Variational Inference for Bayesian Mixtures of Factor Analysers

Zoubin Ghahramani & **Matthew J. Beal**

Gatsby Computational Neuroscience Unit
University College London

{zoubin,m.beal}@gatsby.ucl.ac.uk
http://www.gatsby.ucl.ac.uk

Abstract

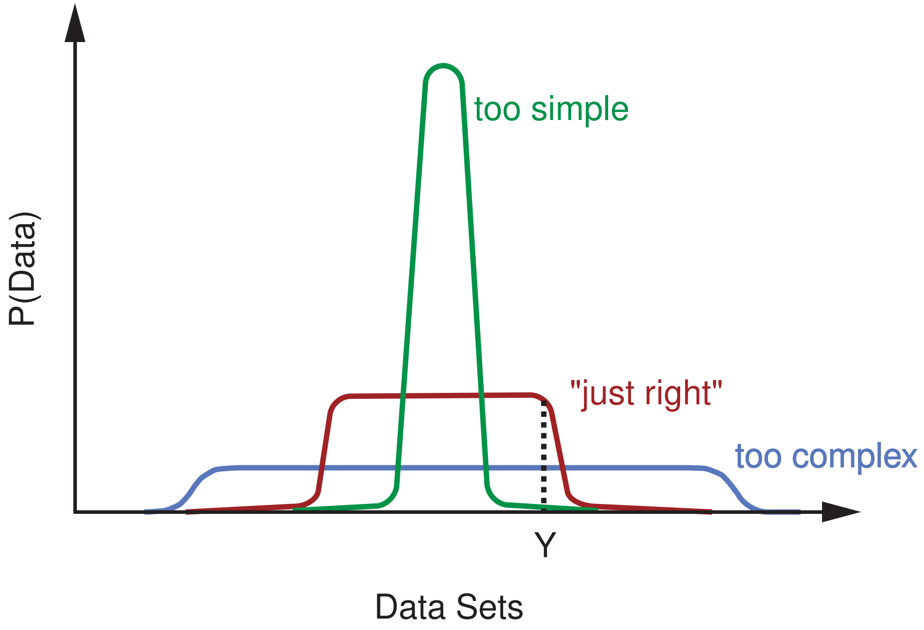
We present an algorithm that infers the model structure of a mixture of factor analysers using an efficient and deterministic variational approximation to full Bayesian integration over model parameters. This procedure can automatically determine the optimal number of components and the local dimensionality of each component (i.e. the number of factors in each factor analyser). Alternatively it can be used to infer posterior distributions over number of components and dimensionalities. Since all parameters are integrated out the method is not prone to overfitting. Using a stochastic procedure for adding components it is possible to perform the variational optimisation incrementally and to avoid local maxima.

Results show that the method works very well in practice and correctly infers the number and dimensionality of nontrivial synthetic examples. We also show how it is possible to sample from the variational approximation to estimate the true evidence, the KL divergence between the variational posterior and the true posterior, and the exact predictive distribution.

Key ideas

- Bayesian learning avoids overfitting and can be used for model selection.
- Using variational methods, Bayesian learning and model selection can be done tractably.
- An example: Mixture of Factor Analysers to discover number of clusters and their intrinsic dimensionalities.
- Importance sampling from variational approximation to estimate:
  - true Evidence,
  - the KL divergence between the variational posterior and the true posterior,
  - the exact predictive density.
- Simulations:
  - three synthetic problems,
  - unsupervised digit classification.

Bayesian Learning & Model Selection



(adapted from D.J.C. MacKay)

- Maximum Likelihood (**ML**) methods do not take into account model complexity.
- Complex models, which can *a priori* model a larger range of data sets, are not penalized - this leads to overfitting.
- Bayesian approach naturally accounts for model complexity.
- Key quantity: the marginal likelihood, or evidence, obtained by averaging over model parameters:

$$\textbf{Evidence}: \quad P(Y) = \int d\theta \; P(Y|\theta)P(\theta).$$

Practical Bayesian approaches

Unfortunately the evidence integral is in general analytically and computationally intractable, and approximations need to be made:

- Laplace approximations:
  - Appeals to Central Limit Theorem making a
  - Gaussian approximation about maximum *a posteriori* parameter estimate.
- Large sample approximations (e.g. BIC).
- Markov chain Monte Carlo methods (MCMC):
  - In the limit of many samples is guaranteed to converge, but
  - Many samples required to ensure accuracy.
  - Hard to assess convergence.
- Variational approximations...

Variational Bayesian Learning

Let θ denote model parameters, and X denote hidden variables. We can lower bound the log evidence using Jensen's inequality twice

$$
\begin{aligned}
\mathcal{L} &\equiv \ln P(Y) \\
&= \ln \int d\theta \; P(Y,\theta) \\
&\geq \int d\theta \; Q(\theta) \ln \frac{P(Y,\theta)}{Q(\theta)} \\
&= \int d\theta \; Q(\theta) \left[\ln P(Y|\theta) + \ln \frac{P(\theta)}{Q(\theta)} \right] \\
&\geq \int d\theta \; Q(\theta) \left[\int dX \; Q(X) \ln \frac{P(X,Y|\theta)}{Q(X)} + \ln \frac{P(\theta)}{Q(\theta)} \right] \\
&\equiv \mathcal{F}(Q(\theta), Q(X))
\end{aligned}
$$

Assumes the factorisation: $P(\theta,X|Y) \approx Q(\theta)Q(X)$.

This has also been referred to as "ensemble learning".

Advantages of Variational method

EM-like optimisation:

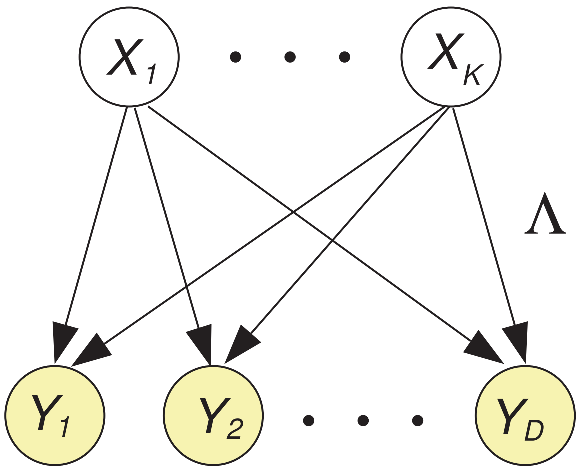
"E-step": Maximise $\mathcal{F}$ w.r.t. $Q(X)$ with $Q(\theta)$ fixed,

"M-step": Maximise $\mathcal{F}$ w.r.t. $Q(\theta)$ with $Q(X)$ fixed.

Finds an approximation to the posterior over parameters $Q(\theta) \approx P(\theta|Y)$ and hidden variables $Q(X) \approx P(X|Y)$

- Maximises a lower bound on the log evidence.
- Convergence can be assessed by monitoring $\mathcal{F}$.
- Global approximation.
- $\mathcal{F}$ transparently incorporates model complexity penalty (i.e. coding cost for all the parameters of the model) so it can be compared across models.
- Forms of $Q(\cdot)$ fall out of free-form extremisation of $\mathcal{F}$ (i.e. not assumed to be Gaussian).
- Often simple modification of the EM algorithm.

Factor Analysis



Linear generative model:

$$y_d = \sum_{k=1}^{K} \Lambda_{dk} \, x_k + \epsilon_d$$

- x_k are independent $\mathcal{N}(0,1)$ Gaussian factors,
- ϵ_d are independent $\mathcal{N}(0, \Psi_{dd})$ Gaussian noise,
- $K < D$.

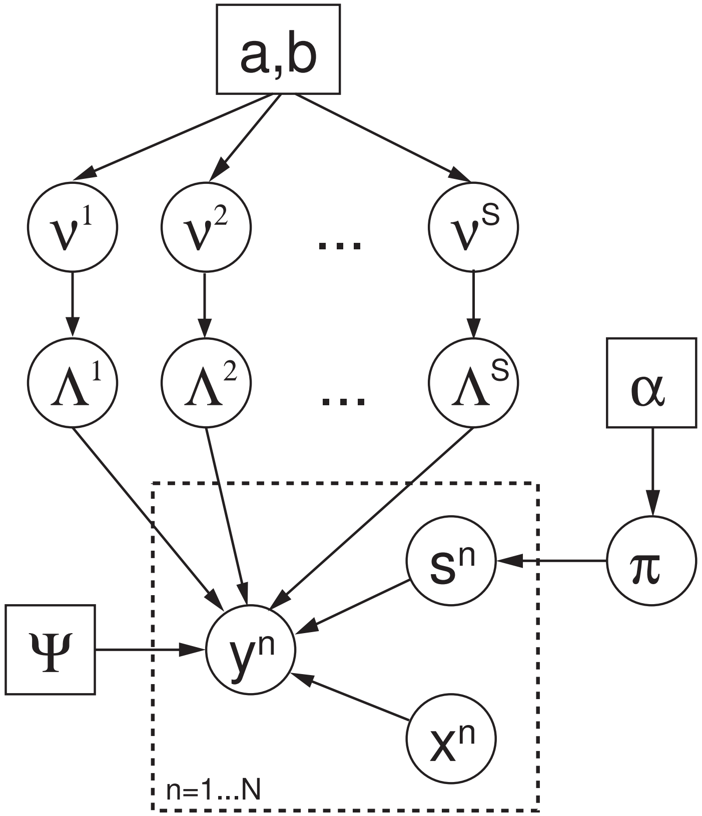
So, $\mathbf{y}$ is Gaussian with:

$$P(\mathbf{y}) = \int P(\mathbf{x}) P(\mathbf{y}|\mathbf{x}) d\mathbf{x} = \mathcal{N}(0, \Lambda\Lambda^\top + \Psi)$$

where Λ is a $D \times K$ matrix, and Ψ is diagonal.

Dimensionality Reduction: Finds a low-dimensional projection of high dimensional data that captures most of the correlation structure of the data.

Mixture of Factor Analysers



Squares represent hyperparameters, circles represent random variables.

a, b govern shape and precision of ν distributions.
ν^s governs size of Λ^s entries (ARD).
α controls spread of $\boldsymbol{\pi}$ mixing proportions.
Ψ is the independent measurement noise.

$$\boldsymbol{\pi} \sim Dirichlet,\ \nu \sim Gamma,\ \Lambda \sim Gaussian,$$
$$s^n \sim Multinomial,\ \mathbf{x}^n \sim Gaussian.$$

Generative model:

$$P(Y) = \int d\boldsymbol{\pi} P(\boldsymbol{\pi}|\alpha) \int d\boldsymbol{\nu} P(\boldsymbol{\nu}|a,b) \int d\Lambda\, P(\Lambda|\boldsymbol{\nu}) \quad \times$$
$$\prod_{n=1}^{N} \left[\sum_{s^n=1}^{S} P(s^n|\boldsymbol{\pi}) \int d\mathbf{x}^n P(\mathbf{x}^n) P(\mathbf{y}^n|\mathbf{x}^n, s^n, \Lambda^s, \Psi) \right].$$

Objective function $\mathcal{F}$

Making use of the approximating factorization

$$P(\mathbf{x}, s, \boldsymbol{\pi}, \boldsymbol{\nu}, \Lambda | Y) \approx Q(\mathbf{x}, s) Q(\boldsymbol{\pi}, \boldsymbol{\nu}) Q(\Lambda)$$

and using Jensen's inequality,

$$\begin{aligned}\mathcal{F} = & \int d\boldsymbol{\pi} Q(\boldsymbol{\pi}) \ln \frac{P(\boldsymbol{\pi}|\alpha)}{Q(\boldsymbol{\pi})} + \sum_{s=1}^{S} \int d\boldsymbol{\nu}^s Q(\boldsymbol{\nu}^s) \left[\ln \frac{P(\boldsymbol{\nu}^s|a,b)}{Q(\boldsymbol{\nu}^s)} + \int d\Lambda^s Q(\Lambda^s) \ln \frac{P(\Lambda^s|\boldsymbol{\nu}^s)}{Q(\Lambda^s)} \right] \\ & + \sum_{n=1}^{N} \sum_{s^n=1}^{S} Q(s^n) \left[\int d\boldsymbol{\pi}\, Q(\boldsymbol{\pi}) \ln \frac{P(s^n|\boldsymbol{\pi})}{Q(s^n)} + \int d\mathbf{x}^n Q(\mathbf{x}^n|s^n) \ln \frac{P(\mathbf{x}^n)}{Q(\mathbf{x}^n|s^n)} \right. \\ & \left. + \int d\Lambda^s Q(\Lambda^s) \int d\mathbf{x}^n Q(\mathbf{x}^n|s^n) \ln P(\mathbf{y}^n|\mathbf{x}^n, s^n, \Lambda^s, \Psi) \right]\end{aligned}$$

- Further factorizations then fall out from the math, and are exact.
- Conjugate priors $\rightarrow$ Conjugate Q distributions.

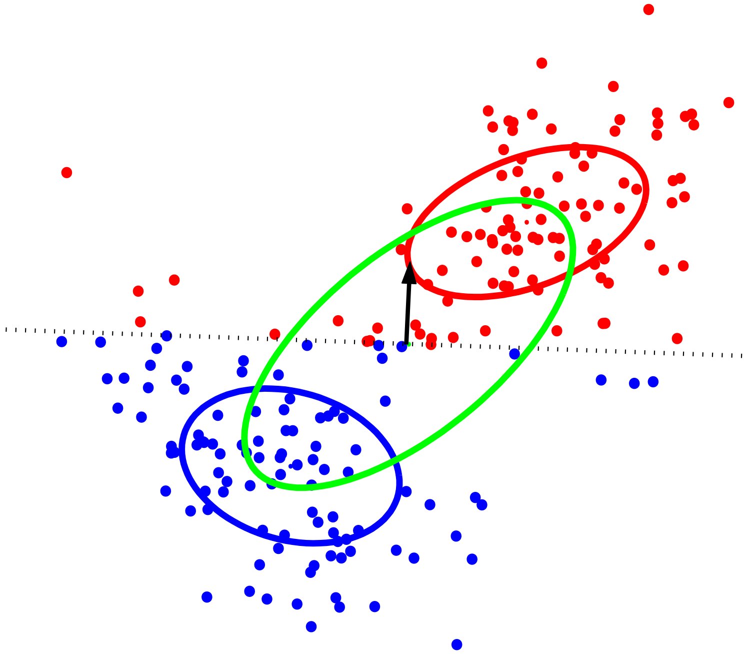
Birth & Death

Death

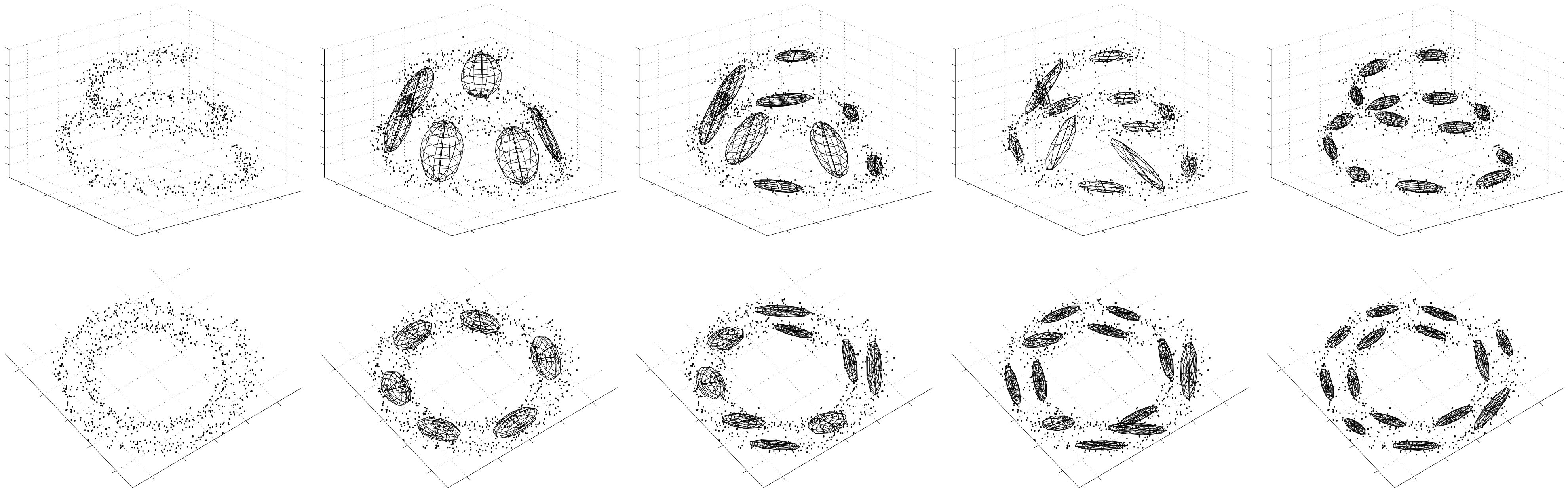
- A component will *die* of natural causes if the local data does not support it:
- It is better to describe the data less accurately with a different component, than waste bits defining the parameters of a more accurate component.

Birth

- "Mitosis": split a component into two.
- This usually causes a *drop* in $\mathcal{F}$ due to complexity, so this birth is accepted only if $\mathcal{F}$ recovers.
- Thus the algorithm can incrementally explore model complexity.
- Intelligent heuristics for:
  - Picking a component to split,
  - The splitting process.



Synthetic problem: Ueda Spiral

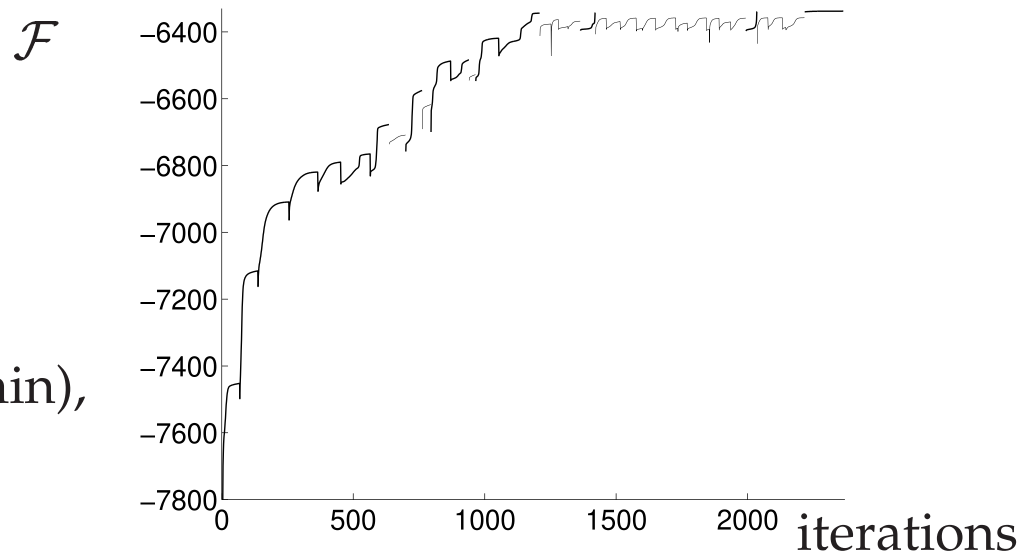


Problems:

- Unknown number of clusters,
- Easily caught in local optima.

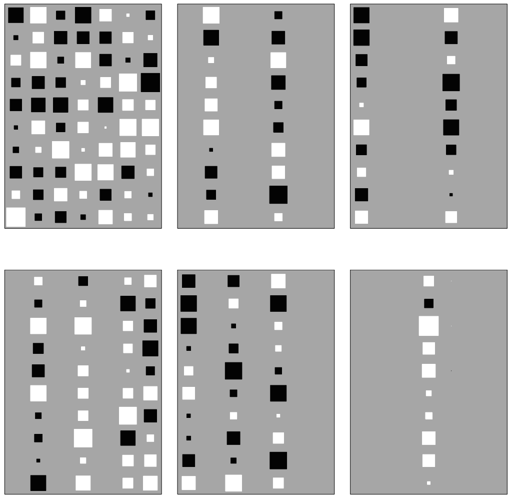
Results:

- Searches over model complexity,
- Algorithm is fast and efficient (~ 1 min),
- Escapes local optima.



Synthetic problem: high dimensional data

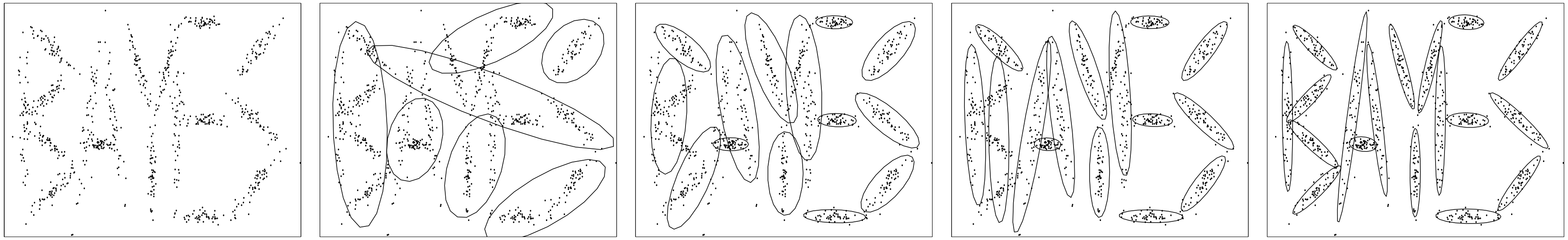
- Six Gaussian clusters defined by n data points within each.
- Place these in a 10-dimensional embedding space.



number of points per cluster	intrinsic dimensionalities 1	7	4	3	2	2
8	2				1	
8	1		2			
16	1	4				2
32	1	6	3	3	2	2
64	1	7	4	3	2	2
128	1	7	4	3	2	2

- Finds the clusters and their dimensionalities efficiently.
- Shows **Ockham's razor** — the complexity of the model reduces in line with the lack of data support.

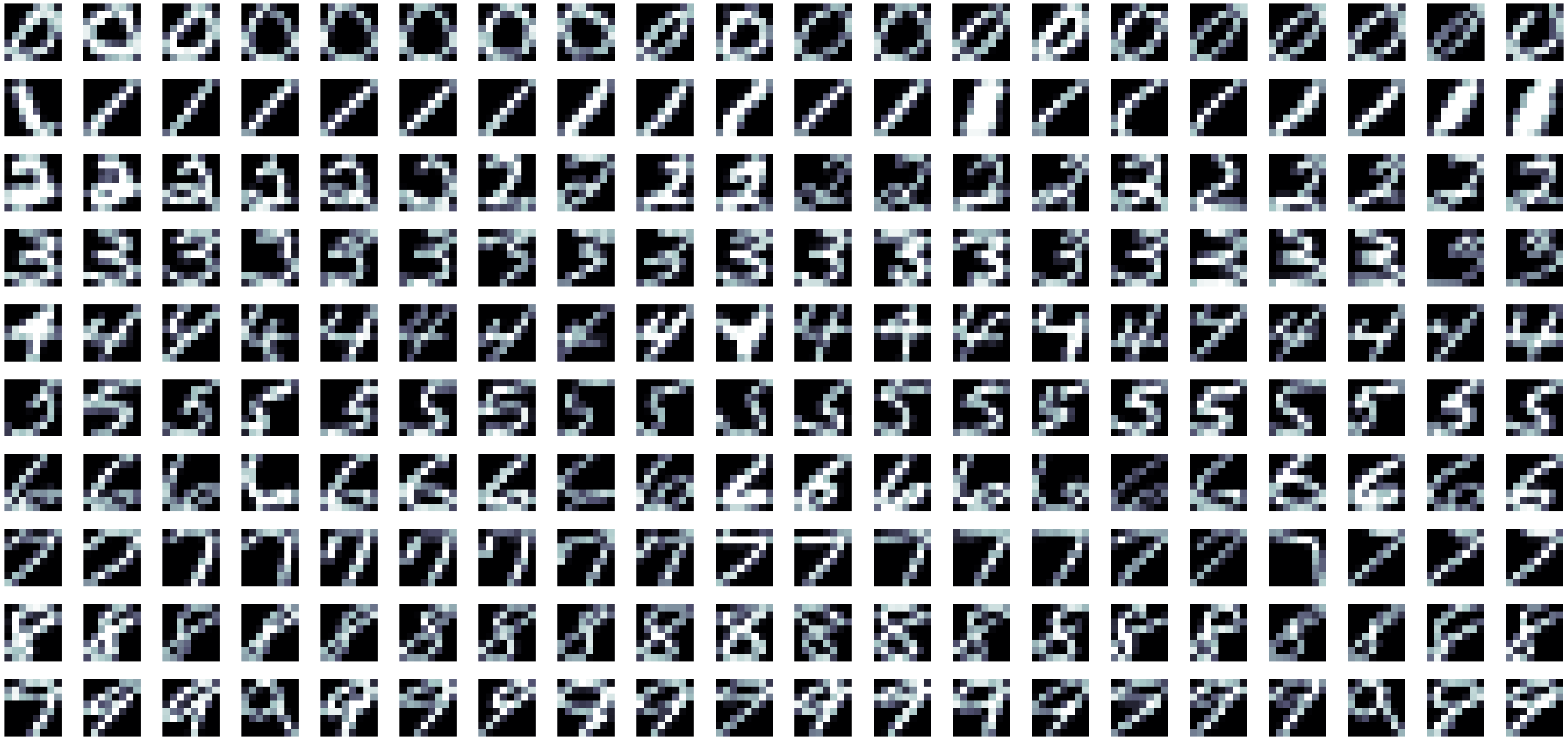
Synthetic problem: BAYES



- Again, number of clusters unknown (but obvious).

Digit Classification

Training data examples:



8x8 Digit results

- Trained on 700 8x8 images of each digit (`CEDAR ROM`).
- Determines the number of clusters (styles) required.
- Also ARD determines the number of deformations in each style.

- The number to the right of each digit image is the cluster's dimensionality.

Classified

True	0	1	2	3	4	5	6	7	8	9
0	687	7	.	2	.	1	3	.	.	.
1	.	699	.	.	.	1	.	.	.	.
2	1	7	671	1	2	.	7	4	7	.
3	.	3	7	629	.	27	1	2	30	1
4	.	3	4	.	609	.	3	14	1	66
5	3	7	5	46	.	618	5	.	16	.
6	.	3	.	.	32	1	664	.	.	.
7	.	2	3	1	3	.	.	589	2	100
8	1	14	1	27	2	43	1	4	603	4
9	.	4	1	.	13	.	.	65	3	614

Training data

Classified

True	0	1	2	3	4	5	6	7	8	9
0	196	1	.	.	.	1	2	.	.	.
1	.	200	.	.	.	.	.	.	.	.
2	.	1	186	1	1	1	2	1	6	1
3	.	2	1	181	.	10	.	.	5	1
4	.	1	.	.	175	.	.	1	.	23
5	1	2	.	13	.	180	.	1	2	1
6	.	1	.	.	5	1	193	.	.	.
7	.	2	1	.	.	.	.	176	.	21
8	.	1	.	5	.	9	.	1	179	5
9	.	4	.	.	3	.	.	17	.	176

Test data

- Each image is classified using hard assignment from most likely cluster.
- Unsupervised classification achieves 8.8% training and 7.9% test error.
- K-means with same number of clusters achieves 12.2% and 13.3% error.

Importance Sampling from the Variational Approximation

Sampling $\theta_m \sim Q(\theta)$ gives us unbiased estimates of:

- The Exact Predictive Density:

$$P(y|Y) = \int d\theta\; P(y|\theta)P(\theta|Y) = \int d\theta\; Q(\theta)P(y|\theta)\frac{P(\theta|Y)}{Q(\theta)}$$
$$\approx \sum_{m=1}^{M} P(y|\theta_m)\,\omega_m$$

  where the importance weights are $\omega_m = \frac{1}{\Omega}\frac{P(\theta_m,Y)}{Q(\theta_m)}$, where Ω s.t. $\sum \omega_{\mathrm{m}} = 1$.

- The True Log Evidence:

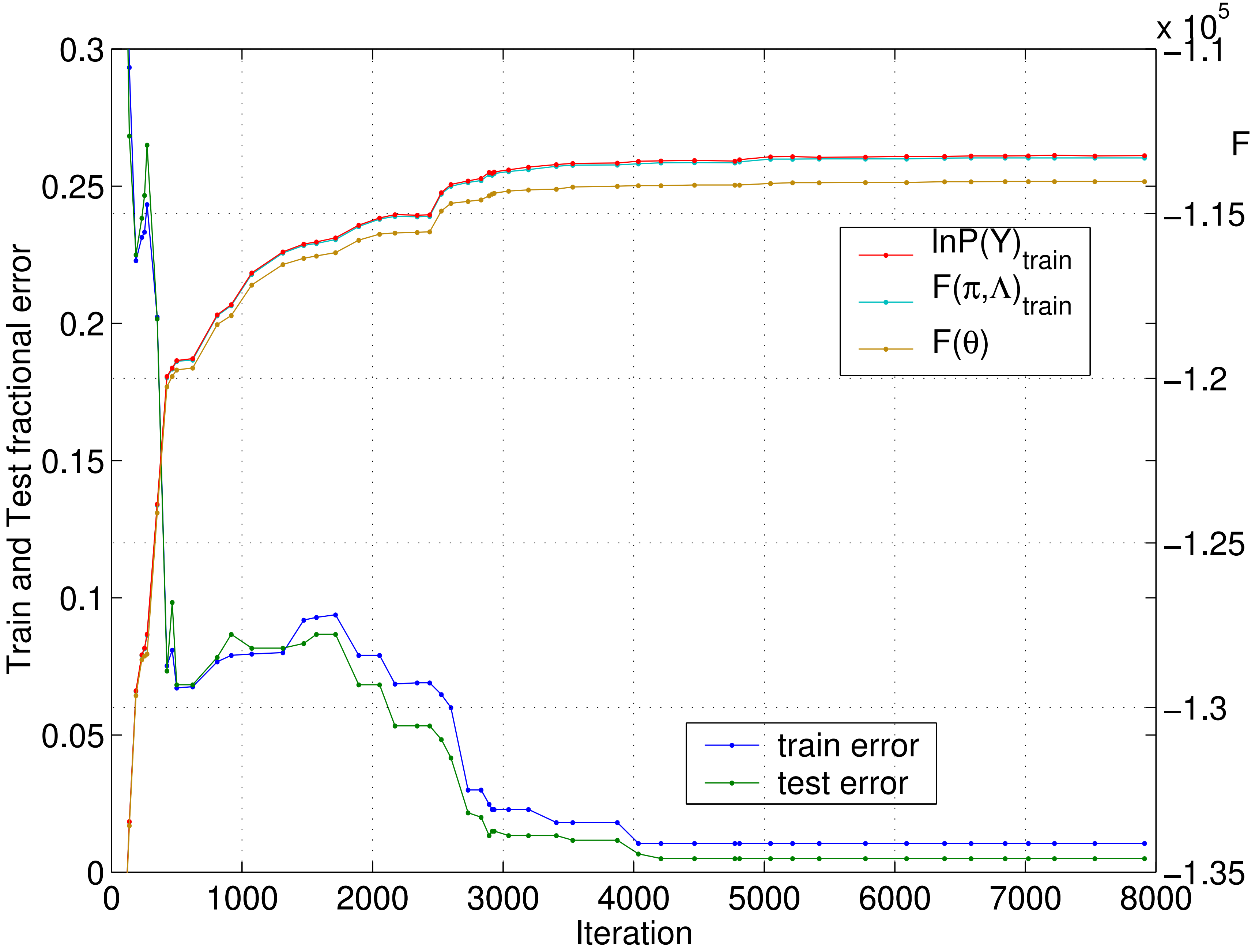
$$\ln P(Y) = \ln \int d\theta\; Q(\theta)\frac{P(\theta,Y)}{Q(\theta)} = \ln\langle\omega\rangle + \ln\Omega.$$

- The KL Divergence:

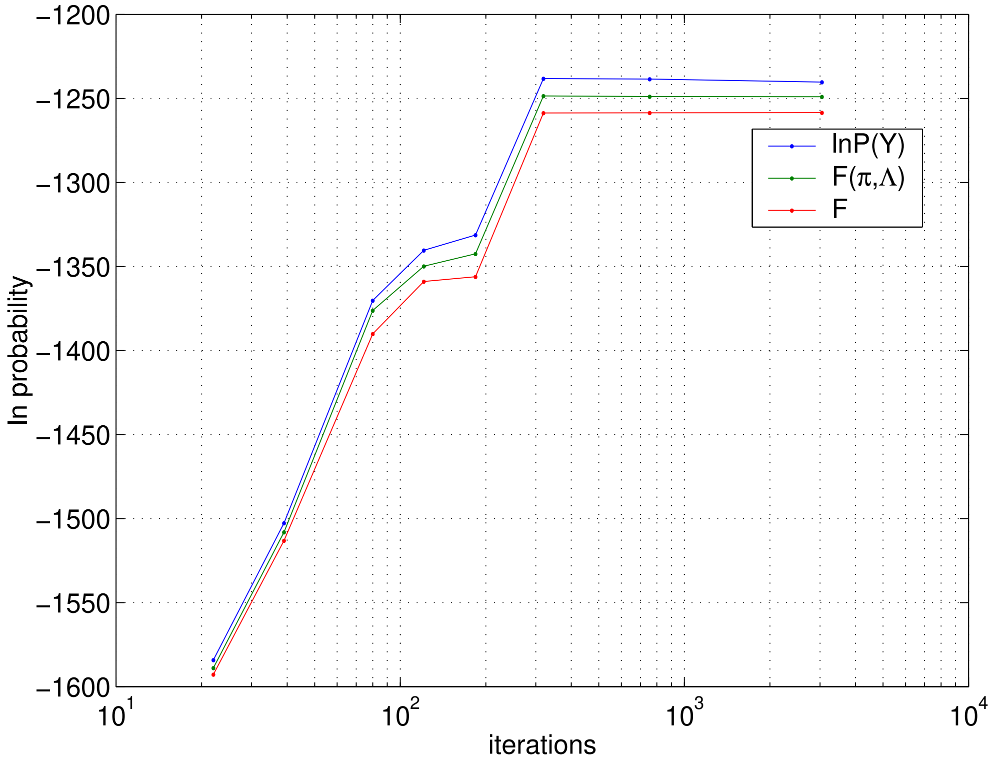
$$\mathrm{KL}(Q(\theta)\|P(\theta|Y)) = \ln\langle\omega\rangle - \langle\ln\omega\rangle.$$

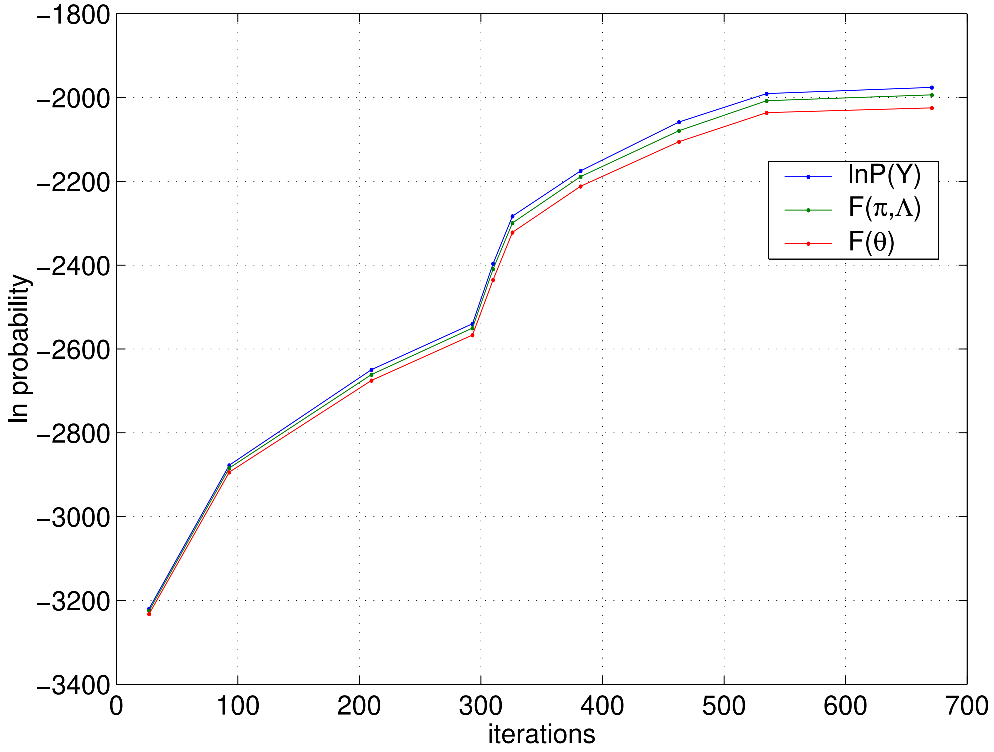
Note: same weights can be used for all three!

Evolution of $\mathcal{F}$, true evidence and KL-divergence



Sampling results for other classes of problem





References

[1] H. Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Proc. 15th Conf. on Uncertainty in Artificial Intelligence*, 1999.

[2] C.M. Bishop. Variational PCA. In *Proc. Ninth Int. Conf on Artificial Neural Networks ICANN99*, **1**:509 - 514, 1999.

[3] Z. Ghahramani and G. E. Hinton. The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1 [`http://www.gatsby.ucl.ac.uk/~zoubin/papers/tr-96-1.ps.gz`], Dept. of Comp. Sci., Univ. of Toronto, 1996.

[4] G. E. Hinton and D. van Camp. Keeping Neural Networks Simple by Minimizing the Description Length of the Weights. In *Sixth ACM Conference on Computational Learning Theory*, Santa Cruz, 1993.

[5] D.J.C. MacKay. Ensemble learning for hidden Markov models. Technical report, Cavendish Laboratory, University of Cambridge, 1997.

[6] R.M. Neal. Assessing relevance determination methods using DELVE. In C. M. Bishop (editor), *Neural Networks and Machine Learning*: 97-129, Springer-Verlag, 1998.

[7] C.E. Rasmussen. The countably infinite Bayesian Gaussian mixture density model. Technical report, Dept. of Math. Modelling, Tech. Univ. Denmark, 1999.

[8] S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Stat. Soc.–Ser. B*, 59(4):731–758, 1997.

[9] S.J. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to Gaussian mixture modeling. *IEEE PAMI*, 20(11):1133–1142, 1998.

[10] S. T. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345, 1999.

[11] M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. *Neural Computation*, 11(2):443–482, 1999.

[12] N. Ueda, R. Nakano, Z. Ghahramani, and G.E. Hinton. SMEM algorithm for mixture models. In *Adv. Neur. Inf. Proc. Sys. 11*. MIT Press, 1999.

[13] S. Waterhouse and D. J. C. MacKay and T. Robinson. Bayesian methods for Mixtures of Experts. In *Adv. Neur. Inf. Proc. Sys. 7*, 1995.

Appendix: Optimal $Q(\cdot)$ distributions under $\mathcal{F}$

$$Q(\mathbf{x}^n|s^n) \sim \mathcal{N}(\overline{\mathbf{x}}^{n,s}, \Sigma^s): \quad \Sigma^{s\,-1} = \langle \Lambda^{sT}\Psi^{-1}\Lambda^s\rangle_{Q(\Lambda^s)} + I, \quad \overline{\mathbf{x}}^{n,s} = \Sigma^s\overline{\Lambda}^{sT}\Psi^{-1}\mathbf{y}^n$$

$$Q(\Lambda_q^s) \sim \mathcal{N}(\overline{\Lambda}_q^s, \Sigma^{q,s}): \quad \Sigma^{q,s\,-1} = \Psi_{qq}^{-1}\sum_{n=1}^N Q(s^n)\langle \mathbf{x}^n\mathbf{x}^{nT}\rangle_{Q(\mathbf{x}^n|s^n)} + \mathrm{diag}\langle \boldsymbol{\nu}^s\rangle_{Q(\boldsymbol{\nu}^s)}$$

$$\overline{\Lambda}_q^s = \left[\Psi^{-1}\sum_{n=1}^N Q(s^n)\mathbf{y}^n\overline{\mathbf{x}}^{n,sT}\Sigma^{q,s}\right]_q$$

$$Q(\boldsymbol{\nu}_l^s) \sim \mathrm{Gamma}(a_l^s, b_l^s): \quad a_l^s = a + \frac{p}{2}, \quad b_l^s = b + \frac{1}{2}\sum_{q=1}^p \langle {\Lambda_{ql}^s}^2\rangle_{Q(\Lambda^s)}$$

$$Q(\boldsymbol{\pi}) \sim \mathrm{Dirichlet}(\omega\mathbf{u}): \quad \omega u_s = \frac{\alpha}{S} + \sum_{n=1}^N Q(s^n)$$

$$\ln Q(s^n) = [\psi(\omega u_s) - \psi(\omega)] + \frac{1}{2}\ln|\Sigma^s| + \langle \ln P(\mathbf{y}^n|\mathbf{x}^n, s^n, \Lambda^s, \Psi)\rangle_{Q(\mathbf{x}^n|s^n), Q(\Lambda^s)} + \mathrm{c}$$

where $\psi(x)$ is the *digamma* function, defined as $\psi(x) = \frac{\partial}{\partial x}\ln\Gamma(x)$.

Note that the optimal distributions $Q(\Lambda^s)$ have block diagonal covariance structure. So even though each Λ^s is a $p \times q$ matrix, its covariance only has $O(pq^2)$ parameters.

Differentiating $\mathcal{F}$ with respect to the parameters, a and b, of the precision prior we get fixed point equations $\psi(a) = \langle \ln\boldsymbol{\nu}\rangle + \ln b$ and $b = a/\langle\boldsymbol{\nu}\rangle$. Similarly the fixed point for the hyperparameters for the Dirichlet prior is $\psi(\alpha) - \psi(\alpha/S) + \sum[\psi(\omega u_s) - \psi(\omega)]/S = 0$.