Variational inference in the conjugate-exponential family

Matthew J. Beal

Work with Zoubin Ghahramani
Gatsby Computational Neuroscience Unit

August 2000

Abstract

Variational inference in the conjugate-exponential family

When learning models from data there is always the problem of over-fitting/generalisation performance. In a Bayesian approach this can be resolved by averaging over all possible settings of the model parameters.

However for most interesting problems, averaging over models leads to intractable integrals and so approximations are necessary. Variational methods are becoming a widespread tool to approximate inference and learning in many graphical models: they are deterministic, generally fast, and can monotonically increase an objective function which transparently incorporates model complexity cost.

I'll provide some theoretical results for the variational updates in a very general family of "conjugate-exponential" graphical models, and show how well-known algorithms (e.g. belief-propagation) can be readily incorporated in the variational updates.

Some examples to illustrate the ideas: determining the most probable number of clusters and their intrinsic latent-space dimensionality in a Mixture of Factor Analysers model; and recovering the hidden state-space dimensionality in a Linear Dynamical System model.

This is work with Zoubin Ghahramani.

Outline

- Briefly review variational Bayesian learning.
- Concentrate on conjugate-exponential models.
- Variational Expectation-Maximisation (VEM).
- Examples in Mixture of Factor Analysers (MFA) and Linear Dynamical Systems (LDS).

Bayesian approach

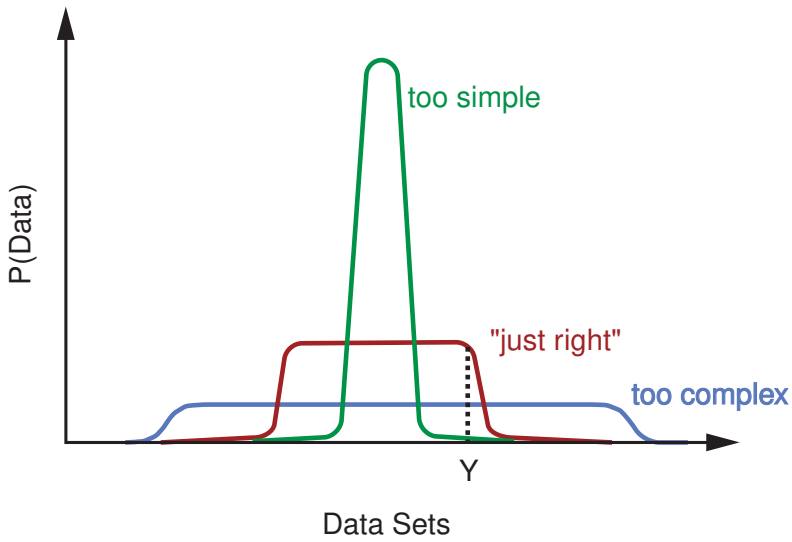
Motivation

- Maximum likelihood (ML) does not penalise complex models, which can *a priori* model a larger range of data sets.
- Bayes does not suffer from overfitting if we integrated out all the parameters.
- We should be able to do principled model comparison.

Method

- Express distributions over all parameters in the model.
- Calculate the *Evidence*

$$P(Y|\mathcal{M}) = \int d\theta \; P(Y|\theta, \mathcal{M}) P(\theta|\mathcal{M}).$$



(adapted from D.J.C. MacKay)

Problem

- These integrals are computationally intractable.

Practical approaches

- Laplace approximations:
  - Appeal to Central Limit Theorem, making a Gaussian approximation about the maximum *a posteriori* parameter estimate.
- Large sample approximations:
  - e.g. BIC.
- Markov chain Monte Carlo (MCMC):
  - Guaranteed to converge in the limit.
  - Many samples required for accurate results.
  - Hard to assess convergence.
- Variational approximations.

The variational method

Let the hidden states be X and the parameters θ,

$$\begin{aligned} \ln P(Y) &= \ln \int dX\, d\theta\; P(Y, X, \theta) \\ &\geq \int dX\, d\theta\; Q(X, \theta) \ln \frac{P(Y, X, \theta)}{Q(X, \theta)}. \end{aligned}$$

Approximate the intractable distribution over hidden states (X) and parameters (θ) with a simpler distribution.

$$\begin{aligned} \ln P(Y) &= \ln \int dX\, d\theta\; P(Y, X, \theta) \\ &\geq \int dX\, d\theta\; Q_X(X) Q_\theta(\theta) \ln \frac{P(Y, X, \theta)}{Q_X(X) Q_\theta(\theta)} \\ &= \mathcal{F}(Q, Y). \end{aligned}$$

Maximisation of this lower bound leads to **EM-like** updates:

$$\begin{aligned} \mathbf{E-step} \quad \mathbf{Q}^*_{\mathbf{X}}(\mathbf{X}) &\propto \exp \langle \ln P(X,Y|\theta) \rangle_{Q_\theta(\theta)} \\ \mathbf{M-step} \quad \mathbf{Q}^*_{\theta}(\theta) &\propto P(\theta) \exp \langle \ln P(X,Y|\theta) \rangle_{Q_X(X)} \end{aligned}$$

Equivalent to minimizing the KL-divergence between the *approximating* and *true* posteriors.

a bit more explanation

$$\begin{aligned}\mathcal{F}(Q,Y) &= \int dX\, d\theta\, Q_X(X) Q_\theta(\theta) \ln \frac{P(X,Y,\theta)}{Q_X(X)Q_\theta(\theta)} \\ &= \int d\theta\, Q_\theta(\theta) \left[\ln \frac{P(\theta)}{Q_\theta(\theta)} + \right. \\ &\qquad \left. \int dX\, Q_X(X) \ln \frac{P(X,Y|\theta)}{Q_X(X)} \right].\end{aligned}$$

e.g. setting a derivative to zero:

$$\frac{\partial \mathcal{F}(Q,Y)}{\partial Q_\theta(\theta)} \propto \ln P(\theta) - 1 - \ln Q_\theta(\theta) + \langle \ln P(X,Y|\theta) \rangle_{Q_X(X)}$$

$$\begin{aligned}\mathbf{E-step} \quad \mathbf{Q_X^*(X)} &\propto \exp \langle \ln P(X,Y|\theta) \rangle_{Q_\theta(\theta)} \\ \mathbf{M-step} \quad \mathbf{Q_\theta^*(\theta)} &\propto P(\theta) \exp \langle \ln P(X,Y|\theta) \rangle_{Q_X(X)}\end{aligned}$$

Question:

What distributions can we make use of in our models?

Conjugate-Exponential models

Consider variational Bayesian learning in models that satisfy:

Condition (1). The complete data likelihood is in the exponential family:

$$P(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) = f(\mathbf{x}, \mathbf{y})\ g(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x}, \mathbf{y})\right\}$$

where $\boldsymbol{\phi}(\boldsymbol{\theta})$ is the vector of *natural parameters*, and $\mathbf{u}$ and f and g are the functions that define the exponential family.

Condition (2). The parameter prior is conjugate to the complete data likelihood:

$$P(\boldsymbol{\theta}|\eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu})\ g(\boldsymbol{\theta})^\eta \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \boldsymbol{\nu}\right\}$$

where η and $\boldsymbol{\nu}$ are hyperparameters of the prior.

We call models that satify conditions **(1)** and **(2)** *conjugate-exponential*.

Exponential family models

Some models that are in the exponential family:

- Gaussian mixtures,
- factor analysis, probabilistic PCA,
- hidden Markov models and factorial HMMs,
- linear dynamical systems and switching state-space models,
- Boltzmann machines,
- discrete-variable belief networks.

Other as yet undreamt-of models can combine Gaussian, Gamma, Poisson, Dirichlet, Wishart, Multinomial and others.

Some which are not in exponential family:

- sigmoid belief networks,
- independent components analysis

Make approximations/changes to move into exponential family...

Theoretical Results

Theorem 1 Given an iid data set $Y = (\mathbf{y}_1, \ldots \mathbf{y}_n)$, if the model satisfies conditions (1) and (2), then at the maxima of $\mathcal{F}(Q, Y)$ (minima of $KL(Q\|P)$):

(a) $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is conjugate and of the form:

$$Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = h(\tilde{\eta}, \tilde{\boldsymbol{\nu}}) g(\boldsymbol{\theta})^{\tilde{\eta}} \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^\top \tilde{\boldsymbol{\nu}}\right\}$$

where $\tilde{\eta} = \eta + n$, $\tilde{\boldsymbol{\nu}} = \boldsymbol{\nu} + \sum_{i=1}^n \overline{\mathbf{u}}(\mathbf{x}_i, \mathbf{y}_i)$, and $\overline{\mathbf{u}}(\mathbf{x}_i, \mathbf{y}_i) = \langle \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i) \rangle_Q$, using $\langle \cdot \rangle_Q$ to denote expectation under Q.

(b) $Q_X(X) = \prod_{i=1}^n Q_{\mathbf{x}_i}(\mathbf{x}_i)$ and $Q_{\mathbf{x}_i}(\mathbf{x}_i)$ is of the same form as the known parameter posterior:

$$\begin{aligned} Q_{\mathbf{x}_i}(\mathbf{x}_i) &\propto f(\mathbf{x}_i, \mathbf{y}_i) \exp\left\{\overline{\boldsymbol{\phi}}(\boldsymbol{\theta})^\top \mathbf{u}(\mathbf{x}_i, \mathbf{y}_i)\right\} \\ &= P(\mathbf{x}_i | \mathbf{y}_i, \overline{\boldsymbol{\phi}}(\boldsymbol{\theta})) \end{aligned}$$

where $\overline{\boldsymbol{\phi}}(\boldsymbol{\theta}) = \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_Q$.

KEY points: (a) the approximate parameter posterior is of the same form as the prior; (b) the approximate hidden variable posterior is of the same form as the hidden variable posterior for **known** parameters.

Variational EM algorithm

Since $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ and $Q_{\mathbf{x}_i}(\mathbf{x}_i)$ are coupled, (a) and (b) do not provide an analytic solution to the minimisation problem.

VE Step: Compute the expected sufficient statistics $\mathbf{t}(Y) = \sum_i \overline{\mathbf{u}}(\mathbf{x}_i, \mathbf{y}_i)$ under the hidden variable distributions $Q_{\mathbf{x}_i}(\mathbf{x}_i)$.

VM Step: Compute the expected natural parameters $\overline{\phi}(\boldsymbol{\theta})$ under the parameter distribution given by $\tilde{\eta}$ and $\tilde{\boldsymbol{\nu}}$.

Properties:

- VE step has same complexity as corresponding E step.
- Reduces to the EM algorithm if $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$. M step then involves re-estimation of θ^*.
- $\mathcal{F}$ increases monotonically, incorporates model complexity penalty.

Belief networks

Corollary 1: Conjugate-Exponential Belief Networks. Let $\mathcal{M}$ be a conjugate-exponential model with hidden and visible variables $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ that satisfy a belief network factorisation. That is, each variable z_j has parents $\mathbf{z}_{p_j}$ and $P(\mathbf{z}|\boldsymbol{\theta}) = \prod_j P(z_j|\mathbf{z}_{p_j}, \boldsymbol{\theta})$. Then the approximating joint distribution for $\mathcal{M}$ satisfies the same belief network factorisation:

$$Q_{\mathbf{z}}(\mathbf{z}) = \prod_j Q(z_j|\mathbf{z}_{p_j}, \tilde{\boldsymbol{\theta}})$$

where the conditional distributions have exactly the same form as those in the original model but with natural parameters $\boldsymbol{\phi}(\tilde{\boldsymbol{\theta}}) = \overline{\boldsymbol{\phi}}(\boldsymbol{\theta})$. Furthermore, with the modified parameters $\tilde{\boldsymbol{\theta}}$, the expectations under the approximating posterior $Q_{\mathbf{x}}(\mathbf{x}) \propto Q_{\mathbf{z}}(\mathbf{z})$ required for the VE Step can be obtained by applying the **belief propagation** algorithm if the network is singly connected and the **junction tree** algorithm if the network is multiply-connected.

Markov networks

Theorem 2: Markov Networks. Let $\mathcal{M}$ be a model with hidden and visible variables $\mathbf{z} = (\mathbf{x}, \mathbf{y})$ that satisfy a Markov network factorisation. That is, the joint density can be written as a product of clique-potentials ψ_j, $P(\mathbf{z}|\boldsymbol{\theta}) = g(\boldsymbol{\theta}) \prod_j \psi_j(C_j, \boldsymbol{\theta})$, where each clique C_j is a subset of the variables in $\mathbf{z}$. Then the approximating joint distribution for $\mathcal{M}$ satisfies the same Markov network factorisation:

$$Q_{\mathbf{z}}(\mathbf{z}) = \tilde{g} \prod_j \overline{\psi}_j(C_j)$$

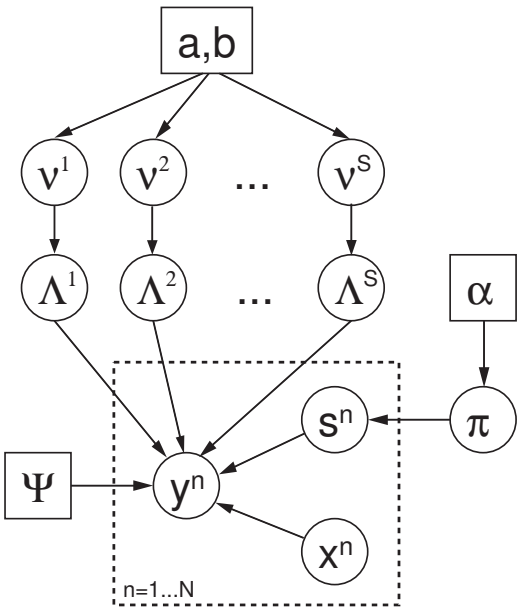
where $\overline{\psi}_j(C_j) = \exp\left\{\langle \ln \psi_j(C_j, \boldsymbol{\theta})\rangle_Q\right\}$ are new clique potentials obtained by averaging over $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, and $\tilde{g}$ is a normalisation constant. Furthermore, the expectations under the approximating posterior $Q_{\mathbf{x}}(\mathbf{x})$ required for the VE Step can be obtained by applying the junction tree algorithm.

Corollary 2: Conjugate-Exponential Markov Networks. Let $\mathcal{M}$ be a conjugate-exponential Markov network over the variables in $\mathbf{z}$. Then the approximating joint distribution for $\mathcal{M}$ is given by $Q_{\mathbf{z}}(\mathbf{z}) = \tilde{g} \prod_j \psi_j(C_j, \tilde{\boldsymbol{\theta}})$, where the clique potentials have exactly the same form as those in the original model but with natural parameters $\boldsymbol{\phi}(\tilde{\boldsymbol{\theta}}) = \overline{\boldsymbol{\phi}}(\boldsymbol{\theta})$.

Mixture of factor analysers

p-dim. vector $\mathbf{y}$ generated from k unobserved independent Gaussian sources, $\mathbf{x}$, then corrupted with Gaussian noise, $\mathbf{r}$, with diagonal covariance matrix Ψ: $\mathbf{y} = \Lambda\mathbf{x} + \mathbf{r}$.

Marginal density of $\mathbf{y}$ is Gaussian with zero mean and covariance $\Lambda\Lambda^T + \Psi$.



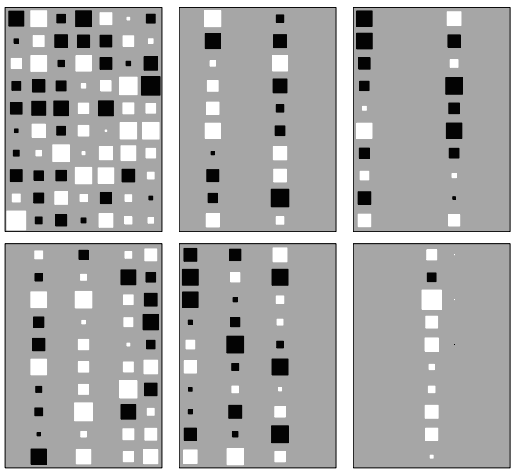
- Complete data likelihood is in exponential family.
- Choose conjugate priors:
  ($\nu \sim$ Gamma, $\Lambda \sim$Gaussian, $\pi \sim$Dirichlet).

VE: posteriors over hidden states calculated as usual.
VM: posteriors over parameters have same form as priors.

Synthetic example

True data: 6 Gaussian clusters with dimensions: (1 7 4 3 2 2) embedded in 10 dimensions.

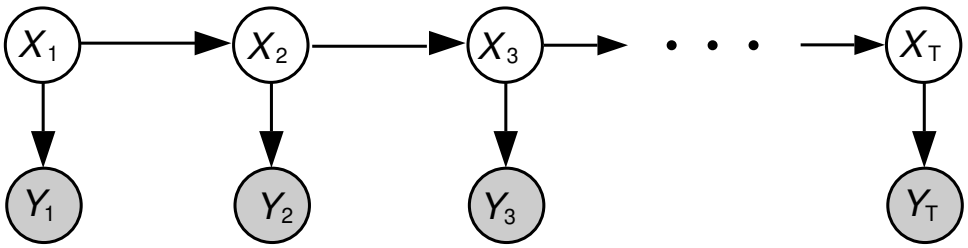
Inferred structure:



- Finds the clusters and their dimensionalities.
- Model complexity reduces in line with lack of data support.

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

Linear Dynamical Systems



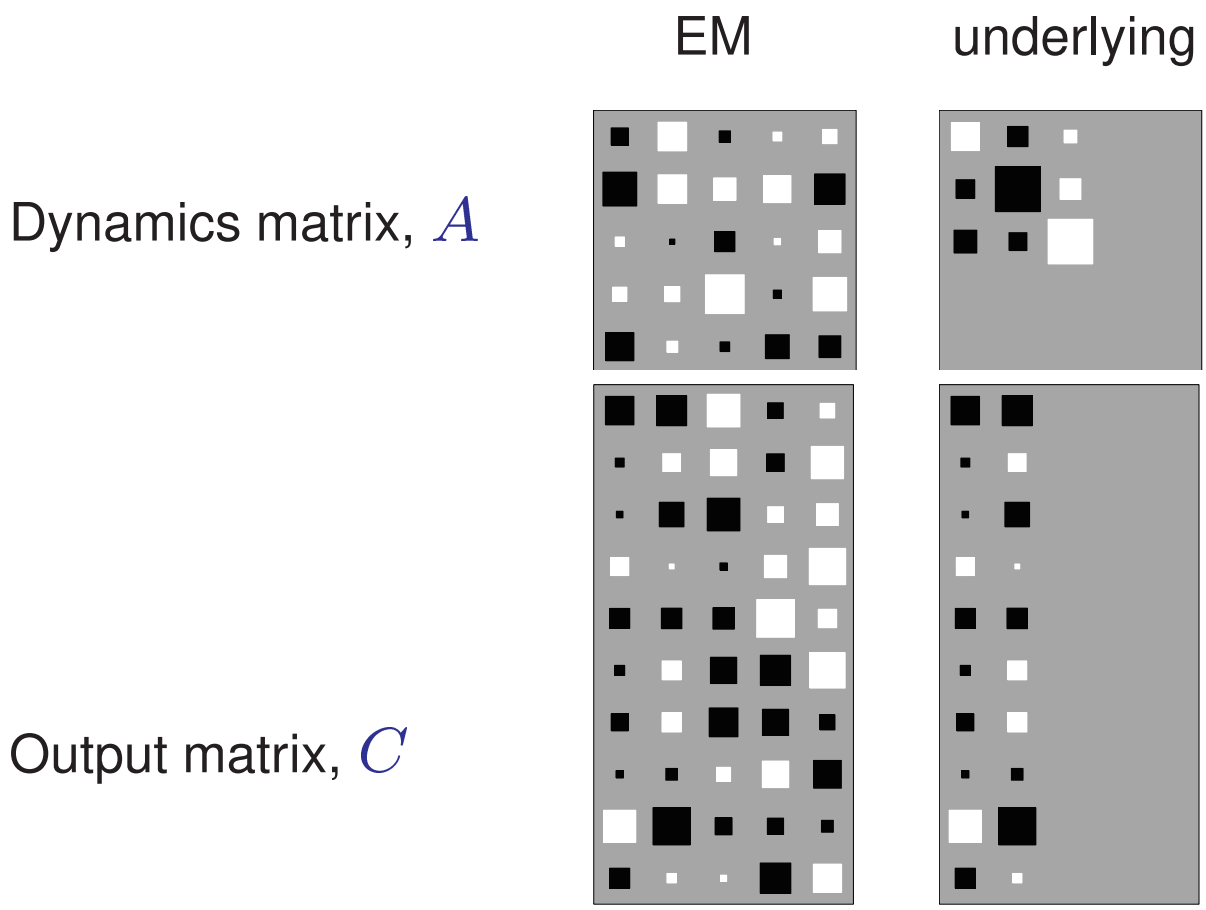
- Sequence of D-dimensional real-valued *observed* vectors $\mathbf{y}_{1:T}$.
- Assume at each time step t, $\mathbf{y}_t$ was generated from a k-dimensional real-valued *hidden* state variable $\mathbf{x}_t$, and that the sequence of $\mathbf{x}_{1:T}$ define a first order Markov process.
- The joint probability is given by

$$P(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}) = P(\mathbf{x}_1)P(\mathbf{y}_1|\mathbf{x}_1) \prod_{t=2}^{T} P(\mathbf{x}_t|\mathbf{x}_{t-1})P(\mathbf{y}_t|\mathbf{x}_t).$$

- If transition and output functions are linear, time-invariant, and noise distributions are Gaussian, this is a Linear-Gaussian state-space model:

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{y}_t = C\mathbf{x}_t + \mathbf{r}_t.$$

Model selection using ARD



- Gaussian priors on the entries in the columns of these matrices.

  $P(\mathbf{a}_i|\boldsymbol{\alpha}) = \mathcal{N}(\mathbf{0}, \mathrm{diag}(\boldsymbol{\alpha})^{-1}), \qquad P(\mathbf{c}_i|\boldsymbol{\beta}) = \mathcal{N}(\mathbf{0}, \mathrm{diag}(\rho_i\boldsymbol{\beta})^{-1})$

- $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of *hyperparameters*, to be optimised during learning. Some $\boldsymbol{\alpha}$,$\boldsymbol{\beta} \to \infty$.

- Empty columns signify **extinct** hidden dimensions for:
  - modelling hidden dynamics,
  - for modelling output covariance structure.

LDS variational approximation

Parameters of the model $\boldsymbol{\theta} = (A, C, \boldsymbol{\rho})$, hidden states $\mathbf{x}_{1:T}$.

$$\ln P(Y) = \ln \int dA\, dC\, d\boldsymbol{\rho}\, d\mathbf{x}_{1:t}\; P(A, C, \boldsymbol{\rho}, \mathbf{x}_{1:T}, \mathbf{y}_{1:T}).$$

Lower bound: Jensen's inequality

$$\ln P(Y) \geq \mathcal{F}(Q, Y) \;=\; \int dA\, dC\, d\boldsymbol{\rho}\, d\mathbf{x}_{1:T}\; Q(A, C, \boldsymbol{\rho}, \mathbf{x}_{1:T}) \cdot \ln \frac{P(A, C, \boldsymbol{\rho}, \mathbf{x}_{1:T}, \mathbf{y}_{1:T})}{Q(A, C, \boldsymbol{\rho}, \mathbf{x}_{1:T})}.$$

Factored approximation to the true posterior

$$Q(A, C, \boldsymbol{\rho}, \mathbf{x}_{1:T}) = Q_A(A) Q_{C\rho}(C, \boldsymbol{\rho}) Q_{\mathbf{x}}(\mathbf{x}_{1:T}).$$

This factorisation falls out from the initial assumptions and the conditional independencies in the model.

Priors:

- Sensor precisions, $\boldsymbol{\rho}$, given conjugate gamma priors.
- Transition and output matrices given conjugate zero-mean Gaussian priors — *ARD*.

Optimal $Q(\cdot)$ distributions: LDS

Complete data likelihood is of *conjugate-exponential* form: iterative analytic maximisations.

Dynamics matrix $Q_A(A)$: $A \sim \mathcal{N}(\mathbf{a}_i; \overline{\mathbf{a}_i}, \Sigma_i^A)$

$$\overline{\mathbf{a}_i} = S^\top \Sigma_i^A, \quad \Sigma_i^A = (\mathrm{diag}(\boldsymbol{\alpha}) + W)^{-1}$$

Noise covariance $Q_\rho(\boldsymbol{\rho})$: $\rho_i \sim Gamma(\rho_i; \tilde{a}, \tilde{b})$

$$\tilde{a} = a + T/2, \quad \tilde{b} = b + g_i/2$$

Output matrix $Q_C(C|\rho_i)$: $C \sim \mathcal{N}(\mathbf{c}_i; \overline{\mathbf{c}_i}, \Sigma_i^C)$

$$\overline{\mathbf{c}_i} = \rho_i U_i \Sigma_i^C, \quad \Sigma_i^C = (\mathrm{diag}(\boldsymbol{\beta}) + W')^{-1}/\rho_i$$

Hidden state $Q_{\mathbf{x}}(\mathbf{x}_{1:T})$: $\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t; \eta_t^{(T)}, \Psi_t^{(T)})$

$$\{\eta_t^{(T)}, \Psi_t^{(T)}\} = \mathrm{KalmanSmoother}(Q_{AC\rho}, \mathbf{y}_{1:T})$$

Hyperparameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$: $1/\boldsymbol{\alpha} = \langle \mathbf{a} \cdot \mathbf{a} \rangle_{Q_A}$
$1/\boldsymbol{\beta} = \langle \mathbf{c} \cdot \mathbf{c} \rangle_{Q_{C\rho}}$

$$S = \sum_{t=2}^{T} \langle \mathbf{x}_{t-1} \mathbf{x}_t^\top \rangle, \quad W = \sum_{t=1}^{T-1} \langle \mathbf{x}_t \mathbf{x}_t^\top \rangle, \quad W' = W + \langle \mathbf{x}_t \mathbf{x}_t^\top \rangle$$

$$g_i = \sum_{t=1}^{T} y_{ti}^2 - U_i(\mathrm{diag}(\boldsymbol{\beta}) + W')U_i^\top, \quad U_i = \sum_{t=1}^{T} y_{ti} \langle \mathbf{x}_t^\top \rangle,$$

Method overview

1. Randomly initialise the approximate posteriors;

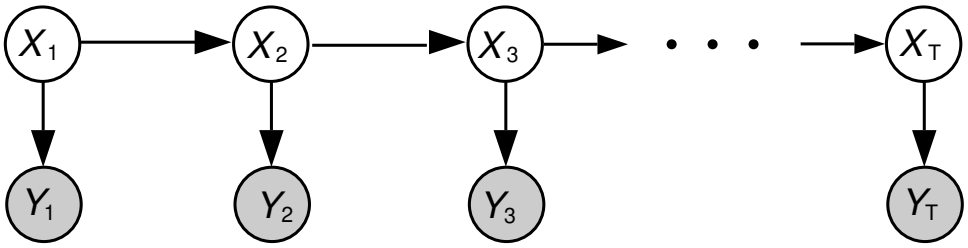
2. Update each posterior according to variational E- & M-steps

   - $Q_A(A) \quad Q_{C\rho}(C|\boldsymbol{\rho}) \quad Q_\rho(\boldsymbol{\rho}) \quad Q_{\mathbf{x}}(\mathbf{x}_{1:T})$;

3. Update the hyperparameters: $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$;

4. Repeat steps 2-3 until convergence.

Variational E Step for LDS



- Conjugate-exponential singly connected belief network,
- (**C1**) use Belief Propagation, which for LDSs is the *Kalman smoother*:

  Forward: infer state $\mathbf{x}_t$ given *past* observations,

  Backward: infer state $\mathbf{x}_t$ given *future* observations.
- Inference using an ensemble of parameters is possible, using just a single setting of the parameters.

  Required averages: $\langle \rho_i \mathbf{c}_i \rangle, \ \langle \rho_i \mathbf{c}_i \mathbf{c}_i^\top \rangle, \ \langle A \rangle, \ \langle A^\top A \rangle$
- The result is a recursive algorithm very similar to the Kalman smoothing propagation.

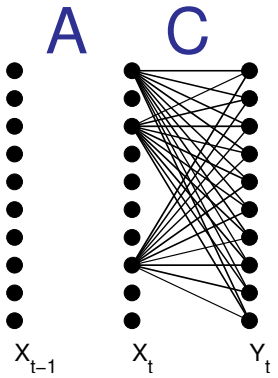
Synthetic experiments

Generated a variety of state-space models and looked the recovered hidden structure.

Observed data is • 10-dim output • 200 time-steps

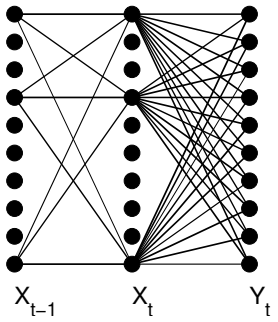
- **True model**: 3-dim static state-space.

Inferred model:



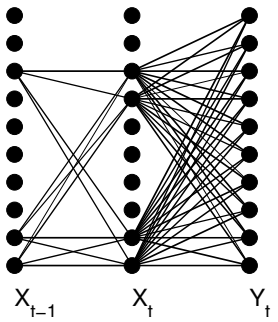
- **True model**: 3-dim dynamical state-space.

Inferred model:



- **True model**: 4-dim state-space, of which 3 dynamical.

Inferred model:

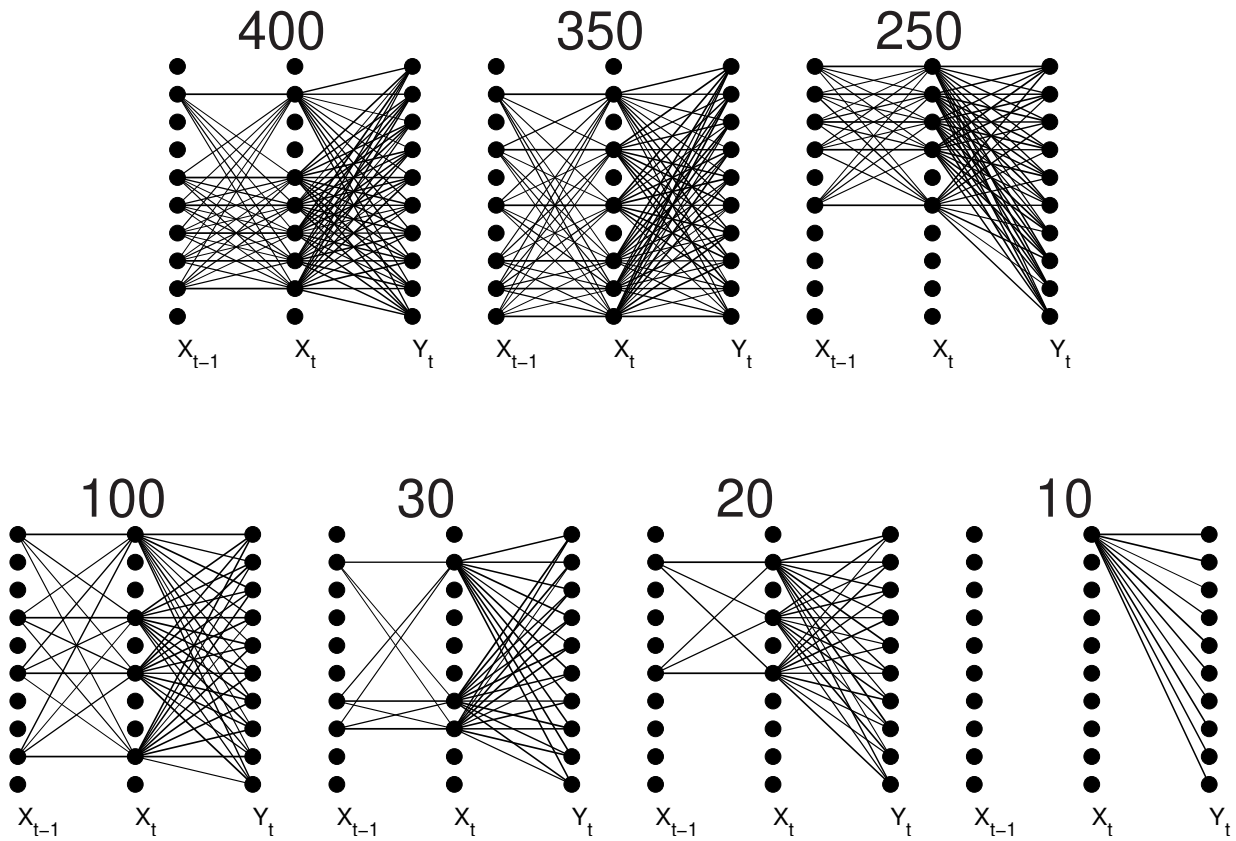


Degradation with data support

True model: • 6-dim state-space • 10-dim output

Complexity of recovered structure decreases with decreasing data support (400 to 10 time-steps).

Inferred models:



Summary

- Bayesian learning avoids overfitting and can be used to learn model structure.
- Tractable learning using variational approximations.
- Conjugate-exponential familes of models.
- Variational EM and integration of existing algorithms.
- Examples learning structure in MFA and LDS models.

Issues & Extensions

- Quality of the variational approximations?
- How best to bring a model into exponential family?
- Extensions to other models, e.g. switching state-space models [Ueda], is straightforward.
- Problems integrating existing algorithms into the VE step.
- Implementation on a general graphical model.

Optimal $Q(\cdot)$ distributions: MFA

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

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

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

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

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

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 $\psi(\alpha) - \psi(\alpha/S) + \sum[\psi(\omega u_s) - \psi(\omega)]/S = 0$.

Sampling from Variational Approximations

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

- The Exact Predictive Density:

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

weights: $\omega_m = \frac{1}{\Omega}\frac{P(\theta_m,Y)}{Q(\theta_m)}$, with Ω s.t. $\sum \omega_m = 1$

- The True Evidence:

$$P(Y|\mathcal{M}) = \int d\theta \; Q(\theta)\frac{P(\theta,Y)}{Q(\theta)} = \langle \Omega\omega \rangle$$

- The KL Divergence:

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

Variational Bayes & Ensemble Learning

- multilayer perceptrons (Hinton & van Camp, 1993)
- mixture of experts (Waterhouse, MacKay & Robinson, 1996)
- hidden Markov models (MacKay, 1995)
- other work by Jaakkola, Barber, Bishop, Tipping, etc

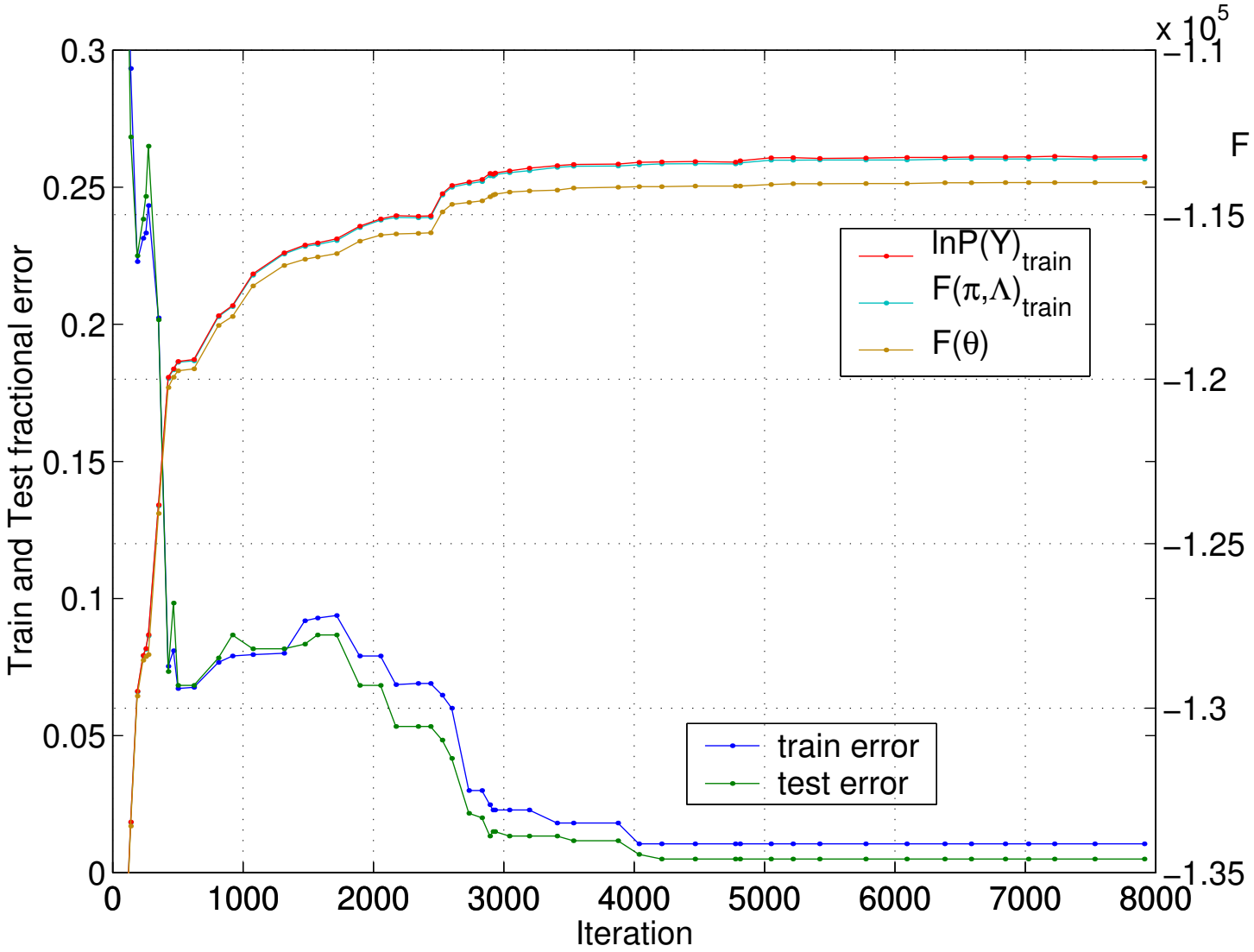
Examples of VB Learning Model Structure

Model learning has been treated with variational Bayesian techniques for:

- mixtures of factor analysers (Ghahramani & Beal, 1999)
- mixtures of Gaussians (Attias, 1999)
- independent components analysis (Attias, 1999; Miskin & MacKay, 2000; Valpola 2000)
- principal components analysis (Bishop, 1999)
- linear dynamical systems (Ghahramani & Beal, 2000)
- mixture of experts (Ueda & Ghahramani, 2000)
- hidden Markov models (Ueda & Ghahramani, in prep)

(compiled by Zoubin)

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



Required expectations for linear-Gaussian state-space models

$$
\begin{aligned}
\text{defining} \quad \mathbf{G}' &= \text{diag}\left(\frac{a + T/2}{b + \mathbf{G}/2}\right) \\
\langle C^\top R^{-1} C \rangle &= \left(\text{diag}\boldsymbol{\beta} + W'\right)^{-1} \cdot \left[p + U\mathbf{G}'U^\top \left(\text{diag}\boldsymbol{\beta} + W'\right)^{-1}\right] \\
\langle C^\top R^{-1} \rangle &= \left(\text{diag}\boldsymbol{\beta} + W'\right)^{-1} U\mathbf{G}' \\
\langle A \rangle &= S^\top \left(\text{diag}\boldsymbol{\alpha} + W\right)^{-1} \\
\langle A^\top A \rangle &= k \left(\text{diag}\boldsymbol{\alpha} + W\right)^{-1} + \langle A \rangle^\top \langle A \rangle
\end{aligned}
$$

where $S = \sum_{t=2}^{T} \left\langle \mathbf{x}_{t-1} \mathbf{x}_t{}^\top \right\rangle$, $W = \sum_{t=1}^{T-1} \left\langle \mathbf{x}_t \mathbf{x}_t{}^\top \right\rangle$, $U_i = \sum_{t=1}^{T} y_{ti} \langle \mathbf{x}_t{}^\top \rangle$

and $W' = W + \left\langle \mathbf{x}_T \mathbf{x}_T{}^\top \right\rangle$.

Kalman smoothing propagation

Forward recursion:

$$
\begin{aligned}
\Sigma_t^* &= \left(\Sigma_t^{-1} + \langle A^\top A \rangle\right)^{-1} \\
\boldsymbol{\mu}_t &= \Sigma_t \left[\langle C^\top R^{-1} \rangle \mathbf{y}_t + \langle A \rangle \Sigma_{t-1}^* \Sigma_{t-1}^{-1} \boldsymbol{\mu}_{t-1}\right] \\
\Sigma_t &= \left[I + \langle C^\top R^{-1} C \rangle - \langle A \rangle \Sigma_{t-1}^* \langle A \rangle^\top\right]^{-1}
\end{aligned}
$$

Backward recursion:

$$
\begin{aligned}
\boldsymbol{\eta}_t &= \Psi_t \Big[\Sigma_t^{-1} \boldsymbol{\mu}_t + \langle A \rangle^\top \left(\Psi_{t+1}^{-1} + \langle A \rangle \Sigma_t^* \langle A \rangle^\top\right)^{-1} \cdot \\
&\qquad \left(\Psi_{t+1}^{-1} \boldsymbol{\eta}_{t+1} - \langle A \rangle \Sigma_t^* \Sigma_t^{-1} \boldsymbol{\mu}_t\right)\Big] \\
\Psi_t &= \left[\Sigma_t^{*-1} - \langle A \rangle^\top \left(\Psi_{t+1}^{-1} + \langle A \rangle \Sigma_t^* \langle A \rangle^\top\right)^{-1} \langle A \rangle\right]^{-1} \\
\text{cov}(\mathbf{x}_t, \mathbf{x}_{t+1}) &= \left[\left(\Psi_{t+1}^{-1} + \langle A \rangle \Sigma_t^* \langle A \rangle^\top\right) \langle A \rangle^{-\top} \Sigma_t^* - \langle A \rangle\right]^{-1}
\end{aligned}
$$