Precision Matrix Modelling for LVCSR

K.C. Sim & M.J.F. Gales

May 2004

Cambridge University Engineering Department

Overview

- Precision Matrix Modelling
  - motivations;
  - structured approximations;
  - examples: STC, EMLLT, SPAM.
- MPE discriminative training
- Implementation Issues
  - required statistics;
  - variance flooring;
  - determination of MPE smoothing constant.
- Initial performance evaluated on CTS and BN English.

Covariance vs. Precision Matrix Modelling

- Standard systems: HMM-based with GMM output distribution:

$$p\left(\boldsymbol{o}_t|\{c_m,\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m\}\right)=\sum_{m=1}^{M}c_m\sqrt{\frac{|\boldsymbol{P}_m|}{(2\pi)^d}}\exp\left(-\frac{(\boldsymbol{o}_t-\boldsymbol{\mu}_m)'\boldsymbol{P}_m(\boldsymbol{o}_t-\boldsymbol{\mu}_m)}{2}\right)$$

- Full covariance matrix modelling: impractical for LVCSR
  - Covariance matrix dominates number of model parameters
- Covariance modelling is computationally expensive for decoding
- Precision matrix model, $\boldsymbol{P}_m$
  - Compact model representation
  - Efficient likelihood calculation

Structured Precision Matrix Approximations

- Structured approximation: linear superposition of symmetric basis

$$\boldsymbol{P}_m = \sum_{i=1}^{n} \lambda_{ii}^{(m)} \boldsymbol{S}_i = \sum_{i=1}^{n} \lambda_{ii}^{(m)} \left(\sum_{r=1}^{R} \lambda_{ii}^{(r)} \boldsymbol{a}_{ir}' \boldsymbol{a}_{ir} \right)$$

  - **"Global"** parameters: basis matrices $\boldsymbol{S}_i$ or basis vectors $\boldsymbol{a}_{ir}$
  - **"Component"** parameters: basis coefficients $\lambda_{ii}^{(m)}$

- Auxiliary function for EM parameters estimation:

$$\mathcal{Q}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = K + \frac{1}{2} \sum_{m=1}^{M} \beta_m \left\{ \log |\boldsymbol{P}_m| - \sum_{i=1}^{n} \lambda_{ii}^{(m)} \operatorname{Tr}(\boldsymbol{S}_i \boldsymbol{W}_m) \right\}$$

where required statistics are

$$\boldsymbol{W}_m = \frac{\sum_{t=1}^{T} \gamma_m(t)(\boldsymbol{o}_t - \boldsymbol{\mu}_m)(\boldsymbol{o}_t - \boldsymbol{\mu}_m)'}{\beta_m} \quad \text{and} \quad \beta_m = \sum_{t=1}^{T} \gamma_m(t)$$

Precision Matrix Model Examples

- **STC**: $R = 1$, $n = d$
  - Equivalent to feature transformation $\boldsymbol{A}$
  - Closed-form update for $\lambda_{ii}^{(m)}$
  - $\boldsymbol{a}_i$ updated efficiently in an iterative row-by-row fashion

- **EMLLT**: $R = 1$, $d < n \leq \frac{d}{2}(d+1)$
  - Extension to STC: *rectangular* transform
  - Closed-form update for $\lambda_{ii}^{(m)}$
  - $\boldsymbol{a}_i$ updated row-by-row using gradient descent method
  - Initialise $\boldsymbol{A}$ by stacking STC/HLDA transforms

- **SPAM**: $R = d$, $1 < n \leq \frac{d}{2}(d+1)$
  - Extension to EMLLT with arbitrary symmetric basis matrices
  - Conjugate gradient descent update for $\lambda_{ii}^{(m)}$
  - Update of basis matrices is *slow* due to positive-definite constraint
  - Initialise $\boldsymbol{S}_i$ by selecting top n singular vector of average inverse covariance statistics

HLDA as a Precision Matrix Model

- Precision matrix expression for HLDA model

$$\boldsymbol{P}_m = \sum_{i=1}^{n} \lambda_{ii}^{(m)} \boldsymbol{a}_i' \boldsymbol{a}_i + \sum_{i=n+1}^{d} \lambda_{ii} \boldsymbol{a}_i' \boldsymbol{a}_i$$

- HLDA useful dimension, $n < d$

- Second summation corresponds to *nuisance* dimensions

- Extension of STC/EMLLT with global tying for *nuisance* coefficients, λ_{ii}
  - λ_{ii} initialised as *inverse* variances of nuisance dimensions
  - λ_{ii} estimated using conjugate gradient method

- Efficient updates for $\boldsymbol{a}_i$ and $\lambda_{ii}^{(m)}$ (c.f. STC)

Minimum Phone Error Criterion

- MPE criterion

$$\mathcal{F}(\mathcal{M}) = \frac{\sum_w p(\mathbf{O}|\mathcal{M}_w)^\kappa P(w)\mathrm{RawAccuracy}(w)}{\sum_w p(\mathbf{O}|\mathcal{M}_w)^\kappa P(w)}$$

- Use weak-sense auxiliary function

$$\mathcal{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \mathcal{Q}^{(n)}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) - \mathcal{Q}^{(d)}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) + \mathcal{Q}^{(sm)}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$$

where,

$$\mathcal{Q}^*(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = K + \frac{1}{2}\sum_{m=1}^{M} \beta_m^* \left\{ \log|\boldsymbol{P}_m| - \sum_{i=1}^{n} \lambda_{ii}^{(m)} \,\mathrm{Tr}(\boldsymbol{S}_i \boldsymbol{W}_m) \right\}$$

- Component specific smoothing function weights, D_m to ensure convexity of auxiliary function

Projected Statistics

- Accumulation of full covariance statistics, $\boldsymbol{W}_m$
  - impractical for LVCSR;
  - only required to initialise and update basis vectors/matrices
- Update of basis coefficients alone requires only the *projected* statistics, $\tilde{w}_i$, $\forall i = \{1, 2, \ldots, n\}$:
  - STC/EMLLT:
$$\tilde{w}_i = \boldsymbol{a}_i \boldsymbol{W}_m \boldsymbol{a}_i'$$
  - SPAM:
$$\tilde{w}_i = \mathrm{Tr}\left(\boldsymbol{S}_i \boldsymbol{W}_m\right) = \sum_{r=1}^{R} \left(\lambda_{ii}^{(r)} \boldsymbol{a}_{ir} \boldsymbol{W}_m \boldsymbol{a}_{ir}'\right)$$
- $\tilde{w}_i$ is a scalar term for each basis, $\boldsymbol{a}_i$ or $\boldsymbol{S}_i$
- MPE training: only update basis coefficients

Implementation Issues

- **Variance flooring**
  - Variance floor – a technique to ensure training robustness.
  - Computationally expensive for structured precision matrix models
  - Apply variance floor on *full covariance statistics*
  - Variance flooring on *projected* statistics possible for STC and EMLLT
  - Non-trivial for SPAM models

- **Determining smoothing constant, D_m, for MPE**
  - D_m is required to ensure convexity of auxiliary function
  - A Quadratic Eigenvalue Problem (QEP)
  - Requires full covariance statistics
  - For STC/EMLLT, possible to solve independent quadratic equations with *projected* statistics
  - For projected statistics with SPAM, use *pseudo* projections:
    - Another set of projected statistics associated with rank-1 projections
    - Examples: identity matrix or STC transforms

Experimental Setup

- Unadapted results
- Conversational telephone speech – English (CTS):
  - Training dataset: `h5etrain03` (296hr)
  - Test dataset: `dev01sub` (3hr) & `eval03` (6hr)
  - CMN, CVN and VTLN are used
  - Basis vectors/matrices: ML trained
- Broadcast News – English (BN):
  - Training dataset: `bnac` (144hr)
  - Test dataset: `dev03` (3hr) & `eval03` (3hr)
- System configurations
  - Front-end: PLP with log energy + 1st, 2nd & 3rd derivatives
  - Approx. 7000 distinct states
  - 16 components and 28 components
  - Trigram language model

Initial Results – CTS

System	# of xforms	Dimension		WER (%)	
		μ	Σ	ML	MPE
HLDA	1	39	39	33.5	29.8
STC		52	52	33.3	29.7
HLDA-PMM		52	39	33.2	29.4
EMLLT		52	78	32.6	29.2
EMLLT	64	52	78	32.0	*28.3*

- 16-comp models trained on `h5etrain03`; evaluated on `dev01sub`
- Modelling mean vectors in 52 dim space gave slight improvement
- HLDA-PMM is 0.3% better than STC; less parameters for HLDA-PMM
- Single-transform EMLLT yields 0.6% absolute WER reduction
- EMLLT with 64 transforms gave 1.5% improvement over HLDA

28 component systems – CTS

- Selected systems for evaluation
  - 28-comp HLDA
  - 16-comp 64-transform EMLLT
  - 28-comp single-transform SPAM

System	# of comps	# of xforms	Dimensions		dev01sub		eval03	
			μ	Σ	ML	MPE	ML	MPE
HLDA	28	1	39	39	32.3	29.1	31.7	28.4
EMLLT	16	64	52	78	32.0	28.3	31.7	28.1
SPAM	28	1	52	39	31.5	28.3	30.8	*27.6*

- SPAM gave 0.8% absolute WER reduction

- 64-transform EMLLT is only 0.3% better than the baseline on `eval03`

Broadcast News English Systems

- Selected 16-comp systems for evaluation on `dev03` and `eval03`
  - SPAM
  - HLDA+SPAM (SPAM within HLDA subspace)

System	`dev03` WER (%)			`eval03` WER (%)		
	ML	MPE	MPE-MAP	ML	MPE	MPE-MAP
HLDA	17.7	15.2	14.9	15.6	13.7	13.6
SPAM	17.0	15.1	–	15.4	13.7	–
HLDA+SPAM	16.9	14.9	*14.6*	15.1	13.4	*13.4*

- SPAM did not yield any gain after MPE training
- MPE HLDA+SPAM is 0.3% better than HLDA, on both `dev03` and `eval03`
- For MPE-MAP, HLDA+SPAM gave 0.3%(`dev03`) and 0.2% (`eval03`)

Summary

- Precision matrix modelling used in LVCSR;
- Successful discriminative MPE training;
- Best model was found to be:
  - *SPAM* for CTS;
  - *HLDA+SPAM* for Broadcast News.
- Candidate system combination branch for BN and CTS;
- Gains retained after MLLR adaptation;
- Further investigations:
  - HLDA+SPAM model for CTS;
  - Dynamic MMI prior for SPAM;
  - SPAM model training using 400h Fisher data.set