Propagation Algorithms for Variational Bayesian Learning

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 provide some theoretical results for the variational updates in a very general family of *conjugate-exponential* graphical models.
- We show how the belief propagation and the junction tree algorithms can be used in the inference step of variational Bayesian learning.
- Applying these results to the Bayesian analysis of linear-Gaussian state-space models we obtain a learning procedure that exploits the Kalman smoothing propagation, *while integrating over all model parameters*.
- We demonstrate how this can be used to infer the hidden state dimensionality of the state-space model in a variety of synthetic problems and one real high-dimensional data set.

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 integrate out all the parameters.
- We should be able to do principled model comparison.

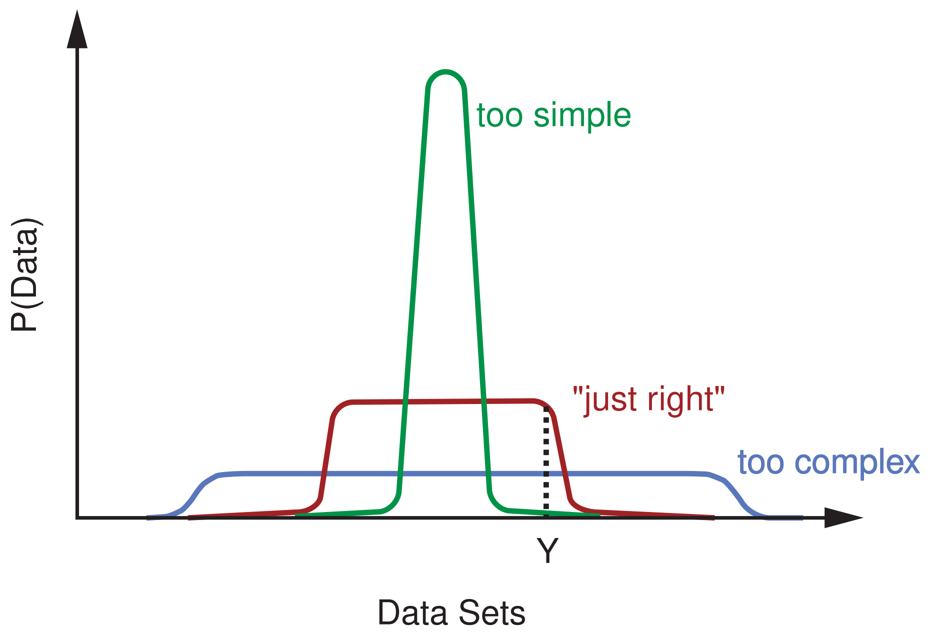
Method

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

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

Problem

- These integrals are computationally intractable.



(adapted from D.J.C. MacKay)

Practical Bayesian methods

- 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 $\mathbf{x}$ and the parameters $\boldsymbol{\theta}$,

$$\ln P(\mathbf{y}) = \ln \int d\mathbf{x}\, d\boldsymbol{\theta}\ P(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) \geq \int d\mathbf{x}\, d\boldsymbol{\theta}\ Q(\mathbf{x}, \boldsymbol{\theta}) \ln \frac{P(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{Q(\mathbf{x}, \boldsymbol{\theta})}.$$

Restrict the arbitrary $Q(\mathbf{x}, \boldsymbol{\theta})$ distribution to a simpler factorised form:

$$\ln P(\mathbf{y}) \geq \int d\mathbf{x}\, d\boldsymbol{\theta}\ Q_{\mathbf{x}}(\mathbf{x}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{P(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})}{Q_{\mathbf{x}}(\mathbf{x}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} = \mathcal{F}(Q, \mathbf{y}).$$

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

$$\begin{aligned} Q^*_{\mathbf{x}}(\mathbf{x}) &\propto \exp \langle \ln P(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} && \text{E}-\text{like step} \\ Q^*_{\boldsymbol{\theta}}(\boldsymbol{\theta}) &\propto P(\boldsymbol{\theta}) \exp \langle \ln P(\mathbf{x}, \mathbf{y} | \boldsymbol{\theta}) \rangle_{Q_{\mathbf{x}}(\mathbf{x})} && \text{M}-\text{like step} \end{aligned}$$

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

Questions:

- **Which forms of model distributions and priors result in tractable updates?**
- **How does this relate formally to EM?**
- **When can we use belief propagation algorithms?**

Conjugate-Exponential (CE) models

Consider variational Bayesian learning in models that satisfy:

Condition (1). The joint probability over *variables* 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 $\phi(\boldsymbol{\theta})$ is the vector of *natural parameters*.

Condition (2). The prior over *parameters* is conjugate to this joint probability:

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

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

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

Conjugate-Exponential examples

In the **CE** family:

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

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

Not in the **CE** family:

- Boltzmann machines (no conjugacy),
- logistic regression (no conjugacy),
- sigmoid belief networks (not exponential),
- independent components analysis (not exponential).

Note: one can often approximate these models with models in the **CE** family.

Theoretical Results

Theorem 1 Given an iid data set $\mathbf{y} = (\mathbf{y}_1, \ldots \mathbf{y}_n)$, if the model is **CE** then:

(a) $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is also conjugate, *i.e.*

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

(b) $Q_{\mathbf{x}}(\mathbf{x}) = \prod_{i=1}^n Q_{\mathbf{x}_i}(\mathbf{x}_i)$ is of the same form as in the E step of regular EM, but using pseudo parameters computed by averaging over $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$

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

KEY points: (a) the approximate parameter posterior is of the same form as the prior. (b) the *approximate* hidden variable posterior, averaging over *all* parameters, is of the same form as the *exact* hidden variable posterior for a single setting of the parameters.

Variational EM algorithm (VEM)

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, so we iterate:

VE Step: Compute the expected sufficient statistics $\mathbf{t}(\mathbf{y}) = \sum_i \overline{\mathbf{u}}(\mathbf{x}_i, \mathbf{y}_i)$ 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 $\boldsymbol{\theta}^*$.
- $\mathcal{F}$ increases monotonically, and incorporates a model complexity penalty.

Belief networks

Corollary 1: If the model is **CE**, with hidden and visible variables $\mathbf{z} = (\mathbf{x}, \mathbf{y})$, and satisfies a belief network factorisation

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

then the approximate joint satisfies the same BN factorisation

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

but with $\boldsymbol{\phi}(\tilde{\boldsymbol{\theta}}) = \overline{\boldsymbol{\phi}}(\boldsymbol{\theta})$, [where ϕ is invertible].

So we can use the **belief propagation** algorithm if the network is singly connected and the **junction tree** algorithm if the network is multiply connected, to obtain the expectations under $Q_{\mathbf{x}}(\mathbf{x})$ required for the VE Step.

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

$$P(\mathbf{z}|\boldsymbol{\theta}) = g(\boldsymbol{\theta}) \prod_j \psi_j(C_j, \boldsymbol{\theta})$$

then the approximating joint 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 $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$.

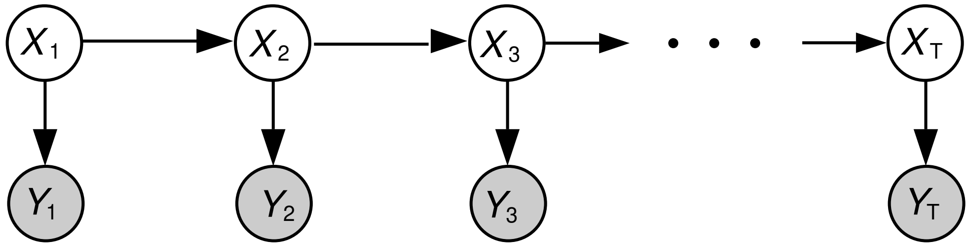
Corollary 2: CE Markov Networks. If $\mathcal{M}$ is **CE** then the approximate joint distribution is

$$Q_{\mathbf{z}}(\mathbf{z}) = \tilde{g} \prod_j \psi_j(C_j, \tilde{\boldsymbol{\theta}})$$

where the clique potentials have exactly the same form as those in $\mathcal{M}$ but with natural parameters $\phi(\tilde{\boldsymbol{\theta}}) = \overline{\phi}(\boldsymbol{\theta})$.

So we can use the **junction tree** algorithm to obtain the expectations under the approximating posterior $Q_{\mathbf{x}}(\mathbf{x})$ required for the VE Step.

CE belief network example: Linear Dynamical Systems



- Assumes $\mathbf{y}_t$ generated from a *hidden* state variable $\mathbf{x}_t$, and that the sequence of $\mathbf{x}_{1:T}$ is Markov:

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

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

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

- Dynamic generalisation of factor analysis.
- Three levels of inference:
  1. Given data, structure and parameters, **Kalman smoothing** $\rightarrow$ hidden state;
  2. Given data and structure, **EM** $\rightarrow$ hidden state and parameter point estimates;
  3. Given data only, **VEM** $\rightarrow$ **model structure** *and* **distributions over parameters** *and* **hidden state**.

LDS variational approximation

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

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

Lower bound: Jensen's inequality

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

Factored approximation to the true posterior

1. Assumed factorisation:

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

2. Additional factorisations (for free!)

$$\begin{aligned} Q(A, C, \boldsymbol{\rho}, \mathbf{x}_{1:T}) &= Q_A(A) Q_{C\rho}(C, \boldsymbol{\rho}) Q_{\mathbf{x}}(\mathbf{x}_{1:T}) \\ &= \left[\prod_{j=1}^{K} Q(\mathbf{a}_j) \right] \left[\prod_{i=1}^{D} Q(\rho_i) Q(\mathbf{c}_i | \rho_i) \right] Q_{\mathbf{x}}(\mathbf{x}_{1:T}) \end{aligned}$$

$\mathcal{F}$ for LDS

$$
\begin{aligned}
\mathcal{F} \;=\; & -\int dA\, Q(A) \ln \frac{Q(A)}{P(A|\boldsymbol{\alpha})} - \int d\boldsymbol{\rho}\, Q(\boldsymbol{\rho}) \frac{Q(\boldsymbol{\rho})}{P(\boldsymbol{\rho}|a,b)} \\
& - \int d\boldsymbol{\rho}\, Q(\boldsymbol{\rho}) \int dC\; Q(C|\boldsymbol{\rho}) \frac{Q(C|\boldsymbol{\rho})}{P(C|\boldsymbol{\rho},\boldsymbol{\beta})} - \int d\mathbf{x}_{1:T}\; Q(\mathbf{x}_{1:T}) \ln Q(\mathbf{x}_{1:T}) \\
& + \int dA\; Q(A) \int d\boldsymbol{\rho}\; Q(\boldsymbol{\rho}) \int dC\; Q(C|\boldsymbol{\rho}) \;\int d\mathbf{x}_{1:T}\; Q(\mathbf{x}_{1:T}) \ln P(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}|A, C, \boldsymbol{\rho})
\end{aligned}
$$

Conjugate Priors

- $P(A, C, \boldsymbol{\rho}) = P(A)P(\boldsymbol{\rho})P(C|\boldsymbol{\rho})$
- Gamma prior on output noise precision:

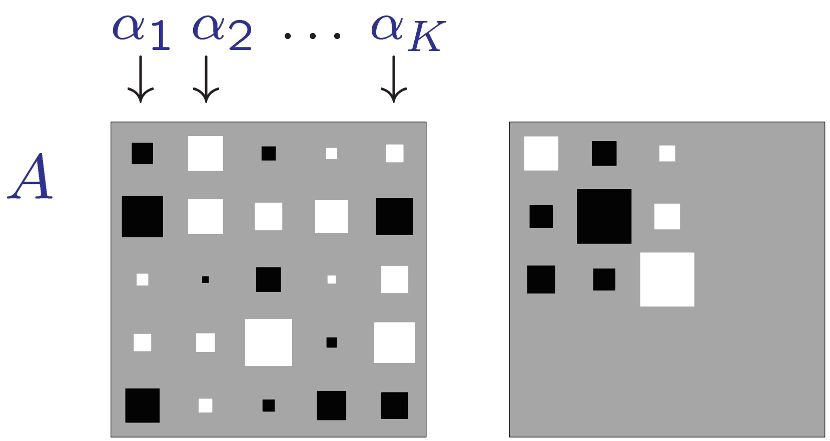
$$P(\boldsymbol{\rho}\,|a, b) = \prod_{i=1}^{D} \frac{b^a}{\Gamma(a)} \rho_i^{a-1} \exp\{-b\rho_i\}.$$

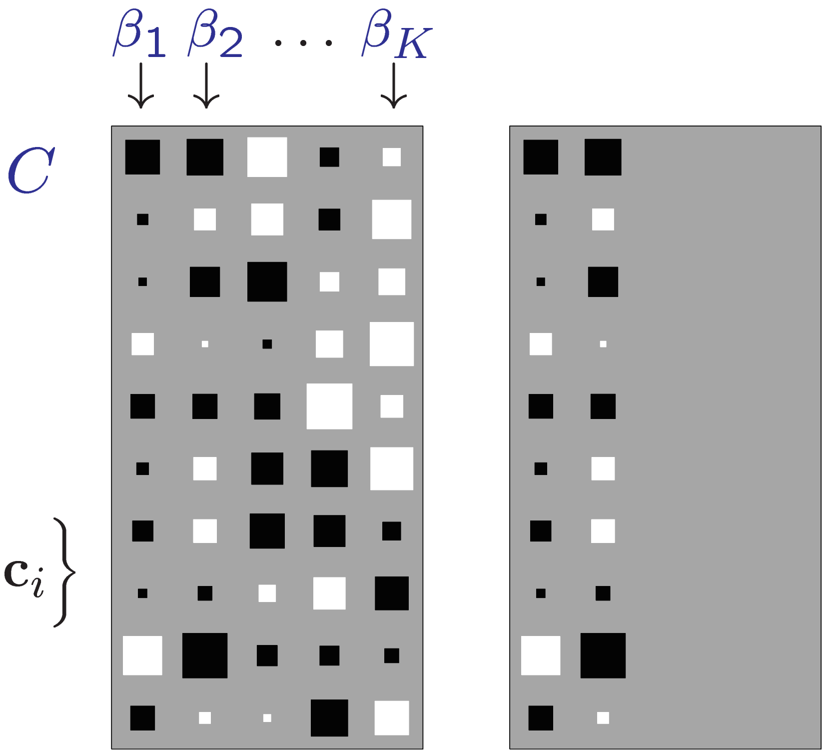
- Gaussian priors **ARD** on **Rows** of transition (A) and output (C) matrices:

$$P(\mathbf{a}_j|\boldsymbol{\alpha}) = \left[\prod_{l=1}^{K} \left(\frac{\alpha_l}{2\pi}\right)^{\frac{1}{2}}\right] \exp\left\{-\frac{1}{2}\mathbf{a}_j^\top \mathrm{diag}(\boldsymbol{\alpha})\mathbf{a}_j\right\}$$

$$P(\mathbf{c}_i|\boldsymbol{\beta}, \rho_i) = \left[\prod_{l=1}^{K} \left(\frac{\rho_i\beta_l}{2\pi}\right)^{\frac{1}{2}}\right] \exp\left\{-\frac{1}{2}\rho_i\mathbf{c}_i^\top \mathrm{diag}(\boldsymbol{\beta})\mathbf{c}_i\right\}$$

- $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of *hyperparameters*.





- During learning, some $\boldsymbol{\alpha}$,$\boldsymbol{\beta} \rightarrow \infty$. Empty columns signify **extinct** hidden dimensions for modelling hidden dynamics and output covariance structure.

Method overview

1. Initialise the approximate posteriors to the model priors.

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

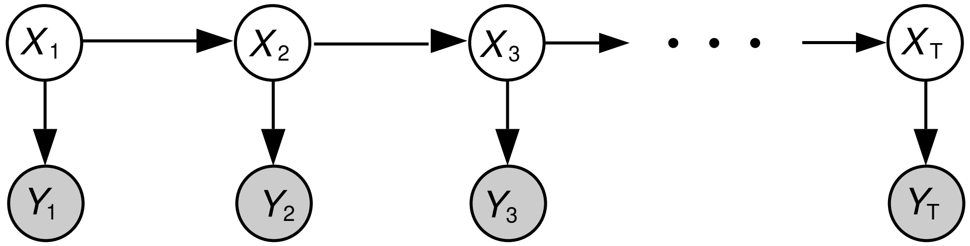
   VE Step: Compute the expected sufficient statistics for the hidden variable distributions: $Q_{\mathbf{x}}(\mathbf{x}_{1:T})$, using the natural parameters.

   VM Step: Compute the expected natural parameters for parameter distributions: $Q_A(A)$, $Q_\rho(\boldsymbol{\rho})$, $Q_{C\rho}(C|\boldsymbol{\rho})$, using the sufficient statistics of the hidden variable distributions.

3. Update the hyperparameters: $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, a and b.

4. Repeat steps 2-3 until convergence.

Variational E Step for LDS



- The LDS is a conjugate-exponential singly-connected belief network,
- (**C1**) use Belief Propagation, which for LDSs is the *Kalman smoother*:
  Forward recursion: infer state $\mathbf{x}_t$ given *past* observations,
  Backward recursion: infer state $\mathbf{x}_t$ given *all* observations.
- Inference using an ensemble of parameters is possible, using just a single value of the natural parameters.
- Required expected natural parameters: $\overline{\phi}(\boldsymbol{\theta}) = \left\{ \langle \rho_i \mathbf{c}_i \rangle, \; \langle \rho_i \mathbf{c}_i \mathbf{c}_i^\top \rangle, \; \langle A \rangle, \; \langle A^\top A \rangle \right\}$
- Result: a natural parameter version of the Kalman smoothing propagation.

Optimal $Q(\cdot)$ distributions: LDS

Dynamics matrix $Q_A(A)$: $A \sim \mathsf{N}(\mathbf{a}_i; \overline{\mathbf{a}_i}, \Sigma_i^A)$ $\quad \overline{\mathbf{a}_i} = S^\top \Sigma_i^A, \quad \Sigma_i^A = (\mathsf{diag}(\boldsymbol{\alpha}) + W)^{-1}$

Noise covariance $Q_\rho(\boldsymbol{\rho})$: $\rho_i \sim \mathsf{Gamma}(\rho_i; \tilde{a}, \tilde{b_i})$ $\quad \tilde{a} = a + T/2, \quad \tilde{b_i} = b + g_i/2$

Output matrix $Q_C(C|\rho_i)$: $C \sim \mathsf{N}(\mathbf{c}_i; \overline{\mathbf{c}_i}, \Sigma_i^C)$ $\quad \overline{\mathbf{c}_i} = \rho_i U_i \Sigma_i^C, \quad \Sigma_i^C = (\mathsf{diag}(\boldsymbol{\beta}) + W')^{-1}/\rho_i$

Hidden state $Q_\mathbf{x}(\mathbf{x}_{1:T})$: $\mathbf{x}_t \sim \mathsf{N}(\mathbf{x}_t; \eta_t^{(T)}, \Psi_t^{(T)})$ $\quad \{\eta_t^{(T)}, \Psi_t^{(T)}\} = \textbf{KalmanSmoother}(Q_{AC\rho}, \mathbf{y}_{1:T})$

Hyperparameters $\alpha_k = \dfrac{K}{\langle A^\top A \rangle_{kk}}, \qquad \beta_k = \dfrac{D}{\langle C^\top \mathsf{diag}(\boldsymbol{\rho}) C \rangle_{kk}}.$

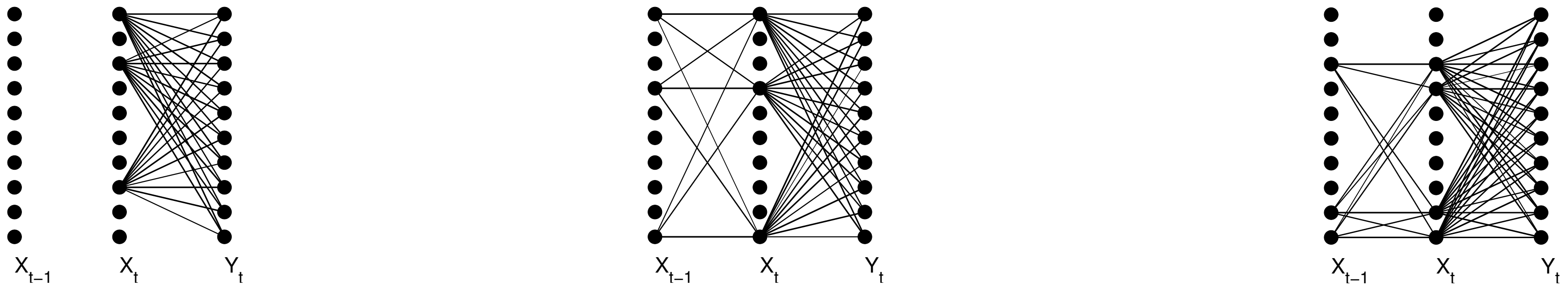
$$S = \sum_{t=2}^{T} \langle \mathbf{x}_{t-1} \mathbf{x}_t^\top \rangle, \quad W = \sum_{t=1}^{T-1} \langle \mathbf{x}_t \mathbf{x}_t^\top \rangle, \quad W' = W + \langle \mathbf{x}_T \mathbf{x}_T^\top \rangle$$

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

Results

Inferring model structure (synthetic):

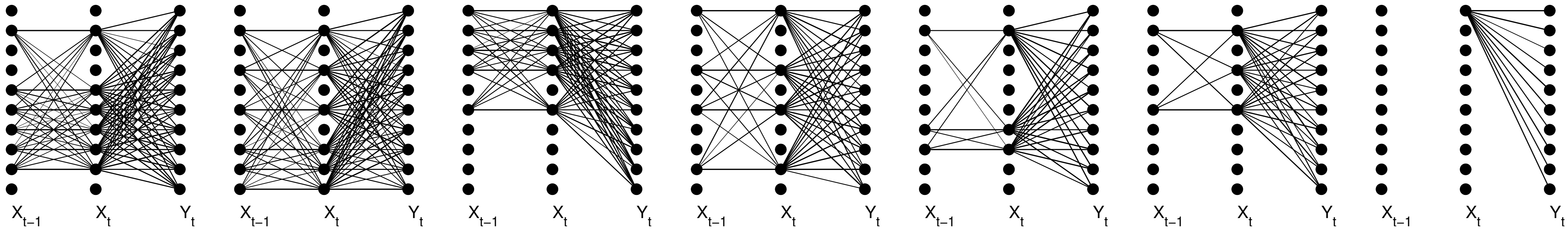
a) FA with 3 factors. b) SSM with 3 state variables. c) SSM with 3 dynamic states, and 1 static state.



Inferred model complexity reduces with less data:

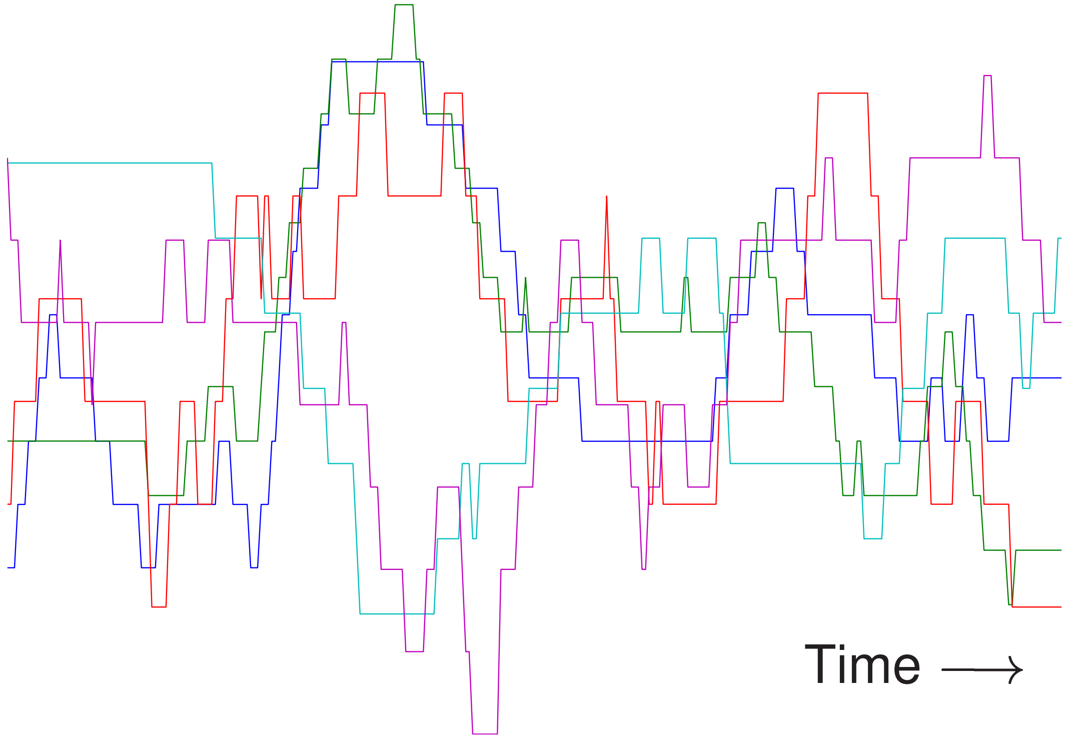
True model: • SSM with 6 state variables, • 10-dimensional observation vector.

400 350 250 100 30 20 10

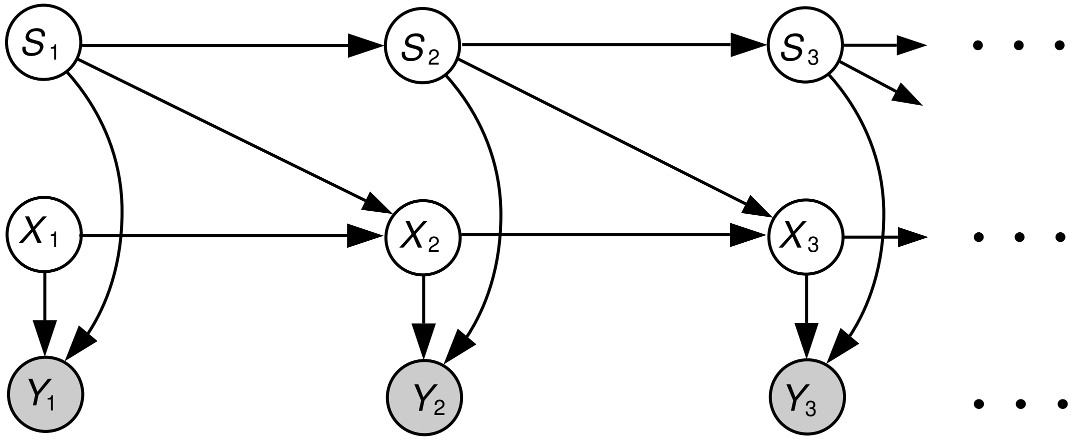


Steel plant data set

- 38 sensors (temperatures, pressures, etc.) sampled at 2Hz from a continuous casting process for 150 seconds.
- Sensors *covaried* and were *temporally correlated*, suggesting an LDS could capture some of its structure.
- **True model**: **???**.
- **Inferred model**: **16** state variables required, of which **14** emitted outputs.



SwSSM: a switching state-space model



Switch transitions: $P(s_t = i|s_{t-1} = j) = T_{ij}$

Hidden state dynamics: $P(\mathbf{x}_t|s_{t-1}, \mathbf{x}_{t-1}) = \mathrm{N}(A_{s_{t-1}}\mathbf{x}_{t-1}, Q_{s_{t-1}})$

Output function: $P(\mathbf{y}_t|s_t, \mathbf{x}_t) = \mathrm{N}(C_{s_t}\mathbf{x}_t, R_{s_t})$

Markov property:
$$P(s_{1:T}, \mathbf{x}_{1:T}, \mathbf{y}_{1:T}) = P(s_1|\pi^{(s)})P(\mathbf{x}_1|\pi^{(\mathbf{x})})P(\mathbf{y}_1|s_1, \mathbf{x}_1) \cdot \prod_{t=2}^{T} P(s_t|s_{t-1})P(\mathbf{x}_t|s_{t-1}, \mathbf{x}_{t-1})P(\mathbf{y}_t|s_t, \mathbf{x}_t)$$

Contains as special cases: mixtures of factor analysers, mixtures of linear dynamical systems, Gaussian-output hidden Markov models, mixtures of Gaussians, ...

is a conjugate-exponential belief network.

Summary

- Bayesian learning avoids overfitting and can be used to learn model structure.
- Tractable learning using variational approximations.
- Conjugate-exponential families of models.
- The Variational EM algorithm
  - Integration of existing propagation algorithms into the E-Step.
  - Simple M-Step in terms of natural parameters.
- Examples learning structure in State-Space Models.
- A general framework for many (but not all) models.

Theoretical Results (detail)

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

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

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

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

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

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

Belief networks (detail)

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 $\phi(\tilde{\boldsymbol{\theta}}) = \overline{\phi}(\boldsymbol{\theta})$. 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 (detail)

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 $\phi(\tilde{\boldsymbol{\theta}}) = \overline{\phi}(\boldsymbol{\theta})$.

1. **H. Attias**. A variational Bayesian framework for graphical models. In *Advances in Neural Information Processing Systems 12*. MIT Press, Cambridge, MA, 2000.
2. **C.M. Bishop**. Variational PCA. In *Proc. Ninth ICANN*, 1999.
3. **S. Früwirth-Schnatter**. Bayesian model discrimination and Bayes factors for linear Gaussian state space models. *J. Royal. Stat. Soc. B*, 57:237–246, 1995.
4. **Z. Ghahramani and M.J. Beal**. Variational inference for Bayesian mixtures of factor analysers. In *Adv. Neur. Inf. Proc. Sys. 12*. MIT Press, Cambridge, MA, 2000.
5. **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.
6. **M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K Saul**. An introduction to variational methods in graphical models. *Machine Learning*, 37:183–233, 1999.
7. **D.J.C. MacKay**. Ensemble learning for hidden Markov models. Technical report, Cavendish Laboratory, University of Cambridge, 1997.
8. **J. Pearl**. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.
9. **R. H. Shumway and D. S. Stoffer**. An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Series Analysis*, 3(4):253–264, 1982.
10. **S. Waterhouse, D.J.C. Mackay, and T. Robinson**. Bayesian methods for mixtures of experts. In *Adv. Neur. Inf. Proc. Sys. 7*. MIT Press, 1995.