Product of Gaussians & Multiple Stream Systems

Sarah Airey & Mark Gales

Cambridge University Engineering Department

{ssa26, mjfg}@eng.cam.ac.uk

INTRODUCTION

- Product of Experts for HMM based speech recognition
  - alternative to standard mixture modelling
  - requires fewer parameters than Mixtures of Experts
  - relates to multiple stream systems

PRODUCT VS MIXTURE OF EXPERTS

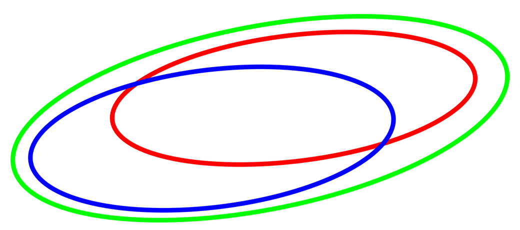
- **Mixture** of S experts $\mathcal{M}^{(1)}, \dots, \mathcal{M}^{(s)}$ has output likelihood:

$$p(\mathbf{o}_t|\mathcal{M}) = \sum_{s=1}^{S} c^{(s)} p(\mathbf{o}_t|\mathcal{M}^{(s)})$$

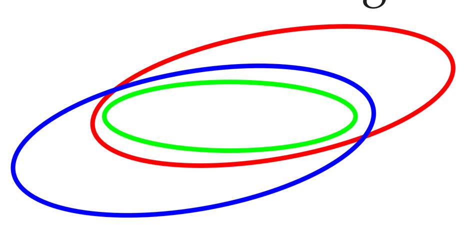
  - $\sum_{s=1}^{S} c^{(s)} = 1$ where $c^{(s)}$ is the prior for expert $\mathcal{M}^{(s)}$
  - Training is straightforward using EM techniques
- **Product** has output likelihood:

$$p(\mathbf{o}_t|\mathcal{M}) = \frac{1}{Z}\prod_{s=1}^{S} p(\mathbf{o}_t|\mathcal{M}^{(s)})$$

  - Normalisation term $Z = \int_{\mathcal{R}^d} \prod_{s=1}^{S} p(\mathbf{o}|\mathcal{M}^{(s)}) d\mathbf{o}$
  - Training is complex; usually normalisation term must be approximated
- Product gives the **intersection** of experts, whereas Mixture gives the **union**



Mixture of Gaussians



Product of Gaussians

PRODUCT OF GAUSSIANS

- Each expert is a Gaussian Mixture Model (GMM)

$$p(\mathbf{o}_t|\mathcal{M}) = \frac{1}{Z}\prod_{s=1}^{S}\left(\sum_{m=1}^{M^{(s)}} c_m^{(s)} \mathcal{N}\left(\mathbf{o}_t; \boldsymbol{\mu}_m^{(s)}, \boldsymbol{\Sigma}_m^{(s)}\right)\right)$$

  - $M^{(s)}$ is the number of components in GMM s
  - $c_m^{(s)}$, $\boldsymbol{\mu}_m^{(s)}$, $\boldsymbol{\Sigma}_m^{(s)}$ are the prior, mean, and variance of component m of GMM s
- Expand product of sums into sum of products:

$$p(\mathbf{o}_t|\mathcal{M}) = \frac{1}{Z}\sum_{m_1=1}^{M^{(1)}} \dots \sum_{m_S=1}^{M^{(S)}} \prod_{s=1}^{S} c_{m_s}^{(s)} \mathcal{N}\left(\mathbf{o}_t; \boldsymbol{\mu}_{m_s}^{(s)}, \boldsymbol{\Sigma}_{m_s}^{(s)}\right)$$

- Product of Gaussians is also Gaussian distributed:

$$p(\mathbf{o}_t|\mathcal{M}) = \frac{1}{Z}\sum_{m_1=1}^{M^{(1)}} \dots \sum_{m_S=1}^{M^{(S)}} c_{\mathbf{m}} K_{\mathbf{m}} \mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_{\mathbf{m}}, \boldsymbol{\Sigma}_{\mathbf{m}})$$

  - a **meta-component** $\mathbf{m}$ is the product of one component from each expert
  - meta-component normalisation term $K_{\mathbf{m}}$ is a function of $\boldsymbol{\mu}_{\mathbf{m}}$,$\boldsymbol{\Sigma}_{\mathbf{m}}$

$$\boldsymbol{\Sigma}_{\mathbf{m}} = \left(\sum_{s=1}^{S} \boldsymbol{\Sigma}_{m_s}^{(s)-1}\right)^{-1}, \quad \boldsymbol{\mu}_{\mathbf{m}} = \boldsymbol{\Sigma}_{\mathbf{m}}\left(\sum_{s=1}^{S} \boldsymbol{\Sigma}_{m_s}^{(s)-1} \boldsymbol{\mu}_{m_s}^{(s)}\right)$$

- Product means and variances can be trained using generalised EM
- **Normalisation** options:
  1. At the **state level**

$$Z = \sum_{m_1=1}^{M^{(1)}} \dots \sum_{m_S=1}^{M^{(S)}} c_{\mathbf{m}} K_{\mathbf{m}}$$

  2. At the **meta-component level**

$$p(\mathbf{o}_t|\mathcal{M}) = \sum_{m_1=1}^{M^{(1)}} \dots \sum_{m_S=1}^{M^{(S)}} c_{\mathbf{m}} \mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_{\mathbf{m}}, \boldsymbol{\Sigma}_{\mathbf{m}})$$

MULTIPLE STREAM SYSTEMS

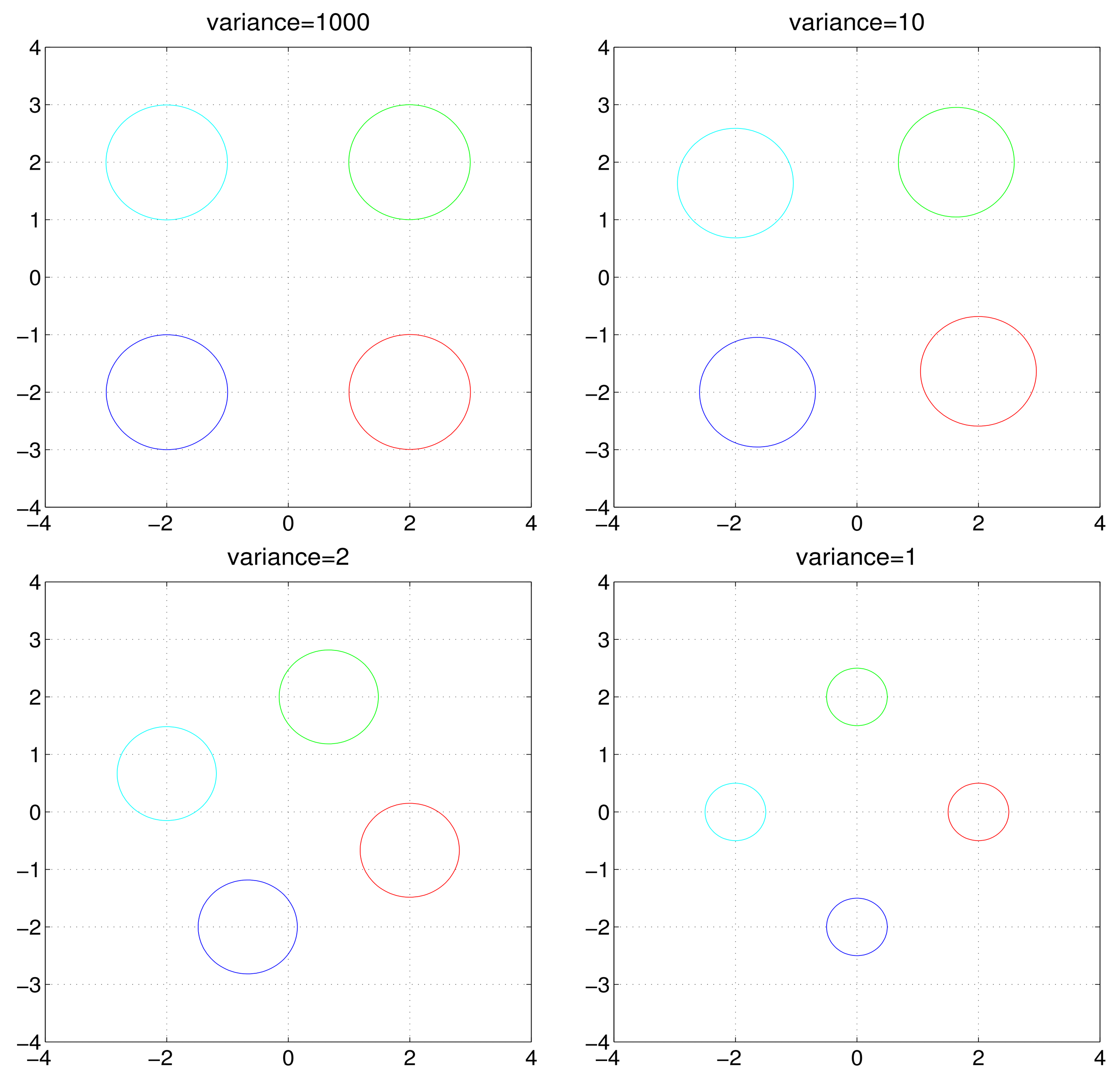
- A special case of Product of Gaussians (PoG)

$$p(\mathbf{o}_t|q_t) = \prod_{s=1}^{S}\left(\sum_{m=1}^{M^{(s)}} c_m^{(s)} \mathcal{N}\left(\mathbf{o}_t^{(s)}; \boldsymbol{\mu}_m^{(s)}, \boldsymbol{\Sigma}_m^{(s)}\right)\right)$$

- PoG is Multiple Stream when off-stream variances are large and means are 0
- Consider 2-expert PoG where off-stream variance σ^2 approaches ∞

$$\boldsymbol{\Sigma}_1^{(1)} = \begin{bmatrix} \boldsymbol{\Sigma}^{(1)} & \mathbf{0} \\ \mathbf{0} & \sigma^2\boldsymbol{I} \end{bmatrix} \times \boldsymbol{\Sigma}_1^{(2)} = \begin{bmatrix} \sigma^2\boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}^{(2)} \end{bmatrix} \implies \boldsymbol{\Sigma}_{[11]'} = \begin{bmatrix} \boldsymbol{\Sigma}^{(1)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}^{(2)} \end{bmatrix}$$

- Products of two GMM experts shift as σ^2 increases



- Initialise PoG from Multiple Stream by padding off-stream elements

DISTRIBUTED REPRESENTATIONS

- Only final system must satisfy:

$$p(\mathbf{o}|\mathcal{M}) \geq 0; \quad \int_{\mathcal{R}^d} p(\mathbf{o}|\mathcal{M}) d\mathbf{o} = 1$$

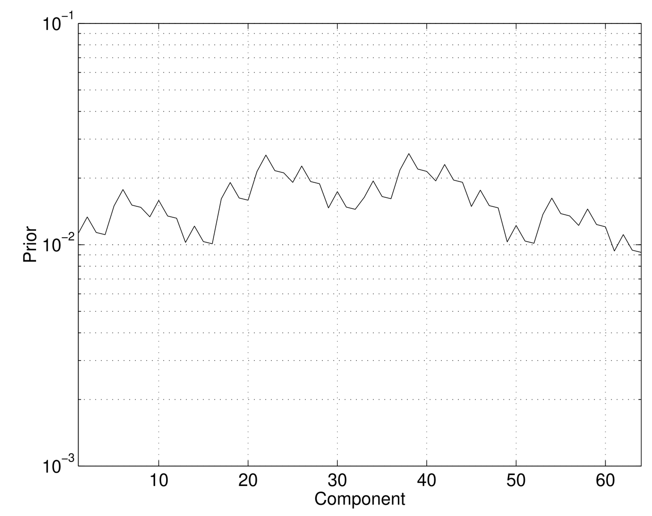
- Relaxes standard constraints for individual expert components on:
  1. **Covariances**: only meta-component covariances must be positive definite
     - Allows **negative variances** on leading diagonal
  2. **Priors**: only meta-component priors must satisfy $\sum_{\mathbf{m}} c_{\mathbf{m}} = 1$, $c_{\mathbf{m}} \geq 0$
     - Allows ML estimated priors based on meta-component likelihoods

SWITCHBOARD EXPERIMENTS

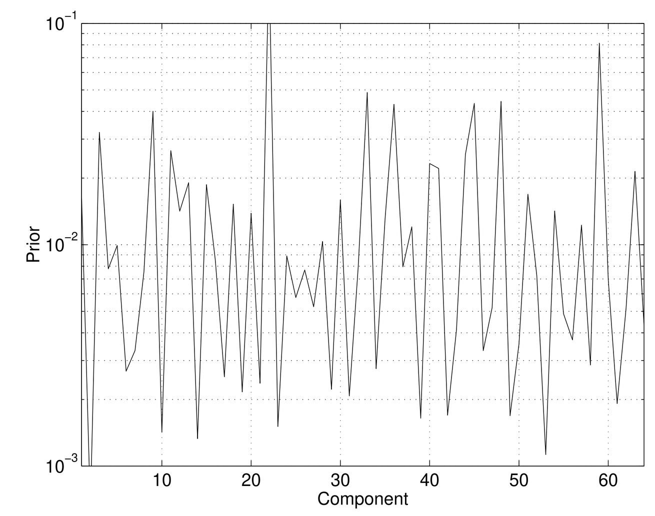
- Baseline setup:
  - 68 hour training data set (h5train00sub)
  - 39 dimensional feature vector (MF-PLP+A+D)
  - VTLN + cepstal mean substraction and variance normalisation
  - gender-independent cross-word triphone diagonal covariance tied-state
  - 3 hour test data subset of 2001 development data (dev01sub)
  - trigram language model
- Standard GMM System compared to 3-stream Multiple Stream System

HMM System	Number of Components					
	2	4	6	8	10	12
Standard GMM	46.1	43.7	42.0	40.7	39.3	39.1
Multiple Stream	46.0	43.8	43.1	42.4	42.1	42.0

  - GMM significantly outperforms MS, especially as components increase
- Using **ML estimated priors** improved Multiple Stream system performance



Producted Priors



ML Estimate Priors

  - Reduced WER from 43.8% to 43.0% for 4-comp Mulitple Stream system
  - All PoGs implemented with ML priors after preliminary experiments
- Allowing **negative variances** improved performance
  - Reduced WER from 44.4% to 43.1% for 2-component PoG system
- 3-stream Product of Gaussian System (ML priors & negative variances)

Number of Comps/Stream	Number of Meta–comps	Word Error Rate (%)
2	8	43.1
3	27	41.3
4	64	40.9

  - PoG outperforms GMMs and MS with the same number of components
  - But PoG requires more parameters

CONCLUSIONS

- Best PoG performs worse than best Mixture Model
- Implement larger PoG systems
- Extend to Products of GMMs and HMMs