of the score space

The score space may simply be viewed as a collection of model-dependent 'features'. If score spaces are always defined through Taylor expansions, the 'parameters' of the score space are the defining scalar function and the order of expansion ϱ. The order of expansion is often limited by computational or analytical complexity, so the choice of scalar function is often the most important influence. For the present, it is sufficient to show how score spaces may be used to enhance class discrimination. Classes which cannot be disambiguated by inspecting likelihoods alone can sometimes be distinguished by inspecting their scores. For example, consider the two-class problem in Figure 4.3. The statistical model $S(\boldsymbol{\theta}_q)$ is a single Gaussian in a 1-component input space with variable mean μ_q but fixed variance. The distribution may be viewed as a poor fit to either of the two classes ω_a or ω_b. The two classes cannot be linearly separated by viewing the log likelihoods $\ln p(\boldsymbol{O}; (\mu_q)_0)$ alone but can in the likelihood score space.

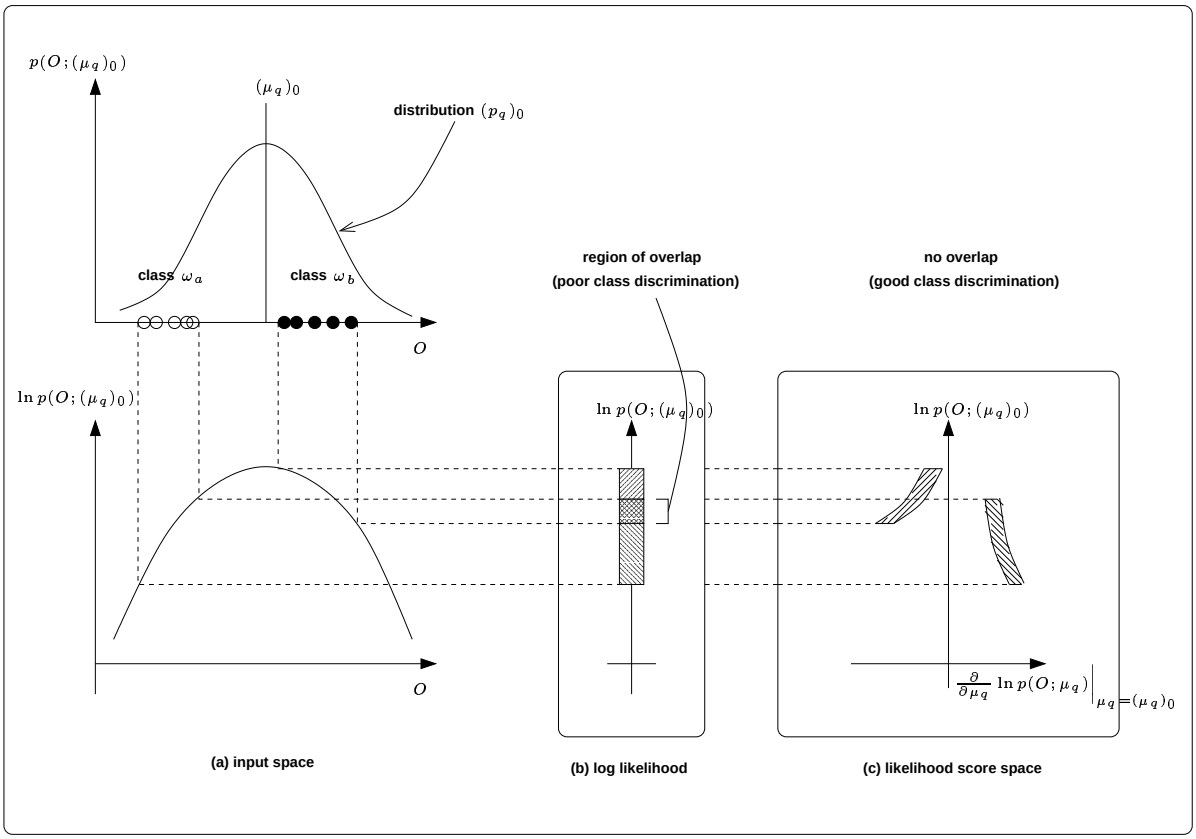


Figure 4.3: Enhancing class discrimination by mapping into a score space

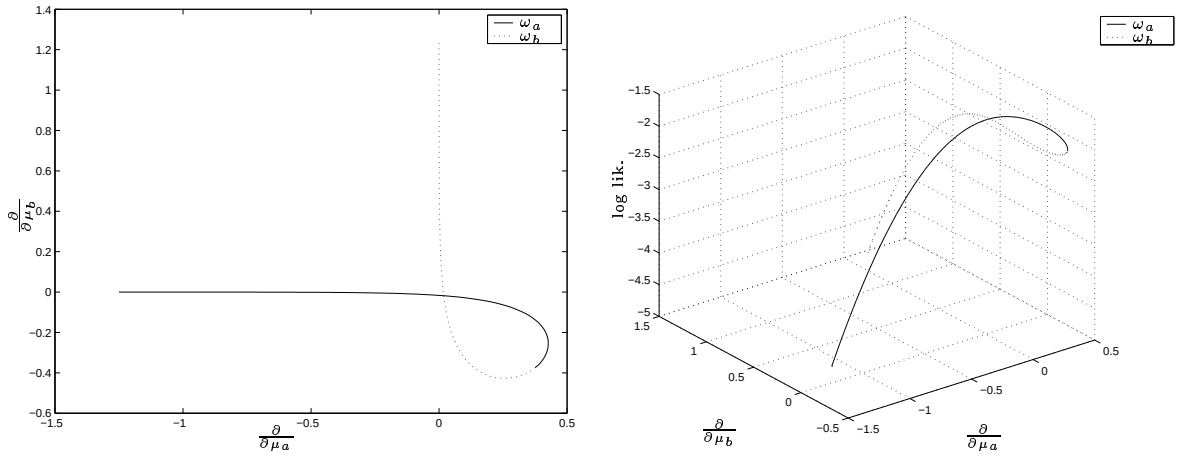
4.3.2 The noninjective nature of the score mapping

A single point in input space $L(\boldsymbol{O})$ can only map to a single point in score space $\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$ since the score mapping φ is deterministic. However two points in input space may map to a single point in score space. The score mapping may be noninjective. Providing the mapping is differentiable, the correspondence between points in $L(\boldsymbol{O})$ and $\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$ imply decision boundaries in score space can be consistently mapped back to decision boundaries in input space, but not vice versa. A single class region in score space may map back to disjoint class regions in input space. However with strong regularisation in score space, it is hopefully possible to maintain the good generalisation ability of the classifier, by a similar argument to that for SVMs with nonlinear kernels.

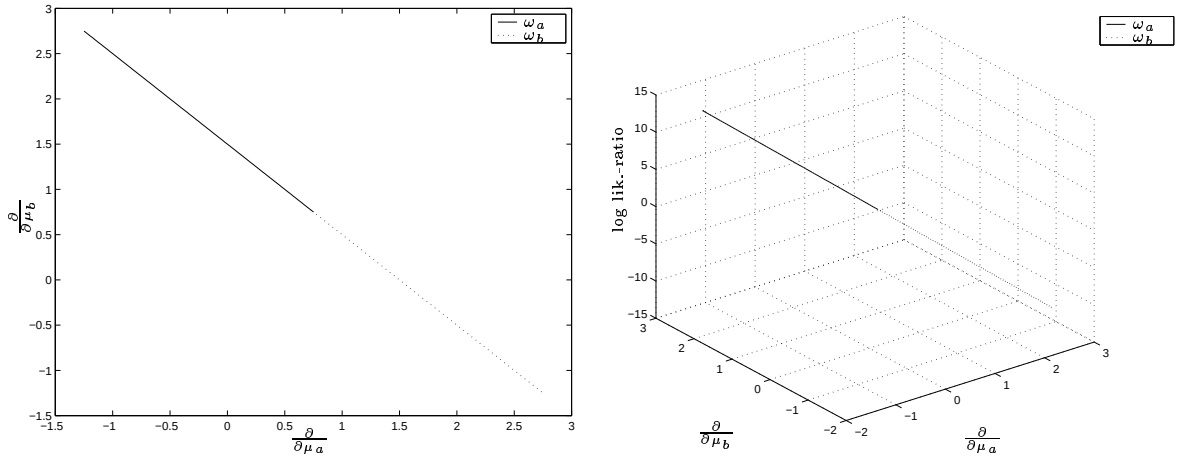
The noninjective nature of the score mapping may aid or impede class discrimination[3]. For example, let $\boldsymbol{O}_a, \boldsymbol{O}_b \in L(\boldsymbol{O})$ where $\boldsymbol{O}_a \neq \boldsymbol{O}_b$. The score mapping φ maps the two samples $\boldsymbol{O}_a$ and $\boldsymbol{O}_b$ to the same point in score space so that $\bar{\boldsymbol{\varphi}}(\boldsymbol{O}_a) = \bar{\boldsymbol{\varphi}}(\boldsymbol{O}_b)$. If the two samples are drawn from the same class, then the inability to distinguish the two scores is not harmful and eliminates within-class variability between the two samples. However if drawn from different classes, then it is impossible to define a decision rule in score space to separate these two samples. Ideally, the score mapping should emphasise or retain between-class variability and eliminate or reduce within-class variability.

A simple example illustrates how the noninjective nature of the score mapping can limit classification performance. Two single Gaussians are located in a 1-component input space with means at -3.0 and 3.0 and with coincident fixed variances at 2.0. The two Gaussians respectively represent classes ω_a and ω_b. Two score mappings and score spaces are defined. The two Gaussians may be given equal weighting and used to form a 2-component GMM modelling both classes. The Gaussian distributions then define a likelihood (2-class) score space $\varphi^{\mathrm{lk(a,b)}}(1, \boldsymbol{\theta}_0)$ where $\boldsymbol{\theta}_0 = ((\mu_a)_0, (\mu_b)_0)^\top$. Alternatively, the two Gaussians may be kept distinct and used to define a likelihood-ratio score space $\varphi^{\mathrm{lr(a,b)}}(1, \boldsymbol{\xi}_0)$ where $\boldsymbol{\xi}_0 = ((\mu_a)_0, (\mu_b)_0)^\top$. The 2-component projections and 3-component score spaces are plotted in

[3]This explanation assumes the classifier in score space has sufficient complexity to take advantage of increased class separability in score space.



(a) likelihood (2-class) score space



(b) likelihood-ratio score space

Figure 4.4: Distributions in score space

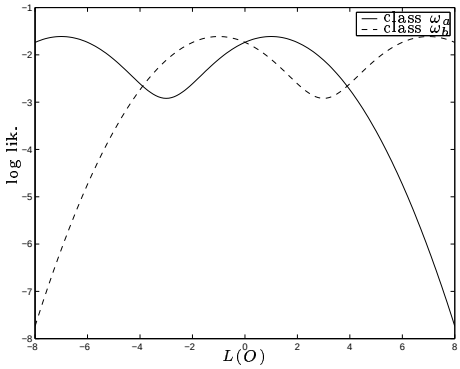
Figures 4.4(a) and 4.4(b). The full line represents scores from class ω_a and the dotted line scores from class ω_b, where, for convenience, class membership is assigned using a MAP decision rule based on the outputs of the single Gaussians and assuming equal class priors. The effect of the component posteriors in the likelihood score space causes the structure in Figure 4.4(a) to fold back on itself and intersect at the point defined by zero gradients. This not only introduces ambiguity at the cross-over point but also an 'XOR-like' classification problem. A single linear discriminant is insufficient for separating the two classes. However there is no folded structure for scores in the likelihood-ratio score space in Figure 4.4(b) and there is better class separation. These remarks can be extrapolated to score spaces defined on multiple component GMMs for each class. This is the fundamental reason why

the likelihood (2-class) and Fisher score spaces, when defined on a distribution modelling two classes, often perform poorly in separating those two classes.

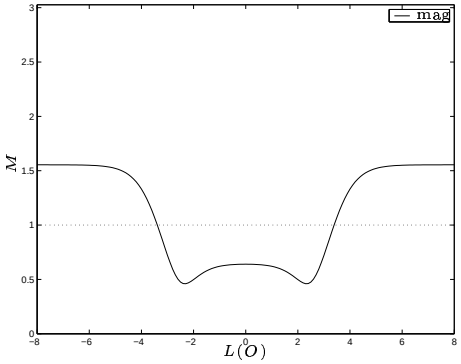
An increase in the complexity of the defining distributions may detriment class discrimination. For example, Figure 4.5(a) illustrates the log likelihoods for two classes ω_a and ω_b which are both modelled by GMMs with 2 mixture components. Class ω_a has components a1 and a2 centred at -7.0 and 1.0, and class ω_b has components b1 and b2 centred at -1.0 and 7.0. All variances coincide at 4.0. By symmetry, the MAP decision boundary is located at the origin $O = 0.0$. The plot detailing the magnification factor in Figure 4.5(b) is explained later. The score subspace defined on unit degree covariant derivatives has 4 components. Six orthogonal 2-component projections of this subspace are plotted in Figure 4.5(c). Again the trajectory is bold for points assigned to class ω_a by the MAP classifier and dotted for points assigned to class ω_b, where class priors are assumed equal. The 2-component projections show folding in score space. Such folding is a characteristic of distant mixture components, whether or not they model the same class. Fortunately, there is a clear linear separation in the projection of derivatives relative to μ_{a2} and μ_{b1}, and this indicates linear separation in the full 4-component subspace. If score spaces are defined on more complicated GMMs, then the resulting score spaces exhibit more folding. This suggests that increasing the complexity of statistical models may yield score spaces with decreasing class discrimination.

4.3.3 The nature of the classifier in score space

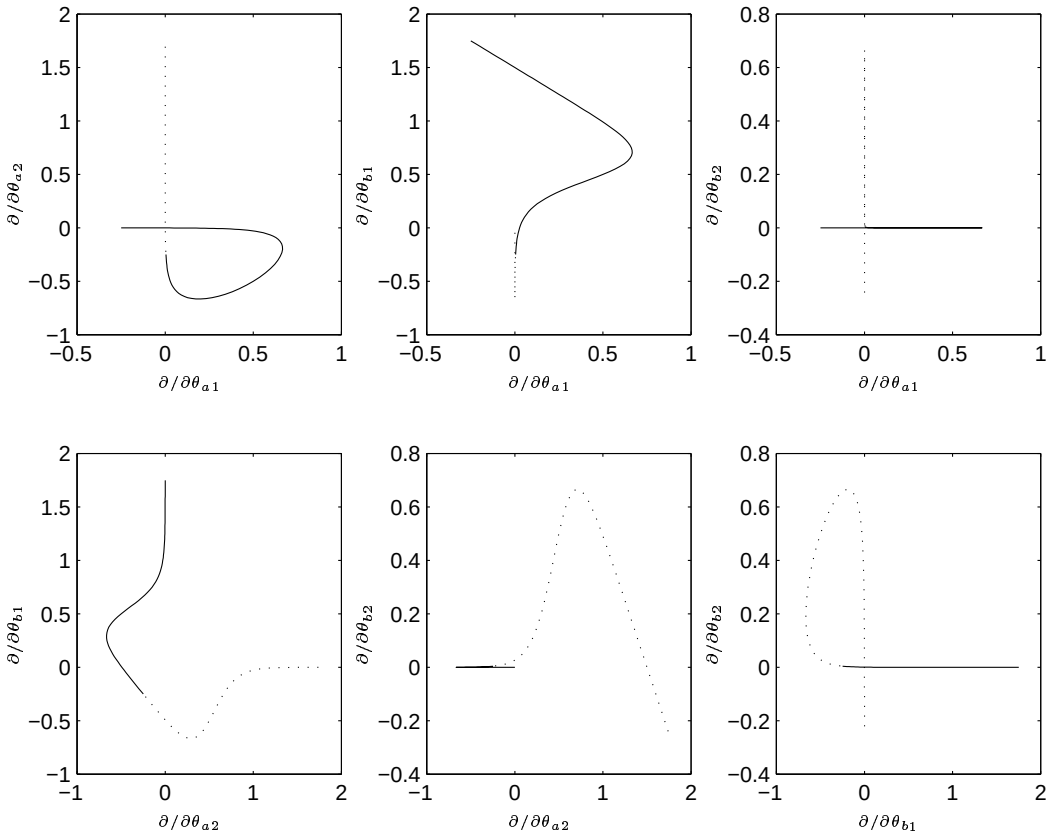
In general, any mapping which increases the separation between classes while simultaneouly decreasing the variability within each class yields a space in which it is potentially easier to distinguish classes. The adjective 'potential' is important since the ability to distinguish classes is dependent on the nature of the separation in the new space and whether the learning algorithm can take advantage of this. For example, a linear discriminant cannot take advantage of a hyperquadric separation. The experiments in this thesis favour linear discriminants in score space because of their regularisation properties and their relation to fibre bundles.



(a) Class log likelihoods



(b) Magnification factor



(c) Orthogonal projections of the unit degree subspace defined on Gaussian means

Figure 4.5: Likelihood-ratio score space based on GMMs with 2 mixture components per class

The generalisation properties of classification algorithms are often linked to the degree of smoothness of their decision boundaries. The degree of smoothness may be measured as inversely proportional to the highest nonzero derivative along the decision boundary. An analysis of the interaction between degrees of smoothness in the score space and input space is possible through defining a scalar function $h(\bar{\boldsymbol{\varphi}})$ such that $h(\bar{\boldsymbol{\varphi}}) = 0$ defines the decision boundary γ_{sc} in score space where $\bar{\boldsymbol{\varphi}} \in \varphi(\boldsymbol{\theta}_0)$. Splitting the score mapping φ into two steps $\varphi(1)$ and $\varphi(2)$ so that $\varphi = \varphi(2) \circ \varphi(1)$, then defining for fixed $\boldsymbol{\theta}_0$,

$$\varphi(1) \quad : \quad \boldsymbol{O} \mapsto \bar{\varsigma}(\boldsymbol{O}; \boldsymbol{\theta}_0) \tag{4.20}$$

$$\varphi(2) \quad : \quad \bar{\varsigma}(\boldsymbol{O}; \boldsymbol{\theta}_0) \mapsto \bar{\boldsymbol{\varphi}}(\boldsymbol{O}; \boldsymbol{\theta}_0) \tag{4.21}$$

$$h \quad : \quad \bar{\boldsymbol{\varphi}} \mapsto h(\bar{\boldsymbol{\varphi}}) \tag{4.22}$$

it is possible to define a scalar function $f = h \circ \varphi(2) \circ \varphi(1)$ where,

$$f \quad : \quad \boldsymbol{O} \mapsto f(\boldsymbol{O}) \tag{4.23}$$

The decision boundary γ_{sc} in score space then induces a decision boundary γ_{ip} in input space at the contour $f(\boldsymbol{O}) = 0$. Defining all covariant derivatives at a point $\boldsymbol{O}_0$,

$$\nabla f \quad = \quad (\nabla h \circ \varphi(2) \circ \varphi(1)) + (h \circ \nabla\varphi(2) \circ \varphi(1)) + (h \circ \varphi(2) \circ \nabla\varphi(1)) \tag{4.24}$$

and similarly for higher order covariant derivatives. If the decision boundary γ_{sc} is wholly contained within $\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$ and is differentiable to maximum order r, then since the decision boundary follows a contour on $h(\bar{\boldsymbol{\varphi}})$, it should be possible to find a scalar function h which is C^r over score space. Providing the function $\bar{\varsigma}(\boldsymbol{O}; \boldsymbol{\theta})$ is C^∞ over $L(\boldsymbol{O})$ for fixed $\boldsymbol{\theta}$ and C^∞ over $L(\boldsymbol{\theta}; S)$ for fixed $\boldsymbol{O}$, then the scalar function f is C^r over $L(\boldsymbol{O})$. However there is no means of deducing the order of differentiability of the zero contour of f, or the order of the maximum nonzero derivative along the zero contour. It is therefore nontrivial to deduce the 'degree of smoothness' for a decision boundary in input space given the 'degree of smoothness' of the boundary in score space. This problem is very similar to the decision boundaries trained in input space by SVMs with nonlinear kernels. Rather than optimise kernel parameters such as GRBF widths, an appropriate distribution p_0, scalar function $\bar{\varsigma}$ and order of expansion ϱ should be selected. The analysis is equally valid for injective or noninjective mappings φ. The restriction that γ_{sc} is contained within the image space $\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$ is unimportant since any points along γ_{sc} outside $\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$ have no projection in $L(\boldsymbol{O})$.

4.3.4 The number of training samples

This section assumes a linear discriminant trained in score space has a weight vector defined within the subspace $\varphi(\text{sp}, \mathcal{O}, \boldsymbol{\theta}_0)$ spanned by the training scores but no projection in the complimentary subspace $\varphi(\text{sp}, \mathcal{O}, \boldsymbol{\theta}_0)^{\perp}$. This assumption often applies for learning algorithms whose solution is defined only on the data. If an unseen score is located wholly within $\varphi(\text{sp}, \mathcal{O}, \boldsymbol{\theta}_0)^{\perp}$, then its classification is no better than random. Therefore, if the dimension of the image space $\dim(\varphi(\mathcal{O}, \boldsymbol{\theta}_0)) \ll \dim(\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0))$, then the generalisation ability of the decision rule is severly limited. Since $\dim(\varphi(\mathcal{O}, \boldsymbol{\theta}_0)) \leq \ell$, the generalisation ability normally decreases with fewer training samples ℓ.

At this stage, it is instructive to review the decision boundaries in score space described in Figure 4.1(b). A decision rule trained in $\varphi(\text{sp}, \mathcal{O}, \boldsymbol{\theta}_0)$ by default assigns a class decision to all points in $\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$. First, it is clear that only decision boundaries which are identical in $\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$ are identical in $L(\boldsymbol{O})$. Hence decision boundaries (1) and (2) in Figure 4.1(b) yield different solutions when mapped back to input space, but (2) and (3) yield identical solutions in input space. Decision boundaries (4) and (2) are identical within $\varphi(\text{sp}, \mathcal{O}, \boldsymbol{\theta}_0)$ but not elsewhere in $\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$, and the decision boundaries differ when mapped back to input space. The subspace $\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$ may be nonlinear and it is possible to train decision boundaries such as (5) which are continuous in $\varphi(\boldsymbol{\theta}_0)$ but not wholly defined within $\varphi(\boldsymbol{O}, \boldsymbol{\theta}_0)$. Such a boundary may correspond to two disjoint decision boundaries in input space.

4.3.5 The magnification induced by the score mapping

There are two important aspects of magnification from input space to score space.

- If a classifier is trained to minimise errors in score space, then the classifier is more sensitive to variations in samples from regions of input space which have higher magnification. If the magnification factor varies uniformly across all of input space, then the relative sensitivity to samples is transferred without loss from input space

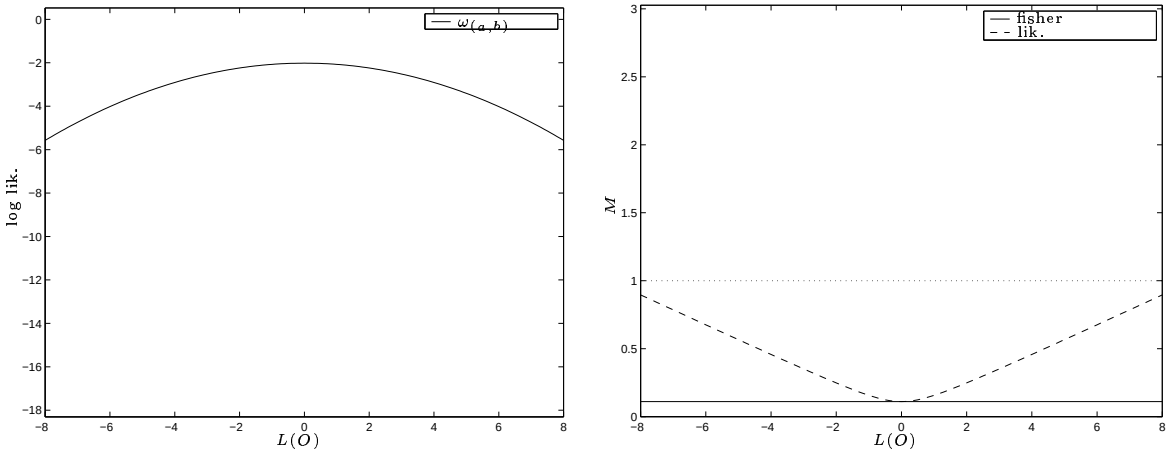
to score space.

- More important is the increase or decrease in separation between closely-spaced samples as they are mapped from input space to score space. High magnification is desirable in the locality of the decision boundary since this usually magnifies the errors of misclassified scores. Minimal error classifiers are then biased towards the correct classification of these scores, reducing the training error rate. However the potential magnification of 'noise' at the decision boundary requires strongly regularised classifiers in score space. Low magnification is desirable in regions far from the decision boundary since this implicitly deweights the contribution of distant scores to the specification of the decision boundary. Magnification factors which are non-symmetric about the decision boundary introduce implicit cost functions.

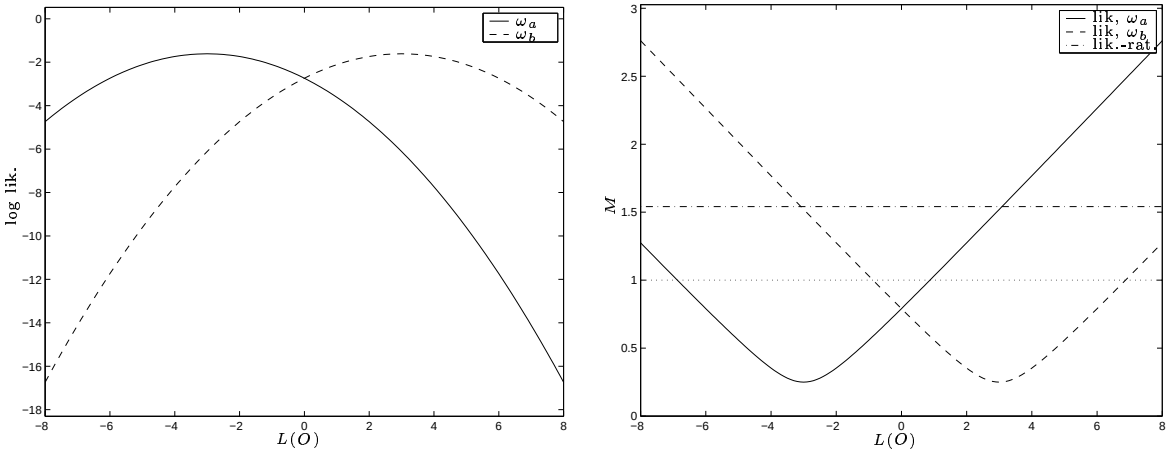
It is instructive to compare the magnification factors from input space for a variety of simple score spaces. A simple example is a 1-component input space $L(O)$ which is populated by two classes of samples, ω_a and ω_b, respectively modelled by Gaussians centred at -3.0 and 3.0 with coincident variances at 4.0. The two Gaussians may either be combined as a single two-mixture component GMD modelling both classes (labelled as class $\omega_{(a,b)}$), or kept distinct as two class-conditional Gaussians. Different score spaces are defined on these distributions. Their log likelihoods 'log lik.' and magnification factors M are plotted to the same scale in Figures 4.6(b) and 4.7(a), with similar plots for other distributions in Figures 4.6(a) and 4.7(b) (Appendix E details expressions for the metrics $g(O)$ required to calculate the magnification factors). A dotted line is drawn at unit magnification to demark regions of compression and expansion from input space[4].

In Figures 4.6 and 4.7, the magnification factor is nondecreasing with distance from the decision boundary in input space, any increase originating from the zeroth order score

[4]For the likelihood score space for class ω_q, the likelihood-ratio score space between classes ω_a and ω_b, and the likelihood (2-class) score space for both classes, the metric tensors in the tangent space to the statistical manifolds at selected points are respectively denoted by $g^{\mathrm{lk(q)}}((\boldsymbol{\theta}_q)_0)$, $g^{\mathrm{lr(a,b)}}(\boldsymbol{\xi}_0)$ and $g^{\mathrm{lk(a,b)}}(\boldsymbol{\theta}_0)$ where the corresponding statistical manifolds are $S(\boldsymbol{\theta}_q)$, $\mathcal{S}(\boldsymbol{\xi})$ and $S(\boldsymbol{\theta})$. Arbitrarily in these plots, $g^{\mathrm{lk(q)}}((\boldsymbol{\theta}_q)_0)$ is set to unity for all classes, $g^{\mathrm{lr(a,b)}}(\boldsymbol{\xi}_0)$ to Identity, and $g^{\mathrm{lk(a,b)}}(\boldsymbol{\theta}_0)$ has leading diagonal components set to unity and off-diagonal elements set to 0.5.

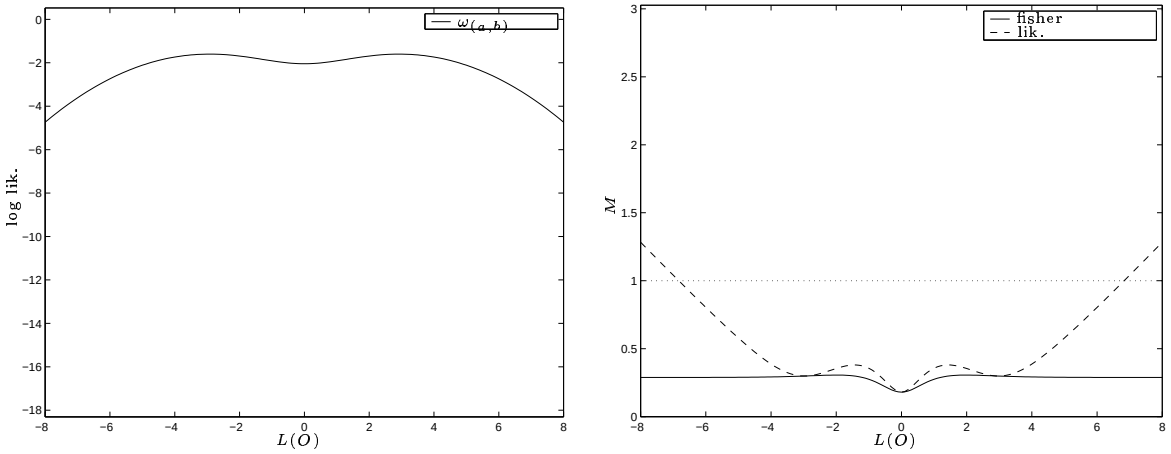


(a) likelihood (2-class) and Fisher score spaces (1-mixture component GMM)

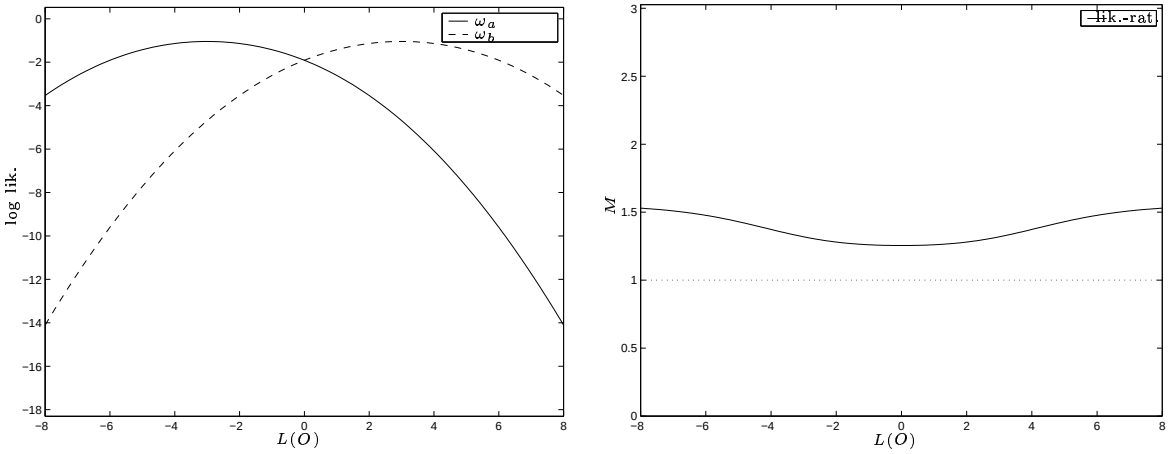


(b) likelihood-ratio score space (1-mixture component GMMs)

Figure 4.6: Log likelihoods and magnification factors for some simple score spaces



(a) likelihood (2-class) and Fisher score spaces (2-mixture component GMM)



(b) likelihood-ratio score space (2-mixture component GMMs)

Figure 4.7: Log likelihoods and magnification factors for some simple score spaces

subspace. The most important region of input space is that directly between the two classes. The likelihood (2-class) score space in Figure 4.6(a) has lowest magnification at the peak of its single defining Gaussian (mean at 0.0, variance at 9.0), just where overlap between the classes is expected. This should be detrimental to the performance of classifiers in such a score space. For reference, the corresponding Fisher score space has uniform magnification across input space since there is no contribution from the zeroth order subspace. The likelihood-ratio score space in Figure 4.6(b) has a uniform magnification since the variances of the two class-conditional Gaussians are tied. Magnification then plays no role in aiding or impeding classification at the decision boundary. In the general case of unequal variances, the magnification varies quadratically with a minimum located at a point in input space dependent on the model parameters. The minimum may be located between or outside the peaks of the two Gaussians (see Appendix E). In Figure 4.7(a), the two Gaussian components are combined with equal weight to form a 2-mixture component GMD modelling both classes. The 'trough' of the magnification profile becomes complicated by the component posteriors. The magnification as $O \to \pm\infty$ is governed by the mixture component whose posterior tends to unity. Figure 4.7(b) shows how the magnification is modified by defining a likelihood-ratio score space on 2-mixture component GMDs for each class. Class ω_a is modelled by a 2-mixture component GMM with components centred at -4.0 and -2.0, and class ω_b by a 2-mixture component GMM with components centred at 2.0 and 4.0. All variances coincide at 4.0. Another example is given in Figure 4.5(b). In general, for score spaces defined on multiple component class-conditional GMMs, the magnification in the region of overlap is perturbed by the effect of component posteriors. The 'trough' in magnification factor near to the decision boundary does not aid class discrimination.

A technique to increase classification performance is proposed in [1] whereby the region of image space near to the decision boundary is artificially magnified. If the mapping is conformal, the relative angles between scores are preserved and if so, it is unlikely that the mapping can linearly separate previously linearly inseparable scores. In [1], the regions of artificial magnification are defined by the locations of support vectors. A similar approach can also be adopted for score spaces. Neglecting the contribution of the zeroth order subspace, a mapping is required which effectively increases the gradients in the

region of overlap between classes. Such an effect may be achieved through retraining Gaussian mixture components nearer to the decision boundary. Variances subsequently narrow and gradients and magnification increase. A training criterion which repositions Gaussian components in this manner is MMIE. However, the situation is complicated by the contribution of the zeroth order subspace to the magnification. An example and details are given in Appendix E.

Appendix E also details how metric tensors induced in 1-component input space can be assembled into metric tensors for d-component input space under various conditions. Unfortunately, the task of calculating the magnification factor for a d-component space is nontrivial due to the determinant in the expression for the magnification factor.

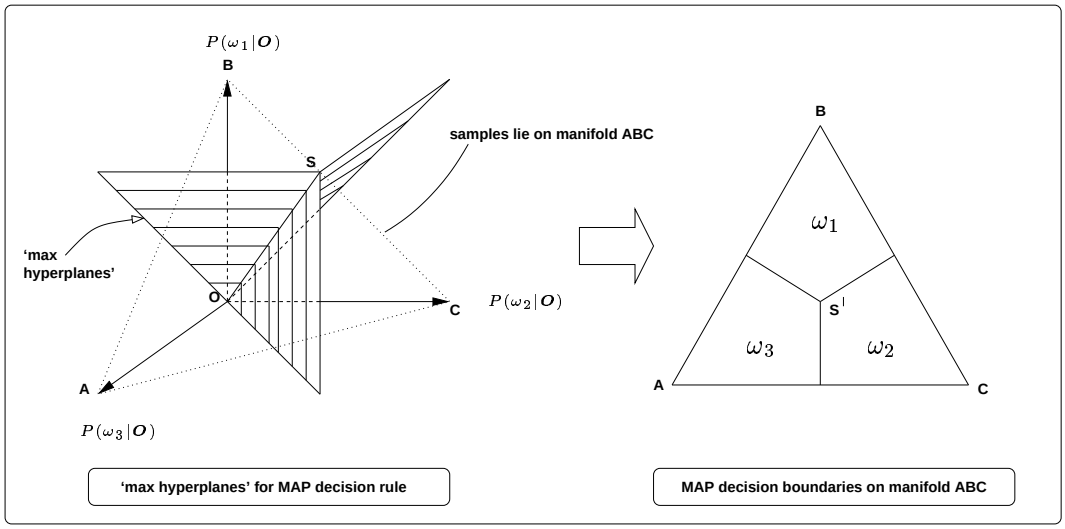
4.3.6 Summary

Overall, improved classification in score space is best served by selecting a score space, such as the likelihood-ratio score space, which yields good between-class separation. The score mapping is often noninjective but this is only detrimental to classification if two different classes map to the same region in score space. The score mapping should also increase magnification near to the decision boundary and decrease it elsewhere, though this must be coupled with a strongly regularised classifier in score space. The classifier should also model the expected separation in score space whether linear or nonlinear, but compromise this with sufficient regularisation.

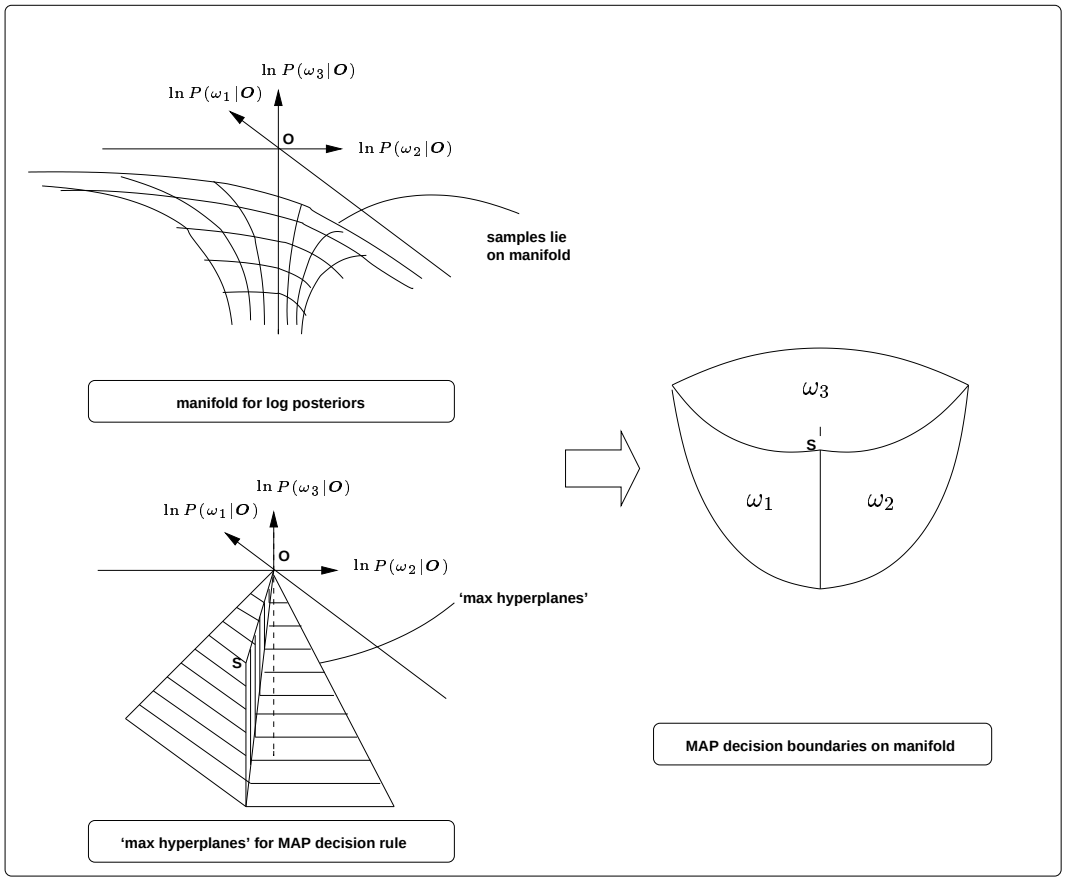
4.4 Multicategory classifiers trained in score spaces

4.4.1 Viewing the MAP decision rule as a score space classifier

An interesting set of score spaces are the zeroth order appended posterior and likelihood score spaces due to their relationship to MAP decision rules. Such a decision rule yields



(a) MAP decision rule in $\varphi^{\mathrm{psl(all)}}(0, \boldsymbol{\xi}_0)$



(b) MAP decision rule in $\varphi^{\mathrm{ps(all)}}(0, \boldsymbol{\xi}_0)$

Figure 4.8: MAP decision rules sketched in zeroth order appended posterior-based score spaces

the lowest probability of error but only when the class distributions and class priors are correct and an appropriate loss function used. For an unlabelled sample $\boldsymbol{O}$, the MAP decision rule assigns the class $\hat{\omega}_{\text{MAP}}(\boldsymbol{O})$ where,

$$\hat{\omega}_{\text{MAP}}(\boldsymbol{O}) \quad = \quad \underset{\omega_q \in L(\omega)}{\text{argmax}}\, P(\omega_q|\boldsymbol{O}) \tag{4.25}$$

and $L(\omega) = \{\omega_1, \ldots \omega_Q\}$ is the set of all competing classes. Without loss in generality the class priors are assumed equal and the notation is otherwise as described for multiple classes in Section 3.4.4. When viewed in terms of the score spaces $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$ and $\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$, this decision rule is defined by a set of Q piecewise linear decision hyperplanes radiating from a central one-dimensional axis described respectively by,

$$P(\omega_1|\boldsymbol{O}) \quad = \quad P(\omega_2|\boldsymbol{O}) = \ldots = P(\omega_Q|\boldsymbol{O})$$

or,

$$\ln P(\omega_1|\boldsymbol{O}) \quad = \quad \ln P(\omega_2|\boldsymbol{O}) = \ldots = \ln P(\omega_Q|\boldsymbol{O})$$

The MAP decision rule for $Q = 3$ is sketched in Figures 4.8(a) and 4.8(b) for the score spaces $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$ and $\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$ where the radiating hyperplanes are labelled as 'max hyperplanes'. Since there is a sum-to-unity constraint for the class posteriors, the projections $\bar{\boldsymbol{\varphi}}^{\text{psl(all)}}(\boldsymbol{O}_l; 0, \boldsymbol{\xi}_0)$ and $\bar{\boldsymbol{\varphi}}^{\text{ps(all)}}(\boldsymbol{O}_l; 0, \boldsymbol{\xi}_0)$ for $\boldsymbol{O}_l \in L(\boldsymbol{O})$ are constrained to lie on a 2-dimensional structure, in general a $(Q-1)$-dimensional structure. Each figure sketches the shape of the decision boundaries.

The $(Q-1)$-dimensional structure is linear in Figure 4.8(a) and nonlinear in Figure 4.8(b) in the frame of the score space. The MAP decision rules are respectively piecewise linear and piecewise nonlinear along these structures. For this reason, a piecewise linear multi-category classifier is sufficient for recovering the MAP decision rule when trained directly in $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$, but insufficient when trained directly in $\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$. Due to model incorrectness, it is useful to train alternative decision rules in $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$ or $\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$. This is investigated in some experiments later in the thesis.

4.4.2 Viewing MLE and MMIE learning machines from score spaces

The training criteria applied to train statistical models in the baseline experiments in this thesis may be described from the perspective of appended zeroth order linear likelihood and linear posterior score spaces $\varphi^{\text{lkl(all)}}(0, \boldsymbol{\xi}_0)$ and $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$. The statistical models are assumed mixture models in this description.

For training, there is a set of ℓ training samples $\mathcal{O} = \{\boldsymbol{O}_1, \ldots \boldsymbol{O}_\ell\}$ and $\omega(\boldsymbol{O}_l) \in L(\omega) = \{\omega_1, \ldots \omega_Q\}$ is the correct class for sample $\boldsymbol{O}_l \in L(\boldsymbol{O})$. The Maximum Likelihood Estimation (MLE) and Maximum Mutual Information Estimation (MMIE) training criteria are described in Section 2.2.2 and select the sets of distributions $\mathcal{P}_0 \in \mathcal{S}(\boldsymbol{\xi})$ to respectively maximise,

$$\mathcal{F}_{\text{MLE}}(\boldsymbol{\xi}) = \sum_{l=1}^{\ell} \ln p(\boldsymbol{O}_l|\omega(\boldsymbol{O}_l)) \tag{4.26}$$

$$\mathcal{F}_{\text{MMIE}}(\boldsymbol{\xi}) = \sum_{l=1}^{\ell} \ln P(\omega(\boldsymbol{O}_l)|\boldsymbol{O}_l) \tag{4.27}$$

where $p(\boldsymbol{O}_l|\omega_q)$ and $P(\omega_q|\boldsymbol{O}_l)$ are respectively the class likelihood and class posterior for sample $\boldsymbol{O}_l$ and class ω_q.

The MLE criterion is best viewed in the score space of linear class likelihoods $\varphi^{\text{lkl(all)}}(0, \boldsymbol{\xi}_0)$. The constraints on the class likelihoods restrict the scores to the positive hyperquadrant of score space. This is illustrated for a simple 3-class problem in Figure 4.9(a). The sketch implies an Identity metric matrix for this space though this is not necessary. For clarity, the projections of the scores on the linear plane $p(\boldsymbol{O}|\omega_1) = 0$ are sketched. The measure of importance for each sample $\boldsymbol{O}$ is the log likelihood of the correct class $\omega(\boldsymbol{O})$. Geometrically, this is the log of the norm measured parallel to the axis for $p(\boldsymbol{O}|\omega(\boldsymbol{O}))$, interepreted as the *likelihood margin of the sample*. The MLE criterion attempts to recover the distributions $\mathcal{P}_0$ which, at least locally, maximise the average likelihood margin across all training samples. This average may be interpreted as the *likelihood margin of the training set*. If each class model is a mixture model then the assignment of samples to mixture components is unknown. The EM algorithm is applied and the likelihood margin

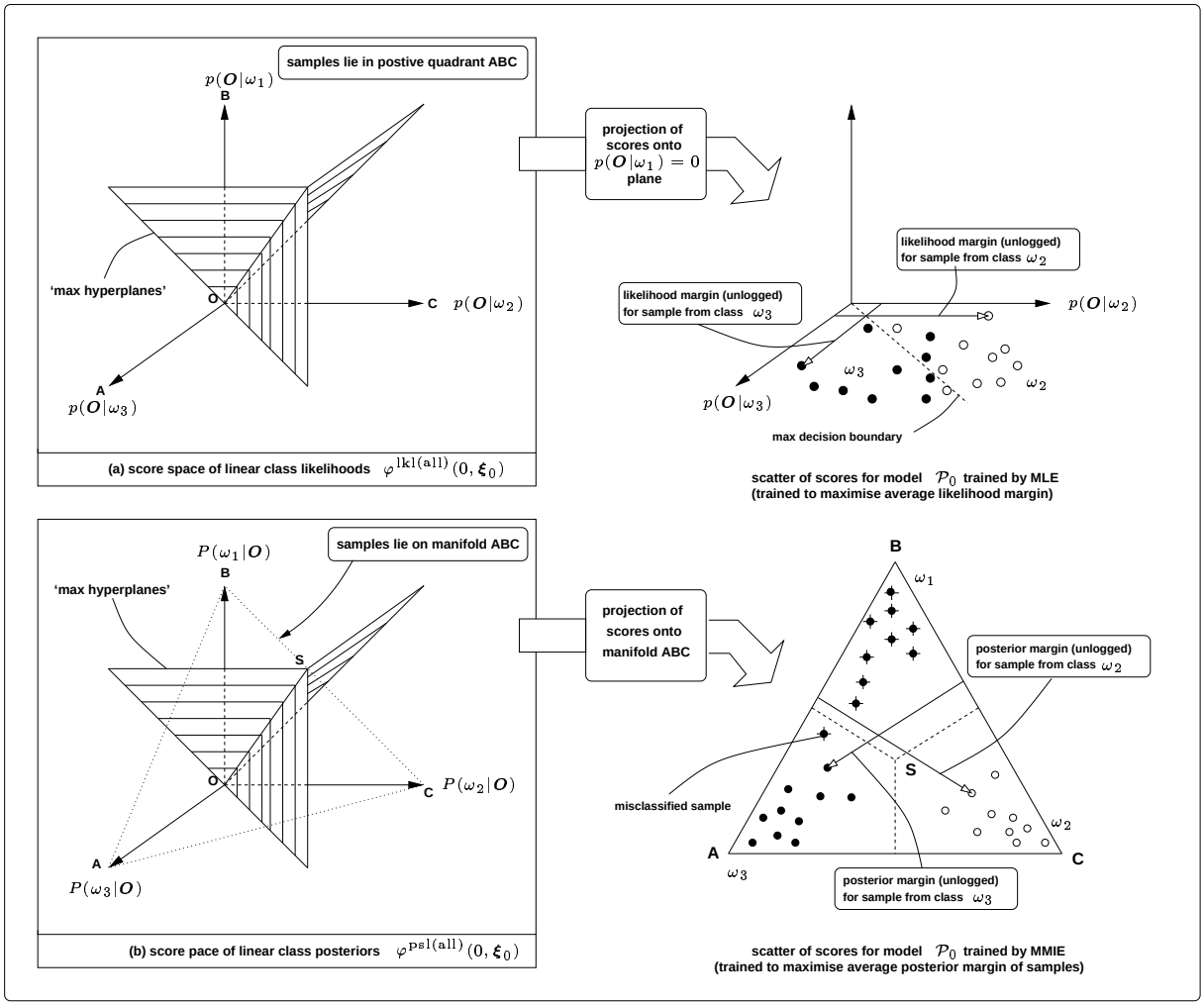


Figure 4.9: Training distributions $\mathcal{P}_0$ using MMIE and MLE from the perspective of score spaces

of the training set is nondecreasing with each EM iteration. Successive iterations modify $\mathcal{P}_0$ in such a way that the scores in score space migrate outwards from the origin in an average sense. Unfortunately, there is no mechanism which prevents scores from migrating outwards in close proximity to the max hyperplanes. Hence small amounts of incorrectness in the distributions $\mathcal{P}_0$ may be sufficient to force unseen scores across onto the false sides of the max hyperplanes. This is a weakness of MLE.

MMIE may be viewed from the appended linear posterior score space $\varphi^{\mathrm{psl(all)}}(0, \boldsymbol{\xi}_0)$. For a Q-class problem, the sum-to-unity constraint and zero-unity bounds on the class posteriors restrict the scores to a closed set on a $(Q-1)$-dimensional plane. This is illustrated for

the 3-class problem in Figure 4.9(b), where without loss in generality an Identity metric matrix is assumed for this space. The scores are restricted within the triangle ABC. In this case the measure of importance is the log posterior of the correct class. For a sample $\boldsymbol{O} \in L(\boldsymbol{O})$, this is the log of the norm measured parallel to the axis $P(\omega(\boldsymbol{O})|\boldsymbol{O})$ and is the *posterior margin of the sample.* It is instructive to view posterior margins in terms of the scaled distances[5] along the trangular plane ABC rather than distances parallel to the axis $P(\omega(\boldsymbol{O})|\boldsymbol{O})$. The distance along the plane can be interpreted as the norm between the score and the edge of the manifold lying opposite to the 'correct vertex' (for example, if $\boldsymbol{O}$ belongs to class ω_3, then the 'correct vertex' is vertex A which lies opposite to edge BC). The log of this distance is the posterior margin of the sample *along the plane.* The average of these posterior margins across all training samples may be interpreted as the scaled *posterior margin of the training set.* The MMIE criterion seeks to estimate distributions $\mathcal{P}_0$ which maximise the posterior margin of the training set. An EM-like framework (e.g. [111]) may be applied which iteratively modifies the distributions $\mathcal{P}_0$ to increase the posterior margin of the training set. However with successive iterations of MMIE, the sum-to-unity constraints force all scores, at least in an average sense, to migrate towards the 'correct vertices' of the $(Q-1)$ dimensional 'hypertriangle'. Consequently, all scores are driven away from the max hyperplanes and training error rates are typically very low. If an unseen sample is mapped onto the manifold, then if the distributions $\mathcal{P}_0$ are sufficiently regularised, then the resulting score should lie far from the max hyperplanes and can be labelled with confidence.

The likelihood and posterior margins for the training set are fundamentally different from the margin of the training set applied to an SVM.

- The margin of the training set for MLE or MMIE is the average of the margins for individual samples, but for the SVM it is the minimum of the SVM margins for the individual samples as measured in a feature space. So MLE and MMIE are 'max-average' techniques, while the SVM is a 'max-min' technique.

- For MLE and MMIE, all samples contribute an error in the calculation of the like-

[5]If the posterior margin is γ_{ps} and the posterior margin along the plane is γ_{psm}, then $\gamma_{\text{ps}} = (\sqrt{2/3})\gamma_{\text{psm}}$.

lihood and posterior margins of the training set. In this respect, the edges of the positive quadrant in $\varphi^{\text{lkl(all)}}(0, \boldsymbol{\xi}_0)$ or of the $(Q-1)$-dimensional 'hypertriangle' in $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$ fulfil a similar role to the canonical hyperplanes in the SVM feature space. However for SVM learning, only the samples which lie on the 'incorrect side' of the canonical hyperplanes contribute an error.

The SVM selects its feature space by an arbitrary but powerful mapping induced by a kernel function, and then learns a linear decision rule in this new feature space. These feature spaces may include as special cases either $\varphi^{\text{lkl(all)}}(0, \boldsymbol{\xi}_0)$ or $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$. In either of these cases the SVM cannot train the score mapping but directs all effort into training a linear discriminant in score space. However MLE and MMIE learning machines direct all effort towards training score mappings into $\varphi^{\text{lkl(all)}}(0, \boldsymbol{\xi}_0)$ or $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$, and then apply max decision rules[6] in these score spaces. Of course, both MLE and MMIE learning machines may be combined with training SVMs in score space. This approach is taken in the experiments in this thesis. Unfortunately SVMs are binary classifiers and schemes such as majority voting or multicategory extensions [109] [48] are required.

4.5 Complexity in the score mapping and score space classifier

There is degeneracy between the score mapping and score space classifier and a score mapping can sometimes be incorporated into the classifier by techniques such as kernelisation. The resultant classifier then operates in the input space rather than the score space. However this thesis prefers to maintain a distinct division between the score mapping and the score space classifier for the following reasons, rather than kernelise all algorithms and specify them in the input space.

- The division permits the application of a greater diversity of learning algorithms and feature selection techniques in score space. For example, kernelisation does

[6] Assuming equal class priors.

not permit the training of score space classifiers specified on second order central moments defined in score space.

- Learning algorithms often use training data multiple times to optimise a criterion. Mapping each sample into the score space where it is stored temporarily incurs a once-only computational burden. This improves computational efficiency over its kernelised form which must repeat calculations each time a sample is accessed[7].

- Kernelisation hides the structure of the score space.

The complexity of the classifier is the union of the complexity of the score mapping and the score space classifier. Since the total complexity is limited by the quantity of training data, then it is sensible to question whether the number of parameters required to specify the distributions in input space and the score space classifier may not be better utilised in increasing the complexity of the distributions in input space, and then applying the conventional MAP decision rule. The approach via the score mapping is favoured for the following reasons.

- Simple score space classifiers can map back to complicated decision boundaries in input space by virtue of the nonlinear score mapping. Given a fixed number of parameters to estimate, it may be difficult to obtain these decision boundaries using a MAP decision rule operating on distributions in input space.

- Strongly regularised classifiers in score space may compensate for some model incorrectness in the distributions defining the score mapping. By a similar argument to SVMs with nonlinear kernels, it is hoped that training a strongly regularised classifier in score space should yield a classifier with good regularisation in input space.

However, decoupling the score mapping from the score space classifier effectively introduces a 'filter' method of feature extraction. Hence the score mapping should be chosen with care considering the factors described earlier in the chapter.

[7]This assumes the functional complexity of the score mapping and corresponding kernel are similar. The argument would be reversed if the kernel function were of a much simpler form than the mapping, for example for SVMs with GRBF kernels.

4.6 Sequence length normalisation

The degeneracy between the score mapping and the score space classifier implies that a particular aim can often be achieved by increasing the complexity of either the mapping or classifier. An example for classifying variable length patterns is normalising the variability due to pattern length when it is considered primarily a within-class variable. For example in speech recognition, the duration of a speech unit depends on factors including context, and speaker manner and accent. This section proposes some methods which attempt to normalise the effects of pattern length by modifying the score mapping. The class of techniques is called *sequence length normalisation* originally proposed by [35]. The proposals concentrate on score spaces defined on the log likelihood scalar field and statistical models which are HMMs (relevant covariant derivatives are detailed in Appendix B.3).

4.6.1 Different forms of sequence length normalisation

Score spaces are here restricted to those defined on zeroth and first order covariant derivatives of the log likelihood scalar field for an HMM. The sample $\boldsymbol{O}_l \in L(\boldsymbol{O})$ is an observation sequence where,

$$\boldsymbol{O}_l = ((\boldsymbol{o}_l)_1, \ldots (\boldsymbol{o}_l)_{T_l}) \tag{4.28}$$

The score space is as detailed in Appendix B.3. For class ω_q and the log likelihood $l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) = \ln p(\boldsymbol{O}_l; \boldsymbol{\theta}_q)$, a member of score space $\bar{\boldsymbol{\varphi}}^{\mathrm{lk(q)}}(\boldsymbol{O}_l; 1, (\boldsymbol{\theta}_q)_0) \in \varphi^{\mathrm{lk(q)}}(1, (\boldsymbol{\theta}_q)_0)$ is,

$$\bar{\boldsymbol{\varphi}}^{\mathrm{lk(q)}}(\boldsymbol{O}_l; 1, (\boldsymbol{\theta}_q)_0) = \begin{bmatrix} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \\ \frac{\partial}{\partial \rho_q(j)} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \\ \frac{\partial}{\partial a_q(j,j)} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \end{bmatrix} \tag{4.29}$$

where the scalar field and covariant derivatives are evaluated at $(\boldsymbol{\theta}_q)_0$. The term $\rho_q(j)$ refers to a parameter from state j so $\rho_q(j) \in \{\boldsymbol{\mu}_{qjk} \forall k, \boldsymbol{\Sigma}_{qjk} \forall k, w_{qjk} : k = \{2, \ldots K\}\}$. The

score mapping is,

$$\varphi^{\mathrm{lk(q)}} : L(\boldsymbol{O}) \quad \rightarrow \quad \varphi^{\mathrm{lk(q)}}(1, \boldsymbol{O}, (\boldsymbol{\theta}_q)_0) \tag{4.30}$$

where $\varphi^{\mathrm{lk(q)}}(1, \boldsymbol{O}, (\boldsymbol{\theta}_q)_0) \subseteq \varphi^{\mathrm{lk(q)}}(1, (\boldsymbol{\theta}_q)_0)$ is the set of 'reachable' points in score space. The log likelihood $l(\boldsymbol{O}_l; \boldsymbol{\theta}_q)$ is calculated for the whole sequence using the forward-backward algorithm [80]. If the log likelihood is approximated by the most probable path, i.e. there is a Viterbi approximation, then the definition of the score space is identical to that in Equation 4.29 except $l(\boldsymbol{O}_l; \boldsymbol{\theta}_q)$ is substituted for $l(\boldsymbol{O}_l; \psi^{\mathrm{vt}}, \boldsymbol{\theta}_q)$ where ψ^{vt} refers to the state-level Viterbi path. Then the member of score space is $\bar{\boldsymbol{\varphi}}^{\mathrm{lkv(q)}}(\boldsymbol{O}_l; 1, (\boldsymbol{\theta}_q)_0) \in \varphi^{\mathrm{lkv(q)}}(1, (\boldsymbol{\theta}_q)_0)$.

The log likelihood and covariant derivatives are sensitive to variations in the length of the sample. However the normalisation of all length information is undesirable since length is sometimes still a cue for distinguishing different classes, for example different units of speech, in addition to being a source of within-class variability. For this reason, only selected components of the score space are normalised, in particular those derivatives with respect to parameters of form $\rho_q(j)$. The covariant derivatives with respect to the transition probabilities $\{a_q(j,j), \forall j\}$ remain unnormalised since transition probabilities relate most directly to duration. The zeroth order derivative also remains unnormalised though this is purely a design option.

The sequence length normalisation considered proposes modified forms of the log likelihood scalar field. The normalisation may be implemeneted as either a direct mapping from $L(\boldsymbol{O})$ to the new score space, or as a composite mapping for example from $L(\boldsymbol{O})$ to $\varphi^{\mathrm{lk(q)}}(1, (\boldsymbol{\theta}_q)_0)$ and thence into the new score space. The proposals include the following.

- $l_{\mathrm{av}}(\boldsymbol{O}_l; \boldsymbol{\theta}_q)$, the average log likelihood per observation: this is the simplest form of normalisation where,

$$l_{\mathrm{av}}(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \quad = \quad \frac{1}{T_l} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \tag{4.31}$$

The effect on the relevant components of score space is a scaling by $1/T_l$. So,

$$\bar{\boldsymbol{\varphi}}^{\mathrm{avlk(q)}}(\boldsymbol{O}_l; 1, (\boldsymbol{\theta}_q)_0) = \begin{bmatrix} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \\ \frac{\partial}{\partial \rho_q(j)} l_{\mathrm{av}}(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \\ \frac{\partial}{\partial a_q(j,j)} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \end{bmatrix} = \begin{bmatrix} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \\ \frac{1}{T_l} \frac{\partial}{\partial \rho_q(j)} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \\ \frac{\partial}{\partial a_q(j,j)} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \end{bmatrix} \tag{4.32}$$

where the scalar field and covariant derivatives are again evaluated at $(\boldsymbol{\theta}_q)_0$. This normalisation may also be implemented as the composite mapping,

$$\varphi^{\mathrm{avlk(q)}} = f \circ \varphi^{\mathrm{lk(q)}} \tag{4.33}$$

where,

$$\varphi^{\mathrm{avlk(q)}} : L(\boldsymbol{O}) \rightarrow \varphi^{\mathrm{avlk(q)}}(1, \boldsymbol{O}, (\boldsymbol{\theta}_q)_0) \tag{4.34}$$

and,

$$f : \begin{cases} x \mapsto \frac{x}{T_l} & \text{if} \quad x \in \{\frac{\partial}{\partial \rho_q(j)} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q), \forall j\} \\ x \mapsto x & \text{otherwise} \end{cases} \tag{4.35}$$

Unfortunately, if the sample is a sequence of quasi stationary segments as typical in speech, then the scaling preserves the relative contributions of each segment to the mapping into score space. This is undesirable, for example when the length of the initial or final segment of a speech unit is lengthened or shortened by coarticulation without effecting the length of the central segments. A normalisation which forces an equal contribution from each quasi stationary segment would be useful for speech and similar signals. Hence the following proposals.

- $l_{\mathrm{nmh}}(\boldsymbol{O}_l; \psi^{\mathrm{vt}}, \boldsymbol{\theta}_q)$, a normalised form of log likelihood encompasing a Viterbi approximation: denoting the most probable state-level path through the HMM by ψ^{vt} then,

$$l_{\mathrm{nmh}}(\boldsymbol{O}_l; \psi^{\mathrm{vt}}, \boldsymbol{\theta}_q) = \sum_{j=1}^{N} \frac{1}{\hat{d}(\psi^{\mathrm{vt}}, qj)} \sum_{\substack{t=1 \\ s(\psi^{\mathrm{vt}}, t) = qj}}^{T_l} \Big\{ \ln b_{qj}((\boldsymbol{o}_l)_t) + \ln a_q(j, s(\psi^{\mathrm{vt}}, t+1)) \Big\} \tag{4.36}$$

where $\hat{d}(\psi^{\mathrm{vt}}, qj)$ is the duration in state j of the model for class ω_q according to path ψ^{vt}, $s(\psi^{\mathrm{vt}}, t)$ is the state at time t according to path ψ^{vt}, and the notation is otherwise as described in Section 2.2.1. The circumflex in the duration term $\hat{d}(\psi^{\mathrm{vt}}, qj)$ relates to parameter update rules and indicates it is determined by the 'old' set of parameters and hence is regarded as a constant when differentiating with respect to model parameters. Hence,

$$\frac{\partial}{\partial \rho_q(s)} l_{\mathrm{nmh}}(\boldsymbol{O}_l; \psi^{\mathrm{vt}}, \boldsymbol{\theta}_q) = \frac{1}{\hat{d}(\psi^{\mathrm{vt}}, qs)} \sum_{\substack{t=1 \\ s(\psi^{\mathrm{vt}}, t) = qs}}^{T_l} \frac{\partial}{\partial \rho_q(s)} \ln b_{qs}((\boldsymbol{o}_l)_t) \quad (4.37)$$

Then for $\bar{\boldsymbol{\varphi}}^{\mathrm{nmh(q)}}(\boldsymbol{O}_l; 1, (\boldsymbol{\theta}_q)_0) \in \varphi^{\mathrm{nmh(q)}}(1, (\boldsymbol{\theta}_q)_0)$,

$$\begin{aligned} \bar{\boldsymbol{\varphi}}^{\mathrm{nmh(q)}}(\boldsymbol{O}_l; 1, (\boldsymbol{\theta}_q)_0) &= \begin{bmatrix} l(\boldsymbol{O}_l; \psi^{\mathrm{vt}}, \boldsymbol{\theta}_q) \\ \vdots \\ \frac{\partial}{\partial \rho_q(j)} l_{\mathrm{nmh}}(\boldsymbol{O}_l; \psi^{\mathrm{vt}}, \boldsymbol{\theta}_q) \\ \vdots \\ \frac{\partial}{\partial a_q(j,j)} l(\boldsymbol{O}_l; \psi^{\mathrm{vt}}, \boldsymbol{\theta}_q) \\ \vdots \end{bmatrix} \\ &= \begin{bmatrix} l(\boldsymbol{O}_l; \psi^{\mathrm{vt}}, \boldsymbol{\theta}_q) \\ \vdots \\ \frac{1}{\hat{d}(\psi^{\mathrm{vt}}, qj)} \frac{\partial}{\partial \rho_q(j)} l(\boldsymbol{O}_l; \psi^{\mathrm{vt}}, \boldsymbol{\theta}_q) \\ \vdots \\ \frac{\partial}{\partial a_q(j,j)} l(\boldsymbol{O}_l; \psi^{\mathrm{vt}}, \boldsymbol{\theta}_q) \\ \vdots \end{bmatrix} \end{aligned} \quad (4.38)$$

This normalisation may be implemented as a composite mapping similar to Equation 4.33. The normalisation equalises the contribution of each state to the covariant derivative with respect to $\rho_q(j)$.

- $l_{\mathrm{nms}}(\boldsymbol{O}_l; \boldsymbol{\theta}_q)$, a normalised form of the log likelihood: this is a 'soft' form of the normalisation detailed in $l_{\mathrm{nmh}}(\boldsymbol{O}_l; \psi^{\mathrm{vt}}, \boldsymbol{\theta}_q)$ where,

$$l_{\mathrm{nms}}(\boldsymbol{O}_l; \boldsymbol{\theta}_q) = \Big(\sum_{j=1}^{N} \frac{1}{\sum_{\tau=1}^{T_l} \hat{\gamma}_{qj}(\tau)} \sum_{t=1}^{T_l} \hat{\gamma}_{qj}(t) \ln b_{qj}((\boldsymbol{o}_l)_t) \Big) + \mathrm{fn}(\{a_q(j,j), \forall j\}) \quad (4.39)$$

where fn$(\cdot)$ is a generic function of its arguments. The term $\hat{\gamma}_{qj}(t)$ is the posterior probability for state j of the model for class ω_q at time t. The circumflex indicates it is calculated using the 'old' parameters in an update rule. Then,

$$\frac{\partial}{\partial \rho_q(s)} l_{\text{nms}}(\boldsymbol{O}_l; \boldsymbol{\theta}_q) = \frac{1}{\sum_{\tau=1}^{T_l} \hat{\gamma}_{qs}(\tau)} \sum_{t=1}^{T_l} \hat{\gamma}_{qs}(t) \frac{\partial}{\partial \rho_q(s)} \ln b_{qs}((\boldsymbol{o}_l)_t) \tag{4.40}$$

Hence for $\bar{\boldsymbol{\varphi}}^{\text{nms(q)}}(\boldsymbol{O}_l; 1, (\boldsymbol{\theta}_q)_0)) \in \varphi^{\text{nms(q)}}(1, (\boldsymbol{\theta}_q)_0)$,

$$\bar{\boldsymbol{\varphi}}^{\text{nms(q)}}(\boldsymbol{O}_l; 1, (\boldsymbol{\theta}_q)_0) = \begin{bmatrix} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \\ \frac{\partial}{\partial \rho_q(j)} l_{\text{nms}}(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \\ \frac{\partial}{\partial a_q(j,j)} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \end{bmatrix} = \begin{bmatrix} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \\ \frac{1}{\sum_{\tau=1}^{T_l} \hat{\gamma}_{qj}(\tau)} \frac{\partial}{\partial \rho_q(j)} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \\ \frac{\partial}{\partial a_q(j,j)} l(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \\ \vdots \end{bmatrix} \tag{4.41}$$

This normalisation can also be implemented as a composite mapping as in Equation 4.33. The sum of state posteriors across the length of the sequence is a 'soft' form of the state duration. The normalisation for the covariant derivatives of parameters of form $\rho_q(j)$ may also be related to the auxiliary function of the EM algorithm applied to the state-conditional likelihood models for the HMM states. The relation is,

$$\frac{\partial}{\partial \rho_q(j)} l_{\text{nms}}(\boldsymbol{O}_l; \boldsymbol{\theta}_q) = \frac{\partial}{\partial \rho_q(j)} Q'_{b_{qj}}(\hat{\boldsymbol{\theta}}_q, \boldsymbol{\theta}_q) \tag{4.42}$$

where $\hat{\boldsymbol{\theta}}_q \in L(\boldsymbol{\theta}_q; S)$ refers to the 'old' set of parameters in the EM parameter update rule and,

$$Q'_{b_{qj}}(\hat{\boldsymbol{\theta}}_q, \boldsymbol{\theta}_q) = \sum_{j=1}^{N} \frac{1}{\hat{p}(\boldsymbol{O}_l; \boldsymbol{\theta}_q) \sum_{\tau=1}^{T_l} \hat{\gamma}_{qj}(\tau)} Q_{b_{qj}}(\hat{\boldsymbol{\theta}}_q, \boldsymbol{\theta}_q) \tag{4.43}$$

and the term $\hat{p}(\boldsymbol{O}_l; \boldsymbol{\theta}_q) = p(\boldsymbol{O}_l; \boldsymbol{\theta}_q)|_{\boldsymbol{\theta}_q = \hat{\boldsymbol{\theta}}_q}$. From Section 6.4.3.1 of [80] [8],

$$Q_{b_{qj}}(\hat{\boldsymbol{\theta}}_q, \boldsymbol{\theta}_q) = \sum_{t=1}^{T_l} \hat{p}(\boldsymbol{O}_l, s(t) = qj; \boldsymbol{\theta}_q) \ln b_{qj}((\boldsymbol{o}_l)_t) \tag{4.44}$$

where $\hat{p}(\boldsymbol{O}_l, s(t) = qj; \boldsymbol{\theta}_q) = p(\boldsymbol{O}_l, s(t) = qj; \boldsymbol{\theta}_q)|_{\boldsymbol{\theta}_q = \hat{\boldsymbol{\theta}}_q}$.

[8] Unlike [80] where $a_q(s(t-1), s(t))$ is defined for $t = \{1, \ldots T_l\}$, the expression $Q_{a_{qj}}(\hat{\boldsymbol{\theta}}_q, \boldsymbol{\theta}_q)$ is consistent with $a_q(s(t), s(t+1))$ for $t = \{1, \ldots T_l\}$ and $a_q(s(T_l), s(T_l+1)) = 1$.

4.6.2 Relation to subsampling

The normalisation techniques described above may be related to subsampling. Subsamplng reduces the data rate while attempting to preserve the salient characteristics of the signal. For the sample $\boldsymbol{O}_l$ subsampling is the mapping,

$$f_{\text{sub}} : \boldsymbol{O}_l \mapsto \tilde{\boldsymbol{O}}_l \tag{4.45}$$

where,

$$\tilde{\boldsymbol{O}}_l = ((\tilde{\boldsymbol{o}}_l)_1, \ldots (\tilde{\boldsymbol{o}}_l)_N) \tag{4.46}$$

for some suitable positive integer N. The mapping of particular interest is of form,

$$f_{\text{sub}} : ((\boldsymbol{o}_l)_{t-c}, \ldots (\boldsymbol{o}_l)_{t+c}) \mapsto (\tilde{\boldsymbol{o}}_l)_j \tag{4.47}$$

where the length of the subsample segment is $(2c+1)$ observations, $t-c \geq 1$, $t+c \leq T_l$, and $1 \leq j \leq N$. The mapping may be simple averaging or some other linear or nonlinear transform. Modifications for endpoints are not of concern here. Two popular methods of subsampling are illustrated in Figure 4.10.

- Fixed-rate subsampling: the subsample segment length is fixed and the subsample segment sometimes called a frame, and the subsampled signal $\tilde{\boldsymbol{O}}_l$ is of fixed rate. This technique is often used as a preprocessing step for a dynamic classifier.
- Fixed-length subsampling: the subsample segment length is variable across sequences but the number of subsample segments per sequence is fixed. Hence the subsampled signal $\tilde{\boldsymbol{O}}_l$ has fixed length for each sample $\boldsymbol{O}_l$. The length of each subsample segment in a sequence may be equal or vary, for example according to the 3:4:3 ratio in [38]. This technique transforms a variable length pattern into a fixed length pattern for a static classifier.

If the sample $\boldsymbol{O}_l$ is a quasi-stationary signal as in speech, then valuable information may be lost if the statistics within each subsample segment vary significantly. However if the subsample segment boundaries are aligned with the transitions from one quasi-stationary

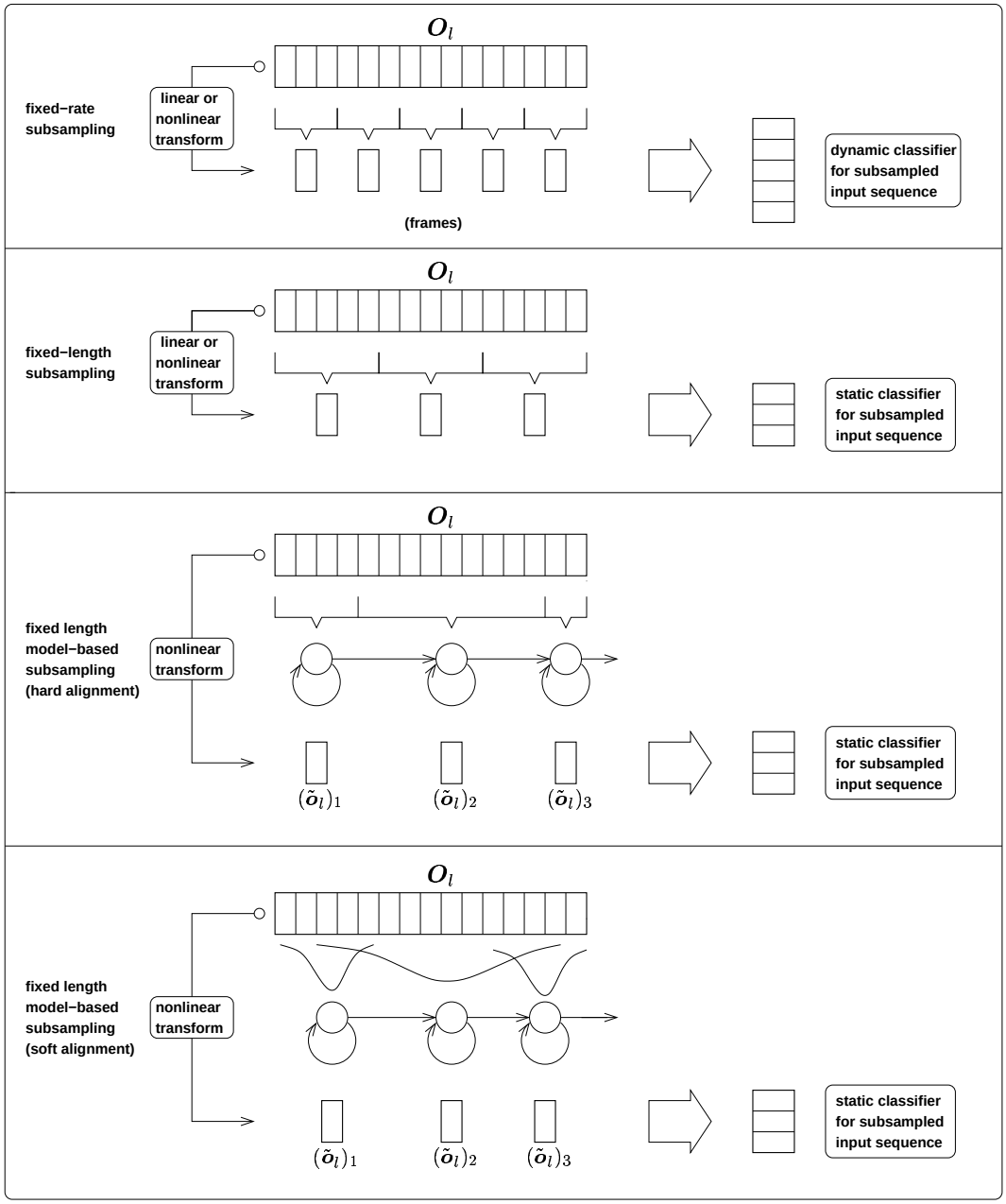


Figure 4.10: Comparing various methods of subsampling

segment of the signal to the next, then the variation of statistics within each subsample segment is minimised. The information lost or 'smoothed out' in subsampling is reduced. Such subsampling is here called *model-based subsampling*. Each quasi-stationary segment in the signal can be modelled by a single state of an HMM. If the number of HMM states is fixed, the model-based subsampling is fixed-length subsampling.

The scalar functions $l_{\text{nmh}}(\boldsymbol{O}_l; \psi^{\text{vt}}, \boldsymbol{\theta}_q)$ and $l_{\text{nms}}(\boldsymbol{O}_l; \boldsymbol{\theta}_q)$ may be viewed as implementing 'hard' and 'soft' forms of such model-based subsampling. First, $l_{\text{nmh}}(\boldsymbol{O}_l; \psi^{\text{vt}}, \boldsymbol{\theta}_q)$ is consistent with the subsampling f_{sub} for $j = \{1, \ldots N\}$ where,

$$\ln b_{qj}((\tilde{\boldsymbol{o}}_l)_j) = \frac{1}{\hat{d}(\psi^{\text{vt}}, qj)} \sum_{\substack{t=1 \\ s(\psi^{\text{vt}}, t) = qj}}^{T_l} \ln b_{qj}((\boldsymbol{o}_l)_t) \tag{4.48}$$

Providing that the transition probabilities of the HMM are modified such that for the forward transitions $j = \{1, \ldots N\}$,

$$\ln a_q(j, j+1) = \frac{1}{\hat{d}(\psi^{\text{vt}}, qj)} \sum_{\substack{t=1 \\ s(\psi^{\text{vt}}, t) = qj}}^{T_l} \ln a_q(j, s(\psi^{\text{vt}}, t+1)) \tag{4.49}$$

then,

$$l_{\text{nmh}}(\boldsymbol{O}_l; \psi^{\text{vt}}, \boldsymbol{\theta}_q)\Big|_{\boldsymbol{\theta}_q = (\boldsymbol{\theta}_q)_0} = l(\tilde{\boldsymbol{O}}_l; \boldsymbol{\theta}_q)\Big|_{\boldsymbol{\theta}_q = \boldsymbol{\theta}'_q} \tag{4.50}$$

where $(\boldsymbol{\theta}_q)_0$ and $\boldsymbol{\theta}'_q$ are the parameterisations respectively prior to and following the modifications to the transition probabilities. Then the subsampling process preserves the normalised log likelihood of the sequence. Unlike subsampling which is not model-based, for example simple averaging, the subsampled signal $\tilde{\boldsymbol{O}}$ always contains typical samples as measured by the state output distributions $b_{qj}((\tilde{\boldsymbol{o}}_l)_j), \forall j$. However the mapping f_{sub} is noninjective, i.e. for the jth subsample segment in $\boldsymbol{O}_l$, there are an infinite number of possible solutions for $(\tilde{\boldsymbol{o}}_l)_j$, corresponding to any point on the relevant contour of the state-conditional likelihood distribution. However another advantage of the normalisation is that,

$$\frac{\partial}{\partial \rho_q(j)} l_{\text{nmh}}(\boldsymbol{O}_l; \psi^{\text{vt}}, \boldsymbol{\theta}_q)\Big|_{\boldsymbol{\theta}_q = (\boldsymbol{\theta}_q)_0} = \frac{\partial}{\partial \rho_q(j)} l(\tilde{\boldsymbol{O}}_l; \boldsymbol{\theta}_q)\Big|_{\boldsymbol{\theta}_q = \boldsymbol{\theta}'_q} \tag{4.51}$$

Hence the subsampling preserves the mapping into score space for the covariant derivatives with respect to state parameters[9] of form $\rho_q(j)$. A similar approach defines a 'soft' form of the mapping f_{sub} consistent with,

$$\ln b_{qj}((\tilde{\boldsymbol{o}}_l)_j) = \frac{1}{\sum_{\tau=1}^{T_l} \hat{\gamma}_{qj}(\tau)} \sum_{t=1}^{T_l} \hat{\gamma}_{qj}(t) \ln b_{qj}((\boldsymbol{o}_l)_t) \tag{4.52}$$

With similar modifications to the forward transition probabilities and similarly to above, the subsampling preserves the normalised log likelihood and the mapping into the score subspace defined on parameters of form $\rho_q(j)$. These 'hard' and 'soft' forms of model-based subsampling are illustrated in Figure 4.10.

4.7 Summary

This chapter has introduced various score spaces based on zeroth and unit degree covariant derivatives of scalar fields, typically those defined on the log likelihoods or log posteriors of samples. Factors affecting classification performance in score space were discussed including the definition of the score space and its possibly noninjective nature, the nature of the score space classifier, the number of training samples, and the relative magnification near to the decision boundary for mapping from input space to score space. Appended zeroth order score spaces were also shown to be a useful means of viewing some training algorithms for GMM classifiers. The chapter also discussed some of the advantages of maintaining a division between the score mapping and score space classifier. Sequence length normalisation was presented as a normalisation technique suitable for variable length patterns.

[9] This property does not require modifying the state transition probabilities and the right hand side of Equation 4.51 may also be evaluated at $\boldsymbol{\theta}_q = (\boldsymbol{\theta}_q)_0$.

Chapter 5

Classifying fixed length patterns

Having introduced score spaces and their characteristics, this chapter applies score spaces to the classification of simple fixed length patterns. The experiments concentrate on artificially generated data and vowel data. For the artificial data the source distributions and optimal classifier are known, whereas the vowel data presents a more demanding task since the source distributions are unknown and almost certainly more complicated than any proposed models. Though simple, these experiments permit an investigation of score spaces, particularly those defined on class posterior scalar fields, without a large computational burden. The experiments are illustrative and are not intended as an exhaustive investigation of different score spaces and classifiers.

5.1 Experimental details

5.1.1 Statistical models and classifiers

Following the notation and definitions in Section 2.2.1, a set of Q class-conditional statistical models is denoted $\mathcal{S}(\boldsymbol{\xi})$ with an estimate $\mathcal{P}_0 \in \mathcal{S}(\boldsymbol{\xi})$. The set of statistical models may be viewed as forming the base manifold for a fibre bundle. Considering the associated

score space as simply another feature space, a set of statistical models may be proposed for score space with estimate $\mathcal{P}_{\text{sc}}$.

In this chapter, statistical models in input space and score space are restricted to GMMs with variable numbers of mixture components, but an identical number for each class. Except for a single comparison in Section 5.3.2, covariance matrices are always diagonalised. A set of distributions $\mathcal{P}_0$ or $\mathcal{P}_{\text{sc}}$ with n mixture components per class is denoted $\mathcal{P}_0(n)$ or $\mathcal{P}_{\text{sc}}(n)$ respectively. GMM classifiers are defined which assign an unlabelled sample to the class whose GMM yields the highest output. Since class priors are assumed equal, GMM classifiers are effectively MAP classifiers. For clarity, this chapter calls a GMM classifier operating in input space a 'MAP classifier' and reserves the term 'GMM classifier' for a score space classifier.

5.1.2 Artificial dataset

There are 4 classes in a 2-component input space. The samples for each class were generated by GMDs with 3 mixture components. These GMDs formed the set of source distributions $\mathcal{P}''$. The locations of the centroids of each Gaussian component in input space are detailed in Figure 5.1, and corresponding weights given alongside in italics. All Gaussian mixture components were given identical covariance matrices set to $3\boldsymbol{I}$, where $\boldsymbol{I}$ is the Identity matrix. In the dataset there are 100 and 500 samples per class for training and testing respectively.

5.1.3 Deterding vowel dataset

The Deterding database [7] describes the steady state portions of 11 vowels in English spoken by 15 speakers, 8 of which were male and 7 female. The speech samples were originally collected for investigating speaker normalisation. As detailed in [7], each speech utterance was low-pass filtered at 4.7kHz and then digitised into 12 bits with a sampling rate of 10kHz. The steady state portion of the vowel in each utterance was then parti-

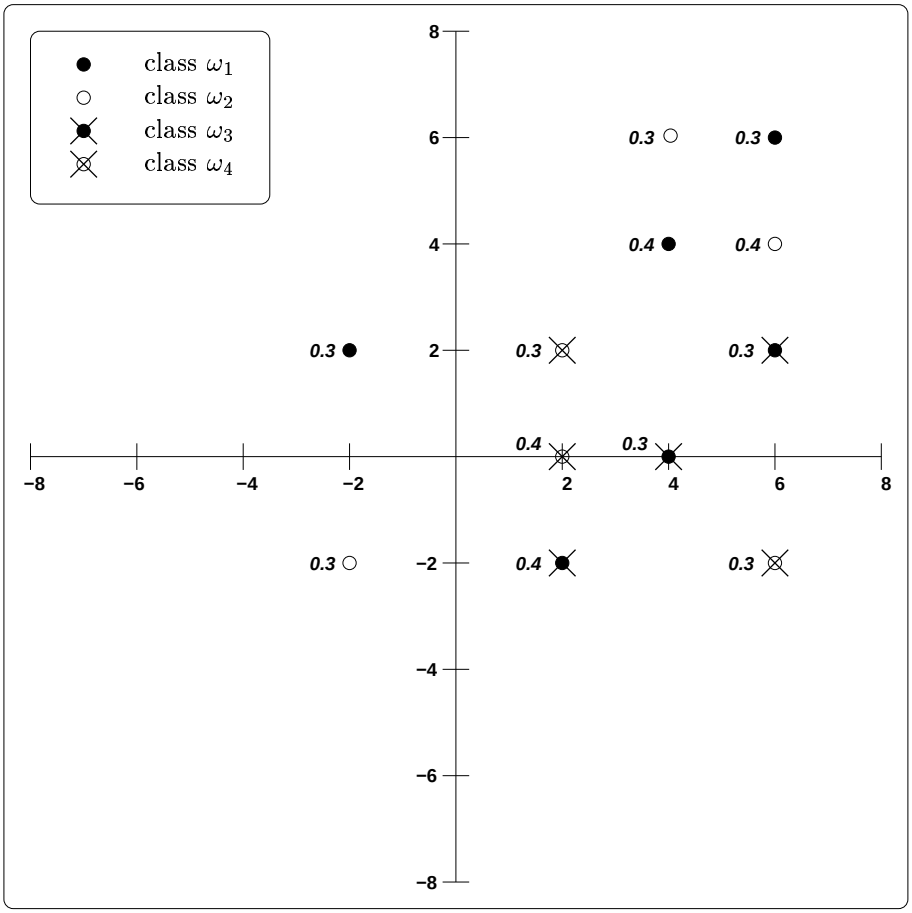


Figure 5.1: The centroids and weights of each Gaussian component for the 4 class-conditional GMDs used to generate the artificial dataset

tioned into six 512 sample Hamming windowed segments and linear prediction reflection coefficients calculated. From this, 10 log area parameters were calculated yielding a vector of 10 components. One utterance therefore yielded one vector or sample. Each speaker spoke 6 utterances per vowel and 11 vowels in total. Hence each speaker provided 66 samples. The database of samples was partitioned into distinct training and test sets. The training set contained 528 samples spoken by four male and four female speakers. The test set contained 462 samples spoken by four male and three female speakers.

5.1.4 Training distributions and classifiers

The whole training set was used to train both $\mathcal{P}_0$ and $\mathcal{P}_{\text{sc}}$. The distributions $\mathcal{P}_0$ defined the MAP classifier and the distributions $\mathcal{P}_{\text{sc}}$ the GMM classifier. MAP and GMM classifiers were trained and tested using HTK [49] [114].

When trained by ML estimation, each class-conditional GMM was trained from a single Gaussian mixture component with mean and covariance set to the global mean and diagonalised global covariance. Next the number of mixture components in each GMM was, if necessary, increased to the required number by mixture splitting [114]. Each application of mixture splitting increased the number of Gaussian components by one, and was followed by retraining the GMMs with 5 iterations of an embedded training version of the Baum-Welch algorithm [114]. Unless otherwise stated, all GMM parameters were updated, i.e. the Gaussian weights, means and covariances. Also unless otherwise stated, the covariances of all Gaussian components in the set of GMMs were either tied and updated, or simply fixed to the diagonalised global covariance. Appropriate training parameters were used.

GMM classifiers were also trained by MMI estimation [1]. The GMM for each class was initialised with the corresponding GMD trained by ML estimation. For each training sample, a numerator lattice was generated with 1 output hypothesis and a denominator

[1] In this thesis the criterion was strictly 'conditional maximum likelihood estimation' since class priors remained fixed during training [111].

lattice with one output hypotheses per class. The lattices were generated using HTK's `HVite` and with the initial ML-trained GMDs. In training the GMMs, either the Gaussian weights and means were updated, or the Gaussian weights, means and covariances. The GMMs were trained with a variable number of iterations of MMI training using HTK's `HERest` modified to support this training technique [111]. The likelihood scale factor κ and E-parameter were both variable, and other MMI parameters were kept fixed at sensible values. Appropriate training parameters were applied.

Testing was performed using HTK's `HVite` Viterbi decoder which implemented the MAP classifier in input space and the GMM classifier in score space.

The calculation of unit degree covariant derivatives was performed by modifications made to SVM^{light} version 3.02. SVM training and testing was performed using SVM^{light} version 4.00 [54] [53]. Some of the processing to form score spaces was performed with MATLAB version 5 [68].

5.2 Experiments on the artificial dataset

In these experiments, GMMs were trained by ML estimation according to the procedure described in Section 5.1.4 and using HTK version 3.1 [49]. The optimal classifier, i.e. the MAP classifier based on the correct distributions $\mathcal{P}''$, yielded 43.75% test error rate and 45.75% training error rate. Such a classifier, assuming a 'faithful' generation process, is guaranteed to yield the lowest probability of error. The training error rate was higher than the test error rate. This was probably an artefact of the small size of the dataset and perhaps because there was some bias in the generation process, for example due to the pseudo-random nature of the implementation. With infinite training and test sets and a 'faithful' generation process, both the training and test error rates should converge to the probability of error.

Different estimates $\mathcal{P}_0$ were obtained by training GMMs of varying complexity and the resulting MAP classifiers compared in Figure 5.2 with respect to test and training error

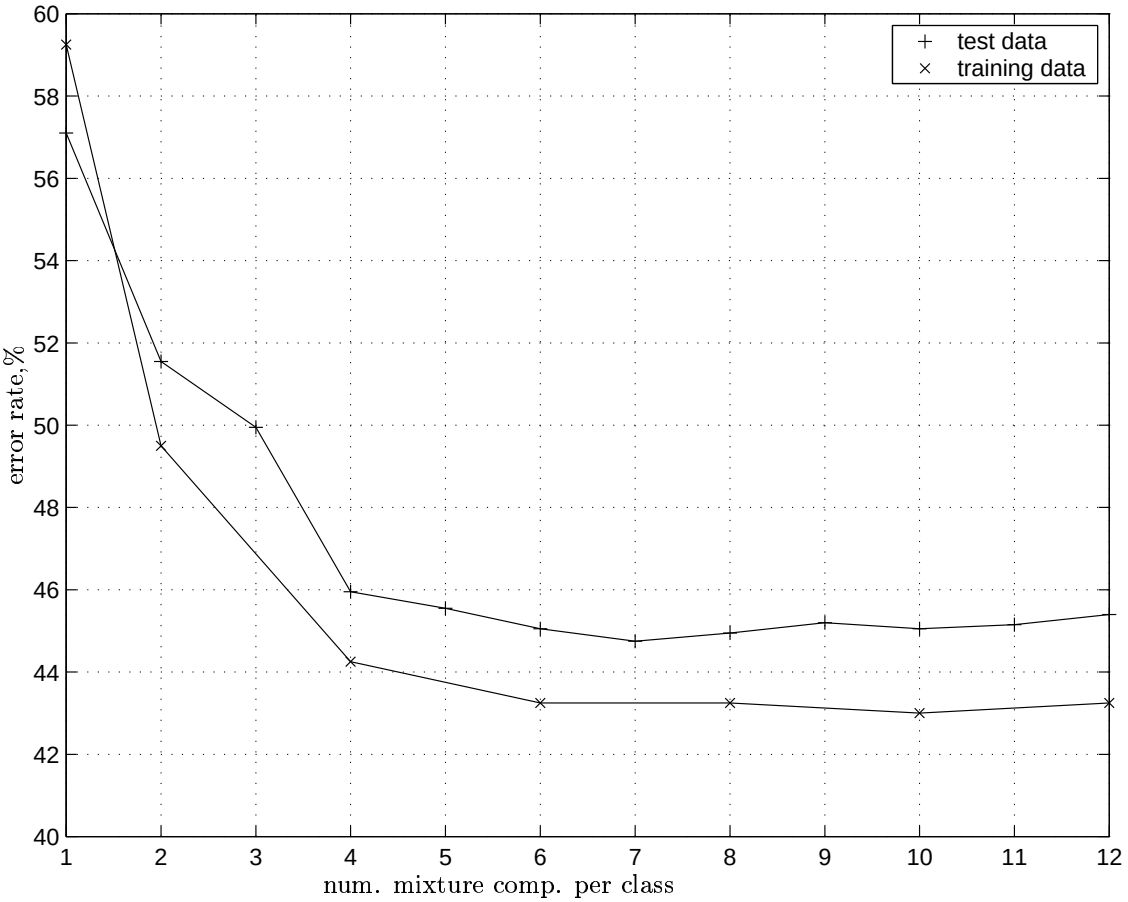


Figure 5.2: Test and training error rates for MAP classifiers defined on class-conditional GMDs with different numbers of mixture components

rates. As expected, training error rates were higher than for the optimal MAP classifier based on the correct distributions $\mathcal{P}''$. Generally the training error rate decreased with increasing complexity while the test error rate showed a slight trough. This illustrates overtraining.

The MAP classifier is a max decision rule in the space of linear class posteriors $\varphi^{\mathrm{psl(all)}}(0, \boldsymbol{\xi}_0)$ or log class posteriors $\varphi^{\mathrm{ps(all)}}(0, \boldsymbol{\xi}_0)$. Since there is no guarantee that this decision rule is optimal when the distributions $\mathcal{P}_0$ are incorrect, alternative GMM classifiers were trained in these two score spaces with the freedom to depart from the max decision rule.

First for reference, the distributions defining the score mapping $\mathcal{P}_0$ were forced to coincide with the correct distributions $\mathcal{P}''$. The max decision rule is then the optimal decision rule

in $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$ or $\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$. Distributions with 1, 2, 4 and 6 mixture components per class were then trained in these score spaces and used to define GMM classifiers. The test error rates for the GMM and MAP classifiers are detailed in Table 5.1. For simple GMM classifiers based on 1 and 2 mixture components per class, classifiers in the space of linear class posteriors yielded lower test error rates than classifiers in the space of log class posteriors. This is expected since, as illustrated in Section 4.4.1, the max decision rule has a piecewise linear projection onto the 3-dimensional structure of scores in $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$ but a piecewise nonlinear projection onto the structure of scores in $\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$. In the latter case it is difficult to model the nonlinear max decision boundaries even with more complicated GMM classifiers. An increase in classifier complexity in the space of linear class posteriors $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$ resulted in a marked increase in test error rate. This is not unexpected since the max decision rule, though only guaranteed to yield the lowest test error rate for infinite tests sets and assuming a 'faithful' generation process, is still expected to be a good decision rule for this test set. It can be sufficiently modelled by distributions with a single mixture component per class, and the extra degrees of freedom from more complicated distributions simply model sample 'noise'.

classifier type	num. mixture comp. per class	score spaces	
		$\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$	$\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$
MAP	-	43.75	43.75
GMM	1	44.20	47.05
GMM	2	44.05	47.75
GMM	4	50.90	48.55
GMM	6	52.35	50.35

Table 5.1: Percentage test error rates for GMM classifiers in score spaces and the MAP classifier in input space (the MAP classifier and score spaces were defined on correct distributions so $\mathcal{P}_0 = \mathcal{P}''$)

It is useful to extract single components from the appended posterior score space and measure class discrimination. Score spaces were defined on single log class posteriors, i.e. score spaces of the form $\varphi^{\text{ps(q)}}(0, \boldsymbol{\xi}_0)$ where $\mathcal{P}_0$ was set to the correct distributions

$\mathcal{P}''$. Table 5.2 details the test error rates yielded by piecewise linear classifiers based on distributions $\mathcal{P}_{\text{sc}}(1)$. The highest test error rates are for score spaces defined on the posteriors for ω_1 and ω_2. Inspection of Figure 5.1 reveals that these classes are the most distant from the global centroid of the training samples. Their posterior distributions were probably not as informative for distinguishing samples located in the confusible region near the global centroid.

class	test error rate,%
ω_1	59.05
ω_2	56.85
ω_3	56.35
ω_4	55.70

Table 5.2: Percentage test error rates for GMM classifiers based on $\mathcal{P}_{\text{sc}}(1)$ trained in the score spaces $\varphi^{\text{ps(q)}}(0, \boldsymbol{\xi}_0)$, $q \in \{1, 2, 3, 4\}$

The test error rates in Table 5.2 are much higher than the equivalent test error rate for a GMM classifier based on $\mathcal{P}_{\text{sc}}(1)$ in the appended posterior score space $\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$ at 47.05% in Table 5.1. This is as expected since the appended posterior score space is defined on all class posteriors rather than a single class posterior. The cost of this extra information is an increase in the size of the score space.

In realistic situations where $\mathcal{P}_0$ is incorrect, the max decision rule implementing the MAP classifier is no longer guaranteed to be optimal. Alternative decision rules may outperform it. For example, $\mathcal{P}_0$ was set to the distribution $\mathcal{P}_0(3)$. As detailed in Figure 5.2, the max decision rule yielded 49.95% test error rate. However a GMM classifier constructed on $\mathcal{P}_{\text{sc}}(1)$ in the score space of log class posteriors $\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$ defined on $\mathcal{P}_0(3)$ yielded a lower test error rate of 48.70%. This illustrates that freeing a score space classifier from a max decision rule is sometimes beneficial when the defining distributions for the score mapping are incorrect.

It is useful to introduce unit degree covariant derivatives into posterior score spaces. Despite the good performance of score spaces based on linear posteriors, covariant derivatives

were only defined on log posteriors. As explained with regard to fibre bundles in Section 3.4.1, defining bundles on log scalar fields ensures that fibres with similar semantics to the base manifold are exponential families. If this property is not required, or if score spaces are simply regarded as alternative feature spaces, then covariant derivatives may be defined on linear posteriors. However the restriction to log scalar fields is enforced in the experiments of this chapter. Posterior score spaces based on zeroth and first degree covariant derivatives were defined on $\mathcal{P}_0(3)$ and $\mathcal{P}_0(7)$ (the MAP classifier defined on $\mathcal{P}_0(7)$ yielded the lowest test error rate for MAP classifiers at 44.75%, while the MAP classifier defined on $\mathcal{P}_0(3)$ provided a suitable contrast since it was simpler but yielded a higher test error rate of 49.95%). GMM classifiers were defined on reduced and generalised appended posterior score spaces[2] $\varphi^{\mathrm{psr(all)}}(1, \boldsymbol{\xi}_0)$ and $\varphi^{\mathrm{psg(all)}}(1, \boldsymbol{\xi}_0)$. Table 5.3 compares performance for GMM classifiers based on $\mathcal{P}_{\mathrm{sc}}(1)$, and the best performance when $\mathcal{P}_{\mathrm{sc}}$ was set to distributions with 1, 2, 4 and 6 mixture components per class.

defining distributions	classifier type	input space	score spaces		
			$\varphi^{\mathrm{ps(all)}}(0, \boldsymbol{\xi}_0)$	$\varphi^{\mathrm{psr(all)}}(1, \boldsymbol{\xi}_0)$	$\varphi^{\mathrm{psg(all)}}(1, \boldsymbol{\xi}_0)$
$\mathcal{P}_0(3)$	GMM ($\mathcal{P}_{\mathrm{sc}}(1)$)	-	48.70	51.55	51.40
	GMM (best)	-	48.70	46.90	45.10
	MAP	49.95	-	-	-
$\mathcal{P}_0(7)$	GMM ($\mathcal{P}_{\mathrm{sc}}(1)$)	-	47.45	47.50	47.90
	GMM (best)	-	47.45	45.75	46.25
	MAP	44.75	-	-	-

Table 5.3: Percentage test error rates for GMM classifiers in posterior score spaces and MAP classifiers in input space

In general, the performance of the GMM classifiers improved as the estimates improved from $\mathcal{P}_0(3)$ to $\mathcal{P}_0(7)$. The results also show performance improvement from introducing covariant derivatives. However the piecewise linear classifiers, as implemented by GMM classifiers operating on distributions $\mathcal{P}_{\mathrm{sc}}(1)$, did not have sufficient complexity to take

[2]The appended posterior score space $\varphi^{\mathrm{ps(all)}}(1, \boldsymbol{\xi}_0)$ contains no more information than the score space $\varphi^{\mathrm{psr(all)}}(1, \boldsymbol{\xi}_0)$. However the presence of repeated components may influence the training algorithm and yield different classifiers.

advantage of the extra discriminative information. The results also suggest that when $\mathcal{P}_0$ is a poor estimate for the correct distributions $\mathcal{P}''$, then posterior score spaces based on $\mathcal{P}_0$ may be used to improve performance over the MAP classifier based on $\mathcal{P}_0$. However when $\mathcal{P}_0$ is a better estimate of $\mathcal{P}''$, then there is probably less to gain by training score space classifiers. Many practical problems at best offer poor estimates $\mathcal{P}_0$ due to lack of training samples, limitations in the estimator, or a simplistic knowledge of the functional form of the source distributions $\mathcal{P}''$. For this reason, score spaces based on covariant derivatives are worthy of further investigation. Experiments should be performed on real data where the source distributions are unknown and most probably more complicated.

Experiments were also performed on posterior score spaces defined on single classes, i.e. score spaces of form $\varphi^{\mathrm{ps(q)}}(1, \boldsymbol{\xi}_0)$. They showed useful class discriminative information existed in these spaces. For example, a test error rate of 46.90% was obtained for a GMM classifier defined on $\mathcal{P}_{\mathrm{sc}}(1)$ in $\varphi^{\mathrm{ps}(2)}(1, \boldsymbol{\xi}_0)$ based on $\mathcal{P}_0(7)$. These score spaces are subspaces within the corresponding reduced appended posterior score space. Experiments were inconclusive as to whether the posterior score spaces defined on single classes outperformed the reduced and generalised appended posterior score spaces. More difficult tasks are required.

5.3 Experiments on the Deterding vowel dataset

In these experiments, GMMs were either trained by ML or MMI estimation as described in Section 5.1.4. ML training and testing were via HTK version 3.1.1 [49]. When GMMs were trained by MMI estimation, then a version of HTK modified to support this training was used (see the Acknowledgments). Further modifications were implemented for both versions permitting larger input spaces.

5.3.1 MAP classifiers

First MAP classifiers were trained with varying numbers of mixture components per class. The covariance matrices were either globally tied and updated (referenced here as GMMs with 'updated covariances'), or kept fixed at the global covariance of the training samples (referenced here as GMMs with 'global covariances'). These effectively implemented different metrics for input space. The resulting test and training error rates are plotted in Figure 5.3. Training error rates decreased to zero for the GMDs with global covariances and virtually to zero for GMDs with updated covariances. This, coupled with the increasing test error rates, illustrates overtraining. The worse performance of GMMs with global covariances may simply be due to their poorer class modelling ability. The lowest test error rates were for classifiers based on GMMs with updated covariances at 41.6% test error rate for 5 mixture components per GMD, and then 42.0% for 3 mixture components per GMD. In the remainder of this section, these two sets of GMDs with updated covariances are labelled as $\mathcal{P}_0(5)$ and $\mathcal{P}_0(3)$ respectively. In [15], 16 component class-conditional GMDs with diagonal covariances were trained and attained a test error rate of 37.9%.

5.3.2 Score spaces defined on zeroth degree covariant derivatives

The distributions $\mathcal{P}_0(3)$ were selected as baseline distributions since the number of parameters, and hence size of score spaces, was smaller than for $\mathcal{P}_0(5)$. As for the experiments on artificial data in Section 5.2, GMM classifiers of varying complexity were constructed in simple zeroth order appended posterior-based score spaces $\varphi^{\mathrm{psl(all)}}(0, \boldsymbol{\xi}_0)$ and $\varphi^{\mathrm{ps(all)}}(0, \boldsymbol{\xi}_0)$, but also in the likelihood equivalents $\varphi^{\mathrm{lkl(all)}}(0, \boldsymbol{\xi}_0)$ and $\varphi^{\mathrm{lk(all)}}(0, \boldsymbol{\xi}_0)$. The performances of the resulting classifiers are plotted in Figure 5.4 and can be compared with the MAP classifier defined by max decision rules in these spaces (since class priors were assumed fixed and equal). Class discrimination is poor in the appended linear likelihood score space $\varphi^{\mathrm{lkl(all)}}(0, \boldsymbol{\xi}_0)$ but much improved in the appended linear posterior score space $\varphi^{\mathrm{psl(all)}}(0, \boldsymbol{\xi}_0)$. This is possibly since class posteriors are inherently normalised for variations in acoustic conditions. However this reasoning is by itself insufficient to explain why the space of log likelihoods $\varphi^{\mathrm{lk(all)}}(0, \boldsymbol{\xi}_0)$ outperforms the space of log posteriors $\varphi^{\mathrm{ps(all)}}(0, \boldsymbol{\xi}_0)$. The MAP

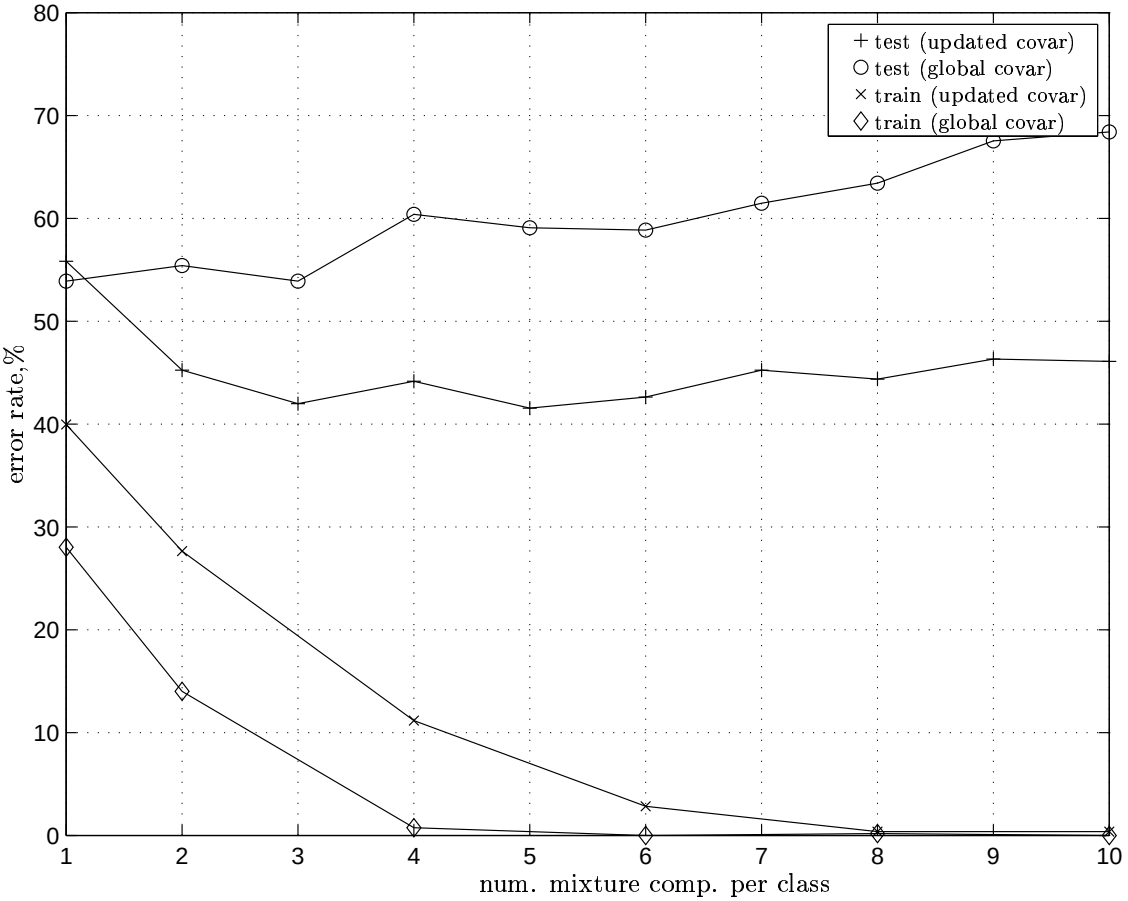


Figure 5.3: Training and test error rates for MAP classifiers in input space as the number of mixture components per class and type of covariance matrices vary

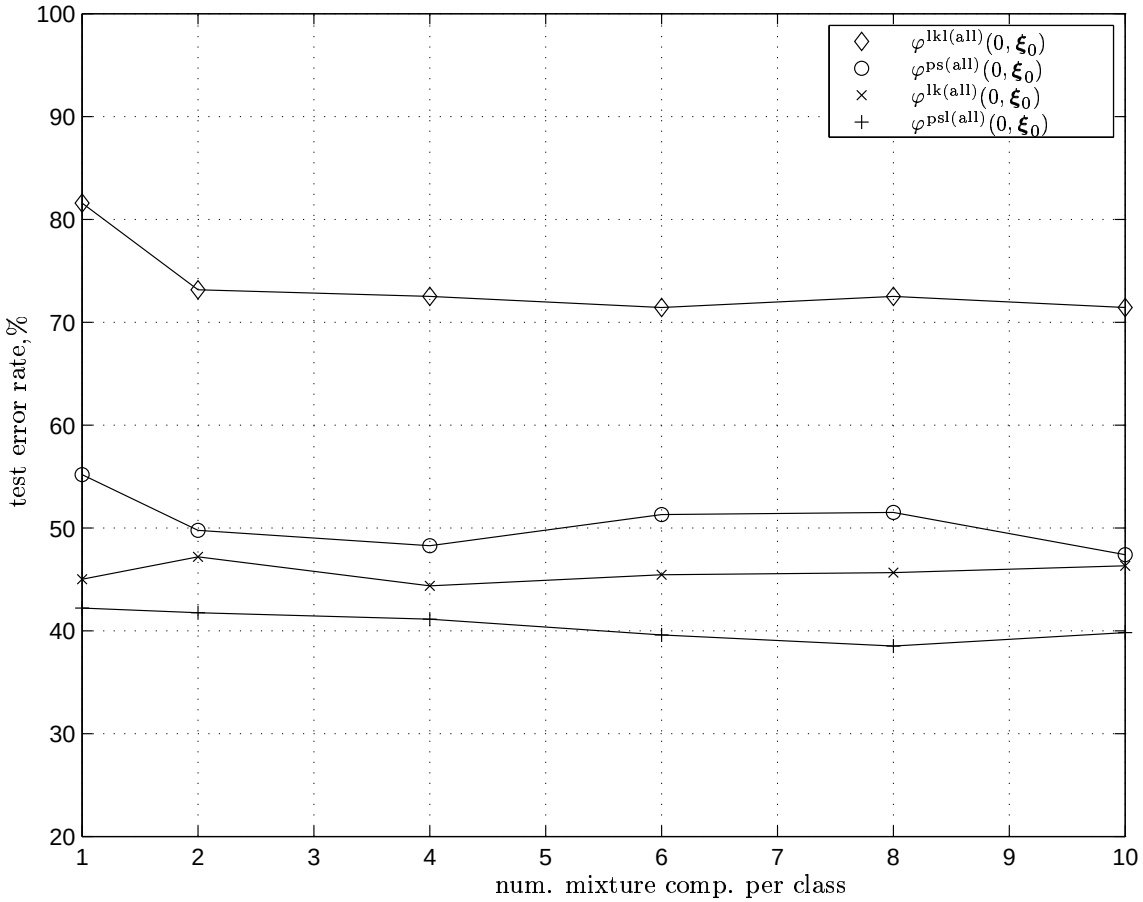


Figure 5.4: Test error rates for GMM classifiers of varying complexity in various zeroth order score spaces based on $\mathcal{P}_0(3)$

error rate of 42.0% as implemented by the max decision rule in these score spaces was lowered to 38.5% for a GMM classifier based on $\mathcal{P}_{\text{sc}}(8)$ trained in $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$. This again illustrates that the max decision rule is not necessarily optimal when the defining distributions are incorrect. Encouragingly the test error rate at 38.5% was lower than those obtained by MAP classifiers in Figure 5.3. Similar experiments on the distributions $\mathcal{P}_0(1)$ and $\mathcal{P}_0(5)$ also confirmed that classifiers constructed in the appended linear posterior score space $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$ outperformed those trained in the linear likelihood equivalent $\varphi^{\text{lkl(all)}}(0, \boldsymbol{\xi}_0)$. However, with regard to log scalar fields, the experiments generally showed that classifiers in the space of log likelihoods $\varphi^{\text{lk(all)}}(0, \boldsymbol{\xi}_0)$ outperformed those in the space of log posteriors $\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$.

Next, techniques were applied to improve performance, either by transforming the scores in the linear posterior score space $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$ based on $\mathcal{P}_0(3)$ or by training alternative classifiers in this space. The best results obtained for each technique are summarised in Table 5.4. The first row gives the lowest test error rate obtained for the unprocessed score space $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$. The next row gives the test error rate for likelihood scaling with a scale factor $\kappa = 0.05$. Next for the third row, the 11 linear class posteriors sum to unity, so covariance matrices in 11-component space are not of full rank and cannot be inverted. For this reason, 10 components were selected according to highest Fisher ratios, and this subspace then transformed using LDA but without further change in the number of components in score space. For the fourth row, the distributions $\mathcal{P}_{\text{sc}}(8)$ were trained using 40 iterations of MMI estimation, and with $\kappa = 1$ and $E = 10$. Only the means and weights of the Gaussian mixture components were updated rather than all the GMM parameters. Although the search-space for all the experimental parameters was not exhaustive, the experiments performed suggest there was little to gain by application of these techniques. The task may be too simple to show the possible benefits from these techniques.

technique	num. mixture comp. in $\mathcal{P}_{sc}$	test error rate, %
none	8	38.5
likelihood scaling	1	39.8
LDA	8	39.4
MMI estimation	8	38.5
MAP classifier in input space		42.0

Table 5.4: Percentage test error rates for applying different techniques to $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$, where the score space is defined on $\mathcal{P}_0(3)$

With regard to covariance modelling, the substitution of diagonal by full covariance matrices in the distributions $\mathcal{P}_{sc}(1)$ in $\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$ demonstrated the expected decrease in test error rate for the resulting GMM classifier from 49.6% to 48.1% for $\mathcal{P}_0(1)$, 55.2% to 50.9% for $\mathcal{P}_0(3)$, and 59.7% to 57.8% for $\mathcal{P}_0(5)$. In practice, the size of score spaces defined by zeroth and first degree covariant derivatives disadvantages the application of full covariance matrices in score space, unless there is a sufficiently large quantity of training data or the distributions $\mathcal{P}_0$ have few parameters.

5.3.3 Score spaces defined on zeroth and first degree covariant derivatives

Next score spaces defined on log scalar fields were augmented by adding unit degree covariant derivatives. Three sets of baseline distributions $\mathcal{P}_0(1)$, $\mathcal{P}_0(3)$ and $\mathcal{P}_0(5)$, all with updated covariances, were chosen.

5.3.3.1 Posterior score spaces defined on single classes

Posterior score spaces defined on single classes are smaller than the corresponding appended score spaces, and the experiments on artificial data showed their good perfor-

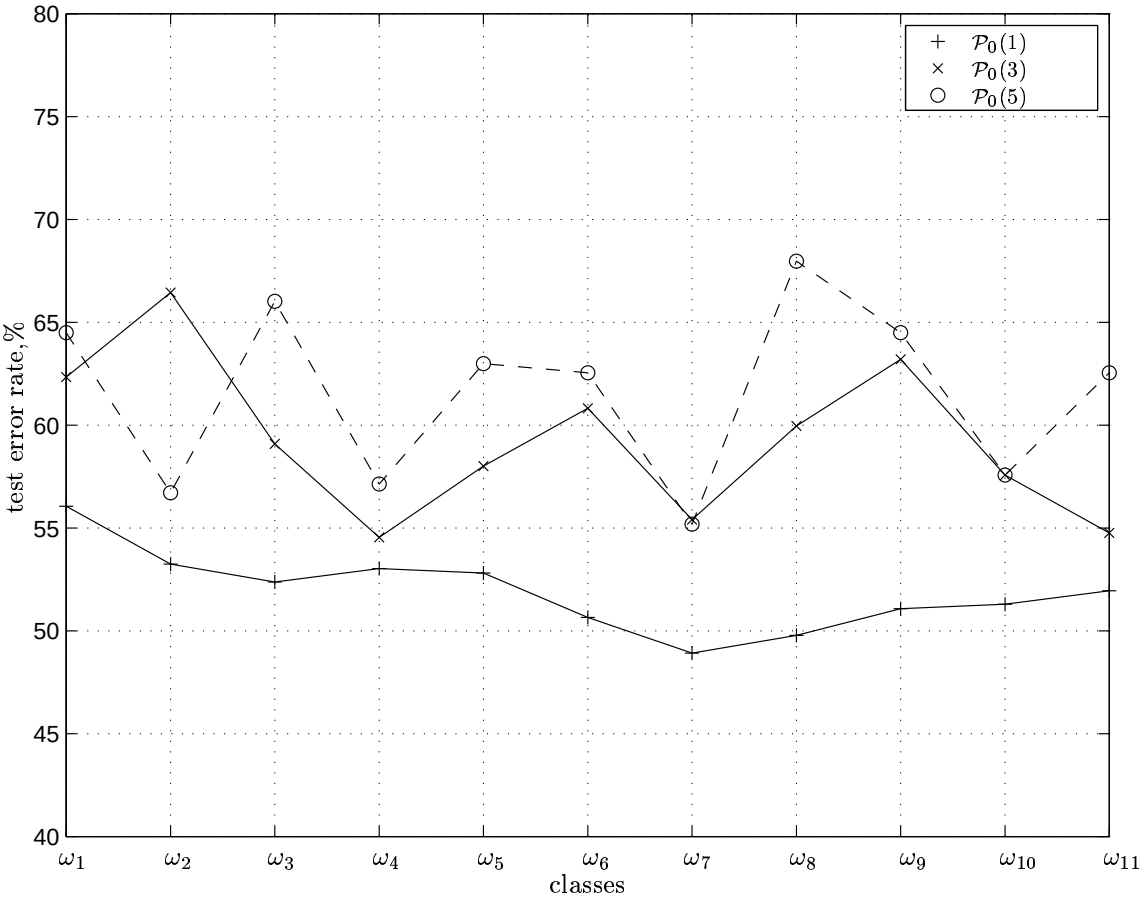


Figure 5.5: Percentage test error rates for GMM classifiers based on $\mathcal{P}_{\rm sc}(1)$ trained in posterior score spaces $\varphi^{\rm ps(q)}(1, \boldsymbol{\xi}_0)$ as the class ω_q and distributions $\mathcal{P}_0$ were varied

mance. So for the three sets of baseline distributions, posterior score spaces $\varphi^{\rm ps(q)}(1, \boldsymbol{\xi}_0)$ were constructed for each class $\omega_q, q = \{1, \ldots 11\}$. In each score space a GMM classifier was trained based on distributions $\mathcal{P}_{\rm sc}(1)$ where covariance matrices were tied and updated. The resulting test error rates are plotted in Figure 5.5. The number of components in the score spaces defined on $\mathcal{P}_0(1)$, $\mathcal{P}_0(3)$ and $\mathcal{P}_0(5)$ were respectively 111, 331 and 551. Since single Gaussians were then trained on each class in score space with tied and updated covariance matrices, there were respectively 1332, 3972 and 6612 parameters per classifier. Since there were only 528 training samples, there was still a high risk of overtraining particularly for classifiers constructed in score spaces defined on $\mathcal{P}_0(3)$ and $\mathcal{P}_0(5)$. In Table 5.5, the best of these GMM classifiers in score space are compared with the MAP classifiers trained directly in input space. There was a significant gain in performance of the score space classifier over the MAP classifier for distributions $\mathcal{P}_0(1)$, but a

significant loss for distributions $\mathcal{P}_0(3)$ and $\mathcal{P}_0(5)$. This was probably due to overtraining. Furthermore, any noise in the parameterisation of $\mathcal{P}_0$ may be accentuated in the mappings to $\varphi^{\mathrm{ps(q)}}(1, \boldsymbol{\xi}_0)$, though there is a hope this can be negated by inherent regularisation in simple score space classifiers. Hence simple distributions $\mathcal{P}_0$ may enable more robust classifiers in $\varphi^{\mathrm{ps(q)}}(1, \boldsymbol{\xi}_0)$. This is further evidenced by the smoothness of the curve for $\mathcal{P}_0(1)$ in Figure 5.5.

defining distributions	test error rate,%	
	MAP classifier	score space classifier
$\mathcal{P}_0(1)$	55.8	48.9
$\mathcal{P}_0(3)$	42.0	54.5
$\mathcal{P}_0(5)$	41.6	55.2

Table 5.5: Percentage test error rates for classifiers based on different distributions $\mathcal{P}_0$ (MAP classifiers and best GMM classifiers based on $\mathcal{P}_{\mathrm{sc}}(1)$ in $\varphi^{\mathrm{ps(q)}}(1, \boldsymbol{\xi}_0)$)

Section 5.2 suggests the classification performance of posterior score spaces for single classes is influenced by the location of the class relative to other classes. Table 5.6 lists distances between the centroids of each class, for both training and test datasets, and the global centroid of the training samples. The distances are calculated with a metric tensor in input space set to the global full covariance across all training samples (see Appendix D.1.2). According to the training data, the most distant class is ω_1 and the third most central is ω_7. In Figure 5.5, the posterior score space based on ω_1 distinguishes different classes poorly while that based on ω_7 generally yielded low test error rates. However a more informative approach using the KL information may be useful. A comparison of the distances between the class centroids in the training and test sets and the global centroid in the training set reveals a mismatch between the training and test sets. This contributes to the poor state-of-the-art test error rates at approximately 30% for this classification task.

class	dataset partition	
	training	test
ω_1	2.06	2.12
ω_2	1.33	1.65
ω_3	1.45	1.32
ω_4	1.56	1.55
ω_5	1.37	2.35
ω_6	1.01	1.63
ω_7	1.11	2.09
ω_8	1.63	2.29
ω_9	1.19	1.50
ω_{10}	1.75	2.03
ω_{11}	0.78	0.86

Table 5.6: Distances between the centroid of each class and the global centroid of the training samples (using the global full covariance of training data as metric tensor)

5.3.3.2 Appended posterior score spaces

Next the posterior score spaces for single classes were appended to yield the reduced appended posterior score space $\varphi^{\mathrm{psr(all)}}(1, \boldsymbol{\xi}_0)$. Reduced score spaces were defined on $\mathcal{P}_0(1)$, $\mathcal{P}_0(3)$ and $\mathcal{P}_0(5)$, and the respective number of components in each was 231, 671 and 1111. In Table 5.7, the simplest GMM classifiers in $\varphi^{\mathrm{psr(all)}}(1, \boldsymbol{\xi}_0)$ are compared with the simplest GMM classifiers in the posterior score spaces $\varphi^{\mathrm{ps(q)}}(1, \boldsymbol{\xi}_0)$ and $\varphi^{\mathrm{ps(all)}}(0, \boldsymbol{\xi}_0)$ (although the classifiers were the simplest available based on distributions of form $\mathcal{P}_{\mathrm{sc}}(1)$, the different sizes of score space imply different numbers of parameters and hence complexity). The results consistently show a decrease in test error rate from the best classifiers in $\varphi^{\mathrm{ps(q)}}(1, \boldsymbol{\xi}_0)$ to the corresponding classifiers in $\varphi^{\mathrm{psr(all)}}(1, \boldsymbol{\xi}_0)$. This implies there is some complimentary information in the different score spaces $\varphi^{\mathrm{ps(q)}}(1, \boldsymbol{\xi}_0), q = \{1 \ldots 11\}$. Also, there was a decrease in test error rate from classifiers in $\varphi^{\mathrm{ps(all)}}(0, \boldsymbol{\xi}_0)$ to corresponding classifiers in $\varphi^{\mathrm{psr(all)}}(1, \boldsymbol{\xi}_0)$, indicating that the unit degree covariant derivatives added extra class discriminative information. However lower test error rates can sometimes be obtained

in $\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$ by increasing the complexity of the GMM classifier in score space. For example, the test error rate for $\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$ defined on $\mathcal{P}_0(3)$ decreased from 55.2% to 47.4% when the number of mixture components per class in $\mathcal{P}_{\text{sc}}$ increased from 1 to 10.

defining distributions	MAP classifier	score spaces for GMM classifiers		
		$\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$	$\varphi^{\text{ps(q)}}(1, \boldsymbol{\xi}_0)$	$\varphi^{\text{psr(all)}}(1, \boldsymbol{\xi}_0)$
$\mathcal{P}_0(1)$	55.8	49.6	48.9	47.2
$\mathcal{P}_0(3)$	42.0	55.2	54.5	47.8
$\mathcal{P}_0(5)$	41.6	59.7	55.2	48.7

Table 5.7: Percentage test error rates for GMM classifiers defined on $\mathcal{P}_{\text{sc}}(1)$ in score spaces, and the MAP classifiers in input space (the results for $\varphi^{\text{ps(q)}}(1, \boldsymbol{\xi}_0)$ are the lowest from any class $\omega_q, q = \{1, \ldots 11\}$)

Various techniques were then applied aimed at increasing classification performance in $\varphi^{\text{psr(all)}}(1, \boldsymbol{\xi}_0)$ based on $\mathcal{P}_0(3)$. However, selecting components of score space with highest Fisher ratios yielded classifiers with higher test error rates (the GMM classifiers were still defined on simple distribututions $\mathcal{P}_{\text{sc}}(1)$ within the subspaces). The results suggest that, for this task, important class discriminative information is lost when components are discarded from $\varphi^{\text{psr(all)}}(1, \boldsymbol{\xi}_0)$. Next the GMM classifiers were defined on distributions $\mathcal{P}_{\text{sc}}(1)$ trained by MMI estimation. This successfully decreased test error rate from 47.8% (for the ML-estimated distributions which initialised the MMI training) to 44.4% (for $\kappa = 1$, $E = 10$, and after 20 iterations of MMI estimation in which just the means and weights of mixture components were updated). This decrease in test error rate is as expected from discriminative training.

Next, the reduced appended posterior score space was generalised to $\varphi^{\text{psg(all)}}(1, \boldsymbol{\xi}_0)$. The change in performance for GMM classifiers in score spaces based on $\mathcal{P}_0(1)$, $\mathcal{P}_0(3)$ and $\mathcal{P}_0(5)$, and $\mathcal{P}_{\text{sc}}$ set to $\mathcal{P}_{\text{sc}}(1)$, was inconclusive. Finally, a hybrid score space $\varphi^{\text{psh(all)}}(1, \boldsymbol{\xi}_0)$ was tested in an attempt to combine the good class discrimination in the zeroth order linear posterior score space $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$ with the extra information in the covariant derivatives of the log class likelihoods. A comparison of performance of GMM classifiers based on $\mathcal{P}_{\text{sc}}(1)$ in different appended posterior score spaces is detailed in Table 5.8.

defining distributions	MAP classifier	score spaces for GMM classifiers		
		$\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$	$\varphi^{\text{psh(all)}}(1, \boldsymbol{\xi}_0)$	$\varphi^{\text{psr(all)}}(1, \boldsymbol{\xi}_0)$
$\mathcal{P}_0(1)$	55.8	53.0	54.5	47.2
$\mathcal{P}_0(3)$	42.0	42.2	43.3	47.8
$\mathcal{P}_0(5)$	41.6	42.0	42.6	48.7

Table 5.8: Percentage test error rates for GMM classifiers based on $\mathcal{P}_{\text{sc}}(1)$ in different score spaces, and the MAP classifier in input space

No firm conclusions can be drawn on the relative merit of the hybrid score space relative to the reduced appended posterior score space. However, performance in the hybrid score space was consistently worse than in the corresponding space of linear class posteriors $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$. The addition of unit degree covariant derivatives here decreased class discrimination. The present task may be too simple to better the simple GMM classifiers in $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$.

5.3.4 Summary

These experiments have considered some simple score spaces. The standard MAP classifier is a max decision rule in the zeroth order score spaces $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$, $\varphi^{\text{ps(all)}}(0, \boldsymbol{\xi}_0)$, $\varphi^{\text{lkl(all)}}(0, \boldsymbol{\xi}_0)$ and $\varphi^{\text{lk(all)}}(0, \boldsymbol{\xi}_0)$ (assuming class priors are fixed and equal). However if the distributions $\mathcal{P}_0$ are incorrect, better performance may sometimes be obtained by training alternative classifiers in these spaces. The score space $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$ based on linear class posteriors was shown to yield better performance than the equivalent score space based on log class posteriors. Unit degree covariant derivatives were introduced into appended score spaces defined on log class posteriors and classification performance improved. The best multicategory classifier obtained by training a classifier in a single score space yielded a test error rate of 38.5%. This was obtained twice, once for a classifier in $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$ defined on $\mathcal{P}_0(3)$ and with $\mathcal{P}_{\text{sc}}(8)$ trained by ML estimation, and additionally by an identical classifier based on the same distribution $\mathcal{P}_{\text{sc}}(8)$ but updated using MMI estimation. The performance at 38.5% is still significantly worse than state-of-the-art classifiers which

attain test error rates nearer 30%.

5.4 Multicategory decisions from binary classifiers

In the previous section, multicategory classifiers were constructed in input space or a single score space. An alternative multicategory classifier constructs a set of binary 1-v-1 classifiers between every pair of classes, and then combines their decisions in a high level multicategory decision rule. Each binary classifier is trained on the two relevant classes. This is inherently suboptimal unless the binary classifiers and high level multicategory decision rule are optimised jointly. There is also the need to resolve conflicting decisions. Despite this, such classifiers are attractive for the following reasons.

- MAP classifiers in input space or GMM classifiers in score space assume a single set of distributions $\mathcal{P}_0$ or $\mathcal{P}_{\text{sc}}$. However a binary classifier has flexibility to assume its own distributions for its relevant pair of classes. A single consistent set of distributions $\mathcal{P}_0$ or $\mathcal{P}_{\text{sc}}$ for all binary classifiers is no longer necessary. Discriminative training methods are expected to improve class discrimination.

- Learning algorithms which are inherently binary in nature, for example SVMs, may be applied.

- Section 3.5.3.3 describes how a linear discriminant constructed in certain score spaces, may under constraints, be related to a MAP classifier defined on distributions in the total space of a fibre bundle. This encourages the subdivision of a muticategory classification task into binary tasks and the application of linear discriminants.

A suitable high level multicategory decision rule is also required. This thesis does not investigate the options available but prefers to use a simple majority voting scheme. For a Q-class problem, $Q(Q-1)/2$ binary classifiers are trained to distinguish between each pair of classes. For an unlabelled sample, the full set of binary classifiers yields $Q(Q-1)/2$

binary decisions on class membership and the majority voting scheme assigns the sample to the class with most frequent occurrence. If two classes are tied after application of the rule, for example classes ω_a and ω_b, then the result of the binary classifier for ω_a-v-ω_b is inspected. If three or more classes are tied, for example classes ω_a, ω_b and ω_c, and if there is still no clear winner after inspection of the relevant binary classifiers ω_a-v-ω_b, ω_a-v-ω_c and ω_b-v-ω_c, then the winner is selected by a resolution technique. A simple resolution technique is random selection, but it is sensible to take the expectation across an infinite number of random selections. An alternative is back-off where unresolved decisions are decided by an alternative classifier, ideally immune to unresolved decisions. For consistent comparison, the alternative classifier should require no additional information beyond that available to the original classifier.

This section describes experiments on the Deterding vowel dataset. Each binary classifier was trained in input space or score space using 96 training samples, 48 samples from each class. Since there were 11 classes, 55 binary classifiers were required. The test data comprised the full 462 samples in the test partition of the Deterding dataset.

5.4.1 Binary classifiers constructed in input space

The 10-component input space was either scaled or unscaled. In scaling, each component was scaled by its standard deviation calculated on the training data of the two relevant classes only. The classifiers were as follows.

- GMM classifier: the distributions $\mathcal{P}_0 = ((p_a)_0, (p_b)_0)$ for each binary problem ω_a-v-ω_b were given complexities of either 1, 3 or 5 components per GMM. The covariance matrices for the Gaussian mixture components were either tied and updated, tied to the global covariance, or left untied.

- SVM classifier: the linear kernel and nonlinear Gaussian Radial Basis Function (GRBF) kernel were used. The GRBF kernel has a width parameter w where,

$$w = F_{\mathrm{GRBF}} M \tag{5.1}$$

and for data-dependency M was set to the geometric mean of the ranges of each component in input space as determined by the training samples [14]. The factor F_{GRBF} was user-defined.

In each experiment the 55 binary classifiers were all given the same complexity or parameter settings. There was no attempt to 'tune' individual binary classifiers. The experiments were therefore purely illustrative. In addition there was only a coarse optimisation of classifier complexity or classifier settings since score spaces are the main emphasis of this thesis. A majority voting scheme was used and unresolved decisions decided by random selection. Test error rates were reported by calculating expectations over an infinite number of random selections.

The MAP, linear SVM and GRBF-SVM classifiers with best performance are summarised in Table 5.9.

classifier type	parameters	input space	test error rate,%
pairwise MAP	GMMs with 3 mixture comp. (diag. global covars.)	-	41.4
SVM	linear kernel C=1	scaled	45.2
SVM	GRBF kernel C=1, $F_{\mathrm{GRBF}} = 1$	unscaled	35.1

Table 5.9: Selected percentage test error rates for a majority voting scheme on binary classifiers trained in the input space for the Deterding vowel dataset

- The best MAP classifier in input space with 5 mixture components per GMM and tied diagonalised covariances yielded 41.6% test error rate. The best classifier based on binary decisions and a majority voting scheme yielded 41.4% test error rate, with 3 mixture components per GMM and tied diagonalised covariances set to the

global diagonalised covariances for each pair of classes. The full set of GMDs for the binary classifiers do not necessarily map back to a single consistent set of GMDs in input space, thereby permitting tied decisions in the majority voting scheme. The exception is when each GMM in a two-class problem is trained on within-class samples only. However using tied and updated covariances or discriminative training violates this exception. The method of tying covariances therefore influences performance.

- In these experiments, SVMs with GRBF or linear kernels were both sensitive to the value of the SVM parameter C. The SVM with GRBF kernel was however more sensitive to the kernel width as specified by the scale factor F_{GRBF}. The need to optimise both the C parameter and kernel width disadvatanges the application of GRBF kernels. A coarse optimisation of C and F_{GRBF} for SVMs with GRBF kernels in unscaled input space yielded a multicategory classifier with a test error rate of 35.1%. The lowest test error rate for SVMs with linear kernels, where C was coarsely optimised for scaled input space, was higher at 45.2%. The better performance of GRBF kernels suggests classes were not linearly separable in input space. Firmer conclusions cannot be made without an exhaustive optimisation of SVM and kernel parameters.

- The performance of SVMs with linear or GRBF kernels depend on scaling the input space, or equivalently on the metric tensor applied to input space. For example, the introduction of scaling to the input space decreased test error rate from 47.9% to 45.2% for SVMs with linear kernels, and increased test error rate from 35.1% to 43.4% for SVMs with GRBF kernels[3]. Scaling effectively applies a metric tensor in input space set to the diagonalised form of the global covariance, where components of input space are for example assumed contravariant components. According to Appendix D.1.2, this metric tensor may be viewed, under assumptions, as maximally noncommittal in some sense. This makes it more suitable than the Identity metric tensor or scaled Identity metric tensor implied by an unscaled input space. However, the experiments showed that scaling was not always beneficial for the GRBF ker-

[3]The SVM trade-off parameter C and GRBF width factor F_{GRBF} were kept fixed; ideally they should have been optimised for the scaled and unscaled spaces.

nel, probably since this kernel is more susceptible to overtraining and the variances of components in the original input space were already fairly similar (the standard deviations of each component in input space varied between 0.5 and 1.2 where standard deviations were calculated over the training data for all 11 classes). Scaling is expected to be more beneficial when component variances vary significantly.

The lowest test error rate at 35.1% outperformed the best multicategory classifier trained in a single score space at 38.5% from Section 5.3. However it was still higher than the 31% test error rate reported in the SVM system in [37]. This is partly because the SVM was not fine-tuned, but also since in their system a different multicategory decision rule was implemented and a different kernel width applied. The majority voting scheme cannot be applied to their system since 1-v-rest SVM classifiers were trained.

The performance of SVMs is sensitive to the C parameter, the kernel and where applicable the kernel parameters. The performance of GMM classifiers is sensitive to the number of mixture components per GMM, the method of tying covariance matrices and the training criterion.

5.4.2 Binary classifiers constructed in score spaces

Next, the experiments were repeated except that for each pair of classes ω_a and ω_b, a binary classifier was trained in likelihood-ratio score space $\varphi^{\mathrm{lr(a,b)}}(1, \boldsymbol{\xi}_0)$ rather than input space. For brevity, the experiments focussed only on the distributions $\mathcal{P}_0(3)$ and tied covariances. Hence a suitable baseline was the MAP classification based on $\mathcal{P}_0(3)$ in input space which yielded a test error rate of 42.0%. Again a majority voting scheme was used except that an unresolved decision was referred back to the MAP classification based on the distributions $\mathcal{P}_0(3)$ defining the score spaces. Hence no extra information was required to resolve decisions. The best test error rates for this approach are summarised in Table 5.10.

classifier type	classifier parameters	score space	test error rate,%
GMM (score space)	$\mathcal{P}_{\mathrm{sc}}(1)$ (diag. global covars.)	-	42.6
SVM (score space)	linear kernel C=0.01	scaled	32.0
MAP (input space)	-	-	42.0

Table 5.10: Selected percentage test error rates for the MAP classifier, and majority voting scheme for binary classifiers trained in likelihood-ratio score spaces (the MAP classifier and score spaces were defined on $\mathcal{P}_0(3)$)

- The performance of the multicategory classifier based on linear SVMs was sensitive to the parameter C, the lowest test error rate at 32.0% yielded with C set to 0.01. The GMM classifier, where the GMMs for each two-class problem had covariances set to the global diagonalised covariance (where global refers to the two relevant classes only) yielded a higher test error rate[4] at 42.6%. With regard to test error rates, the classifier at 32.0% outperformed the best multicategory classifiers constructed in a single score space at 38.5% from Section 5.3, and from a set of binary classifiers in input space at 35.1% from Section 5.4.1.

- Linear classifiers were the focus of these experiments since they possess good regularisation properties and, under constraints, can be related to MAP classifiers defined on distributions in the total space of fibre bundles. The linear classifiers were implemented by SVMs with linear kernels or GMM classifiers defined on single mixture component GMMs with tied covariances. Nevertheless, a comparison with nonlinear classifiers is useful. More complicated GMM classifiers were constructed in the likelihood-ratio score spaces $\varphi^{\mathrm{lr(a,b)}}(1, \boldsymbol{\xi}_0)$ for each binary problem. Increasing the number of mixture components per GMM from 1 to 3 to 5 yielded respective test error rates of 42.6%, 43.1% and 47.2% for the final multicategory classifier. The

[4] Limitations in the GMM training precluded the use of GMMs with tied and updated covariances.

increase in test error rate was probably due to overtraining since each likelihood-ratio score space was relatively large at 61 components. A multicategory classifier based on SVMs with GRBF kernels was trained in scaled and unscaled score spaces, and yielded test error rates of 38.7% and 40.3% respectively. This was worse than corresponding test error rates of 36.8% and 37.9% for corresponding classifiers based on linear kernels. Little can be gained from this comparison without an exhaustive search of SVM and kernel parameters. The experiments illustrate the benefit of scaling for SVMs.

Overall, the results show that constructing a full set of SVM classifiers in likelihood-ratio score spaces is promising. It is not possible, through the experimental results, to distinguish the effect of discriminatively training pairs of classes and the effect of applying SVMs.

5.5 Discussion of results on the Deterding dataset

In this chapter, a single set of GMMs trained directly in input space implemented a MAP classifier and yielded test error rates as low as 41.6%. When a full set of binary GMM classifiers were trained in input space and a majority voting scheme applied with random selection of undecided samples, the test error rate remained approximately the same at 41.4%. These GMMs had a sensible initialisation method. Score spaces were then introduced. If multicategory GMM classifiers were constructed in single score spaces rather than in the input space, lower test error rates were obtained down to 38.5%. However, if SVM classifiers were constructed in likelihood-ratio score spaces and a majority voting scheme applied with a sensible back-off scheme for unresolved samples, then a test error rate of 32.0% was obtained.

The documentation supplied with the Deterding vowel dataset [7] lists test error rates in [82] obtained using a variety of techniques. The best performance was 44% test error rate given by a nearest neighbour classifier, although it should be noted that the list of results

are based on single trials and therefore dependent on initialisation. Other leading results include a test error rate of 38.3% obtained in [47] using a discriminant adaptive nearest neighbour technique. A lower test error rate of $30.2 \pm 0.3\%$ was obtained for the Separable Mixture Model (SMM) in [99] where the style (speaker) and content (vowel) were modelled by a bilinear model embedded within a GMM. A test error rate of 30% was obtained for an approach with Relevance Vector Machines in [45]. Of particular interest is the SVM classifier described in [37] which attained a test error rate of 31%. In their approach, SVMs were trained with GRBF kernels and width w set to a single universal value. A mixture-of-experts paradigm was used to obtain a single decision on class membership from the set of binary 1-v-rest classifiers. Tests using SVMs with standard GRBF and polynomial kernels in input space are described in [15] in which the GRBF kernel outperformed the polynomial kernel. A test error rate of 33.9% was yielded for a multicategory scheme involving 1-v-all classifiers, and 30.0% for a scheme utilising 1-v-1 classifiers[5]. Their baseline test error rate with GMMs was 37.9% obtained with 16 component class-conditional GMMs. The best result with score spaces at 32.0% is a little worse, but comparable, to the best speaker-independent results known to the author for this task at $30.2 \pm 0.3\%$, 30.0% and 30%.

5.6 Summary

Experiments in this chapter contrasted the performance of some simple score spaces in classification. The score space of linear class posteriors $\varphi^{\mathrm{psl(all)}}(0, \boldsymbol{\xi}_0)$ gave very good performance. In this space, the max decision rule implements the MAP classifier. However the max, and hence MAP, decision rule is not necessarily optimal when the distributions $\mathcal{P}_0$ upon which the classifier is defined are incorrect. Alternative decision rules in this score space may yield lower test error rates. Larger score spaces were then introduced which included unit degree covariant derivatives. For certain score spaces and linear discriminants, these permit the recovery of MAP classifiers based on distributions which are located in the total space of a fibre bundle defined on the original statistical models. Alternatively,

[5] Unfortunately, no indication is given of how tied decisions were resolved in the majority voting scheme among 1-v-1 classifiers.

the unit degree covariant derivatives can simply be viewed as a method to add extra class discriminative 'features' into score space. The typically large size of such score spaces promotes the application of strongly regularised score space classifiers.

The multicategory classification task was then subdivided into a set of binary 1-v-1 classification tasks and a simple majority voting scheme applied. This permitted the application of SVMs. When SVM classifiers were trained in likelihood-ratio score spaces, the multicategory classification was only a little worse than state-of-the-art systems. A similar approach is adopted in the next chapter.

Chapter 6

Classifying variable length patterns

This chapter applies score spaces to the classification of variable length patterns. The patterns are isolated letter utterances drawn from a speech database. Speech is a naturally occurring source of variable length patterns.

6.1 Description of the ISOLET dataset

The ISOLET database [16] consists of utterances of isolated letters drawn from the american english alphabet. There are 26 letters. In total there are 7800 utterances spoken by 150 speakers, with two utterances per letter per speaker. The dataset is split into five equal subsets labelled `isolet1` to `isolet5`. Each subset contains the utterances from 15 male speakers and 15 female speakers, with no overlap in speakers between subsets. In these experiments, `isolet1`, `isolet2`, `isolet3` and `isolet4` were retained as training data, and `isolet5` as test data. An important subset of letters is the E-set which comprises the letters {`B,C,D,E,G,P,T,V,Z`}. Each utterance was supplied in a preprocessed form with long periods of silence removed from either side of each isolated letter utterance. However approximately 80ms of silence was retained immediately preceding and following each letter [16]. The speech was recorded at 16kHz sample rate.

Rather than extract specialised features, the dataset was coded in a manner consistent with large vocabulary speech recognition tasks. The data was encoded using the `HCopy` tool in HTK version 3.0 [114]. The speech was encoded at a frame rate of 100 frames per second. Each frame was extracted using a Hamming window of 25.6ms in length giving a frame overlap of approximately 150%. Each frame was then processed using a mel-scale filterbank with 20 channels, a preemphasis coefficient of 0.97 and a cepstral liftering coefficient of 22. The first 12 Mel-Frequency Cepstral Coefficients (MFCCs), ignoring the zeroth order coefficient, were extracted. A term describing the log signal energy was also extracted using the default HTK energy normalisation, scaling and silence floor [114]. The 12 MFCCs and log energy term, collectively termed the *static* parameters, were then augmented by first and second order time derivatives respectively called the *delta* and *acceleration* parameters. This yielded a 39 element feature vector for each frame. Delta parameters were calculated using a linear regression over the static parameters of the preceding two frames and following two frames. Similarly, acceleration parameters were calculated using a linear regression over the delta parameters of the preceding two and following two frames. In this chapter, this encoding is abbreviated to `MFCC_E_D_A`. For consistency with the rest of this thesis, each frame is called an 'observation'. An utterance, which is a sequence of observations, is known as a 'sample' (this should not be confused with the digital samples in the original speech waveform files). The input space $L(\boldsymbol{O})$ was the space of observation sequences or samples.

6.2 Baseline input space classifiers

The statistical models $\mathcal{S}(\boldsymbol{\xi})$ were a set of continuous density HMMs. Each HMM topology was constrained to left-to-right with no skips. Each letter in the alphabet was modelled by an HMM with 10 emitting states, and silence by an HMM with 1 emitting state. State-conditional likelihoods were modelled by GMMs with diagonal covariance matrices.

A series of baseline experiments was performed measuring the change in performance with HMM complexity under different training regimes. Complexity was defined as the

number of mixture or Gaussian components per state. The different training regimes were as follows, where there were 6240 training samples for the full alphabet task and 2160 training samples for the E-set task.

- ML estimation: Initial models with 1 mixture component per state were obtained by flat-starting (the Gaussian components were parameterised with the global mean and variance of the training samples). The training samples were then used to update all the model parameters using an embedded version of the Baum-Welch algorithm [114]. The number of mixture components per state was gradually increased using mixture splitting [114]. Following each split, 20 iterations of Baum-Welch reestimation were implemented. Suitable weight and variance floors were also used. This training regime was implemented using `HCompV`, `HHEd` and `HERest` in HTK version 3.0 [114] [49].

- MMI-5 estimation: The initial models for MMI estimation were the ML models of the same complexity yielded by the previous training regime. These ML models were used to obtain numerator and denominator lattices[1]. The ML models were then reestimated to maximise mutual information using 5 iterations of the Extended Baum-Welch algorithm [111]. A suitable weight floor and other MMI parameters were used. All MMI parameters were kept fixed except for the likelihood scale factor κ which was varied for optimal performance for each complexity. This training regime was implemeted using development versions of MMI and lattice generation code within the HTK environment (see the Acknowledgments).

- MMI-20 estimation: This is identical to the MMI-5 training regime except that there were 20 Extended Baum-Welch reestimation iterations instead of 5.

To narrow the search space for the experiments, there was no attempt to optimise the number of training iterations for each trained model. The 20 iterations for ML training was a design parameter even though there was a risk of overtraining some models. The greater computational cost involved in optimising the likelihood scale factor and producing

[1]In these experiments, MMI estimation was only implemented for E-set classification, so only 9 competing paths were required in each denominator lattice.

lattices for MMI training narrowed the application of this training technique to HMM complexities ranging from 1 to 6 mixture components per state rather than the 1 to 20 mixture components per state for ML training. The choice of ML and MMI estimation was made to contrast maximum likelihood and discriminative training techniques, while MMI-5 and MMI-20 were selected to investigate some generalisation issues for MMI estimation.

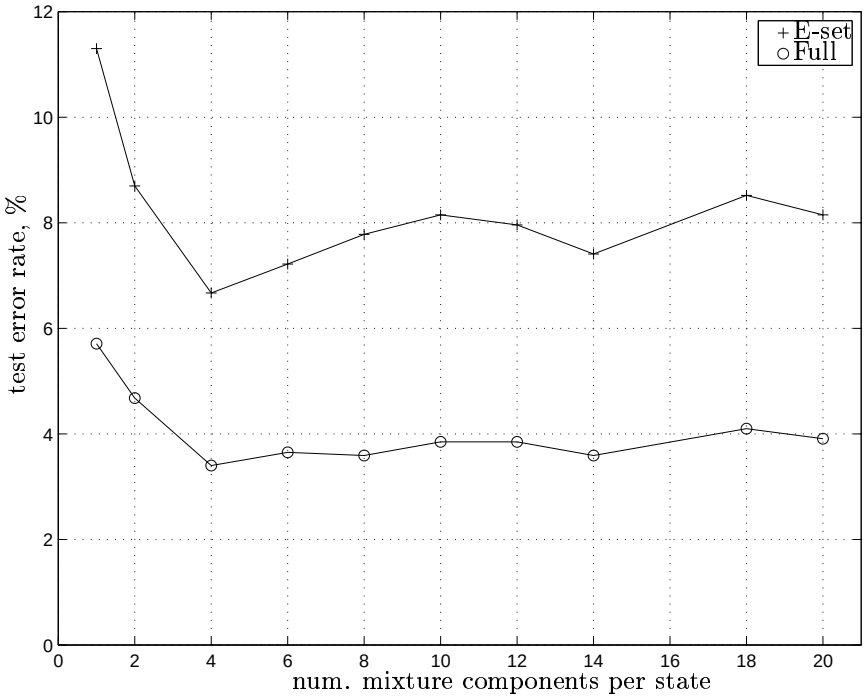
In total there were 1560 test samples for the full alphabet task and 540 test samples for the E-set task. In testing a particular sample, all hypotheses were permitted which conformed to a `silence-letter-silence` format, where `letter` was any letter from either the full alphabet or E-set as required. The sample was then classified according to the most likely hypothesis yielded by the token-passing implementation of the Viterbi algorithm [114]. This effectively summarised the likelihood of all possible paths for a hypothesis with that of the single most probable path for that hypothesis. The classification results reported in this chapter are test or training error rates measured as the percentage of letters recognised incorrectly[2]. Testing was implemented using `HVite` and `HResults` from HTK version 3.0 [49]. The classifier is a MAP classifier using a Viterbi approximation for each hypothesis.

McNemar's test [40], a measure of statistical significance, was applied to yield confidence levels as described in Appendix G.2. The test simply ascertained the confidence that test error rates yielded by two classifiers on the same test set differed not just because of chance effects. The only assumption required independence between errors. The application of nonequal class priors forces modifications to the test, but in these experiments class priors were assumed equal.

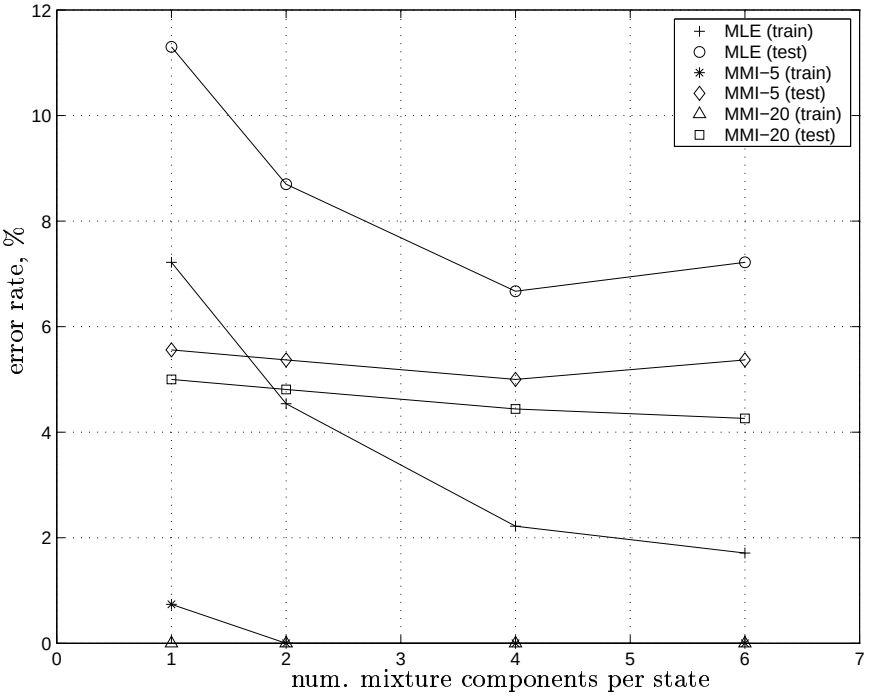
Figure 6.1(a) details the variation of test error rate with complexity for models trained by ML estimation for the E-set and full alphabet tasks. Some remarks may be made.

- The observed optimal complexity for this training regime and task was 4 components per state. Models of lower complexity probably do not possess functionality to model sufficient within-class characteristics such as gender and accent. Models of higher complexity have too many parameters for robust estimation and instead reflect

[2]Silence was implicit in the hypothesis and its successful recognition was ignored.



(a) ML-trained models



(b) E-set: Models trained by ML estimation
and MMI (after 5 and 20 iterations) estimation

Figure 6.1: Baseline error rates for the MAP classifier operating on different sets of distributions in input space

pecularities of the training data which do not exist in the test set. With more training data, more complex models are expected to give better performance.

- For computational reasons, a subset of letters was selected as the principal subset for investigation. A large proportion of the errors in the full alphabet task were between members of the E-set. For this reason, the experiments in this chapter focus on E-set classification. For example, for the HMM system with 4 components per state, 74% of false classifications were for test samples drawn from the E-set. The test set probability of error given an E-set letter was 7.2%, whereas given a non E-set letter it was 1.4%. The confusion matrix for the E-set letters within the full alphabet task is not necessarily identical to that for the E-set task alone. Unless otherwise stated, the 'E-set task' refers to the classification of E-set letters as E-set letters.

- Closer inspection of the test set performance for ML training indicated a tendency to overtrain with 20 iterations. Early-stopping often yielded a set of models with better test set performance. However this design parameter, while not necessarily optimal, still yielded a reasonable system compared to those of other researchers (see Section 6.5).

Figure 6.1(b) details the variation of test error rate with complexity for the three different training regimes on the E-set task. Training error rates are also included for reference. Complexity was restricted to 1, 2, 4 and 6 mixture components per state. Some remarks may be made.

- As expected, training error rates were lower than test error rates because the training and test sets have different statistics and a classifier more accurately reflects statistics in its training set.

- Due to its discriminative approach, lower test error rates were achieved for models estimated by MMI. Test error rates for ML estimation varied down to 6.7%, but for MMI estimation varied down to 4.3%. In all but one case, the test set performance of the MMI-trained models was statistically significant in comparison to ML-trained models of the same complexity to a confidence level of 95%. The exception for the

models with 4 components per state under the MMI-5 training regime was at a 94% confidence level.

- The MMI-20 training regime yielded models with consistently better test set performance than the MMI-5 training regime, even when both yielded models which perfectly separated the training samples. This indicates that prolonging MMI training, at least in this case, repositioned class decision boundaries at better locations. This was in accordance with increasing the average posterior margin calculated over the training samples. However the statistical significance between the test set performance of the MMI-20 and MMI-5 models was much poorer, ranging between 55% and 62% confidence levels for complexities of 1, 2 and 4 components per state. Only at 6 components per state was the confidence level acceptable at 93%.

The superior performance of MMI estimation is reflected in the current speech recognition literature [111]. However MMI estimation requires fine-tuning the likelihood scale factor κ. For 1, 2, 4 and 6 components per state, κ was respectively 1/30, 1/30, 1/50 and 1/60. Of course, values below unity artificially increase the confusion between classes in the space of linear class posteriors, i.e. in $\varphi^{\text{psl(all)}}(0, \boldsymbol{\xi}_0)$ where $\boldsymbol{\xi}_0$ is the set of current HMM parameters.

There is another viable training regime for MMI estimation. Rather than train all class-conditional statistical models within a single framework, i.e. under 'multiclass MMI estimation', each pair of class-conditional models for a binary problem can be trained independently of all other pairs. While these two approaches yield identical sets of models for ML estimation, this is not the case for MMI estimation. For a complexity of 2 mixture components per state, the pairwise MMI approach yielded a test error rate[3] of 4.8% after 4 iterations and $\kappa = 1/70$. This differed from the 5.4% test error rate at a confidence level of 42% obtained for models of identical complexity but trained after 4 iterations of multiclass MMI estimation. Although the pairwise approach may give still further improvements in performance by careful selection of likelihood scale factors for each pair,

[3]Random selection was applied for unresolved decisions. Prior to this, the test error rate was between 4.6% and 5.0%.

notation	score subspace
`w`	covariant derivatives with respect to mixture component weights
`m`	covariant derivatives with respect to mixture component means
`v`	covariant derivatives with respect to mixture component variances
`t`	covariant derivatives with respect to HMM self-transition probabilities
`l`	log class likelihood
`r`	log class likelihood-ratio
`p`	log class posterior

Table 6.1: Abbreviations for score subspaces

the greater computational cost incurred by pairwise training may not justify any increase in performance obtained. For this reason, multiclass MMI estimation was pursued in the remainder of the experiments.

6.3 Score space classifiers

A collection of software was used for the experiments on score space classifiers. The calculation of scores within score space was implemented by modifying SVM^{light} version 3.02. SVM training and testing was using SVM^{light} version 4.00 [54] [53]. The calculation of MLE linear discriminants was implemented in MATLAB version 5 [68]. As described in Section 2.3.1.1, the MLE linear discriminant is here identical to the MSE linear discriminant since there are equal numbers of training samples for each of the classes. The distributions used to define the mapping into score spaces were those calculated by the ML or MMI training regimes for the baseline experiments.

The experiments in this section focus on E-set classification and HMMs with complexity from 1 to 6 mixture components per state. Classifiers were constructed in a variety of score spaces based on these HMMs and as defined in Section 4.1. Where relevant the abbreviations in Table 6.1 are used. Linear discriminants were chosen since they may under constraints be related to classifiers defined on points in the total space of fibre

bundles[4] and they have good regularisation properties. For simplicity for SVMs, a single parameter $C = 100$ was used for all experiments. The covariant derivatives in all the score spaces in these experiments were defined with respect to the Gaussian component means only. The other parameters of the statistical models were ignored. This was for simplicity in comparison, and is sensible since Gaussian means are generally known to contain most of the class discrimination between HMMs. The effect of introducing other parameters into the definition of the score space is detailed later in Section 6.3.4. For the present, it is sufficient to regard this restriction as a good compromise between the size of the score space and adequate performance.

6.3.1 'Normalisation' in score space

There are two 'normalisation' techniques which may be applied to score spaces and which also illustrate the degenerancy between the score mapping and score space classifier.

- The first is the metric matrix in score space. This may either be viewed as embedded in the score mapping or a property of the 'distance-calculating algorithm' in the classifier. For the SVM classifier, the metric matrix applied in these experiments was the diagonal form of the global covariance matrix calculated on the training data. The covariance matrix was calculated individually for each binary problem using the samples mapped into the corresponding score space. Replacing this with an Identity metric matrix in the `mr` score space increased the E-set test error rate for models with 2 mixture components per state from 5.0% to 10.7% with a confidence level of 100%. The trend is similar to that expected for a GMM classifier which has all its covariances set to Identity rather than tied to the global covariance. The diagonal form of the global covariance matrix is a sensible choice of metric matrix as explained in Appendix D.1.2. Also as explained in Section 3.6, this metric matrix is here not constrained to give unit scaling for the zeroth order score subspace.

[4] However the decision rule is based on normalised log likelihoods of form $l_{\text{nms}}(\boldsymbol{O}_l; \boldsymbol{\theta}_q)$ (see Section 4.6); furthermore there was no attempt to enforce the constraints on the linear discriminant's weight vector, so there was no guarantee it related to valid distributions (see Section 3.5.3.3).

- The second is sequence length normalisation. In these experiments the 'soft' form of the normalised log likelihood $l_{\text{nms}}(\boldsymbol{O}_l; \boldsymbol{\theta}_q)$ detailed in Section 4.6 was applied. This normalisation may be viewed as embedded in the score mapping based on the normalised log likelihood. Alternatively, it may be viewed as a 'feature-space transformation' applied to the score space defined on the log likelihood, and may possibly therefore be viewed a property of the classifier. Sequence length normalisation is expected to improve performance whenever letters occur which are spoken at a variety of speaking rates. For the `mr` score space, there was 0% confidence in the change in test error rate on the E-set and full alphabet tasks when sequence length normalisation was neglected. However for the `wmvtr` score space and E-set task, an absence of sequence length normalisation increased test error rate from 4.1% to 5.6% at a confidence level of 90%. These experiments are inconclusive as to the merits of sequence length normalisation. The results suggest that the reduction in within-class variability caused by sequence length is offset by the loss in duration information in the `mr` likelihood-ratio score space. This is not the case in the `wmvtr` score space where duration is preserved in the derivatives of the self-transition probabilities. Despite this, sequence length normalisation is retained in all experiments for the purpose of comparison.

The metric matrix and sequence length normalisation described above are applied in the remaining experiments in this chapter. For convenience, the term 'likelihood score space' and its mathematical notation, for example $\varphi^{\text{lk(q)}}(1, (\boldsymbol{\theta}_q)_0)$, are retained, and sequence length normalisation and its modified form of the log likelihood implied. Similar implications follow for other score spaces.

6.3.2 Comparing classification algorithms in score space

Next it is important to verify, at least in part, whether SVMs are a good choice of classifier for these experiments, as suggested in the experiments of the previous chapter. For this reason, alternative MLE discriminants were trained. These were implemented either by setting the covariance matrix to the diagonal form of the weighted within-class covariance,

or to the diagonal form of the global covariance. Covariances were calculated only on the two relevant classes for each binary problem. A comparison is detailed in Table 6.2 where test error rates are presented for the E-set task and likelihood-ratio `mr` score spaces defined on ML trained models. Confidence levels are given comparing the classifiers with baseline MAP classifiers operating on the same models.

The SVM classifiers consistently outperformed the MLE (weighted covariance) classifiers, which in turn outperformed, in all but one case, the MLE (global covariance) classifiers. These trends were also reflected in identical experiments but where the ML trained models were substituted for those trained under the MMI-5 and MMI-20 training regimes. The results emphasise the good generalisation properties of SVM classifiers and that their training is discriminative. The lack of robustness in the MLE linear classifiers was probably due to insufficient training samples. For each binary problem, there were 480 samples available for the GMM classifier to estimate 4683 parameters, two 1561-component class means and one 1561-component variance. For the SVM, there were 1562 parameters for the weight and bias.

Setting the classifier covariance matrix to the diagonal weighted within-class covariance yielded better performance than setting it to the diagonal global covariance. These may be interpreted as two different metrics which the MLE discriminant assumes for score space. Both metrics are maximally noncommittal in the sense and within the assumptions of Appendix D.1.2, but the former in an average sense with respect to each class, and the latter with respect to the two classes of scores in score space. The former yields better performance by the same argument as a GMM classifier often performs better when its covariance matrix is tied to the average within-class covariance rather than the global covariance. While it is possible to apply other parametric and nonparametric classifiers, the main focus of the chapter is the investigation of score spaces rather than their interaction with different classifiers. The remaining experiments in the chapter solely apply SVMs.

num. mixture comp. per model	input MAP classifier	score space classifier		
		SVM	MLE (wtd)	MLE (glb)
1	11.3	6.9 (100)	9.8 (78)	10.4 (50)
2	8.7	5.0 (99)	7.0 (77)	6.9 (83)
4	6.7	5.4 (74)	6.5 (0)	7.8 (66)
6	7.2	5.4 (89)	6.7 (30)	7.0 (0)

Table 6.2: Percentage test error rates for different classifiers for the likelihood-ratio `mr` score space defined by ML-trained models (confidence levels relative to MAP classifier using same models are given in brackets)

6.3.3 Comparing score spaces

Next, some of the score spaces detailed in Section 4.1 were compared. All experiments applied the metric matrix, sequence length normalisation and SVMs as detailed above. The score spaces were defined using the baseline models from ML training. Test error rates are presented in Table 6.3. Confidence levels relative to the baseline models are given, where useful, in brackets. For a binary problem the two classes are denoted ω_a and ω_b, and class ω_a was arbitrarily chosen as the first of the two classes according to alphabetical ordering. The best performance was for the likelihood-ratio score space defined on models with 2 mixture components per state at 5.0% test error rate comparing to a best MAP classifier performance at 6.7%. A number of remarks may be made.

- The results show poorest performance for the Fisher score space and likelihood (2-class) score space. For each binary classification task, both score spaces were defined on a single statistical model describing the two relevant classes. As detailed previously, the score mapping may be noninjective in nature. This impedes class discrimination if two regions of input space, typical of two competing classes, map to the same region of score space. Such is the case for these score spaces. As the number of mixture components per state increases, so more of input space is mapped to the same region of score space, which is expected to increase class confusion in

num. mixture comp. per model	input space	score space			
	MAP classifier	Fisher $\varphi^{\mathrm{fs(a,b)}}(\boldsymbol{\theta}_0)$ `m`	likelihood (2-class) $\varphi^{\mathrm{lk(a,b)}}(1,\boldsymbol{\theta}_0)$ `ml`	likelihood $\varphi^{\mathrm{lk(a)}}(1,(\boldsymbol{\theta}_a)_0)$ `ml`	lik.-ratio $\varphi^{\mathrm{lr(a,b)}}(1,\boldsymbol{\xi}_0)$ `mr`
1	11.3	6.3	5.9	7.6	6.9 (100)
2	8.7	10.2	9.4	6.3	5.0 (99)
4	6.7	23.5	23.3	7.6	5.4 (74)
6	7.2	31.3	31.1	7.8	5.4 (89)

Table 6.3: E-set: Percentage test error rates for different score spaces for ML-trained models of different complexity (confidence levels relative to the MAP classifier using the same models in input space are given in brackets)

score space and decrease the performance of score space classifiers. Performance in the Fisher score space was worse than in the likelihood (2-class) score space based on the same distributions. The only difference in score spaces was the inclusion of the log likelihood in the latter. The log likelihood is a useful feature which is nonlinearly related to its unit degree covariant derivatives with respect to the Gaussian means. As such it furnishes information which cannot be extracted by a linear classifier acting on these covariant derivatives alone. Its inclusion is here beneficial for class discrimination.

- The likelihood (1-class) score space has limited ability to distinguish samples in input space since the model parameters defining the score space represent one class only. Viewing these parameters as 'triangulation points' for input space, there are insufficient reference points to clearly describe regions of input space outside the scatter of this class. By adding derivatives defined on model parameters for another class, the set of triangulation points stretches across a more expansive region of input space and is far more expressive. For this reason, the likelihood-ratio score space yields better class discrimination than the likelihood score space based on the model for a single class alone. A linear discriminant in an appended likelihood score space

for two classes $\varphi^{\mathrm{lk(all)}}(1, \boldsymbol{\xi}_0)$ can mimic a linear discriminant in the likelihood-ratio score space. However it has an extra degree of freedom in combining the class log likelihoods. The likelihood-ratio score space is preferred for its relation to simple likelihood-ratio tests.

- For the likelihood-ratio score space, the optimal complexity for the defining distributions was 2 mixture components per state. This was a lower complexity than the 4 mixture components per state identified for the best MAP classifier operating on the same models. Although unwise to draw inferences from a single experiment, this may simply be because the size of a score space defined on distributions with 4 mixture components per state is much larger and more susceptible to overtraining when there is a sparsity of training scores.

Experiments were performed with the posterior score spaces[5] $\varphi^{\mathrm{ps(a)}}(1, \boldsymbol{\xi}_0)$ and $\varphi^{\mathrm{ps(b)}}(1, \boldsymbol{\xi}_0)$, with score mappings defined on HMMs with 2 mixture components per state at respective test error rates of 10.6% and 10.9%. The posterior score space $\varphi^{\mathrm{ps(a)}}(1, \boldsymbol{\xi}_0)$ is identical to the likelihood-ratio score space $\varphi^{\mathrm{lr(a,b)}}(1, \boldsymbol{\xi}_0)$ except for the substitution of the log class posterior $\ln P(\omega_a|\boldsymbol{O})$ for the log likelihood ratio $\ln p(\boldsymbol{O}|\omega_a) - \ln p(\boldsymbol{O}|\omega_b)$, and the subsequent weighting of each covariant derivative by $P(\omega_b|\boldsymbol{O})$. The log posterior is nonlinearly related to the individual log class likelihoods, and as such the posterior score space furnishes information which is not available to a linear discriminant operating in the likelihood-ratio score space. Unfortunately, the class-conditional models in these experiments were well-trained and despite the similarities in the E-set letters, class posteriors were very close to zero or unity. Unfortunately a trained classifier will then effectively mainly base its class decision on the value of the class posterior[6] and limit the use of information from covariant derivatives. There is no reason why the posterior score spaces $\varphi^{\mathrm{ps(a)}}(1, \boldsymbol{\xi}_0)$ and $\varphi^{\mathrm{ps(b)}}(1, \boldsymbol{\xi}_0)$ should yield the same performance.

[5] As an exception, problem *DvT* for $\varphi^{\mathrm{ps(a)}}(1, \boldsymbol{\xi}_0)$ required $C = 1000$ since there was none or slow convergence for $C = 100$.

[6] The posterior was not necessarily identical to that used by the MAP classifier since class likelihoods were calculated using all paths through the model rather than the single Viterbi path.

An SVM classifier was trained in the `mr` likelihood-ratio score space defined on ML estimated models for the full alphabet task. The complexity of the ML models was 2 mixture components per state. The classifier achieved a test error rate of 2.95%, whereas a MAP classifier based on the same models achieved 4.68% test error rate. The two classifiers differed with a 100% confidence level. The best MAP classifier for the full alphabet task achieved 3.40% test error rate for ML trained models with 4 mixture components per state. The SVM classifier differed from this at a 57% confidence level. The following details regard experimental implementation.

- The use of a single parameter C is unlikely to be equally acceptable to all score-spaces, particularly those of different sizes. The experimental results were therefore subject to assuming a fixed C parameter.

- Each log likelihood[7] in the `r` subspace which was used to define the covariant derivatives for the remaining components in score space was for the whole utterance, i.e. for the full `silence-letter-silence` hypothesis. It is also possible to constrain this log likelihood to the segment of speech recognised as `letter` and not as `silence`. Using the log likelihood of the whole utterance does indeed introduce some unwanted within-class variation to the `r` subspace since the log likelihood varies with the length of silence. However restricting the log likelihood to the speech segment requires a 'hard' decision on the silence/speech boundaries. This decision may in turn degrade performance. Furthermore, if the log likelihood is calculated over `silence-letter-silence`, then a 'soft' silence/letter segmentation is incorporated into the `wmvt` subspace by virtue of the letter state posteriors. Restricting the log likelihood to only the `letter` segment forces the incorporation of a 'harder' segmentation into the `wmvt` subspace. Therefore, it seems sensible to use the log likelihood of the full `silence-letter-silence` utterance in the definition of the score spaces. Furthermore, since little discrimination is expected between letters in the covariant derivatives with respect to the silence model parameters, these subspaces were always ignored in these experiments. This is fortunate since the calculation of the covariant derivatives with respect to self-transition probabilities detailed in Appendix B.3

[7] Or more exactly, 'soft' form of normalised log likelihood.

num. mixture comp. per model	ML		MMI-5		MMI-20	
	input MAP	lik.-ratio $\varphi^{\mathrm{lr(a,b)}}(1, \boldsymbol{\xi}_0)$	input MAP	lik.-ratio $\varphi^{\mathrm{lr(a,b)}}(1, \boldsymbol{\xi}_0)$	input MAP	lik.-ratio $\varphi^{\mathrm{lr(a,b)}}(1, \boldsymbol{\xi}_0)$
1	11.3	6.9 (100)	5.6	5.9 (15)	5.0	5.9 (89)
2	8.7	5.0 (99)	5.4	4.1 (79)	4.8	4.3 (0)
4	6.7	5.4 (74)	5.0	5.0 (0)	4.4	5.4 (60)
6	7.2	5.4 (89)	5.4	4.4 (56)	4.3	5.4 (66)

Table 6.4: E-set: Percentage test error rates for MAP classifiers and SVM score space classifiers (`mr` score space) defined on models of varying complexity trained under different training regimes (confidence levels relative to the corresponding MAP classifier using the same models are given in brackets)

does not permit repeated states. So in the experiments, score spaces were always restricted to the covariant derivatives with respect to letter model parameters only, but with log likelihoods calculated over the `silence-letter-silence` hypothesis.

It is also useful to present results for equivalent experiments but with score spaces defined on models trained by MMI estimation. Only the likelihood-ratio score space was considered since this score space yielded the best performance for ML trained models. Experimental results are detailed alongside the ML results in Table 6.4. It was difficult to substantiate some of the comparisons in terms of statistical significance. The MMI estimated models were very discriminative in comparison to ML estimated models of the same complexity, for example MMI-20 yielded test error rates as low as 4.3%. Constructing classifiers in likelihood-ratio score spaces sometimes worsened performance. In these cases, covariant derivatives simply added 'classification noise' to the log likelihood-ratio[8].

[8]Or more exactly, the ratio formed from the 'soft' form of the normalised log likelihoods.

score space	num. components	test error rate, %
r	1	8.5
v	1560	7.4 (55)
m	1560	5.2 (91)
mr	1561	5.0 (0) <99>
mv	3120	5.0 (0) ≪0≫
wmv	3140	4.4 (75)
mvt	3140	4.4 (0) [75]
wmvt	3160	4.1 (50)
wmvtr	3161	4.1 (0) <100> {67}

Table 6.5: E-set: Comparing SVM test error rates for different subspaces of score space (confidence levels: $(\cdot)$=relative to classifier on row above, $[\cdot]$=relative to mv, $\{\cdot\}$=relative to mr, $< \cdot >$=relative to r, $\ll \cdot \gg$=relative to m)

6.3.4 Feature selection in score space

In many pattern classification tasks, performance is improved by selecting a linear subspace of the original feature space thereby eliminating features which possess little class discriminative information. For computational reasons, experiments were limited to a 'filter' method of feature selection. Any conclusions drawn are only for this particular dataset and task. Linear subspaces were first selected by parameter type, a form of feature selection by expert knowledge. Test error rates are detailed in Table 6.5 with various confidence levels presented within delimiters.

Generally, adding new components or features increased classification performance. Both the mr and wmvtr score spaces were better than the r score space to respectively 99% and 100% confidence levels on test set performance. This illustrates that score space classifiers are indeed useful. The m score space was better than the v score space at a 91% confidence level. Intuitively, Gaussian means are more descriptive of particular classes than Gaussian variances, so their covariant derivatives should enable better class discrimination.

Variances simply give information pertaining to the variability of trajectories in acoustic space. To a 67% confidence level, the `wmvtr` score space was better than the `mr` score space. Class discriminative information therefore existed in the `wvt` subspace which was complimentary to that in the `mr` score space. Results suggest most of this complimentary information originated in the `t` or `w` subspaces.

Any advantage gained from adding extra components into the score space is normally precluded by the 'curse of dimensionality'. However the linear SVM can still return a robust and competitive solution even in large score spaces, but often with greater computational cost. For this reason performance must sometimes be compromised by the size of the score space. Although increased computational cost from increased score space size was not too problematic in these experiments, most of the experiments in this chapter nevertheless focus on the `mr` score space. The `t` or `w` score subspaces were not tested alone since they are unlikely to capture sufficient charactistics to distinguish classes.

Next, a data-driven method for feature selection was adopted, both in the `mr` and `wmvtr` score spaces. Each feature was ranked according to its Fisher ratio. Subspaces were then formed from components with highest Fisher ratios, either subspaces of fixed size or formed from all components with ratios above a threshold. The test error rates for SVM classifiers constructed in these subspaces are detailed in Figure 6.2. When a threshold was used in the value of the ranking criterion, the test error rate is plotted against the average size of score subspace across all 36 binary classifiers for the E-set task.

The performance curves describe minima in test error rate typical of feature selection. As the size of the score space decreases, performance improves as 'noisy' components are discarded but then worsens as useful components are discarded. The lowest test error rates for the `mr` and `wmvtr` score spaces were 3.2% at an average of 354 components per classifier and 3.3% at 500 components per classifier. These compare favourably with respective test error rates of 5.0% and 4.1% for the full score spaces at respective confidence levels of 94% and 50%. These results are subject to a single value of $C = 100$. No attempts were made to tune this parameter for each size of score space. Improved performance is expected from nonlinear feature extraction techniques, since scores typically occupy

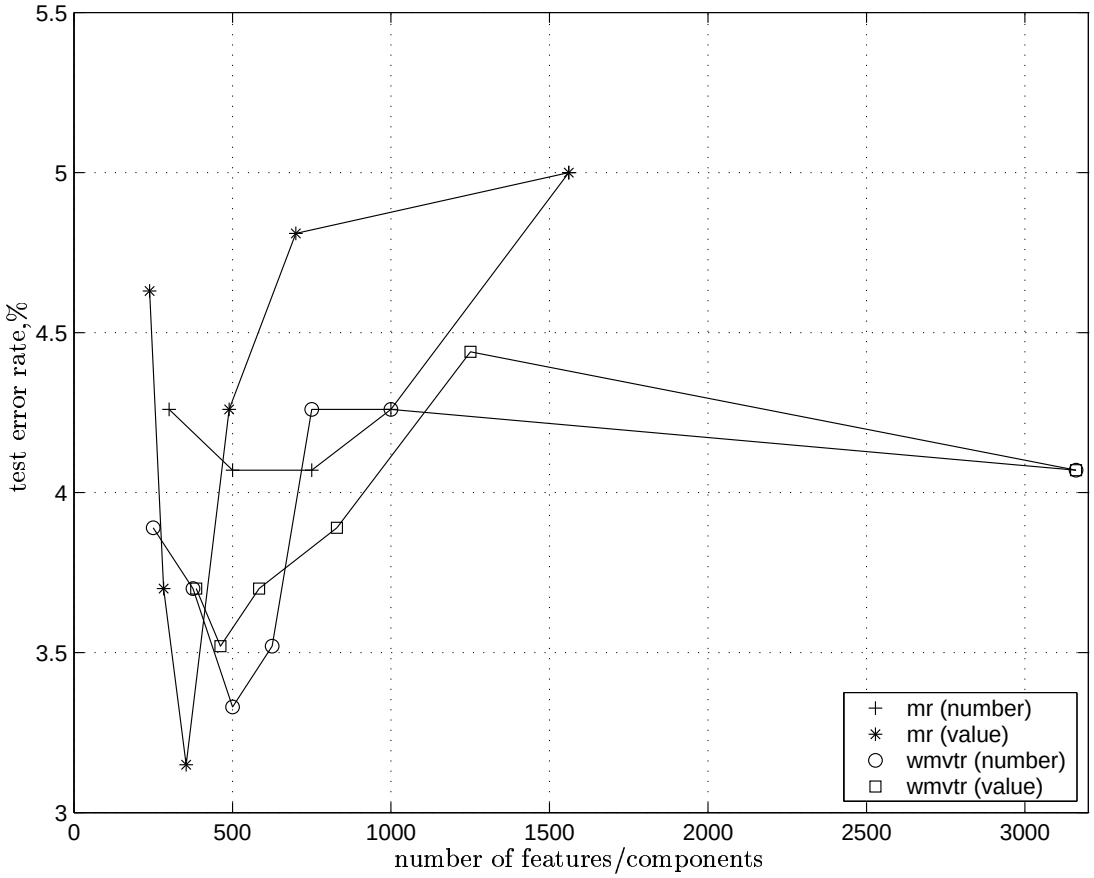


Figure 6.2: Comparing test error rates as the size of score spaces was varied using the Fisher ratio (thresholds either in the 'value' of the Fisher ratio in which case the average number of components per score space is plotted, or in the 'number' of components per score space)

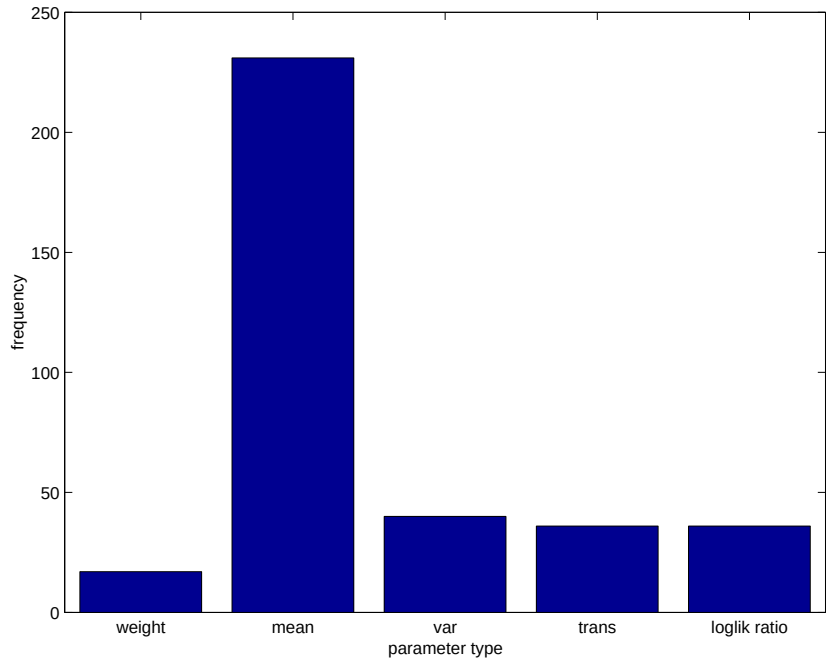
nonlinear structures with lower dimensionality than the size of the score space. However such techniques are often computationally expensive. Encouragingly in [101], experiments were conducted with Fisher score spaces defined on discrete HMMs modelling amino acid sequences, and it was remarked that most of the class discrimination was contained in relatively few components of Fisher score space.

A feature selection experiment was implemented in the `wmvtr` score space defined on the ML models, with 2 mixture components per state, for the full alphabet task. The 500 most discriminative components from each score space were selected. A test error rate of 2.12% was achieved. This was a decrease from 2.37% for a similar classifier in the full `wmvtr` score space, at a confidence level of 60%. The classifier at 2.12% outperformed MAP classifiers operating in the input space using ML models with 2 and 4 mixture components per state, with respective test error rates at 4.68% and 3.40%, and at respective confidence levels of 100% and 99%.

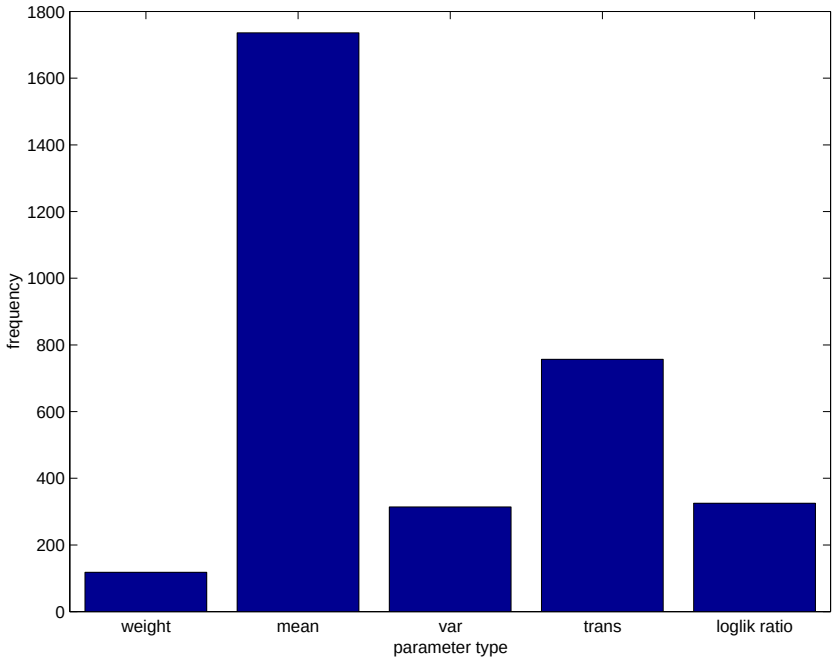
6.4 The importance of different HMM parameters for discriminating letters

The `wmvtr` score spaces were processed to furnish rankings according to the Fisher ratio. The 10 components with highest Fisher ratio for each binary problem were extracted and used to form sets of 'most discriminative features' for the training samples. There were two discriminative feature sets F_{eset} and F_{full} drawn respectively from the E-set and full alphabet tasks and containing respectively 360 and 3250 components. The feature sets were analysed and results presented in Figures 6.3 to 6.7. The first of each pair in Figures 6.3 to 6.6 describes F_{eset} and the second describes F_{full}.

- Figure 6.3 shows simple histograms of the frequency of the covariant derivatives with respect to each parameter type in the sets of most discriminative features. The Gaussian means have a greater representation than the Gaussian variances since they define the typical trajectories of each letter in the `MFCC_E_D_A` representation of

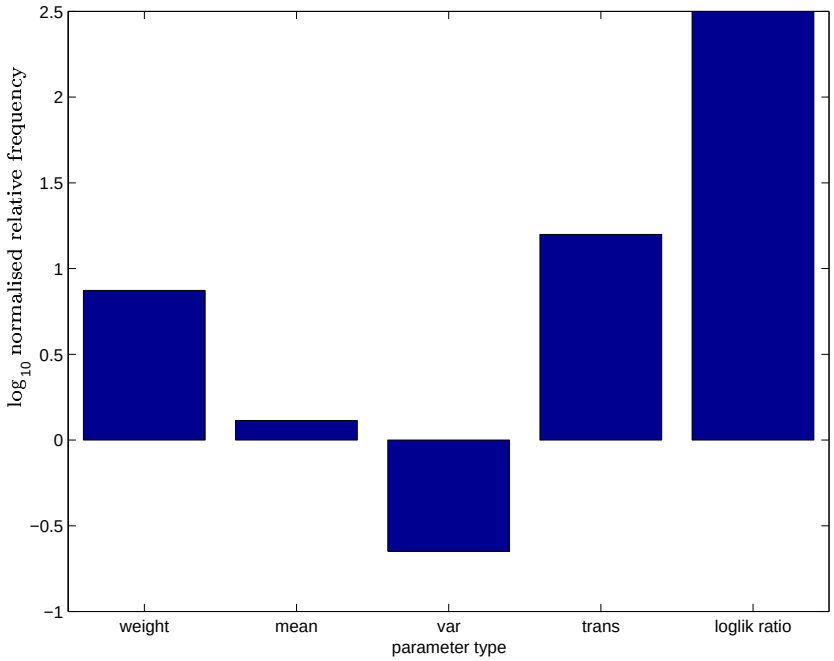


F_{eset}

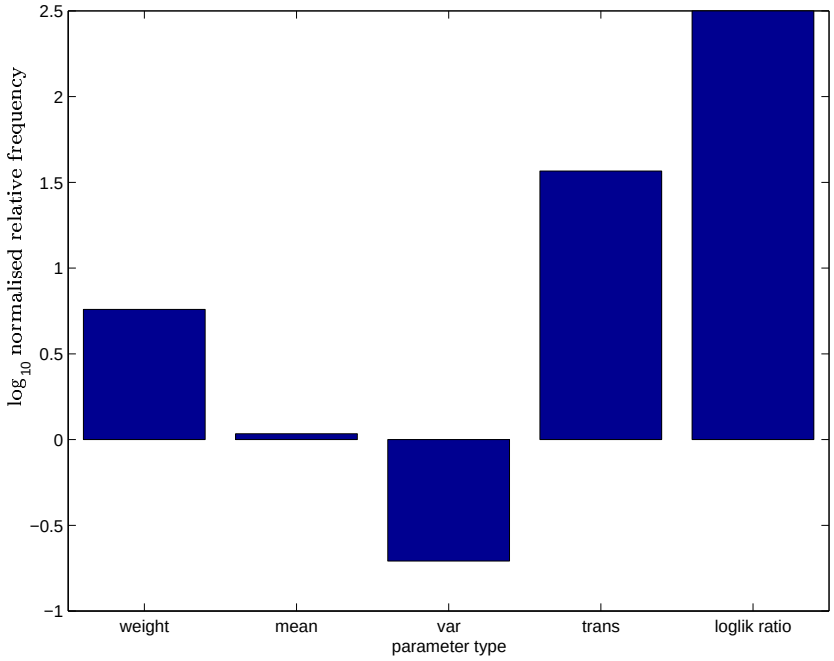


F_{full}

Figure 6.3: Frequency of each parameter type in the most discriminative feature sets

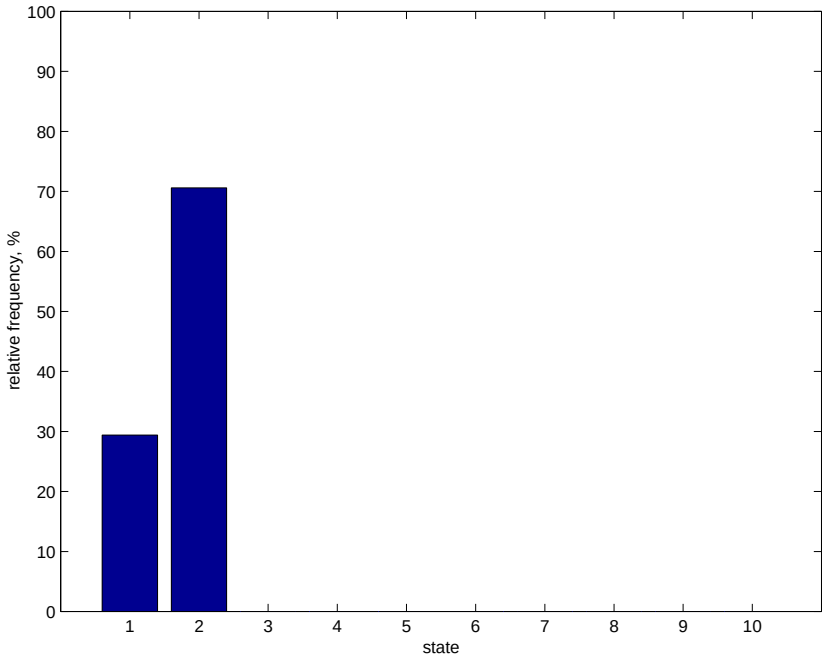


F_{eset}

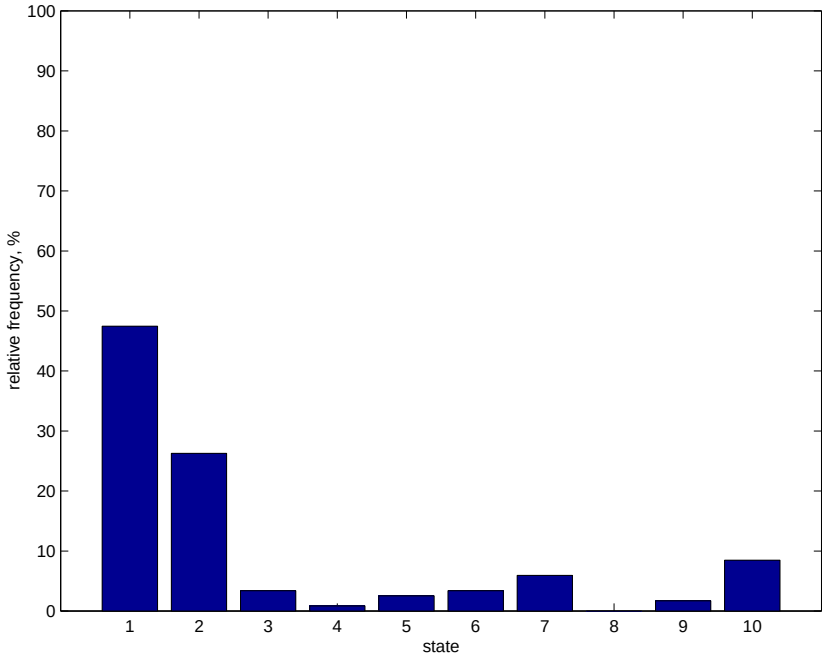


F_{full}

Figure 6.4: Normalised relative frequency of each parameter type in the most discriminative feature sets

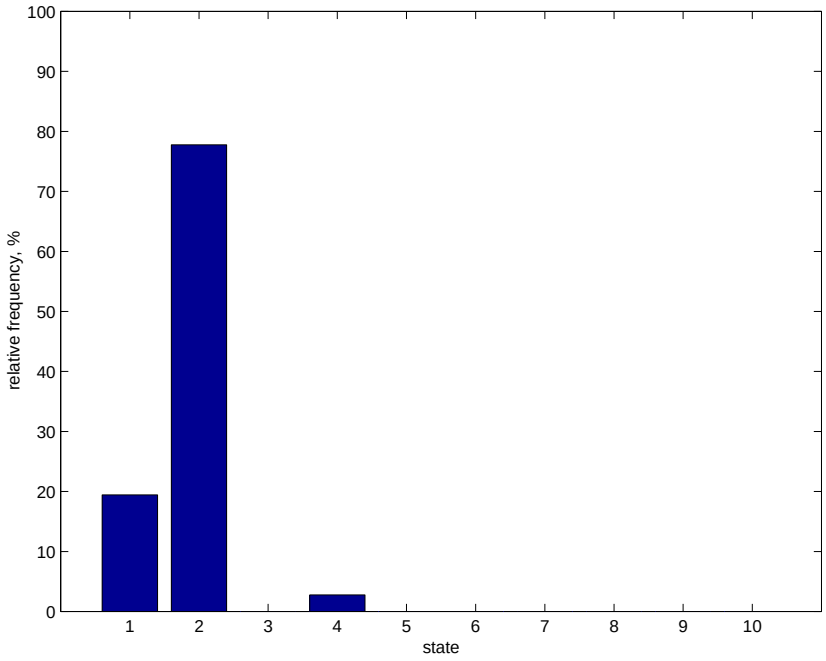


F_{eset}

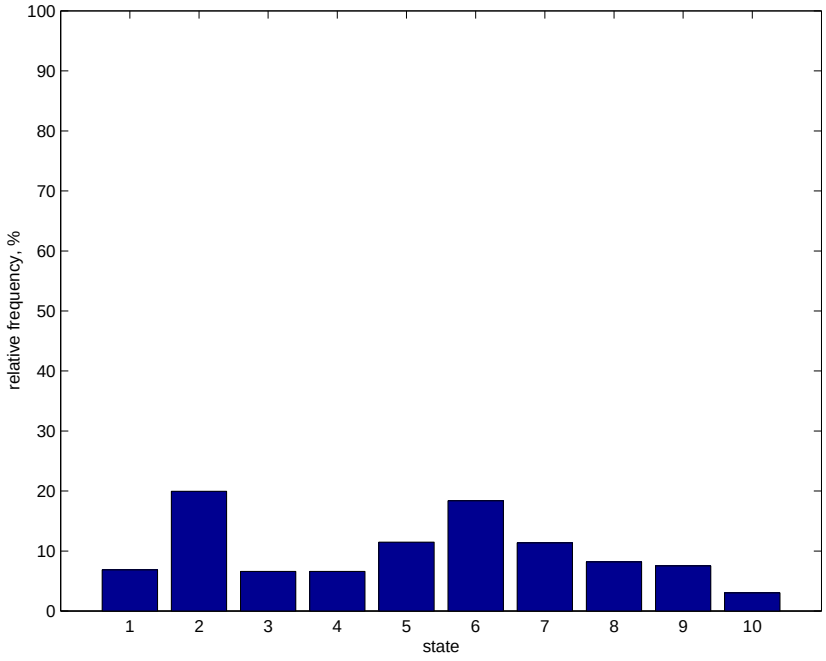


F_{full}

Figure 6.5: Relative frequency of weight terms for each state in the most discriminative feature sets

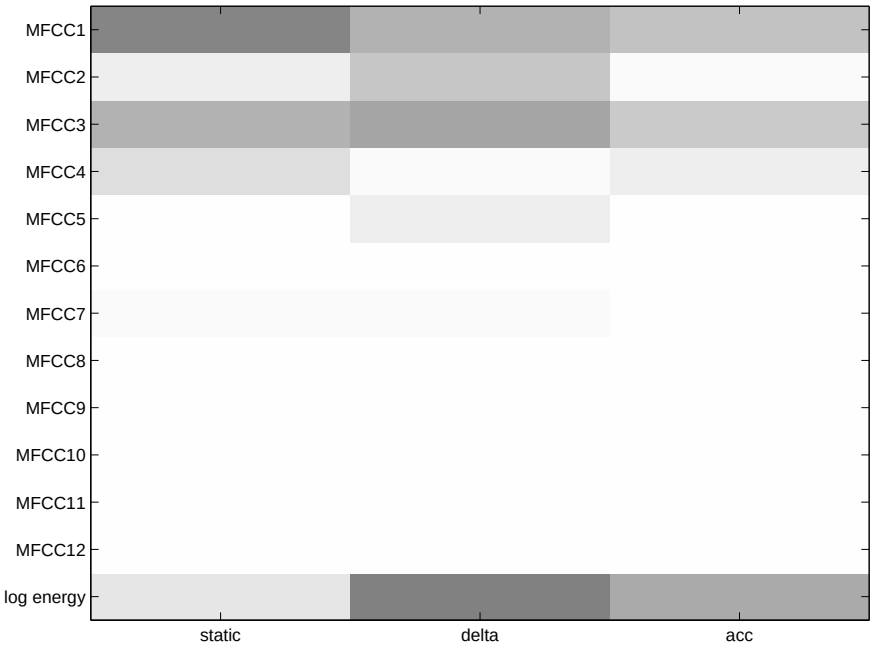


F_{eset}

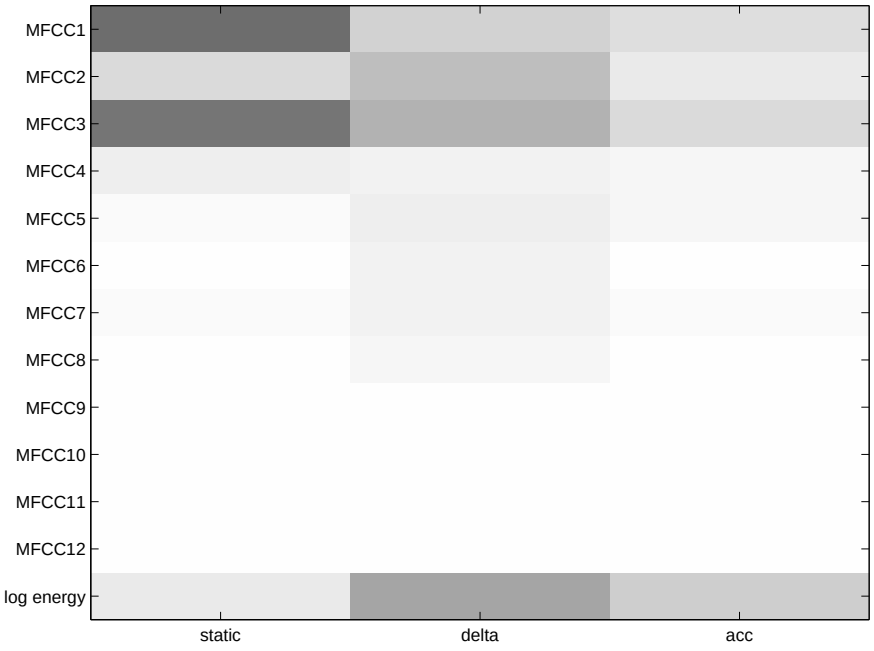


F_{full}

Figure 6.6: Relative frequency of self-transition probability terms for each state in the most discriminative feature sets



F_{eset}



F_{full}

Figure 6.7: Relative frequency of different MFCC and log energy terms in the most discriminative feature sets

acoustic space rather than the variability about the typical trajectories.

- Since 3120 of 3161 features are covariant derivatives of Gaussian means or variances, a greater representation of these parameters is expected in F_{eset} and F_{full}. An attempt to normalise for this yields the histograms in Figure 6.4. Rather than plot f, the frequency of a parameter type, the log normalised relative frequency l is plotted where,

$$l = \log_{10}\left(\frac{f}{n} \times \frac{\sum n}{\sum f}\right) \tag{6.1}$$

where the number of components in score space $\sum n = 3161$ and,

$$\sum f = \begin{cases} 360 & \text{for} \quad F_{\text{eset}} \\ 3250 & \text{for} \quad F_{\text{full}} \end{cases} \tag{6.2}$$

According to the normalised relative frequencies, the log likelihood-ratio and covariant derivatives of the self-transition probabilities and mixture component weights are the most class discriminative features. The log likelihood-ratio is a powerful feature since it summarises the modelling ability of all the parameters within each HMM. The derivatives of self-transition probabilities are informative since they contain duration information. Since all E-set letters are of similar duration, the covariant derivatives of self-transition probabilities have a smaller representation in the normalised relative frequency histogram of the E-set than of the full alphabet. In part, these normalised histograms are unfairly biased against the features derived from Gaussian means and variances. There is probably only a small subset of these features which are discriminative, and their importance is unfairly penalised by the remaining features of the same type which never occur within the sets F_{eset} or F_{full}.

- Figure 6.5 details the variation of features defined on mixture component weights across their defining states. All weight-dependent features in F_{eset} belong to the 1st and 2nd emitting states. This is sensible since most of the class distinguishing characteristics for the E-set originate at the beginning of the letters. The spread of weights across states is much more diffuse for the full alphabet task since there is greater acoustic variety between letters. Overall, the initial states and final state are important for class discrimination. Such states model the transitions between silence and the letter, or vice versa.

- Figure 6.6 details the variation of features defined on self-transition probabilities across their defining states. For the E-set task, the features are concentrated on the initial two states, reflecting the trend in the weight-dependent features. As for the weight-dependent features, there is a greater spread for the full alphabet task. However, excepting the 2nd state, the greatest representation is for the central 5th, 6th and 7th states. These states model the duration aspects of the central part of each letter. This must be a valuable cue for distinguishing non E-set letters, particularly those with short sharp utterances or longer ones.

- Figure 6.7 plots the relative frequency of each MFCC term and log energy term in F_{eset} and F_{full}, distributed across the zeroth, first and second order derivatives with respect to time (respectively the static, delta and acceleration parameters). The darker is the rectangle in the image, the greater is its frequency of occurrence. The plots for the E-set and full alphabet tasks are in broad agreement in that low order MFCCs possess more class discriminative information than higher order MFCCs, and the log energy term is also valuable. This is consistent with general knowledge in speech recognition. In both cases, the most important parameters for the log energy terms are the delta, then acceleration then static. Since the static log energy term is dependent on loudness and channel conditions, the dynamic aspects of the log energy term, as encapsulated in its delta and acceleration parameters, must be more reliable cues for distinguishing letters. The plot for the full alphabet task has a slightly higher representation of middle order MFCCs in the delta stream. This may be a consequence of the greater durational variability in the full alphabet than in the E-set. The results suggest that discarding all covariant derivatives relative to MFCC8 to MFCC12 in the definition of the score space should not seriously degrade performance. This suggests another method of feature selection based on expert knowledge.

task	feature space	classifier type	brief description	test error rate,%
E-set	input	MAP	ML, 4 mix. comp.	6.7
E-set	input	MAP	MMI-20, 6 mix. comp.	4.3
E-set	score space	SVM	subspace of `mr`	3.2
full	input	MAP	ML, 4 mix. comp.	3.40 (7.2)
full	score space	SVM	subspace of `wmvtr`	2.12 (3.5)

Table 6.6: The best performing classifiers from the experiments in this chapter (E-set subset results given in brackets)

6.5 Comparison with other ISOLET classifiers

The aim of these experiments has been to illustrate various properties of score spaces rather than fine-tune performance. However it is useful to compare the results with those of other techniques. The best performing classifiers from this section are summarised in Table 6.6 with a brief description.

First, it is important to establish whether the HMM-based MAP classifiers used as baselines in the full alphabet experiments are state-of-the-art for the 39-component MFCC-based feature space. Similar feature spaces are applied in the following research though model topologies differ. For the same input feature space, [55] report a best ML baseline of 3.01%, and a best performance of 2.95% for HMMs trained using Frame Discrimination. Neglecting the acceleration parameters and adopting a 26-component feature vector, [103] achieves a best ML baseline of 3.91%, and a test error rate of 3.40% following MMI training. For a 39-component feature vector based on subband features, [72] describes a ML baseline of 3.3% test error rate. The ML baseline in this chapter is reasonable at 3.40%.

The best performing classifier that the author has found on the ISOLET task was the HMM-based classifier in [56]. Their classifier yielded 2.0% test error rate on the full alphabet task, the E-set subset of which recorded 2.8% test error rate. The SVM results in Table 6.6 are a little worse. It is important to note that [56] report E-set results

within the full alphabet task, rather than on a dedicated E-set task. Hence their result may be regarded as an upper bound on the test error rate for the dedicated task. Their system included an endpoint detection algorithm and utilised signal modelling techniques to extract new features. The results reported in this chapter did not use endpoint detection explicitly but utilised a separate silence model. The next best system the author has found is in [22]. It has no endpoint detection and yielded a test error rate of 2.6% for the full alphabet task. The score space classifiers described in this chapter are therefore competitive with state-of-the-art systems. A summary of some results on the ISOLET task and other alphabet and alphadigit tasks is contained in [67]. One of the original neural network-based classifiers built for this task was described in [30] and [29], reporting an E-set test error rate of 5% and full alphabet rate of 4%.

6.6 Summary

HMMs are an important type of statistical model appropriate for variable length patterns and may be used to define score spaces. Experiments were performed illustrating how static classifiers such as SVMs can be applied to dynamic classification tasks via score spaces. The experiments concentrated on the classification of isolated letter utterances. The performance of HMMs was improved by training SVM score space classifiers. Feature selection and a careful selection of metric matrix for score space were shown to increase performance. Experiments also suggested that sequence length normalisation is more useful when there are other means of retaining duration information in the score space.

Chapter 7

Conclusions and future work

7.1 Conclusions

Since the nature of a data source is usually more complicated than the statistical model proposed for it, the performance of inference algorithms defined on statistical models is suboptimal due to model incorrectness. There is therefore good sense in augmenting a statistical model to form a fibre bundle in the space of probability distributions. Each fibre is an exponential family which contains as a subfamily local approximates to a distribution in the original statistical model. The fibre bundle 'captures' a greater variety of distributions and can potentially furnish better estimates for the data source. Vector bundles and score spaces can be defined as 'tools' to train distributions in fibre bundles. Various training criteria were proposed including a maximum likelihood approach and discriminative techniques, including an implicit estimation through training a linear discriminant in score space. 'Fibre hopping' was also proposed as a maximum likelihood technique which involves repeated augmentation of the fibre bundle. This thesis has developed the relations between such fibre bundles, vector bundles, score spaces and training techniques. Fibre bundles may also be defined as structures within the space of scalar functions and hence only include distributions as a structural subset. These bundles permit a formalisation of truncated Taylor expansions of scalar fields defined on statistical models. Essentially a

truncated Taylor expansion is the evaluation of the scalar field at points within the total space of the bundle defined with the statistical model or models as its base manifold. Taking a more general view, the mapping from input space to score space may be viewed as a model-dependent feature extraction process. A simple model or classifier can then be constructed in the score space. It is possible to view this score space as that induced by a kernel. An example of such a kernel is the Fisher kernel.

The classification performance of score spaces is primarily influenced by the choice of the defining scalar field. This influences performance through the noninjective nature of the mapping from input space to score space, and the magnification the mapping induces near to the decision boundary. Performance is also affected by the number of training samples available to the classifier in score space and the nature and properties of that classifier. Sequence length normalisation was described to reduce unwanted within-class variability due to the length of patterns. In particular, techniques were developed for signals which may be modelled as contiguous quasi-stationary segments.

A common classification technique for fixed or variable length patterns proposes a set of statistical models, one for each class, and selects the class with the maximum posterior probability. This yields the lowest probability of error but only if the statistical models and class priors are correct and an appropriate loss function is selected. This thesis encourages an approach where the patterns are mapped into a fixed length score space. Static classifiers with good regularisation properties, for example SVMs, can then be applied to train decision rules discriminatively and counteract some model incorrectness. For certain score spaces and solutions, a linear classifier trained in score space can be related to a maximum a-posteriori decision rule operating on distributions in the fibre bundle defined on the original statistical models. The experiments in this thesis demonstrate the promise of this approach.

Experiments on fixed length patterns were conducted on an artificial dataset and a small vowel dataset. Various score spaces and classifiers were investigated and compared. The most promising classifier on the vowel data was formed by training 1-v-1 linear SVMs and combining their decisions in a majority voting scheme. Each SVM was trained in a

likelihood-ratio score space defined on the statistical models for the two relevant classes. This classifier outperformed the best GMM classifiers trained directly in the input space, and was comparable to state-of-the-art systems for this vowel task.

Experiments were also performed for variable length patterns on an isolated letter speech classification task. Different score spaces were compared within the framework of training 1-v-1 linear classifiers and combining their individual decisions in a majority voting scheme. Again linear SVMs trained in the likelihood-ratio score space yielded the best performance. Feature selection improved performance by eliminating 'noisy' features. The results obtained were comparable to state-of-the-art systems on this task. Furthermore, an objective analysis of the relative importance of different features in score space for discriminating letters reinforced common knowledge as to which parameters of HMMs are the most useful for discrimination.

7.2 Future work

There are a number of interesting avenues of research, both theoretical and applied. Some of these are listed below with some helpful, though not a comprehensive list of, citations. The development of fibre bundles and score spaces suggests some applications which were not implemented in this thesis. These include,

- training a distribution in the total space of the fibre bundle to maximise likelihood, either by defining a score space with $\tau = 1$ rather than $\tau = \tau^{\text{tay}}$ or incorporating the normalisation term into the optimisation process,
- extending the maximum likelihood approach to implement 'fibre hopping' and ascertaining that the likelihood of the training set is nondecreasing, and that the solution is prone to overtrain with limited training data,
- deriving expressions for mapping into a score space which includes second degree covariant derivatives, and comparing distributions or classifiers trained in such score spaces with those trained using the simpler score spaces in this thesis,

- investigating the application of distance measures within the total space of the fibre bundle, for example the distance along geodesics of the fibres.

With regard to improving the performance of classifiers in score spaces, there are a number of possible directions of research.

- The application of SVMs to multicategory classifiers defined in score spaces has been constrained to training 1-v-1 linear SVMs and combining their outputs in a majority voting scheme. It would be interesting to apply multicategory kernel methods [109] [26] to score spaces, and also techniques aimed at combining binary decisions more effectively for multicategory tasks [69] [18].

- The interaction between different nonlinear SVMs and score spaces has not been investigated in this thesis. Since the data in score spaces is expected to have a nonlinear spread, properly regularised nonlinear SVMs may yield better classifiers. However it may be difficult to relate nonlinear discriminants to implicitly trained distributions in fibre bundles.

- The components in score space are typically highly correlated and scores within score space typically occupy nonlinear structures of much lower dimension than the size of the score space. The application of nonlinear feature extraction techniques [98] [86] may be promising.

- Score spaces have only been defined on HMMs with state-conditional likelihoods modelled by GMMs. It may be worthwhile to derive score spaces on other statistical models such as other Linear Gaussian Models [85] [84]. Discrete output HMMs are applied in protein classification, and it may be interesting to continue the research in [52] by applying some of the concepts from this thesis.

- Score spaces defined on the covariant derivatives of linear posteriors were not investigated in this thesis, and their comparison with other score spaces would be useful.

- In the experiments in this thesis, SVMs were the only discriminative training technique applied to binary subdivisions of a multicategory task for nontrivial score

spaces. Experiments with other discriminatively-trained binary classifiers would be useful to discern the effect of introducing SVMs.

- It may be useful to apply 'spherical normalisation' [106], which was found in [106] to be particularly appropriate for likelihood-ratio score spaces (this normalisation aims to yield a better conditioned Hessian matrix for the SVM quadratic programming solution).

With regard to the application to continuous speech recognition, there are some significant difficulties.

- The framework developed in this thesis does not accomodate the union of score space classifiers defined on contiguous segments of speech. For example, a classifier trained to distinguish segments ω_a and ω_b cannot be used to construct a classifier to distinguish the double segments (ω_a, ω_b) and (ω_b, ω_a). Any method to achieve this opens up the possibility of recognising a series of contiguous segments using classifiers trained to distinguish individual segments.

- The SVM framework developed in this thesis requires SVM binary score space classifiers to be constructed to distinguish pairs of competing hypotheses. In the experiments in this thesis, the competing hypotheses were vowels or isolated letters. In continuous speech recognition, the competing hypotheses are usually at least sentences. The number of competing hypotheses is prohibitive, unless restricted to an N-best list and the technique applied to rescoring (i.e. re-ranking the members of the N-best list). However even for the N-best list, there is usually insufficient data to train a classifier to distinguish two hypotheses, since there is usually only one example of each hypothesis. This problem can be circumvented by simulating data through a model of the sentence, for example that provided by concatenating HMMs for constituent units of speech [79]. It is also possible to use the constituent HMMs to estimate a classifier in score space without simulating data where the mapping from the parameters of the HMMs to the parameters of the score space classifier is deterministic. Though promising, this latter technique must necessarily make assumptions about the distribution of scores in score space.

Overall, there are significant challenges to the application of fibre bundles and score spaces to continuous speech recognition, and any advantages gained from this approach should be weighed against the advantages of alternative approaches such as those described in Section 2.5. However the application of fibre bundles and score spaces to tasks where the number of competing hypotheses is small, as in the experiments in this thesis, is promising.

Appendix A

Exponential families of distributions

An exponential family of distributions may be written as,

$$S(\boldsymbol{\alpha}) \;=\; \{p(\boldsymbol{O};\boldsymbol{\alpha}) \;\;|\;\; \boldsymbol{\alpha} \in L(\boldsymbol{\alpha};S)\} \tag{A.1}$$

where $L(\boldsymbol{\alpha};S)$ is an open set describing valid distributions and from Section 2.3 of [3],

$$p(\boldsymbol{O};\boldsymbol{\alpha}) \;=\; \exp\{C(\boldsymbol{O}) + \sum_{r=1}^{\varrho} \sum_{j_1 \ldots j_r} \alpha^{j_1 \ldots j_r} F_{j_1 \ldots j_r}(\boldsymbol{O}) - D(\boldsymbol{\alpha})\} \tag{A.2}$$

and,

$$D(\boldsymbol{\alpha}) \;=\; \ln \int \exp\{C(\boldsymbol{O}) + \sum_{r=1}^{\varrho} \sum_{j_1 \ldots j_r} \alpha^{j_1 \ldots j_r} F_{j_1 \ldots j_r}(\boldsymbol{O})\} \mathrm{d}\boldsymbol{O} \tag{A.3}$$

The summation over $j_1 \ldots j_r$ implies all possible permutations with $j_i \in \{1, \ldots n\}, \forall i$ for some positive integer n. Each parameter $\alpha^{j_1 \ldots j_r}$ is called a *natural parameter* and $F_{j_1 \ldots j_r}(\boldsymbol{O})$ is its *sufficient statistic*.

An example of an exponential family of distributions is that of univariate Gaussians where $n = 2,\ \varrho = 1$ [3],

$$\begin{aligned} \alpha^1 &= \frac{\mu}{\sigma^2} \\ \alpha^2 &= -\frac{1}{2\sigma^2} \\ F_1(O) &= O \end{aligned}$$

$$
\begin{aligned}
F_2(O) &= O^2 \\
C(O) &= 0 \\
D(\boldsymbol{\alpha}) &= \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2)
\end{aligned}
\tag{A.4}
$$

with the single constraint $\alpha^2 < 0$. In more familiar guise,

$$
p(O;\boldsymbol{\alpha}) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\{-\frac{1}{2\sigma^2}(O-\mu)^2\} \tag{A.5}
$$

Appendix B

The Taylor expansion along the manifold

B.1 Expressions for the Taylor expansion along the manifold

This section gives an expression for the Taylor expansion along a manifold. The manifold has global coordinate system $[\theta^i]$ and is denoted $S(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the coordinate vector. The manifold is of dimension n and is as described in Section 3.2. An affine connection is assumed described by connection coefficients $\Gamma^k_{ij}, \forall k, i, j$. A scalar field $\varsigma_l : p \mapsto \varsigma(\boldsymbol{O}_l; p)$ varies over $S(\boldsymbol{\theta})$ and, under the coordinate chart $\psi : p \mapsto \boldsymbol{\theta}$ yields the scalar field $\bar{\varsigma}_l : \boldsymbol{\theta} \mapsto \bar{\varsigma}(\boldsymbol{O}_l; \boldsymbol{\theta})$.

The value of $\bar{\varsigma}_l$ at point $\boldsymbol{\theta}'$ may be recovered by a Taylor expansion about point $\boldsymbol{\theta}_0$, assuming $\bar{\varsigma}_l$ is analytic at $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}'$ lies within its convergence domain. Then,

$$(\bar{\varsigma}_l)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} = \sum_{r=0}^{\infty} \frac{1}{r!} \sum_{j_1 \ldots j_r} (\bar{\varsigma}_l)_{;j_1;\ldots;j_r}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \prod_{i=1}^{r} (\theta'^{j_i} - \theta_0^{j_i}) \tag{B.1}$$

where the summation over $j_1 \ldots j_r$ is for all possible permutations where $j_i = \{1, \ldots n\}, \forall i$. From the more general expression for the covariant derivatives of tensors in mixed compo-

nent form in Section 403.B of [50], the covariant derivative of a fully covariant component of a tensor under the affine connection is,

$$(\bar{\varsigma}_l)_{;j_1;\ldots;j_r} = \frac{\partial}{\partial\theta^{j_r}}(\bar{\varsigma}_l)_{;j_1;\ldots;j_{(r-1)}} - \sum_{l=1}^{(r-1)}\sum_{m=1}^{n}\Gamma^m_{j_r j_l}(\bar{\varsigma}_l)_{;j_1;\ldots;j_{(l-1)};m;j_{(l+1)};\ldots;j_{(r-1)}} \quad \text{(B.2)}$$

where the index m has replaced the index j_l in the covariant derivative in the second term. The expression for $(\bar{\varsigma}_l)_{;j_1;\ldots;j_r}$ is a function of the covariant derivatives of form $(\bar{\varsigma}_l)_{;j_1;\ldots;j_{(r-1)}}$. This suggests a recursive though nontrivial relationship for covariant derivatives of ever-increasing degree.

An approximation to the Taylor expansion results from truncating the series in Equation B.1. A ϱth order approximation is,

$$(\bar{\varsigma}_l)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} = \sum_{r=0}^{\varrho}\frac{1}{r!}\sum_{j_1\ldots j_r}(\bar{\varsigma}_l)_{j_1;\ldots;j_r}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\prod_{i=1}^{r}(\theta'^{j_i}-\theta_0^{j_i}) \quad \text{(B.3)}$$

The approximation is exact if the scalar field $\bar{\varsigma}_l$ on $S(\boldsymbol{\theta})$ has highest nonzero covariant derivative of degree r where $r \leq \varrho$. At least in the case where $[\theta^i]$ is a Euclidean coordinate system for the manifold, then the approximation error E is given in Sections 109.E and 109.J of [50] and Section 3.6.2 of [81] as,

$$E = \frac{1}{(\varrho+1)!}\sum_{j_1\ldots j_{(\varrho+1)}}(\bar{\varsigma}_l)_{;j_1;\ldots;j_{(\varrho+1)}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_\epsilon}\prod_{i=1}^{(\varrho+1)}(\theta'^{j_i}-\theta_0^{j_i}) \quad \text{(B.4)}$$

where,

$$\boldsymbol{\theta}_\epsilon = \boldsymbol{\theta}_0 + \epsilon(\boldsymbol{\theta}' - \boldsymbol{\theta}_0) \quad \text{(B.5)}$$

for some $0 < \epsilon < 1$. An upper bound in E is given by differentiating E with respect to $\boldsymbol{\theta}_\epsilon$ (see Section 3.6.2 of [81]). An interesting investigation of the Taylor expansion of a metric tensor g along a manifold is detailed in [43].

B.2 Dependence of the Taylor expansion on the coordinate system of the manifold

The Taylor expansion is not a property of the statistical manifold alone but also of the coordinate system [107] (see the Acknowledgments). To illustrate this, a manifold S of

dimension n is given two global coordinate systems $[\theta^i]$ and $[u^i]$ with respective coordinate charts,

$$\psi_1 : S \rightarrow L(\boldsymbol{\theta}; S) \tag{B.6}$$

$$\psi_2 : S \rightarrow L(\boldsymbol{u}; S) \tag{B.7}$$

Both $L(\boldsymbol{\theta}; S)$ and $L(\boldsymbol{u}; S)$ are open sets in $\mathbb{R}^n$. The mapping from one coordinate space to another is χ where,

$$\chi = \psi_2 \circ \psi_1^{-1} : L(\boldsymbol{\theta}; S) \rightarrow L(\boldsymbol{u}; S) \tag{B.8}$$

which may be expressed in terms of n component-level functions,

$$\chi = (\chi^1, \ldots \chi^n) \tag{B.9}$$

A scalar field ς is defined over S where $\varsigma : S \rightarrow \mathbb{R}$. Then two scalar fields are defined over the coordinate spaces,

$$\bar{\varsigma_1} : L(\boldsymbol{\theta}; S) \rightarrow \mathbb{R} \tag{B.10}$$

$$\bar{\varsigma_2} : L(\boldsymbol{u}; S) \rightarrow \mathbb{R} \tag{B.11}$$

where,

$$\bar{\varsigma_1} = \varsigma \circ \psi_1^{-1} \tag{B.12}$$

$$\bar{\varsigma_2} = \varsigma \circ \psi_2^{-1} \tag{B.13}$$

so,

$$\bar{\varsigma_1} = \bar{\varsigma_2} \circ \chi \tag{B.14}$$

If the change in the $[\theta^i]$ coordinate system is $\boldsymbol{h}_1 = \boldsymbol{\theta'} - \boldsymbol{\theta}_0$ where $\boldsymbol{\theta}_0 = \psi_1(p_0)$ and $\boldsymbol{\theta'} = \psi_1(p')$, then the equivalent change in the $[u^i]$ coordinate system is $\boldsymbol{h}_2 = \boldsymbol{u'} - \boldsymbol{u}_0$ where,

$$\boldsymbol{u'} = \chi(\boldsymbol{\theta'}) \tag{B.15}$$

$$\boldsymbol{u}_0 = \chi(\boldsymbol{\theta}_0) \tag{B.16}$$

Of course $\boldsymbol{\theta}', \boldsymbol{\theta}_0 \in L(\boldsymbol{\theta}; S)$ and $\boldsymbol{u}', \boldsymbol{u}_0 \in L(\boldsymbol{u}; S)$. Then in component form,

$$\begin{aligned}(h_2)^i &= (u'^i - u_0^i) \\ &= \int_{u_0^i}^{u'^i} \mathrm{d}u^i \\ &= \sum_{k=1}^{n} \int_{\theta_0^k}^{\theta'^k} \chi^i_{;k}\Big|_{\boldsymbol{\theta}} \mathrm{d}\theta^k \end{aligned} \tag{B.17}$$

In linear algebraic form,

$$\boldsymbol{h}_2 = \int_{\boldsymbol{\theta}_0}^{\boldsymbol{\theta}'} \underline{\boldsymbol{J}}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{B.18}$$

where $\underline{\boldsymbol{J}}(\boldsymbol{\theta})$ is defined as follows, where all covariant derivatives are evaluated at $\boldsymbol{\theta}$,

$$\underline{\boldsymbol{J}}(\boldsymbol{\theta}) = \begin{bmatrix} \chi^1_{;1} & \cdots & \cdots & \chi^1_{;n} \\ \vdots & \vdots & \vdots & \vdots \\ \chi^n_{;1} & \cdots & \cdots & \chi^n_{;n} \end{bmatrix} \tag{B.19}$$

The bold font with an underbar denotes a matrix formed from components of a type $(1,1)$ tensor. In general $\boldsymbol{h}_2$ cannot be expressed as a simple function of $\boldsymbol{h}_1$ unless $\underline{\boldsymbol{J}}(\boldsymbol{\theta})$ is invariant along the path of integration in the coordinate space between $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}'$, or $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are infinitessimals. In either of these cases, in component form,

$$(h_2)^i = \sum_{k=1}^{n} \chi^i_{;k}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (h_1)^k \tag{B.20}$$

In linear algebraic form,

$$\boldsymbol{h}_2 = \underline{\boldsymbol{J}}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \boldsymbol{h}_1 \tag{B.21}$$

Only if the frame of reference between $[\theta^i]$ and $[u^i]$ is such that there is no curvature then covariant derivatives of χ^i reduce to partial derivatives of the corresponding order,

$$\chi^i_{;j_1;\ldots;j_r} = \frac{\partial^r \chi^i}{\partial\theta^{j_1} \ldots \partial\theta^{j_r}} \tag{B.22}$$

Then the matrix $\underline{\boldsymbol{J}}(\boldsymbol{\theta})$ reduces to the familiar form of the Jacobian matrix [108].

To establish the task, let the scalar fields $\bar{\varsigma}_1$ and $\bar{\varsigma}_2$ be expanded about respective points $\boldsymbol{\theta}_0$ and $\boldsymbol{u}_0$ and used to evaluate scalar fields at $\boldsymbol{\theta}'$ and $\boldsymbol{u}'$ respectively. The analysis assumes that both $\bar{\varsigma}_1$ and $\bar{\varsigma}_2$ are analytic at $\boldsymbol{\theta}_0$ and $\boldsymbol{u}_0$ respectively, and that $\boldsymbol{\theta}'$ and $\boldsymbol{u}'$ lie within

the relevant convergence domains. Then explicitly noting the point of evaluation of each scalar field or its covariant derivative,

$$\bar{\varsigma}_1\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} = \bar{\varsigma}_1\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \sum_{i=1}^{n}(\bar{\varsigma}_1)_{;i}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(h_1)^i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\bar{\varsigma}_1)_{;i;j}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(h_1)^i(h_1)^j + \ldots \tag{B.23}$$

$$\bar{\varsigma}_2\Big|_{\boldsymbol{u}=\boldsymbol{u}'} = \bar{\varsigma}_2\Big|_{\boldsymbol{u}=\boldsymbol{u}_0} + \sum_{i=1}^{n}(\bar{\varsigma}_2)_{;i}\Big|_{\boldsymbol{u}=\boldsymbol{u}_0}(h_2)^i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\bar{\varsigma}_2)_{;i;j}\Big|_{\boldsymbol{u}=\boldsymbol{u}_0}(h_2)^i(h_2)^j + \ldots \tag{B.24}$$

If the Taylor expansion is invariant to the coordinate system, then the scalar value yielded by each term should be identical irrespective of the coordinate system. To test this proposal, the Taylor expansion for $\bar{\varsigma}_1|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}$ in Equation B.23 is processed and an attempt made to map it into the expansion for $\bar{\varsigma}_2|_{\boldsymbol{u}=\boldsymbol{u}'}$ in Equation B.24. The terms $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ are assumed infinitessimal so that the relation in Equation B.20 is valid. First applying Equation B.14 to the zeroth order term,

$$\begin{aligned}\bar{\varsigma}_1\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &= (\bar{\varsigma}_2 \circ \chi)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\ &= \bar{\varsigma}_2\Big|_{\boldsymbol{u}=\boldsymbol{u}_0}\end{aligned} \tag{B.25}$$

The value of the zeroth order term is invariant to a change in coordinate system. Next applying the chain rule for the first order term, where for brevity all covariant derivatives of χ^i are assumed evaluated at $\boldsymbol{\theta}_0$,

$$\begin{aligned}\sum_{i=1}^{n}(\bar{\varsigma}_1)_{;i}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(h_1)^i &= \sum_{i=1}^{n}(\bar{\varsigma}_2 \circ \chi)_{;i}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(h_1)^i \\ &= \sum_{i=1}^{n}(\bar{\varsigma}_2)_{;i}\Big|_{\boldsymbol{u}=\boldsymbol{u}_0}\sum_{k=1}^{n}\chi^i_{;k}(h_1)^k \\ &= \sum_{i=1}^{n}(\bar{\varsigma}_2)_{;i}\Big|_{\boldsymbol{u}=\boldsymbol{u}_0}(h_2)^i\end{aligned} \tag{B.26}$$

The value of the first order term is likewise invariant to the coordinate system. The evaluation of the second order term is slightly more complicated due to the differentiation of a product and the double application of the chain rule. Again assuming the covariant derivatives of χ^i are evaluated at $\boldsymbol{\theta}_0$, the second order term in the Taylor expansion less the scale factor of $1/2$ is,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(\bar{\varsigma}_1)_{;i;j}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(h_1)^i(h_1)^j$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\Big(\sum_{k=1}^{n}(\bar{\varsigma}_2)_{;k}\Big|_{\boldsymbol{u}=\boldsymbol{u}_0}\chi^{k}_{;i}\Big)_{;j}(h_1)^i(h_1)^j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\Big(\sum_{m=1}^{n}(\bar{\varsigma}_2)_{;k;m}\Big|_{\boldsymbol{u}=\boldsymbol{u}_0}\chi^{m}_{;j}\chi^{k}_{;i} + (\bar{\varsigma}_2)_{;k}\Big|_{\boldsymbol{u}=\boldsymbol{u}_0}\chi^{k}_{;i;j}\Big)(h_1)^i(h_1)^j \\
&= \sum_{k=1}^{n}\sum_{m=1}^{n}(\bar{\varsigma}_2)_{;k;m}\Big|_{\boldsymbol{u}=\boldsymbol{u}_0}(h_2)^m(h_2)^k + \sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(\bar{\varsigma}_2)_{;k}\Big|_{\boldsymbol{u}=\boldsymbol{u}_0}\chi^{k}_{;i;j}(h_1)^i(h_1)^j
\end{aligned}
\tag{B.27}
$$

Only if the second of these two terms is zero, for example if $\chi^{k}_{;i;j} = 0, \forall k, i, j$, is the value of the second order term in the Taylor expansion invariant to the coordinate system. The second term is not generally zero and invariance to coordinate systems does not hold in general.

Hence in general if $\varrho > 1$, the value of a ϱth order Taylor expansion approximation varies with the coordinate system. However if the two scalar fields $\bar{\varsigma}_1$ and $\bar{\varsigma}_2$ are analytic about $\boldsymbol{\theta}_0$ and $\boldsymbol{u}_0$ respectively, and both $\boldsymbol{\theta}'$ and $\boldsymbol{u}'$ are within the relevant convergence domains, then the Taylor expansions coincide as $\varrho \to \infty$. The Taylor expansion is not therefore a property of the manifold but also of the coordinate system. This has important consequences for fibre bundles with fibres defined on Taylor expansions.

B.3 Covariant derivatives for selected scalar fields and statistical models

This section presents expressions for the covariant derivatives of log likelihood scalar fields over the statistical manifolds relevant to the experiments in this thesis, namely HMMs with state-conditional likelihoods modelled by GMMs (GMMs are viewed as single state HMMs).

There are Q class-conditional models $S(\boldsymbol{\theta}_q), q = \{1, \ldots Q\}$. For brevity covariant derivatives are presented in linear algebraic rather than component form. The notation for HMMs is as given in Section 2.2.1. With a slight abuse of notation let $\nabla_{\boldsymbol{\mu}_{qjk}}$ denote

the operator of covariant differentiation with respect to $\sum_{i=1}^{d} \partial/\partial(\mu_{qjk})^i$. Using a similar abbreviation for other operators of covariant differentiation, then for $\bar{\varsigma}(q)_l = \ln p(\boldsymbol{O}_l; \boldsymbol{\theta}_q)$,

$$\nabla_{\boldsymbol{\theta}_q} \bar{\varsigma}(q)_l = \begin{bmatrix} \nabla_{w_{q12}} \bar{\varsigma}(q)_l \\ \vdots \\ \nabla_{w_{qNK}} \bar{\varsigma}(q)_l \\ \nabla_{\boldsymbol{\mu}_{q11}} \bar{\varsigma}(q)_l \\ \vdots \\ \nabla_{\boldsymbol{\mu}_{qNK}} \bar{\varsigma}(q)_l \\ \nabla_{\mathrm{vec}(\boldsymbol{\Sigma}_{q11})} \bar{\varsigma}(q)_l \\ \vdots \\ \nabla_{\mathrm{vec}(\boldsymbol{\Sigma}_{qNK})} \bar{\varsigma}(q)_l \\ \nabla_{a_q(1,1)} \bar{\varsigma}(q)_l \\ \vdots \\ \nabla_{a_q(N,N)} \bar{\varsigma}(q)_l \end{bmatrix} \tag{B.28}$$

where $\nabla_{\boldsymbol{\theta}_q} \bar{\varsigma}(q)_l$ is implicitly evaluated at a realisation of $\boldsymbol{\theta}_q$. Then from Appendix A of [92], where $\forall j, k$ unless otherwise stated[1],

$$\nabla_{w_{qjk}} \bar{\varsigma}(q)_l = \sum_{t=1}^{T} \left\{ \frac{\gamma_{qjk}(t)}{w_{qjk}} - \frac{\gamma_{qj1}(t)}{w_{qj1}} \right\} \quad k = \{2, \ldots K\} \tag{B.29}$$

$$\nabla_{\boldsymbol{\mu}_{qjk}} \bar{\varsigma}(q)_l = \sum_{t=1}^{T} \gamma_{qjk}(t) \boldsymbol{\Sigma}_{qjk}^{-1} (\boldsymbol{o}_l(t) - \boldsymbol{\mu}_{qjk}) \tag{B.30}$$

$$\begin{aligned} \nabla_{\mathrm{vec}(\boldsymbol{\Sigma}_{qjk})} \bar{\varsigma}(q)_l = \frac{1}{2} \sum_{t=1}^{T} \gamma_{qjk}(t) \Big\{ & -\Big(\mathrm{vec}(\boldsymbol{\Sigma}_{qjk}^{-1})\Big)^{\top} \\ & + \Big((\boldsymbol{o}_l(t) - \boldsymbol{\mu}_{qjk})^{\top} \boldsymbol{\Sigma}_{qjk}^{-1}\Big) \otimes \Big((\boldsymbol{o}_l(t) - \boldsymbol{\mu}_{qjk})^{\top} \boldsymbol{\Sigma}_{qjk}^{-1}\Big) \Big\}^{\top} \end{aligned} \tag{B.31}$$

$$\nabla_{a_q(j,j)} \bar{\varsigma}(q)_l = \sum_{t=1}^{T} \left\{ \frac{\gamma_{qj}(t)}{a_q(j,j)} - \frac{1}{T a_q(j,j)(1 - a_q(j,j))} \right\} \quad j = \{1, \ldots N\} \tag{B.32}$$

The notation 'vec' refers to the vec operator, $\otimes$ to the Kronecker product of matrices, and $\gamma_{qjk}(t)$ is the component posterior at time t given the sample $\boldsymbol{O}_l$. These derivations are applicable for each model $S(\boldsymbol{\theta}_q)$, where each HMM is left-to-right with no skips and there is no tying of parameters within or across the Q models (see Appendix A of [92]). The

[1] The following derivations assume an Identity metric tensor in parameter space, so contravariant and covariant components coincide and there is no need to distinguish them or introduce the bar notation.

order of the components of $\nabla_{\boldsymbol{\theta}_q}\bar{\varsigma}(q)_l$ can be freely rearranged. Once $\nabla_{\boldsymbol{\theta}_q}\bar{\varsigma}(q)_l$ is defined then it is straightforward to define covariant derivatives for related scalar fields. Letting,

$$\nabla_{\boldsymbol{\xi}}\bar{f}_l = \begin{bmatrix} \nabla_{\boldsymbol{\theta}_1}\bar{f}_l \\ \vdots \\ \nabla_{\boldsymbol{\theta}_Q}\bar{f}_l \end{bmatrix} \tag{B.33}$$

then,

- the log likelihood scalar field $\bar{f}_l = \ln p(\boldsymbol{O}_l;\boldsymbol{\theta}_a)$: Equation B.33 applies where,

$$\nabla_{\boldsymbol{\theta}_q}\bar{f}_l = \begin{cases} \nabla_{\boldsymbol{\theta}_a}\bar{\varsigma}(a)_l & \text{if} \quad q = a \\ \mathbf{0} & \text{if} \quad q \neq a \end{cases} \tag{B.34}$$

- the log likelihood ratio scalar field $\bar{f}_l = \ln p(\boldsymbol{O}_l;\boldsymbol{\theta}_a) - \ln p(\boldsymbol{O}_l;\boldsymbol{\theta}_b)$: Equation B.33 applies where,

$$\nabla_{\boldsymbol{\theta}_q}\bar{f}_l = \begin{cases} \nabla_{\boldsymbol{\theta}_a}\bar{\varsigma}(a)_l & \text{if} \quad q = a \\ -\nabla_{\boldsymbol{\theta}_b}\bar{\varsigma}(b)_l & \text{if} \quad q = b \\ \mathbf{0} & \text{otherwise} \end{cases} \tag{B.35}$$

- the log class posterior scalar field $\bar{f}_l = \ln P(\omega_a|\boldsymbol{O}_l)$: Equation B.33 applies where according to Appendix B of [95],

$$\nabla_{\boldsymbol{\theta}_q}\bar{f}_l = \alpha(q)\nabla_{\boldsymbol{\theta}_q}\bar{\varsigma}(q)_l \tag{B.36}$$

where,

$$\alpha(q) = \begin{cases} 1 - P(\omega_a|\boldsymbol{O}_l) & \text{if} \quad q = a \\ -P(\omega_q|\boldsymbol{O}_l) & \text{if} \quad q \neq a \end{cases} \tag{B.37}$$

B.4 Variations on the appended posterior score space

The appended posterior score space $\varphi^{\text{ps(all)}}(1,\boldsymbol{\xi}_0)$ from Section 4.1 has many repeated components in the unit degree covariant derivatives. The repeated components may be discarded without loss in information yielding the *reduced appended posterior score space*

$\varphi^{\mathrm{psr(all)}}(1, \boldsymbol{\xi}_0)$. For $\boldsymbol{O}_l \in L(\boldsymbol{O})$, adopting the linear algebraic expressions for covariant derivatives in Section B.3, and implicitly assuming class posteriors are evaluated at $\boldsymbol{\xi}_0$, the score is,

$$\bar{\boldsymbol{\varphi}}^{\mathrm{psr(all)}}(\boldsymbol{O}_l; 1, \boldsymbol{\xi}_0) = \begin{bmatrix} \ln P(\omega_1|\boldsymbol{O}_l) \\ \vdots \\ \ln P(\omega_Q|\boldsymbol{O}_l) \\ (1 - P(\omega_1|\boldsymbol{O}_l))\nabla_{\boldsymbol{\theta}_1} \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_1)_0) \\ \vdots \\ (1 - P(\omega_Q|\boldsymbol{O}_l))\nabla_{\boldsymbol{\theta}_Q} \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_Q)_0) \\ -P(\omega_1|\boldsymbol{O}_l)\nabla_{\boldsymbol{\theta}_1} \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_1)_0) \\ \vdots \\ -P(\omega_Q|\boldsymbol{O}_l)\nabla_{\boldsymbol{\theta}_Q} \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_Q)_0) \end{bmatrix} \tag{B.38}$$

A more general form of this appended linear space is the *generalised appended posterior score space* $\varphi^{\mathrm{psg(all)}}(1, \boldsymbol{\xi}_0)$, where for $\boldsymbol{O}_l \in L(\boldsymbol{O})$,

$$\bar{\boldsymbol{\varphi}}^{\mathrm{psg(all)}}(\boldsymbol{O}_l; 1, \boldsymbol{\xi}_0) = \begin{bmatrix} \ln P(\omega_1|\boldsymbol{O}_l) \\ \vdots \\ \ln P(\omega_Q|\boldsymbol{O}_l) \\ -P(\omega_1|\boldsymbol{O}_l)\nabla_{\boldsymbol{\theta}_1} \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_1)_0) \\ \vdots \\ -P(\omega_Q|\boldsymbol{O}_l)\nabla_{\boldsymbol{\theta}_Q} \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_Q)_0) \\ \nabla_{\boldsymbol{\theta}_1} \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_1)_0) \\ \vdots \\ \nabla_{\boldsymbol{\theta}_Q} \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_Q)_0) \end{bmatrix} \tag{B.39}$$

Finally, a *hybrid appended posterior score space* $\varphi^{\mathrm{psh(all)}}(1, \boldsymbol{\xi}_0)$ is also proposed which combines linear posteriors with the unit degree covariant derivatives of the appended likelihood

score space. For $\boldsymbol{O}_l \in L(\boldsymbol{O})$,

$$\bar{\boldsymbol{\varphi}}^{\mathrm{psh(all)}}(\boldsymbol{O}_l; 1, \boldsymbol{\xi}_0) = \begin{bmatrix} P(\omega_1|\boldsymbol{O}_l) \\ \vdots \\ P(\omega_Q|\boldsymbol{O}_l) \\ \nabla_{\boldsymbol{\theta}_1} \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_1)_0) \\ \vdots \\ \nabla_{\boldsymbol{\theta}_Q} \ln p(\boldsymbol{O}_l; (\boldsymbol{\theta}_Q)_0) \end{bmatrix} \tag{B.40}$$

Appendix C

Linear spaces

C.1 A summary of linear spaces

This section gives some brief details concerning linear spaces. A *linear space* or *vector space* L is briefly a space whose members may be added or multiplied by scalars to yield members of the same space (see Section 256.A of [50]). It becomes a *metric space* when endowed with a distance function, also called a metric g (see Section 273 of [50]), and the distance between $X_1, X_2 \in L$ is denoted $g(X_1, X_2)$. The metric space (L, g) becomes more useful if a norm $||X||$ for $X \in L$ is defined. This is then a *normed linear space* and $g(X_1, X_2) = ||X_1 - X_2||$. If the metric space is complete, the space becomes a *Banach space* (see Section 39 of [50]). Scalar products do not exist in all linear spaces but strictly only in *pre-Hilbert spaces* (see Section 199.B of [50]). If the pre-Hilbert space inherits the properties of a Banach space, then a norm exists and the space is complete with respect to the distance $||X_1 - X_2||$. Then the linear space is a *Hilbert space* (see Section 199 of [50]) and,

$$||X|| \quad = \quad \sqrt{(X, X)} \tag{C.1}$$

An *affine space* may then be defined with the Hilbert space as the *standard vector space* and with a basis and origin specified (see Section 9 of [50]). The metric can then be numerically evaluated as a *metric tensor* also in this thesis denoted by g.

C.2 Linear algebraic representation of tensor spaces

Tensor spaces effectively enable bilinear mappings to be viewed as linear mappings (see Section 256 of [50]). A *tensor* is a member of tensor space. This thesis adopts a linear algebraic representation of tensors in vector/matrix form. For example the objects X and a,

$$X = \sum_{i=1}^{n} X^i e_i = \sum_{i=1}^{n} X_i e^i \tag{C.2}$$

$$a = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} e^i \otimes e^j = \sum_{i=1}^{n}\sum_{j=1}^{n} a^{ij} e_i \otimes e_j = \sum_{i=1}^{n}\sum_{j=1}^{n} a^i_j e_i \otimes e^j \tag{C.3}$$

where e^i and e_i are respectively the ith elements of the contravariant and covariant basis for the linear space and $\otimes$ is the tensor product. Then tensors of type $(1,0)$ and $(2,0)$ are respectively cast into column and matrix format in bold font where,

$$\boldsymbol{X} = \begin{bmatrix} X^1 \\ X^2 \\ \vdots \\ X^n \end{bmatrix} \tag{C.4}$$

$$\boldsymbol{A} = \begin{bmatrix} a^{11} & a^{12} & \dots & a^{1n} \\ a^{21} & a^{22} & \dots & a^{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a^{n1} & a^{n2} & \dots & a^{nn} \end{bmatrix} \tag{C.5}$$

Next, tensors of type $(0,1)$ and $(0,2)$ are written in bold font with a bar,

$$\bar{\boldsymbol{X}} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \tag{C.6}$$

$$\bar{\boldsymbol{A}} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \tag{C.7}$$

Finally a mixed tensor of type $(1,1)$ is written in bold font with an underbar,

$$\underline{\boldsymbol{A}} = \begin{bmatrix} a_1^1 & a_2^1 & \dots & a_n^1 \\ a_1^2 & a_2^2 & \dots & a_n^2 \\ \vdots & \vdots & \vdots & \vdots \\ a_1^n & a_2^n & \dots & a_n^n \end{bmatrix} \tag{C.8}$$

As a special case, the metric tensor g yields the matrices $\boldsymbol{G}$ and $\bar{\boldsymbol{G}}$ where $\boldsymbol{G} = \bar{\boldsymbol{G}}^{-1}$. In this thesis these are called *metric matrices* and inherit the properties of the metric tensor g and hence are symmetric and positive definite. The scalar product between the objects X_1 and X_2 may be written as,

$$(\boldsymbol{X}_1, \boldsymbol{X}_2) = \boldsymbol{X}_1^\top \bar{\boldsymbol{G}} \boldsymbol{X}_2 \tag{C.9}$$

$$= \boldsymbol{X}_1^\top \bar{\boldsymbol{X}}_2 = < \boldsymbol{X}_1, \bar{\boldsymbol{X}}_2 > \tag{C.10}$$

$$= \bar{\boldsymbol{X}}_1^\top \boldsymbol{X}_2 = < \bar{\boldsymbol{X}}_1, \boldsymbol{X}_2 > \tag{C.11}$$

$$= \bar{\boldsymbol{X}}_1^\top \boldsymbol{G} \bar{\boldsymbol{X}}_2 = (\bar{\boldsymbol{X}}_1, \bar{\boldsymbol{X}}_2) \tag{C.12}$$

The scalar product may be expressed as a function $(\cdot, \cdot)$ where its two arguments are members of the same tensor space, or as a function $< \cdot, \cdot >$ where the two arguments are members of dual tensor spaces.

If the linear space is a tensor space of rank greater than 2, then it is still possible to denote its members in column format, but an isomorphism to a unit rank tensor space is first required. Given a linear space of dimension n and a tensor space defined on this of type $(r, 0)$, then each tensor in this space has n^r distinct components. However if a tensor is additionally constrained to be fully symmetric, then the number of distinct components is reduced to,

$${}^{n+r-1}C_r = \frac{(r+n-1)!}{r!(n-1)!} \tag{C.13}$$

which is the number of combinations possible when making r selections from n objects with replacement (see Section 24.3 of [81]).

Appendix D

Metric tensors

As detailed in Appendix C.1, given an affine space defined on a Hilbert space, the metric or distance function is completely defined by a *metric tensor*. For example, L is a tensor space of type $(1,0)$ with n components, with a metric tensor g of type $(0,2)$. For $X_1, X_2 \in L$, their scalar product or bilinear form is,

$$\begin{aligned}(X_1, X_2) &= \sum_{i=1}^{n} (X_1)^i (X_2)_i \\ &= \sum_{i=1}^{n} (X_1)^i \sum_{j=1}^{n} g_{ij} (X_2)^j \end{aligned} \tag{D.1}$$

Referring to Appendix C.1, the distance between X_1 and X_2 is $g(X_1, X_2)$ where,

$$\begin{aligned} g(X_1, X_2) &= \sqrt{(X_1 - X_2, X_1 - X_2)} \\ &= \Big\{ \sum_{i=1}^{n} \sum_{j=1}^{n} (X_1 - X_2)^i g_{ij} (X_1 - X_2)^j \Big\}^{\frac{1}{2}} \end{aligned} \tag{D.2}$$

D.1 Properties of metric tensors

By definition, the metric tensor is invariant to changes in coordinate system. If the Hilbert space L can be expressed in terms of two coordinate systems $[\theta^i]$ and $[\theta'^i]$, then the covariant components of the metric tensor g in $[\theta^i]$ transform to those of g' in $[\theta'^i]$ according to (see

for example [81]),

$$g_{ij} = \sum_{k=1}^{n}\sum_{l=1}^{n}\frac{\partial\theta'^{k}}{\partial\theta^{i}}\frac{\partial\theta'^{l}}{\partial\theta^{j}}g'_{kl} \tag{D.3}$$

In addition, desirable properties for the metric tensor include,

- its functional form is invariant to a change in the coordinate system,
- it is maximally noncommittal in some sense,
- it is invariant to sufficient statistics.

The third of these properties is only here relevant for metric tensors defined in tangent spaces of statistical manifolds. This application is the focus of the discussion.

D.1.1 Invariance of the functional form to the coordinate system

In certain applications, it is desirable that the functional form of the metric tensor is also invariant to a change in the coordinate system. In this case, g_{ij} and g'_{kl} in Equation D.3 must share the same functional form. An example is the Fisher metric tensor (see Equation D.16) applied to the tangent space of a statistical manifold. On the contrary, the metric tensor implied in the Natural kernel [73] does not necessarily fulfill this constraint in which case the functional form is tied to a particular coordinate system, else it is not a tensor.

D.1.2 Maximally noncommittal

The metric tensor specifies the contribution of each component of the unit rank tensor space to the norm of a member of that space. It is fully specified, for example, once the relation between the tensor space and its embedding space is known. However, if expert knowledge is unavailable specifying the relative importance of each component and the correlations between them, it is sensible to choose a metric tensor which makes the least

assumptions concerning these factors. This embodies the principles of Occam's Razor in which the simplest model is selected fulfilling the necessary constraints but making the least claims about unknown knowledge.

For example, consider a tensor space L of type $(1,0)$ populated with tensors, and with a tensor metric g of type $(0,2)$. Two tensors exist which are the global second order moment with contravariant component, where $X \in L$,

$$c^{ij} = \int\int X^i X^j p(X^i, X^j) \mathrm{d}X^i \mathrm{d}X^j \tag{D.4}$$

and the global second order central moment or covariance with contravariant component,

$$v^{ij} = \int\int (X^i - \mu^i)(X^j - \mu^j) p(X^i, X^j) \mathrm{d}X^i \mathrm{d}X^j \tag{D.5}$$

where,

$$\mu^i = \int X^i p(X^i) \mathrm{d}X^i \tag{D.6}$$

and $p(\cdot)$ and $p(\cdot,\cdot)$ are distributions. The fully covariant forms are similarly defined on X_i and X_j. Both c and v may be used as metric tensors and are both maximally noncommittal in some sense. The contribution to the expected square of the norm $E\{||X||^2\}$ from the ordered pair (X^i, X^j) under the metric g, abbreviated to $C(g, X^i, X^j)$, is,

$$\begin{aligned} C(g, X^i, X^j) &= \int\int X^i g_{ij} X^j p(X^i, X^j) \mathrm{d}X^i \mathrm{d}X^j \\ &= g_{ij} \int\int X^i X^j p(X^i, X^j) \mathrm{d}X^i \mathrm{d}X^j \\ &= g_{ij} c^{ij} \end{aligned} \tag{D.7}$$

The expected contribution from X^i is then, noting symmetry,

$$\begin{aligned} C(c, X^i, \cdot) &= 2 \sum_{j=1}^{n} C(g, X^i, X^j) - C(g, X^i, X^i) \\ &= 2 \sum_{j=1}^{n} g_{ij} c^{ij} - g_{ii} c^{ii} \end{aligned} \tag{D.8}$$

It is sensible to equate the contributions from each component, for example arbitrarily to unity. Then since by symmetry $c^{ij} = c^{ji}$,

$$2 \sum_{j=1}^{n} g_{ij} c^{ji} - g_{ii} c^{ii} = 1 \tag{D.9}$$

For ease of analysis, assuming c is diagonal,

$$g_{ii}c^{ii} = 1 \tag{D.10}$$

A solution is when $g = c$. Hence the second order moment c is maximally noncommittal with respect to the contribution of each component to $E\{||X||^2\}$ but under the assumption of a diagonal moment c. An alternative metric tensor may be chosen so that the contribution from each pair X^i and X^j to $E\{||X-\mu||^2\}$, denoted by $C(g, \mu, X^i, X^j)$ is,

$$\begin{aligned} C(g, \mu, X^i, X^j) &= \int\int (X^i - \mu^i) g_{ij} (X^j - \mu^j) p(X^i, X^j) \mathrm{d}X^i \mathrm{d}X^j \\ &= g_{ij} \int\int (X^i - \mu^i)(X^j - \mu^j) p(X^i, X^j) \mathrm{d}X^i \mathrm{d}X^j \\ &= g_{ij} v^{ij} \end{aligned} \tag{D.11}$$

Following a similar analysis to that above and assuming v is diagonal, then setting the metric to v yields a metric which is maximally noncommittal in the same sense as c except with respect to $E\{||X-\mu||^2\}$ rather than $E\{||X||^2\}$.

There are good reasons for preferring as metric tensor the second order central moment or covariance v rather than the second order moment c.

- Both the metric tensors c and v are maximally noncommittal but respectively assume c and v are diagonal. Diagonality in v assumes components are decorrelated. A sufficient condition for diagonality in c is the assumption that components are decorrelated and have zero mean. The metric tensor v is preferred since it makes simpler assumptions about the distribution of data. It is also useful to note that statistical independence of components, not linear independence, implies decorrelation.

- The second order moment c is biased, in terms of sensitivity, towards those components with large nonzero mean. For example, for $X_1, X_2 \in L$,

$$(X_1, X_2) = \sum_{i=1}^{n} \sum_{j=1}^{n} (X_1)^i g_{ij} (X_2)^j \tag{D.12}$$

  If X_1 and X_2 are varied by small perturbations ΔX_1 and ΔX_2, the change in the scalar product is,

$$\Delta(X_1, X_2) = (X_1 + \Delta X_1, X_2 + \Delta X_2) - (X_1, X_2)$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\sum_{j=1}^{n}(X_1+\Delta X_1)^i g_{ij}(X_2+\Delta X_2)^j - \sum_{i=1}^{n}\sum_{j=1}^{n}(X_1)^i g_{ij}(X_2)^j \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}(\Delta X_1)^i g_{ij}(\Delta X_2)^j + (X_1)^i g_{ij}(\Delta X_2)^j + (\Delta X_1)^i g_{ij}(X_2)^j
\end{aligned}
\tag{D.13}
$$

The contribution to $\Delta(X_1, X_2)$ from the components $(X_1)^i$ and $(X_2)^j$ is scaled by g_{ij}. The smaller is the magnitude of g_{ij}, the less sensitive is the scalar product to small changes in either $(X_1)^i$ or $(X_2)^j$. Setting g to c implies that if μ_i and μ_j, the means of the ith and jth covariant components in L respectively, are nonzero and large, then g_{ij} is large and unfairly increases sensitivity of the scalar product to variations in $(X_1)^i$ and $(X_2)^j$, or X_2^i and X_1^j. Setting g to v counteracts this bias to components with large nonzero mean. The scalar product is then sensitive to small changes in each component of X_1 and X_2 in proportion to the component covariances which is more sensible.

The two definitions for the metric tensor yield different values for scalar products unless the mean in linear space $\boldsymbol{\mu} = \mathbf{0}$. Expressing the scalar product in linear algebraic form, where $\boldsymbol{\Sigma}_{\text{glob}}$ is the global covariance matrix and where $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{\Sigma}_{\text{glob}}$ are formed from fully contravariant components, yields the Mahanalobis distance [25],

$$
(\boldsymbol{X}_1, \boldsymbol{X}_2) \;=\; \boldsymbol{X}_1^\top \boldsymbol{\Sigma}_{\text{glob}}^{-1} \boldsymbol{X}_2 \tag{D.14}
$$

These concepts can be extended to multiple classes of data. Within-class second order moments or central moments are calculated for each class. A metric tensor is defined as a weighted form of these class-conditional moments, where the weights are either in accordance with expert knowledge or data-dependent priors. The resulting metrics can be interpreted as maximally noncommittal in the same sense as described above but averaged over all classes. In linear algebraic notation, and letting $\boldsymbol{\Sigma}_{\text{wtd}}$ denote the average within-class covariance matrix formed from fully contravariant components,

$$
(\boldsymbol{X}_1, \boldsymbol{X}_2) \;=\; \boldsymbol{X}_1^\top \boldsymbol{\Sigma}_{\text{wtd}}^{-1} \boldsymbol{X}_2 \tag{D.15}
$$

D.1.3 Invariance to sufficient statistics

This property of the metric tensor is only here applicable for tangent spaces to statistical manifolds. Variance to sufficient statistics ties a metric tensor to a particular set of sufficient statistics. Invariance to sufficient statistics subsumes invariance to bijective front-end processing schemes, since bijective mappings yield sufficient statistics (see Section 2.2 of [3]).

An example of particular importance to this thesis is the Fisher metric. According to [3] restating a result by Chentsov [13], the Fisher metric, up to an arbitrary scaling factor, is the only metric for tangent space which is invariant to sufficient statistics. The Fisher metric tensor is [3],

$$g_{ij}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \int \frac{\partial}{\partial \theta^i} \ln p(\boldsymbol{O};\boldsymbol{\theta}) \frac{\partial}{\partial \theta^j} \ln p(\boldsymbol{O};\boldsymbol{\theta}) p(\boldsymbol{O};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{O} \tag{D.16}$$

The Fisher metric is both a second order moment and second order central moment since the first order moment is zero. From the factorisation theorem [25] [3],

$$\ln p(\boldsymbol{O};\boldsymbol{\theta}) = \ln q(\mathcal{F}(\boldsymbol{O});\boldsymbol{\theta}) + \ln h(\boldsymbol{O}) \tag{D.17}$$

where a factorisation is selected where $q(\mathcal{F}(\boldsymbol{O});\boldsymbol{\theta})$ is a distribution over the space of sufficient statistics $L(\mathcal{F}(\boldsymbol{O}))$. Then, temporarily denoting the Fisher metric tensor defined on the distribution in $L(\boldsymbol{O})$ as $g_{ij}(\boldsymbol{\theta};\boldsymbol{O})$,

$$\begin{aligned}
g_{ij}(\boldsymbol{\theta};\boldsymbol{O}) &= \int \frac{\partial}{\partial \theta^i} \ln q(\mathcal{F}(\boldsymbol{O});\boldsymbol{\theta}) \frac{\partial}{\partial \theta^j} \ln q(\mathcal{F}(\boldsymbol{O});\boldsymbol{\theta}) p(\boldsymbol{O};\boldsymbol{\theta}) \mathrm{d}\boldsymbol{O} \\
&= E_{p(\boldsymbol{O};\boldsymbol{\theta})}\Big\{ \frac{\partial}{\partial \theta^i} \ln q(\mathcal{F}(\boldsymbol{O});\boldsymbol{\theta}) \frac{\partial}{\partial \theta^j} \ln q(\mathcal{F}(\boldsymbol{O});\boldsymbol{\theta}) \Big\} \\
&= E_{q(\mathcal{F}(\boldsymbol{O});\boldsymbol{\theta})}\Big\{ \frac{\partial}{\partial \theta^i} \ln q(\mathcal{F}(\boldsymbol{O});\boldsymbol{\theta}) \frac{\partial}{\partial \theta^j} \ln q(\mathcal{F}(\boldsymbol{O});\boldsymbol{\theta}) \Big\} \\
&= \int \frac{\partial}{\partial \theta^i} \ln q(\mathcal{F}(\boldsymbol{O});\boldsymbol{\theta}) \frac{\partial}{\partial \theta^j} \ln q(\mathcal{F}(\boldsymbol{O});\boldsymbol{\theta}) q(\mathcal{F}(\boldsymbol{O});\boldsymbol{\theta}) \mathrm{d}\mathcal{F}(\boldsymbol{O}) \\
&= g_{ij}(\boldsymbol{\theta};\mathcal{F}(\boldsymbol{O}))
\end{aligned} \tag{D.18}$$

The tensor is numerically invariant to definition on a distribution in $L(\boldsymbol{O})$ or a distribution in $L(\mathcal{F}(\boldsymbol{O}))$. The equivalence between expectations over $p(\boldsymbol{O};\boldsymbol{\theta})$ and $q(\mathcal{F}(\boldsymbol{O});\boldsymbol{\theta})$ follows at least for surjective mappings between $L(\boldsymbol{O})$ and $L(\mathcal{F}(\boldsymbol{O}))$ which are bijective, one-to-many

or many-to-one. For bijective mappings, it is useful to note that,

$$p(\boldsymbol{O};\boldsymbol{\theta}) \;=\; q(\mathcal{F}(\boldsymbol{O});\boldsymbol{\theta})\frac{\partial \mathcal{F}(\boldsymbol{O})}{\partial \boldsymbol{O}} \tag{D.19}$$

D.2 Metric tensors for tangent spaces

The Fisher metric is often applied to the tangent space of a statistical manifold since it is invariant to expressing distributions in terms of different sufficient statistics. However if the first order moment in tangent space is nonzero, for example if $p(\boldsymbol{O}) \neq p(\boldsymbol{O};\boldsymbol{\theta})$ then the metric tensor is a second order moment. As explained above, the covariance in tangent space is more desirable and is applied as the metric for tangent spaces in this thesis[1]. Then,

$$g_{ij}(\boldsymbol{\theta}) \;=\; \int \Big(\frac{\partial}{\partial \theta^i}\ln p(\boldsymbol{O};\boldsymbol{\theta}) - \mu_i\Big)\Big(\frac{\partial}{\partial \theta^j}\ln p(\boldsymbol{O};\boldsymbol{\theta}) - \mu_j\Big)p(\boldsymbol{O})\mathrm{d}\boldsymbol{O} \tag{D.20}$$

where,

$$\mu_i \;=\; \int \frac{\partial}{\partial \theta^i}\ln p(\boldsymbol{O};\boldsymbol{\theta})p(\boldsymbol{O})\mathrm{d}\boldsymbol{O} \tag{D.21}$$

Unfortunately, the metric is only invariant to sufficient statistics for various distributions and sufficient statistics and no longer in the general case. However the advantages of the covariance metric tensor are here viewed as outweighing this restriction. In a similar manner, it is more sensible to apply a covariance rather than a second order moment in the calculation of the Mahanalobis distance in Equation D.14.

[1] The covariance reduces to the Fisher metric if $p(\boldsymbol{O}) = p(\boldsymbol{O};\boldsymbol{\theta})$.

Appendix E

Metrics induced on the input manifold

Following the analysis in [1] as referenced in Section 4.2, this appendix assumes the input space $L(\boldsymbol{O})$ is a Riemannian manifold defined by a metric tensor in its tangent space. The appendix details the metric tensors induced by different score mappings. As detailed in [1], these metric tensors may be used to calculate magnification factors from input space to score space in the vicinity of decision boundaries. The score mappings detailed here are those given by some simple statistical models.

As a primary analysis, a 1-component input space $L(O)$ is selected and populated by a set of Q classes, each distributed according to a Gaussian. The class-conditional Gaussian $\mathcal{N}(O; \mu_q, v_q)$ for class ω_q has mean μ_q and variance v_q and $\boldsymbol{\theta}_q = (\mu_q, v_q)^\top$. Both μ_q and v_q are implicitly contravariant components of the vector $\boldsymbol{\theta}_q$. The parameters for the Q statistical models are summarised by $\boldsymbol{\xi} = (\boldsymbol{\theta}_1^\top, \ldots \boldsymbol{\theta}_Q^\top)^\top$. A single statistical model is denoted by $S(\boldsymbol{\theta}_q)$ and the set of models by $\mathcal{S}(\boldsymbol{\xi})$. The score spaces contain zeroth and unit degree covariant derivatives as detailed in Section 4.1. Covariant derivatives are defined on the Gaussian means only. For a statistical model $S(\boldsymbol{\theta}_q)$, the metric matrix $\bar{\boldsymbol{A}}$ for the

dual of score space has the restricted form defined by,

$$\bar{\boldsymbol{A}} \;=\; \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \bar{\boldsymbol{G}} \end{bmatrix} \tag{E.1}$$

where each vector of zeros has the appropriate shape and where $\bar{\boldsymbol{G}}$ is the metric matrix for the dual of tangent space to the statistical manifold. The components of $\bar{\boldsymbol{G}}$ are the fully covariant components $g_{ij}^{\text{sup}}(\boldsymbol{\theta}_q)$, where the superscript 'sup' identifies the score mapping. Also $[g_{ij}^{\text{sup}}(\boldsymbol{\theta}_q)]^{-1} = (g^{ij})^{\text{sup}}(\boldsymbol{\theta}_q)$, i.e. the matrices formed by the fully covariant and fully contravariant forms of the metric tensor are related by the numerical inverse. A (1×1) metric tensor is regarded as a fully covariant component so $[g_{11}^{\text{sup}}(\boldsymbol{\theta}_q)]^{-1} = 1/g_{11}^{\text{sup}}(\boldsymbol{\theta}_q) = (g^{11})^{\text{sup}}(\boldsymbol{\theta}_q)$. Then $g_{11}^{\text{sup}}(\boldsymbol{\theta}_q) = (g^{11})^{\text{sup}}(\boldsymbol{\theta}_q) = 1$ only if the natural basis vector $\partial/\partial(\mu^1)_q$ is of unit length viewed from its embedding space. If a metric tensor is referenced without indexing its components, for example $g^{\text{sup}}(\boldsymbol{\theta}_q)$, then this refers to the whole metric tensor. The metric tensors induced on the input manifold by different score spaces are defined as follows.

- *Likelihood score space*, $\varphi^{\text{lk(q)}}(1, (\boldsymbol{\theta}_q)_0)$: This is a 2-component score space with a (1×1) metric tensor in the tangent space to the statistical manifold, which at point $(\boldsymbol{\theta}_q)_0$ is $g_{11}^{\text{lk(q)}}((\boldsymbol{\theta}_q)_0)$. The metric tensor induced on the input manifold has a single fully covariant component,

$$g_{11}^{\text{lk(q)}}(O; (\boldsymbol{\theta}_q)_0) \;=\; \frac{1}{v_q^2}\Big((O-\mu_q)^2 + [g_{11}^{\text{lk(q)}}((\boldsymbol{\theta}_q)_0)]^{-1}\Big) \tag{E.2}$$

- *Likelihood-ratio score space*, $\varphi^{\text{lr(a,b)}}(1, \boldsymbol{\xi}_0)$: This is a 3-component score space since $\mathcal{S}(\boldsymbol{\xi})$ is defined with two models so $Q = 2$. The metric tensor in the tangent space to $\mathcal{S}(\boldsymbol{\xi})$ at point $\boldsymbol{\xi}_0$ is of size (2×2) with components $g_{ij}^{\text{lr(a,b)}}(\boldsymbol{\xi}_0)$. Since the two class-conditional models are statistically independent, the off-main diagonal elements are assumed zero. So,

$$g_{ij}^{\text{lr(a,b)}}(\boldsymbol{\xi}_0) \;=\; \begin{cases} g_{11}^{\text{lk(a)}}((\boldsymbol{\theta}_a)_0) & \text{if} \quad i = j = 1 \\ g_{11}^{\text{lk(b)}}((\boldsymbol{\theta}_b)_0) & \text{if} \quad i = j = 2 \\ \quad 0 & \text{if} \quad i \neq j \end{cases} \tag{E.3}$$

Then, the metric induced on the input manifold is described by the single fully covariant component,

$$g_{11}^{\mathrm{lr(a,b)}}(O;\boldsymbol{\xi}_0) \;=\; g_{11}^{\mathrm{lk(a)}}(O;(\boldsymbol{\theta}_a)_0) + g_{11}^{\mathrm{lk(b)}}(O;(\boldsymbol{\theta}_b)_0) - \frac{2}{v_a v_b}(O-\mu_a)(O-\mu_b) \tag{E.4}$$

- *Appended likelihood score space*, $\varphi^{\mathrm{lk(all)}}(1,\boldsymbol{\xi}_0)$: For a Q-class problem, this is a $2Q$-component score space. There is a $Q \times Q$ metric tensor $g^{\mathrm{lk(all)}}(\boldsymbol{\xi}_0)$ at point $\boldsymbol{\xi}_0$ in the tangent space to the manifold $\mathcal{S}(\boldsymbol{\xi})$. Assuming statistical independence between each class-conditional model then it is reasonable that,

$$g_{ij}^{\mathrm{lk(all)}}(\boldsymbol{\xi}_0) \;=\; \begin{cases} g_{11}^{\mathrm{lk(i)}}((\boldsymbol{\theta}_i)_0) & \text{if} \quad i=j \\ 0 & \text{if} \quad i \neq j \end{cases} \tag{E.5}$$

  Then the metric tensor induced on the input manifold has component,

$$g_{11}^{\mathrm{lk(all)}}(O;\boldsymbol{\xi}_0) \;=\; \sum_{q=1}^{Q} g_{11}^{\mathrm{lk(q)}}(O;(\boldsymbol{\theta}_q)_0) \tag{E.6}$$

The score spaces based on multiple component GMMs are complicated by the log-of-a-sum terms. However a simple analysis is possible for a 2 mixture component GMM where the two mixture components are trained to seperately model classes ω_a and ω_b, but the two variances are tied so that $v_a = v_b = v$. The new statistical model is simply denoted by $S(\boldsymbol{\theta})$.

- *Likelihood (2-class) score space*, $\varphi^{\mathrm{lk(a,b)}}(1,\boldsymbol{\theta}_0)$: This is a 3-component score space and there is a (2×2) metric tensor in tangent space to $S(\boldsymbol{\theta})$ at $\boldsymbol{\theta}_0$ denoted by $g^{\mathrm{lk(a,b)}}(\boldsymbol{\theta}_0)$. The off-main diagonal elements of this tensor are not necessarily zero. If $\gamma_a(O)$ and $\gamma_b(O)$ are respectively the posteriors of components $\mathcal{N}(O;\mu_a,v_a)$ and $\mathcal{N}(O;\mu_b,v_b)$ for sample O, then assuming the mixture components have equal weight, the component posteriors are sigmoidal so,

$$\begin{aligned} \gamma_a(O) \;&=\; \frac{\mathcal{N}(O;\mu_a,v_a)}{\mathcal{N}(O;\mu_a,v_a)+\mathcal{N}(O;\mu_b,v_b)} \\ &=\; \frac{1}{1+B\exp\{-AO\}} \qquad \text{(E.7)} \\ \gamma_b(O) \;&=\; 1-\gamma_a(O) \qquad \text{(E.8)} \end{aligned}$$

where,

$$A = \frac{(\mu_a - \mu_b)}{v} \tag{E.9}$$

$$B = \exp\Big\{\frac{1}{2v}(\mu_a^2 - \mu_b^2)\Big\} \tag{E.10}$$

The metric tensor induced on the input manifold then has the component,

$$g_{11}^{\mathrm{lk(a,b)}}(O;\boldsymbol{\theta}_0) = \frac{1}{v^2}\Big(\gamma_a(O)(O-\mu_a) + \gamma_b(O)(O-\mu_b)\Big)^2 + \sum_{i=a,b}\sum_{j=a,b}\frac{\partial^2}{\partial O\partial\mu_i}\ln p(O)\frac{\partial^2}{\partial O\partial\mu_j}\ln p(O)[g_{n(i)n(j)}^{\mathrm{lk(a,b)}}(\boldsymbol{\theta}_0)]^{-1} \tag{E.11}$$

where $n(i)$ and $n(j)$ are the appropriate indeces which correspond to their class identifier arguments, and,

$$\frac{\partial^2}{\partial O\partial\mu_a}\ln p(O) = \frac{1}{v}\Big(\frac{AB\exp\{-OA\}(O-\mu_a)}{(1+\exp\{-OA\}B)^2} + \gamma_a(O)\Big) \tag{E.12}$$

$$\frac{\partial^2}{\partial O\partial\mu_b}\ln p(O) = \frac{1}{v}\Big(-\frac{AB^{-1}\exp\{OA\}(O-\mu_b)}{(1+\exp\{OA\}B^{-1})^2} + \gamma_b(O)\Big) \tag{E.13}$$

- *Fisher score space*, $\varphi^{\mathrm{fs(a,b)}}(\boldsymbol{\theta}_0)$: This score space is identical to the likelihood (2-class) score space above except for the omission of the zeroth degree covariant derivative. Then,

$$g_{11}^{\mathrm{fs(a,b)}}(O;\boldsymbol{\theta}_0) = \sum_{i=a,b}\sum_{j=a,b}\frac{\partial^2}{\partial O\partial\mu_i}\ln p(O)\frac{\partial^2}{\partial O\partial\mu_j}\ln p(O)[g_{n(i)n(j)}^{\mathrm{lk(a,b)}}(\boldsymbol{\theta}_0)]^{-1} \tag{E.14}$$

The likelihood-ratio score space is the most important for the experiments in this thesis. When defined on class-conditional Gaussian distributions, then the component of the metric tensor induced on the input manifold varies quadratically along input space. This is evident by rewriting the component of the metric tensor in Equation E.4 using Equation E.2 to give,

$$g_{11}^{\mathrm{lr(a,b)}}(O;\boldsymbol{\xi}_0) = \Big(\frac{(O-\mu_a)}{v_a} - \frac{(O-\mu_b)}{v_b}\Big)^2 + \frac{[g_{11}^{\mathrm{lk(a)}}((\boldsymbol{\theta}_a)_0)]^{-1}}{v_a^2} + \frac{[g_{11}^{\mathrm{lk(b)}}((\boldsymbol{\theta}_b)_0)]^{-1}}{v_b^2} \tag{E.15}$$

Then,

$$g_{11}^{\mathrm{lr(a,b)}}(O;\boldsymbol{\xi}_0) \geq \frac{[g_{11}^{\mathrm{lk(a)}}((\boldsymbol{\theta}_a)_0)]^{-1}}{v_a^2} + \frac{[g_{11}^{\mathrm{lk(b)}}((\boldsymbol{\theta}_b)_0)]^{-1}}{v_b^2} \tag{E.16}$$

and the minimum occurs at,

$$O = \frac{(\mu_a v_b - \mu_b v_a)}{(v_b - v_a)} \tag{E.17}$$

The location of this minimum is dependent on the model parameters. The value of $g_{11}^{\mathrm{lr(a,b)}}(O;\boldsymbol{\xi}_0)$ differs at the peaks of the two Gaussians, i.e. at $O = \mu_a$ and $O = \mu_b$ unless $v_a = v_b$. If the two Gaussians have equal variance, $v_a = v_b = v$, then the value of this component is constant at,

$$g_{11}^{\mathrm{lr(a,b)}}(O;\boldsymbol{\xi}_0) = \frac{[g_{11}^{\mathrm{lk(a)}}((\boldsymbol{\theta}_a)_0)]^{-1}}{v^2} + \frac{[g_{11}^{\mathrm{lk(b)}}((\boldsymbol{\theta}_b)_0)]^{-1}}{v^2} + \frac{(\mu_a - \mu_b)^2}{v^2} \tag{E.18}$$

It is useful to relate the variation of the value of this component to the magnification factor. Since the value is positive and the metric tensor is of size (1×1), then the magnification factor $M(O)$ is,

$$M(O) = \sqrt{g_{11}^{\mathrm{lr(a,b)}}(O;\boldsymbol{\xi}_0)} \tag{E.19}$$

For the case of unequal variances, the minimum magnification increases as the two Gaussians narrow and all other variables are kept unchanged. From Equation E.15 and assuming the Gaussian means μ_a and μ_b are fixed, then magnification is guaranteed to increase at locations between the peaks of the two Gaussians, if either or both v_a and v_b decrease. If the two Gaussians are trained by MMIE, then it is expected that the two Gaussians will be drawn towards the midpoint between the two classes as their variances narrow. In this case, magnification is again guaranteed to increase between the two Gaussian peaks, but providing v_a decreases at a greater rate than $(O - \mu_a)$ and v_b decreases at a greater rate than $(O - \mu_b)$. With this constraint, distributions trained via MMIE are expected to increase magnification induced by the score mapping near to the anticipated decision boundary between the two classes.

The metric tensor $g^{\mathrm{lk(a,b)}}(O;\boldsymbol{\theta}_0)$ can be used to construct a metric tensor for likelihood-ratio score space based on 2-mixture component GMMs for each class ω_a and ω_b, but where the variances of mixture components within each class are tied. The parameters for the model for class ω_a are then $\boldsymbol{\theta}_a = (\mu_{a1}, \mu_{a2}, v_a)^\top$ where μ_{a1} and μ_{a2} denote the Gaussian means and v_a the tied variance. A similar pattern follows for class ω_b. The resulting metric

tensor $g^{\mathrm{lr(a,b)}}(O;\boldsymbol{\xi}_0)$ has the fully covariant component,

$$\begin{aligned} g_{11}^{\mathrm{lr(a,b)}}(O;\boldsymbol{\xi}_0) &= g_{11}^{\mathrm{lk(a1,a2)}}(O;(\boldsymbol{\theta}_a)_0) + g_{11}^{\mathrm{lk(b1,b2)}}(O;(\boldsymbol{\theta}_b)_0) \\ &\quad -\frac{2}{v_a v_b}\Big(\gamma_{a1}(O)(O-\mu_{a1}) + \gamma_{a2}(O)(O-\mu_{a2})\Big) \\ &\qquad \Big(\gamma_{b1}(O)(O-\mu_{b1}) + \gamma_{b2}(O)(O-\mu_{b2})\Big) \end{aligned} \tag{E.20}$$

where $g_{11}^{\mathrm{lk(a1,a2)}}(O;(\boldsymbol{\theta}_a)_0)$ and $g_{11}^{\mathrm{lk(b1,b2)}}(O;(\boldsymbol{\theta}_b)_0)$ are detailed in Equation E.11, but where 'a' and 'b' are respectively replaced by '$a1$' and '$a2$' and then '$b1$' and '$b2$'.

The metric tensors detailed for 1-component input space can be used as 'building blocks' for the metric tensors for d-component input space. This is under the assumptions that covariant derivatives are defined with respect to Gaussian means only, the metric tensor for tangent space to $S(\boldsymbol{\theta}_q)$ or $\mathcal{S}(\boldsymbol{\xi})$ is diagonal, and each component in d-component input space is statistically independent according to the statistical model applied. These assumptions are not always fulfilled, for example statistical independence is not consistent with using GMMs with two or more distinct mixture components to model data in input space. As an example, a d-component input space is proposed and a statistical model $S(\boldsymbol{\theta}_q)$ which is a single Gaussian with mean $\boldsymbol{\mu}_q$ and diagonal covariance $\boldsymbol{\Sigma}_q$. The mean and covariance are both expressed in fully contravariant form so $\boldsymbol{\mu}_q = (\mu^1 \dots \mu^d)^\top$ and the (i,j)th component of $\boldsymbol{\Sigma}_q$ is v^{ij}, and $v^{ij} = 0$ if $i \neq j$. Then the metric tensor induced on the input manifold by the mapping into the likelihood score space for class ω_q is,

$$g_{ij}^{\mathrm{lk(q)}}(\boldsymbol{O};(\boldsymbol{\theta}_q)_0) = \begin{cases} g_{11}^{\mathrm{lk(q)}}(O^i;\mu^i,v^{ii}) & \text{if} \quad i=j \\ \frac{(O^i-\mu^i)}{v^{ii}}\frac{(O^j-\mu^j)}{v^{jj}} & \text{if} \quad i\neq j \end{cases} \tag{E.21}$$

where $g_{11}^{\mathrm{lk(q)}}(O^i;\mu^i,v^{ii})$ is as detailed in Equation E.2 but here applied to the ith component of input space. A similar analysis may be pursued for other score spaces.

Appendix F

Fibre bundles

F.1 General description

The general description of fibre bundles and vector bundles in Section F.1 is summarised from Section 155 of [50], and [108]. First, a general fibre bundle η may be described by,

$$\eta = (E, f, S, F, G) \tag{F.1}$$

where,

- E is a topological space called the *total space*,
- S is a topological space called the *base space*,
- f is a continuous mapping $f : E \to S$ called the *projection*,
- F is a topological space called the *fibre*,
- G is an effective left tranformation group of F called the *bundle group*.

In this thesis, S has a global coordinate system and the bundle is described by the *trivialisation* $S \times F$. The bundle group is then the group of transition functions which map

from one trivialisation to another. An example of a fibre bundle in this thesis is $\eta_{\breve{p}}$ which exists in the space of scalar functions.

An important variety of fibre bundle for this thesis is the *vector bundle* where the fibre is a real vector space, i.e. any linear space over the field of real numbers $\mathbb{R}$. The real vector bundle η_{vec} of bundle rank δ may be summarised by,

$$\eta_{\text{vec}} = (E, f, S, \mathbb{R}^\delta, \text{GL}(\delta, \mathbb{R})) \tag{F.2}$$

where E, f and S are as described above, the fibre F is $\mathbb{R}^\delta$ and the bundle group is $\text{GL}(\delta, \mathbb{R})$, the *general linear group* of degree δ over $\mathbb{R}$. This is essentially the group of all $\delta \times \delta$ invertible matrices over $\mathbb{R}$ (see Section 63.B of [50]).

F.2 Summary of notation for fibre bundles

The fibre bundle $\eta_{\breve{p}}$ lies within the space of scalar functions $L(\breve{p})$. The structure of a fibre anchored at point $p_0 \in S(\boldsymbol{\theta})$ with coordinate vector $\boldsymbol{\theta}_0$ is as follows.

- $\breve{S}(\tau, \boldsymbol{\alpha})$: the fibre without an interpretation but with coordinates $\tau \in L(\tau) = \mathbb{R}$ and $\boldsymbol{\alpha} \in L(\boldsymbol{\alpha}) = \mathbb{R}^\delta$ and where $\delta = \dim(L(\boldsymbol{\alpha})) = \infty$.

- $\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$: this is identical to $\breve{S}(\tau, \boldsymbol{\alpha})$ except it is given an interpretation in terms of its anchor point $\boldsymbol{\theta}_0$ and the scalar function $\bar{\varsigma}$ used to define the fibre, where $\bar{\varsigma}$ initially varies over $L(\boldsymbol{O}) \times L(\boldsymbol{\theta}; S)$. A fixed $\boldsymbol{O}_l \in L(\boldsymbol{O})$ then yields the scalar field $\bar{\varsigma}_l$ over $L(\boldsymbol{\theta}; S)$. The definitions of both $\bar{\varsigma}$ and $\bar{\varsigma}_l$ can be extended from the coordinate space of the base manifold $S(\boldsymbol{\theta})$ to the coordinate space of the fibre bundle.

- $\breve{S}(\tau, \boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0)$: a submanifold of $\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$ where $\boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; \varrho)$, and where $L(\boldsymbol{\alpha}; \varrho)$ ensures all components of $\boldsymbol{\alpha}$ are zero which are isomorphic to $\alpha^{j_1 \cdots j_r}, r > \varrho$. As $\varrho \to \infty$, $\breve{S}(\tau, \boldsymbol{\alpha}; \varrho, \bar{\varsigma}, \boldsymbol{\theta}_0) \to \breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$.

- $\breve{S}(\tau, \boldsymbol{\alpha}; \boldsymbol{O}, \bar{\varsigma}, \boldsymbol{\theta}_0)$: a submanifold of $\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$ corresponding to the points 'reachable' from any $\boldsymbol{O}_l \in L(\boldsymbol{O})$ through the mapping defined on the scalar field $\bar{\varsigma}_l$.

If τ is kept constant, then the τ transfers from the left to the right of the semicolon. For example,

$$\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)\Big|_{\tau=\tau'} \equiv \breve{S}(\boldsymbol{\alpha}; \tau, \bar{\varsigma}, \boldsymbol{\theta}_0)\Big|_{\tau=\tau'} \tag{F.3}$$

Further constraints on the form of $\boldsymbol{\alpha}$ give the following submanifolds.

- $\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$: a submanifold of $\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}_0)$ where $\boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ which implies for $j_i = \{1, \ldots n\}, \forall i$,

$$\alpha^{j_1 \ldots j_r} = \prod_{i=1}^{r} \theta'^{j_i} - \theta_0^{j_i} \tag{F.4}$$

- $\breve{S}(\tau, \boldsymbol{\alpha}; \mathrm{cd}, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$: a submanifold of $\breve{S}(\tau, \boldsymbol{\alpha}; \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$ where $\boldsymbol{\alpha} \in L(\boldsymbol{\alpha}; \mathrm{cd}, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$, and where $L(\boldsymbol{\alpha}; \mathrm{cd}, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0) \subseteq L(\boldsymbol{\alpha}; \boldsymbol{\theta}', \boldsymbol{\theta}_0)$. This implies $\boldsymbol{\theta}'$ is within the convergence domain of $\bar{\varsigma}_l$ about $\boldsymbol{\theta}_0$.

An identical form of notation is applied to the corresponding denormalisation $\tilde{S}(\tau, \boldsymbol{\alpha})$ and statistical model $S(\boldsymbol{\alpha})$. Similar notation is also applied to the subspaces within the fibres of the vector bundle η_{vec}. For example, there is an isomorphism, $\widehat{\breve{L}}^{1(1,0)}(\tau, \boldsymbol{\alpha}; \mathrm{cd}, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0) \cong \breve{S}(\tau, \boldsymbol{\alpha}; \mathrm{cd}, \bar{\varsigma}, \boldsymbol{\theta}', \boldsymbol{\theta}_0)$.

F.3 Intersecting fibres

It is possible that two fibres, defined at different points on the base manifold of the same fibre bundle, intersect. A knowledge of the intersection of fibres is valuable, for example in ascertaining whether the selection of a point on the base manifold for a fibre seriously restricts the set of possible solutions for estimating points in the bundle. This has consequences for 'fibre hopping' described in Section 3.5.3.1. Let two points on the base manifold $S(\boldsymbol{\theta})$ have coordinate vectors $(\boldsymbol{\theta}_1)_0$ and $(\boldsymbol{\theta}_2)_0$. Without loss in generality, the analysis is restricted to fibres within $L(p)$. Two fibres are extended and points selected,

$$p_1' \in S(\boldsymbol{\alpha}_1; \varrho_1, \bar{\varsigma}, (\boldsymbol{\theta}_1)_0) \tag{F.5}$$

$$p_2' \in S(\boldsymbol{\alpha}_2; \varrho_2, \bar{\varsigma}, (\boldsymbol{\theta}_2)_0) \tag{F.6}$$

These two points coincide in terms of their output over $L(\boldsymbol{O})$ if the following equality holds $\forall \boldsymbol{O}_l \in L(\boldsymbol{O})$,

$$\sum_{r=0}^{\varrho_1} \sum_{j_1 \ldots j_r} (T_1)_{j_1 \ldots j_r}(\boldsymbol{O}_l)(\alpha_1)^{j_1 \ldots j_r} - D(\boldsymbol{\alpha}_1) \quad = \quad \sum_{r=0}^{\varrho_2} \sum_{j_1 \ldots j_r} (T_2)_{j_1 \ldots j_r}(\boldsymbol{O}_l)(\alpha_2)^{j_1 \ldots j_r} - D(\boldsymbol{\alpha}_2) \tag{F.7}$$

where the notation is as described for the Taylor expansion in Section 3.3. The two points p'_1 and p'_2 are distributions but only have the same semantic meaning as points on the base manifold if the scalar function is the log likelihood, i.e. $\bar{\varsigma}_l = \ln p(\boldsymbol{O}_l; \boldsymbol{\theta})$.

Appendix G

Error analysis

This section contains a brief analysis of experimental errors and the measure of statistical significance applied in this thesis.

G.1 Sources of errors

Possible sources of error and error estimates are listed, split where appropriate between those for the experiments on the toy/Deterding vowel dataset and the ISOLET dataset. The list is not exhaustive and the error estimates are subjective, so the analysis is simply a guide. In this list, the term 'verification by inspection' implies the source code was carefully inspected in a 'line-by-line' manner. The term 'inspection by effect' implies the source code was principally verified by its effect on experimental results. Hence the list is as follows.

1. Programming errors: Minimal risk expected for implementing algorithms since the trends in the experimental results were largely explainable.

    - Toy/Deterding: 2% error: MATLAB code verified by inspection, C code verified by effect.

- ISOLET: 2% error: C code verified by effect.

2. Numerical errors: Errors arising in implementation.

   - Toy/Deterding: 1% error: C code verified principally by effect, MATLAB code verified by inspection. MATLAB provides robust functionality.
   - ISOLET: 1% error: C code verified principally by effect. The task is data intensive and sufficient robustness for all situations is unlikely.

3. Experimental settings: The large number of experiments introduces the possibility of errors in the experimental parameters.

   - Toy/Deterding: 5% error with regard to individual experiments; 1% error with regard to experimental trends.
   - ISOLET: 1% error with regard to individual experiments; $< 1\%$ error with regard to experimental trends.

4. Shell script errors: Shell scripts were used significantly, both to control the experiments and to prepare and process data files. The scripts were often highly complicated and performed many operations that could otherwise be embedded directly into C code. The shell scripts minimised the degree of human interaction or preparation or analysis of data files. The scripts contained frequent error checks and were written in bash, awk and perl.

   - Toy/Deterding: 2% error: Scripts verified by inspection.
   - ISOLET: 4% error: Scripts verified by effect.

5. NFS and file system errors: The experiments were computationally expensive and the scripts required the creation and processing of many small files. The burden on the file server was intense.

   - Toy/Deterding: $< 1\%$ error: Shell scripts should have detected some errors. The effect on experimental trends should be minimal.
   - ISOLET: 1%: Each experiment was split into chunks and processed simultaneously on different processors. There was also a known and observed, but

deemed low-risk, problem in which two data streams written to two large files were accidentally interchanged. Again the shell script should have detected some errors, and the effect on experimental trends should again be minimal.

6. Errors in detailing, tabulating, plotting and analysing results: 1%

Errors can be divided into systematic and random errors [110]. Systematic errors have a systematic influence on experimental results, whereas random errors have a random and unpredictable effect. Since computers are deterministic, most errors encountered can be repeated, though some programming errors or numerical errors may not. The errors detailed above may include both types of errors. The division of errors into systematic and random errors does not affect their analysis with respect to propagation. According to [110], the error ΔZ for a quantity $Z = XY$, assuming the errors ΔX and ΔY are independent, is estimated as,

$$\left(\frac{\Delta Z}{Z}\right)^2 = \left(\frac{\Delta X}{X}\right)^2 + \left(\frac{\Delta Y}{Y}\right)^2 \tag{G.1}$$

Assuming the errors listed above are independent and are propagated in this manner, the error estimate for each toy/Deterding experimental result is 6% and for the experimental trends is 3%, and the error estimate for each ISOLET experimental result is 5% and for the experimental trends is also 5% (in these calculations, an error estimate of $< 1\%$ is taken as 1%).

G.2 McNemar's test

A more objective analysis of errors via confidence levels is applied to the ISOLET experiments in this thesis, but not to the experiments on the artificial or the Deterding vowel data which are intended for illustrating concepts. The confidence levels are based on McNemar's test as detailed and referenced in [40]. Each confidence level gives a measure of belief that the two relevant classifiers have different probabilities of error, and that any observed difference in their test error rates is not due solely to random effects. The confidence level is calculated solely on the noncommon errors made by the two classifiers. If

the two classifiers have the same probability of error, then noncommon errors may result from random influences, but the number of noncommon errors should be evenly balanced between the two classifiers. The probability P of the difference being due to chance effects is measured, assuming errors are i.i.d. among test samples, using a binomial distribution and a percentage confidence level for a sincere difference of $100(1-P)$ reported[1]. This test is useful for both large and small datasets. However it should be noted that the confidence level is only dependent on the number of noncommon errors, not on the total number of errors or on the total number of samples in the test set. The confidence level is therefore less reliable if the number of noncommon errors in absolute terms is small, or if test samples are not representative of the underlying distribution. Unfortunately the small test sets for the ISOLET task introduce unreliability to the values of confidence levels. Furthermore, the sources of error listed for the ISOLET experiments in Section G.1 include sources of both random and systematic errors, while the confidence levels only accomodate sources of random error. Despite this, the confidence levels reported for the ISOLET experiments give some quantitative and objective, and so useful though insufficient, analysis of errors.

[1] As detailed in [40], McNemar's original χ^2 test applies a normal approximation to the binomial distribution and a continuity correction factor.

Bibliography

[1] S. Amari and S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12:783–789, 1999.

[2] S.-I. Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, 1998.

[3] S.-I. Amari and H. Nagaoka. *Methods of Information Geometry*. Oxford University Press, 2000. (Translations of Mathematical Monographs, Volume 191, American Mathematical Society).

[4] L.R. Bahl, P.F. Brown, P.V. de Souza, and R.L. Mercer. Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition. In *Proceedings*, pages 49–52. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1986.

[5] I. Bazzi and D. Katabi. Using Support Vector Machines for Spoken Digit Recognition. In *Proceedings*. International Conference on Spoken Language Processing, 2000.

[6] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[7] C.L. Blake and C.J. Merz. UCI Repository of machine learning databases, 1998. `http://www.ics.uci.edu/~mlearn/MLRepository.html`, University of California, Irvine, School of Information and Computer Science.

[8] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.

[9] C.J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[10] W.M. Campbell. Generalized Linear Discriminant Sequence Kernels for Speaker Recognition. In *Proceedings*, pages 161–164. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002.

[11] W.M. Campbell. A SVM/HMM System for Speaker Recognition. In *Proceedings*, pages 209–212. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003.

[12] S. Chakrabartty and G. Cauwenberghs. Forward-Decoding Kernel-Based Phone Sequence Recognition. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*. The MIT Press, 2003.

[13] N.N. Chentsov (Čencov). *Statistical Decision Rules and Optimal Inference*. Translations of Mathematical Monographs Volume 53. American Mathematical Society, 1982.

[14] K.K. Chin. Private Communication (1998). (while K.K.Chin was studying towards a MPhil in Computer Speech and Language Processing, Cambridge University).

[15] P. Clarkson and P.J. Moreno. On the Use of Support Vector Machines for Phonetic Classification. In *Proceedings*, pages 585–588. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999.

[16] R. Cole, Y. Muthusamy, and M. Fanty. The ISOLET spoken letter database. Technical Report CSE 90-004, Oregon Graduate Institute of Science and Technology, March 1990.

[17] C. Cortes, P. Haffner, and M. Mohri. Rational Kernels. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*. The MIT Press, 2003.

[18] K. Crammer and Y. Singer. On the Learnability and Design of Output Codes for Multiclass Problems. *Machine Learning*, 47(2):201–233, 2002.

[19] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000 (reprinted).

[20] J.R. Deller, J.G. Proakis, and J.H.L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan Publishing Company, 1993.

[21] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977.

[22] J. di Martino. On the Use of High Order Derivatives for High Performance Alphabet Recognition. In *Proceedings*, pages 953–956. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002.

[23] P. Ding, Z. Chen, Y. Liu, and B. Xu. Asymmetrical Support Vector Machines and Applications in Speech Processing. In *Proceedings*, pages 73–76. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002.

[24] X. Dong, W. Zhaohui, and Y. Yingchun. Exploiting Support Vector Machines in Hidden Markov Models for Speaker Verification. In *Proceedings*, pages 1329–1332. International Conference on Spoken Language Processing, 2002.

[25] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[26] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 681–687. The MIT Press, 2002.

[27] Emacs. GNU Emacs software. `http://www.gnu.org/software/emacs/emacs.html`.

[28] Y. Ephraim and L.R. Rabiner. On the Relations Between Modeling Approaches for Speech Recognition. *IEEE Transactions on Information Theory*, 36(2):372–380, 1990.

[29] M. Fanty and R. Cole. Speaker-Independent English Alphabet Recognition: Experiments with the E-Set. In *Proceedings*, pages 1361–1364. International Conference on Spoken Language Processing, 1990.

[30] M. Fanty and R. Cole. Spoken Letter Recognition. In R.P. Lippmann, J.E. Moody, and D.S. Touretzky, editors, *Advances in Neural Information Processing Systems 3*, pages 220–226. Morgan Kaufmann Publishers, 1991.

[31] S. Fine, J. Navrátil, and R.A. Gopinath. A hybrid GMM/SVM approach to speaker identification. In *Proceedings*, pages 417–420. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001.

[32] S. Fine, J. Navrátil, and R.A. Gopinath. Enhancing GMM Scores using SVM "Hints". In *Proceedings*, pages 1757–1760. Eurospeech, 2001.

[33] S. Fine, G. Saon, and R.A. Gopinath. Digit Recognition in Noisy Environments via a Sequential GMM/SVM System. In *Proceedings*, pages 49–52. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002.

[34] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Electrical Science Series. Academic Press, 1972.

[35] M.J.F. Gales. Private Communication (PhD supervision, 1999-2003). (Machine Intelligence Laboratory, Cambridge University Engineering Department).

[36] M.J.F. Gales. Maximum Likelihood Multiple Subspace Projections for Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, 10(2):37–47, 2002.

[37] A. Ganapathiraju, J. Hamaker, and J. Picone. Support Vector Machines for Speech Recognition. In *Proceedings*, pages 2923–2926. International Conference on Spoken Language Processing, 1998.

[38] A. Ganapathiraju, J. Hamaker, and J. Picone. Hybrid SVM/HMM Architectures for Speech Recognition. In *Proceedings*. International Conference on Spoken Language Processing, 2000.

[39] J.-L. Gauvain and C.-H. Lee. Maximum *a Posteriori* Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, 1994.

[40] L. Gillick and S.J. Cox. Some Statistical Issues in the Comparison of Speech Recognition Algorithms. In *Proceedings*, pages 532–535. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1989.

[41] S.J. Godsill and P.J.W. Rayner. *Digital Audio Restoration*. Springer-Verlag, 1998.

[42] R.A. Gopinath. Some Thoughts on Kernel PCA and LDA, April 2000. (from `http://www.research.ibm.com/people/r/rameshg/publications.htm`).

[43] D.T. Guarrera, N.G. Johnson, and H.F. Wolfe. The Taylor Expansion of a Riemannian Metric, July 2002. (from `http://mail.rochester.edu/~nj001i/reu2002.html`).

[44] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1/2/3):389–422, 2002.

[45] J.E. Hamaker, J. Picone, and A. Ganapathiraju. A Sparse Modeling Approach to Speech Recognition Based on Relevance Vector Machines. In *Proceedings*, pages 1001–1004. International Conference on Spoken Language Processing, 2002.

[46] P.E. Hart. The Condensed Nearest Neighbor Rule. *IEEE Transactions on Information Theory*, 14(3):515–516, 1968.

[47] T. Hastie and R. Tibshirani. Discriminant Adaptive Nearest Neighbor Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(6):607–616, 1996.

[48] C.-W. Hsu and C.-J. Lin. A Comparison of Methods for Multiclass Support Vector Machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[49] HTK3. HTK3 webpage. `http://htk.eng.cam.ac.uk`.

[50] S. Iyanaga and Y. Kawada, editors. *Encyclopedic Dictionary of Mathematics*. The MIT Press, 1977. (By the Mathematical Society of Japan).

[51] T. Jaakkola, M. Diekhans, and D. Haussler. A Discriminative Framework for Detecting Remote Protein Homologies. *Journal of Computational Biology*, 7(1/2):95–114, 2000.

[52] T.S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 487–493. The MIT Press, 1999.

[53] T. Joachims. SVM^{light} Support Vector Machine webpage. `http://svmlight.joachims.org`.

[54] T. Joachims. Making Large-Scale Support Vector Machine Learning Practical. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. The MIT Press, 1998.

[55] S. Kapadia. *Discriminative Training of Hidden Markov Models*. PhD thesis, Cambridge University Engineering Department, March 1998.

[56] M. Karnjanadecha and S.A. Zahorian. Signal Modeling for Isolated Word Recognition. In *Proceedings*, pages 293–296. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999.

[57] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation*, 13(3):637–649, 2001.

[58] R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1/2):273–324, 1997.

[59] G.A. Korn and T.M. Korn. *Mathematical Handbook for Scientists and Engineers*. McGraw-Hill, 2nd edition, 1968.

[60] B. Krishnapuram and L. Carin. Support Vector Machines for Improved Multiaspect Target Recognition Using the Fisher Kernel Scores of Hidden Markov Models. In

Proceedings, pages 2989–2992. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002.

[61] N. Kumar and A.G. Andreou. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication*, 26(4):283–297, 1998.

[62] J.T.-Y. Kwok. Moderating the Outputs of Support Vector Machine Classifiers. *IEEE Transactions on Neural Networks*, 10(5):1018–1031, 1999.

[63] J.T.-Y. Kwok. The Evidence Framework Applied to Support Vector Machines. *IEEE Transactions on Neural Networks*, 11(5):1162–1173, 2000.

[64] W. Lee, C.C. Sekhar, K. Takeda, and F. Itakura. Recognition of Continuous Speech Segments of Monophone Units Using Support Vector Machines. In *Proceedings*, pages 2653–2656. International Conference on Spoken Language Processing, 2002.

[65] C.J. Leggetter. *Improved Acoustic Modelling for HMMs using Linear Transformations*. PhD thesis, Cambridge University Engineering Department, February 1995.

[66] B.T. Logan. *Adaptive Model-Based Speech Enhancement*. PhD thesis, Cambridge University Engineering Department, July 1998.

[67] P.C. Loizou and A.S. Spanias. High-Performance Alphabet Recognition. *IEEE Transactions on Speech and Audio Processing*, 4(6):430–445, 1996.

[68] MATLAB. The MathWorks website. `http://www.mathworks.com`.

[69] E. Mayoraz and M. Moreira. On the Decomposition of Polychotomies into Dichotomies. In *Proceedings*, pages 219–226. International Conference on Machine Learning, Morgan Kaufmann, 1997.

[70] S. Mika. *Kernel Fisher Discriminants*. PhD thesis, University of Technology, Berlin, October 2002.

[71] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, and G. Rätsch. Kernel PCA and De-Noising in Feature Spaces. In M.S. Kearns, S.A. Solla, and D.A. Cohn,

editors, *Advances in Neural Information Processing Systems 11*, pages 536–542. The MIT Press, 1999.

[72] H.J. Nock. *Techniques For Modelling Phonological Processes In Automatic Speech Recognition*. PhD thesis, Cambridge University Engineering Department, May 2001.

[73] N. Oliver, B. Schölkopf, and A.J. Smola. Natural Regularization from Generative Models. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 51–60. The MIT Press, 2000.

[74] Pico. The University of Washington's Pine Information Center. `http://www.washington.edu/pine`.

[75] J.C. Platt. John Platt's Home Page. `http://research.microsoft.com/~jplatt`.

[76] J.C. Platt. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 41–64. The MIT Press, 1998.

[77] J.C. Platt. Probabilities for SV Machines. In A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. The MIT Press, 2000.

[78] J.T. Pohlmann. Factor Analysis Glossary, EPSY 580B - Factor Analysis Seminar. Department of Educational Psychology and Special Education, Southern Illinois University, `http://www.siu.edu/~epse1/pohlmann/factglos`.

[79] H. Printz and P.A. Olsen. Theory and practice of acoustic confusability. *Computer Speech and Language*, 16:131–164, 2002.

[80] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall Signal Processing Series. PTR Prentice-Hall, 1993.

[81] K.F. Riley, M.P. Hobson, and S.J. Bence. *Mathematical methods for physics and engineering*. Cambridge University Press, 1997.

[82] A.J. Robinson. *Dynamic Error Propagation Networks*. PhD thesis, Cambridge University Engineering Department, February 1989.

[83] A.J. Robinson. An Application of Recurrent Nets to Phone Probability Estimation. *IEEE Transactions on Neural Networks*, 5(2):298–305, 1994.

[84] A.-V.I. Rosti and M.J.F. Gales. Generalised Linear Gaussian Models. Technical Report CUED/F-INFENG/TR.420, Cambridge University Engineering Department, November 2001.

[85] S. Roweis and Z. Ghahramani. A Unifying Review of Linear Gaussian Models. *Neural Computation*, 11(2):305–345, 1999.

[86] S.T. Roweis and L.K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290:2323–2326, 2000.

[87] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen. Maximum Likelihood Discriminant Feature Spaces. In *Proceedings*, pages 1129–1132. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000.

[88] C. Saunders, J. Shawe-Taylor, and A. Vinokourov. String Kernels, Fisher Kernels and Finite State Automata. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*. The MIT Press, 2003.

[89] H. Shimodaira, K.-I. Noma, M. Nakai, and S. Sagayama. Support Vector Machine with Dynamic Time-Alignment Kernel for Speech Recognition. In *Proceedings*, pages 1841–1844. Eurospeech, 2001.

[90] N. Smith. Support Vector Machines applied to Speech Pattern Classification, MPhil thesis, Cambridge University Engineering Department, August 1998.

[91] N. Smith and M. Gales. Speech Recognition using SVMs. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 1197–1204. The MIT Press, 2002.

[92] N. Smith, M. Gales, and M. Niranjan. Data-Dependent Kernels in SVM Classification of Speech Patterns. Technical Report CUED/F-INFENG/TR.387, Cambridge University Engineering Department, April 2001 (updated April 2002).

[93] N. Smith and M. Niranjan. Data-Dependent Kernels in SVM Classification of Speech Patterns. In *Proceedings*. International Conference on Spoken Language Processing, 2000.

[94] N.D. Smith and M.J.F. Gales. Using SVMs and Discriminative Models for Speech Recognition. In *Proceedings*, pages 77–80. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002.

[95] N.D. Smith and M.J.F. Gales. Using SVMs to Classify Variable Length Speech Patterns. Technical Report CUED/F-INFENG/TR.412, Cambridge University Engineering Department, April 2002 (updated June 2002).

[96] P. Sollich. Probabilistic Methods for Support Vector Machines. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 349–355. The MIT Press, 2000.

[97] P. Sollich. Bayesian Methods for Support Vector Machines: Evidence and Predictive Class Probabilities. *Machine Learning*, 46(1/2/3):21–52, 2002.

[98] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290:2319–2323, 2000.

[99] J.B. Tenenbaum and W.T. Freeman. Separating Style and Content with Bilinear Models. *Neural Computation*, 12(6):1247–1283, 2000.

[100] C.W. Therrien. *Decision Estimation and Classification: An Introduction to Pattern Recognition and Related Topics*. John Wiley & Sons, 1989.

[101] K. Tsuda, M. Kawanabe, and K.-R. Müller. Clustering with the Fisher Score. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*. The MIT Press, 2003.

[102] K. Tsuda, M. Kawanabe, G. Rätsch, S. Sonnenburg, and K.-R. Müller. A New Discriminative Kernel From Probabilistic Models. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 977–984. The MIT Press, 2002.

[103] V. Valtchev. *Discriminative Methods in HMM-based Speech Recognition.* PhD thesis, Cambridge University Engineering Department, March 1995.

[104] V.N. Vapnik. *The Nature of Statistical Learning Theory (Second Edition).* Statistics for Engineering and Information Science. Springer-Verlag, 2000.

[105] V. Wan and S. Renals. Evaluation of Kernel Methods for Speaker Verification and Identification. In *Proceedings*, pages 669–672. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002.

[106] V. Wan and S. Renals. SVMSVM: Support Vector Machine Speaker Verification Methodology. In *Proceedings*, pages 221–224. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003.

[107] M. Webber. Private Communication (2003). (Machine Intelligence Laboratory, Cambridge University Engineering Department).

[108] E. Weisstein. Eric Weisstein's World of Mathematics (MathWorld^{TM}). `http://mathworld.wolfram.com`.

[109] J. Weston and C. Watkins. Multi-class Support Vector Machines. Technical Report CSD-TR-98-04, Department of Computer Science, Royal Holloway, University of London, May 1998.

[110] F.L.H. Wolfs. Error Analysis, Appendix B of Lab Manuals, 1996-1997. Department of Physics and Astronomy, University of Rochester, `http://teacher.nsrl.rochester.edu/PHY_LABS/AppendixB/AppendixB.html`.

[111] P.C. Woodland and D. Povey. Large scale discriminative training of hidden Markov models for speech recognition. *Computer Speech and Language*, 16(1):25–47, 2002.

[112] S.J. Wright. *Primal-Dual Interior-Point Methods.* Society for Industrial and Applied Mathematics, 1997.

[113] XFig. XFig and related software. `http://www.xfig.org`.

[114] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book (for HTK Version 3.0)*, July 2000. (copyright 1995-1999 Microsoft Corporation).