Variational Methods

Zoubin Ghahramani

zoubin@cs.cmu.edu

http://www.gatsby.ucl.ac.uk

Statistical Approaches to Learning and Discovery
Carnegie Mellon University

April 2003

The Expectation Maximization (EM) algorithm

Given observed/visible variables $\mathbf{y}$, unobserved/hidden/latent/missing variables $\mathbf{x}$, and model parameters θ, **maximize the likelihood** w.r.t. θ.

$$\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) = \log \int p(\mathbf{x}, \mathbf{y}|\theta) d\mathbf{x},$$

where we have written the marginal for the visibles in terms of an integral over the joint distribution for hidden and visible variables.

Using *Jensen's inequality*, any distribution[1] over hidden variables $q(\mathbf{x})$ gives:

$$\mathcal{L}(\theta) = \log \int q(\mathbf{x}) \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{q(\mathbf{x})} d\mathbf{x} \geq \int q(\mathbf{x}) \log \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{q(\mathbf{x})} \, d\mathbf{x} = \mathcal{F}(q, \theta),$$

defining the $\mathcal{F}(q, \theta)$ functional, which is a lower bound on the log likelihood.

In the EM algorithm, we alternately optimize $\mathcal{F}(q, \theta)$ wrt q and θ, and we can prove that this will never decrease $\mathcal{L}(\theta)$.

[1] s.t. $q(\mathbf{x}) > 0$ if $p(\mathbf{x}, \mathbf{y}|\theta) > 0$.

The E and M steps of EM

The lower bound on the log likelihood:

$$\mathcal{F}(q,\theta) = \int q(\mathbf{x}) \log \frac{p(\mathbf{x},\mathbf{y}|\theta)}{q(\mathbf{x})} d\mathbf{x} = \int q(\mathbf{x}) \log p(\mathbf{x},\mathbf{y}|\theta) d\mathbf{x} + \mathcal{H}(q),$$

where $\mathcal{H}(q) = -\int q(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{x}$ is the entropy of q. We iteratively alternate:

基于分布 来max

E step: maximize $\mathcal{F}(q,\theta)$ wrt distribution over hidden variables given the parameters:

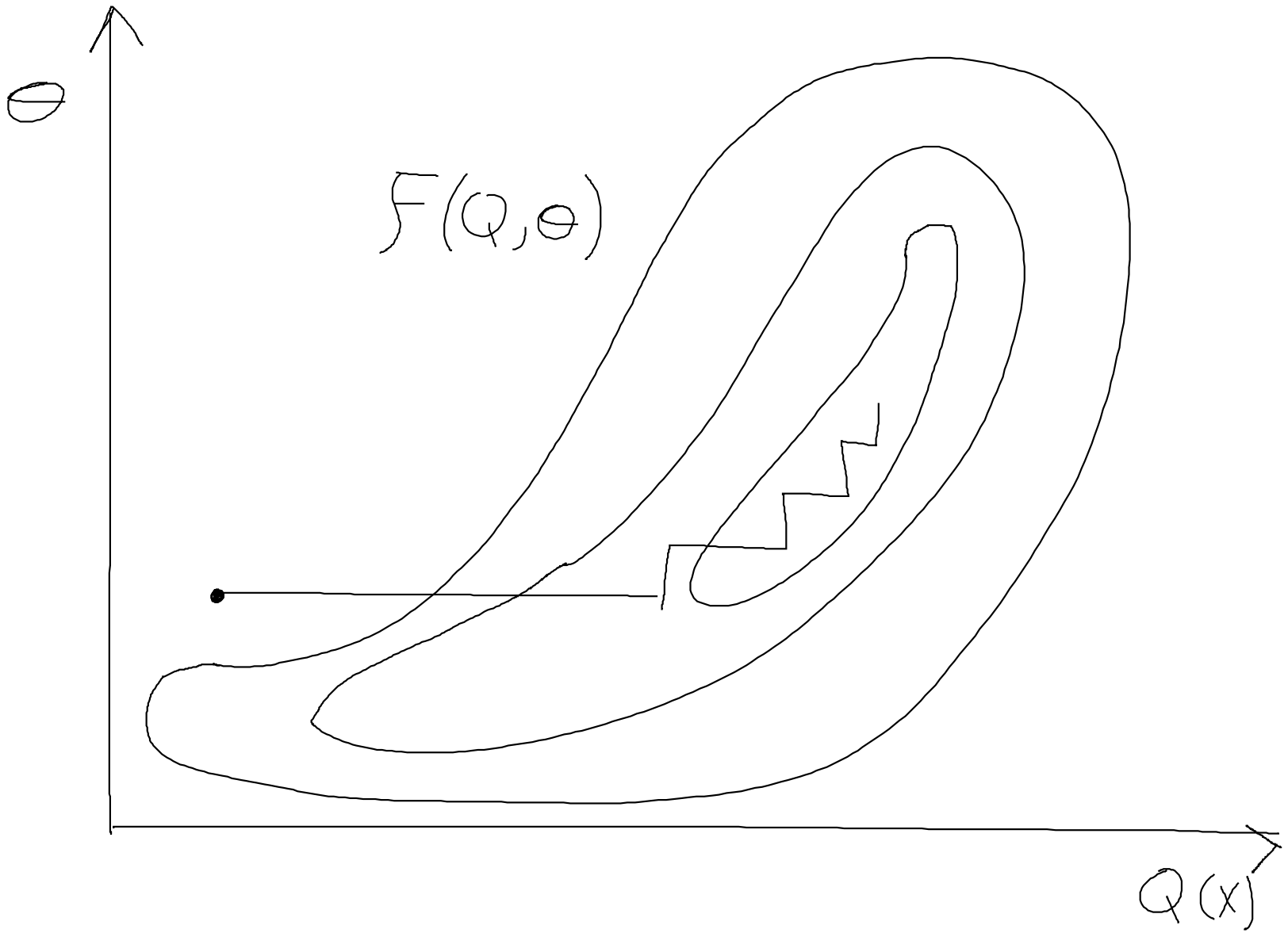
$$q^{(k)}(\mathbf{x}) := \underset{q(\mathbf{x})}{\operatorname{argmax}} \; \mathcal{F}\big(q(\mathbf{x}), \theta^{(k-1)}\big).$$

M step: maximize $\mathcal{F}(q,\theta)$ wrt the parameters given the hidden distribution:

$$\theta^{(k)} := \underset{\theta}{\operatorname{argmax}} \; \mathcal{F}\big(q^{(k)}(\mathbf{x}), \theta\big) = \underset{\theta}{\operatorname{argmax}} \; \int q^{(k)}(\mathbf{x}) \log p(\mathbf{x},\mathbf{y}|\theta) d\mathbf{x},$$

which is equivalent to optimizing the expected complete-data likelihood $p(\mathbf{x},\mathbf{y}|\theta)$, since the entropy of $q(\mathbf{x})$ does not depend on θ.

EM as Coordinate Ascent in $\mathcal{F}$



The EM algorithm never decreases the log likelihood

The difference between the log likelihood and the bound:

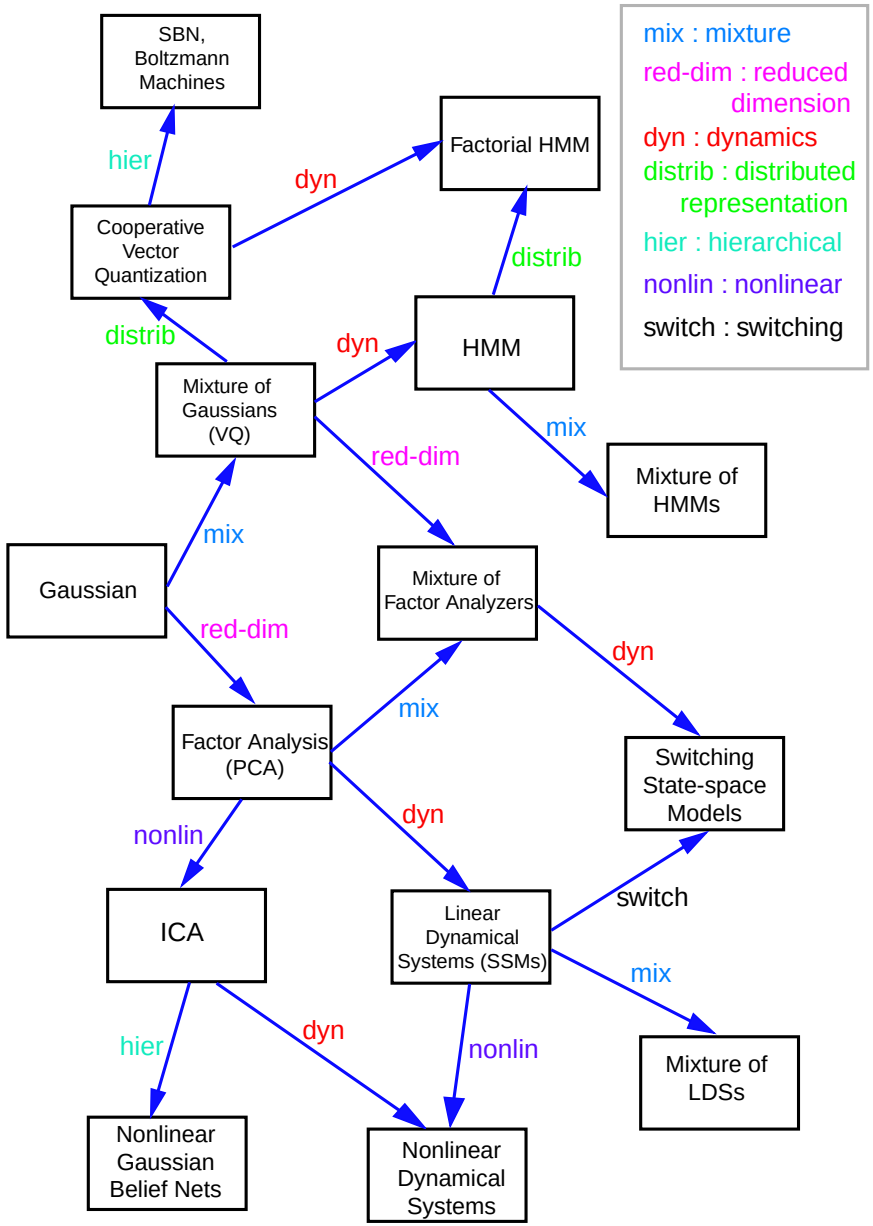
$$\begin{aligned}
\mathcal{L}(\theta) - \mathcal{F}(q,\theta) &= \log p(\mathbf{y}|\theta) - \int q(\mathbf{x}) \log \frac{p(\mathbf{x},\mathbf{y}|\theta)}{q(\mathbf{x})} d\mathbf{x} \\
&= \log p(\mathbf{y}|\theta) - \int q(\mathbf{x}) \log \frac{p(\mathbf{x}|\mathbf{y},\theta)p(\mathbf{y}|\theta)}{q(\mathbf{x})} d\mathbf{x} \\
&= -\int q(\mathbf{x}) \log \frac{p(\mathbf{x}|\mathbf{y},\theta)}{q(\mathbf{x})} d\mathbf{x} = \mathrm{KL}\big(q(\mathbf{x}), p(\mathbf{x}|\mathbf{y},\theta)\big),
\end{aligned}$$

This is the Kullback-Liebler divergence; it is non-negative and zero if and only if $q(\mathbf{x}) = p(\mathbf{x}|\mathbf{y},\theta)$ (thus this is the E step). Although we are working with a bound on the likelihood, the likelihood is non-decreasing in every iteration:

$$\mathcal{L}\big(\theta^{(k-1)}\big) \underset{\text{E step}}{=} \mathcal{F}\big(q^{(k)}, \theta^{(k-1)}\big) \underset{\text{M step}}{\leq} \mathcal{F}\big(q^{(k)}, \theta^{(k)}\big) \underset{\text{Jensen}}{\leq} \mathcal{L}\big(\theta^{(k)}\big),$$

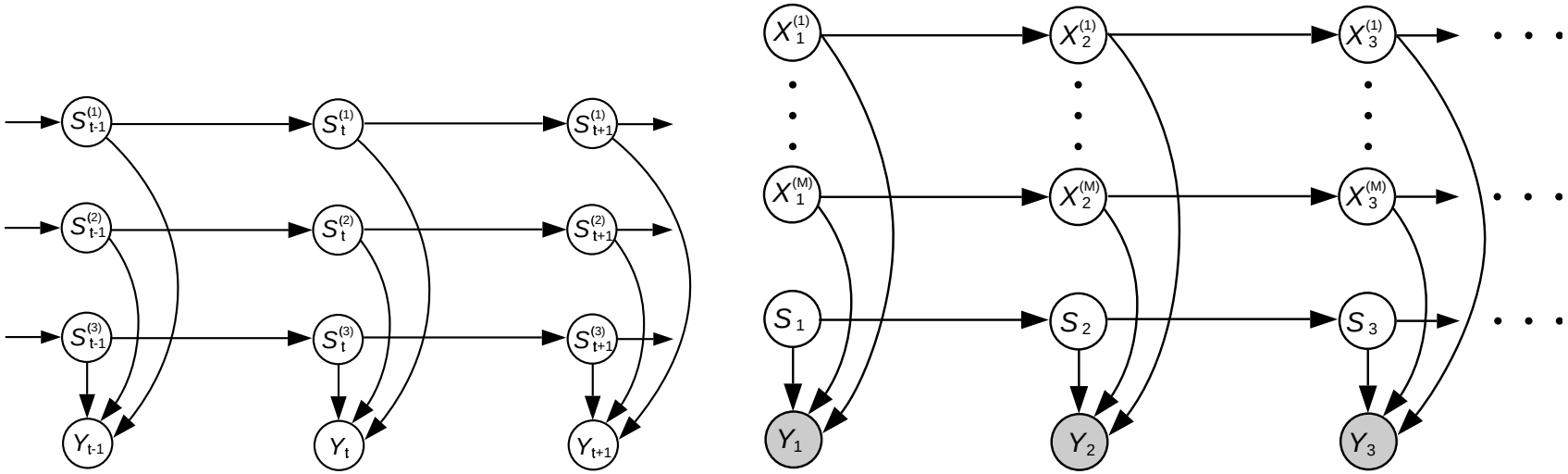
where the first equality holds because of the E step, and the first inequality comes from the M step and the final inequality from Jensen. Usually EM converges to a local optimum of $\mathcal{L}$ (although there are exceptions).

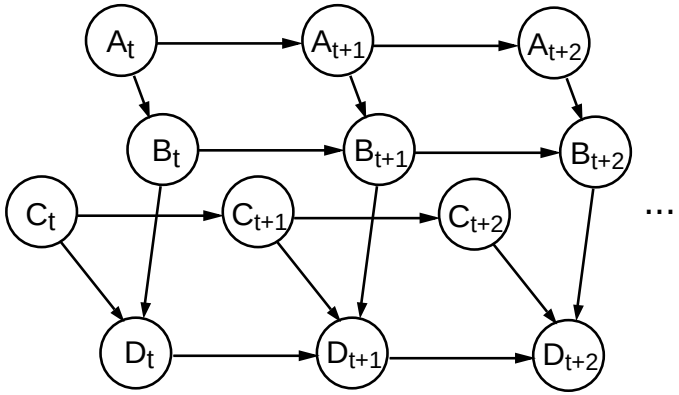
A Generative Model for Generative Models



Example: Dynamic Bayesian Networks

Factorial Hidden Markov Models, Switching State Space Models, and Nonlinear Dynamical Systems, Nonlinear Dynamical Systems, ...

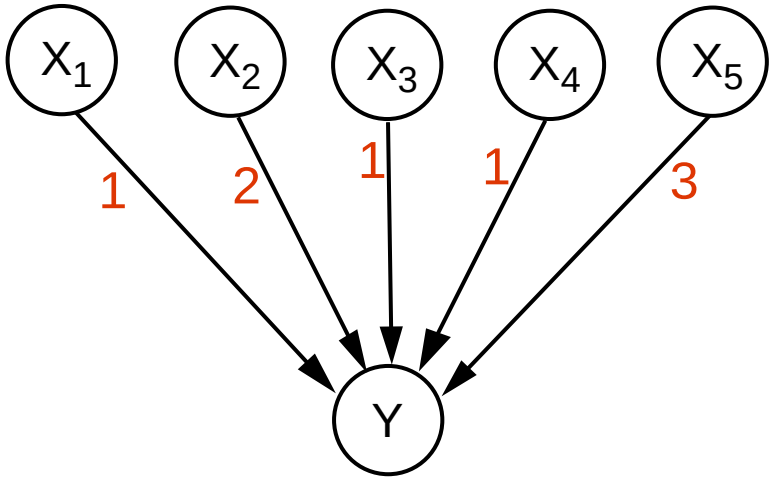




Intractability

For many models of interest, exact inference is not computationally feasible.
This occurs for two (main) reasons:

- distributions may have complicated forms (non-linearities in generative model)
- "explaining away" causes coupling from observations: observing the value of a child induces dependencies amongst its parents



$$Y = X_1 + 2\,X_2 + X_3 + X_4 + 3\,X_5$$

We can still work with such models by using *approximate inference* techniques to estimate the latent variables.

Variational Appoximations

Assume your goal is to maximize likelihood $\ln p(\mathbf{y}|\theta)$.
Any distribution $q(\mathbf{x})$ over the hidden variables defines a lower bound on $\ln p(\mathbf{y}|\theta)$:

$$\ln p(\mathbf{y}|\theta) \geq \sum_{\mathbf{x}} q(\mathbf{x}) \ln \frac{p(\mathbf{x}, \mathbf{y}|\theta)}{q(\mathbf{x})} = \mathcal{F}(q, \theta)$$

Constrain $q(\mathbf{x})$ to be of a particular *tractable* form (e.g. factorised) and maximise $\mathcal{F}$ subject to this constraint

- **E-step:** Maximise $\mathcal{F}$ w.r.t. q with θ fixed, subject to the constraint on q, equivalently minimize:

$$\ln p(\mathbf{y}|\theta) - \mathcal{F}(q, \theta) = \sum_{\mathbf{x}} q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x}|\mathbf{y}, \theta)} = \mathsf{KL}(q \| p)$$

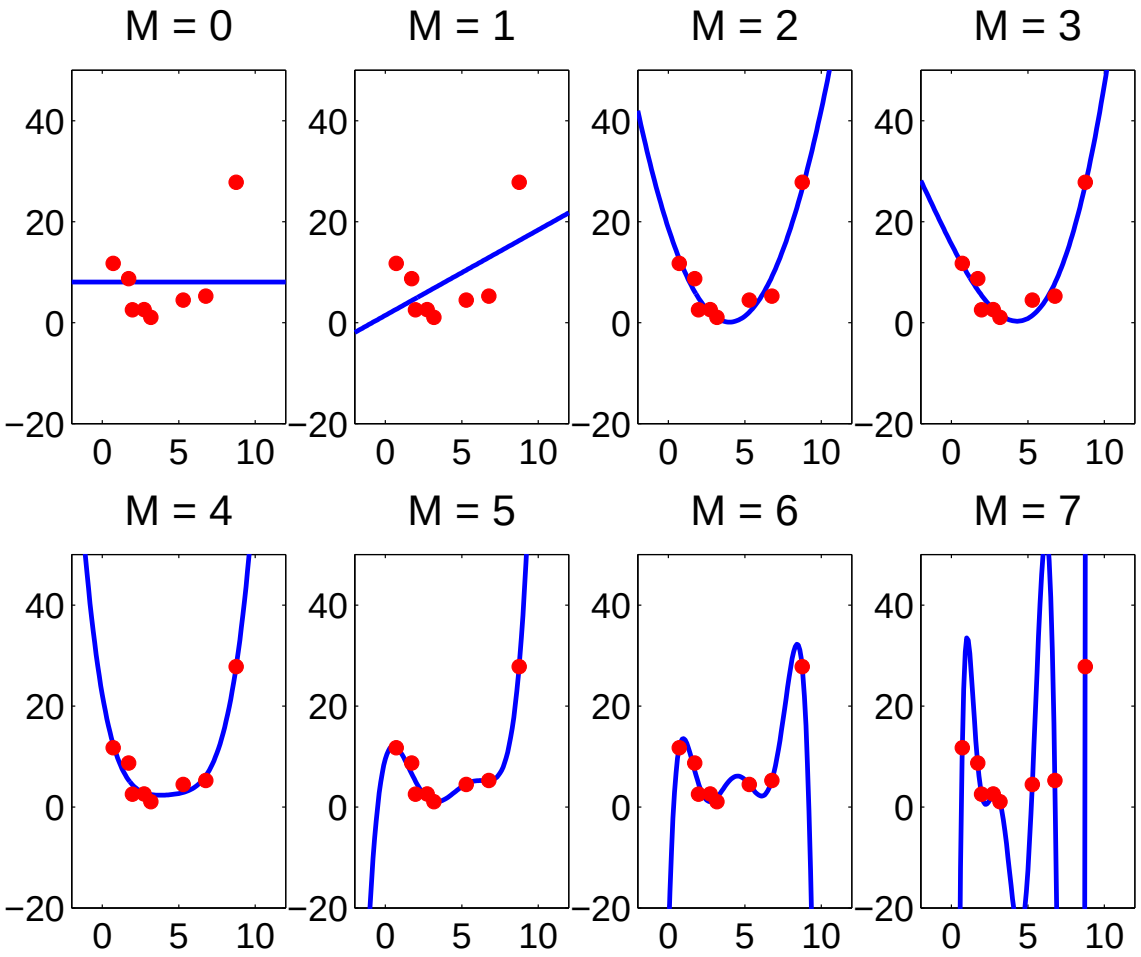
  The inference step therefore tries to find q closest to the exact posterior distribution.

- **M-step:** Maximise $\mathcal{F}$ w.r.t. θ with q fixed

(related to *mean-field approximations*)

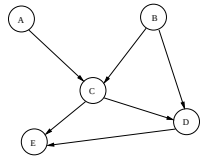
Variational Approximations for Bayesian Learning

Model structure and overfitting: a simple example



Learning Model Structure

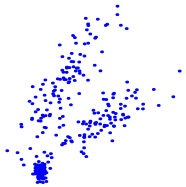
- **Conditional Independence Structure**
  What is the structure of the graph (i.e. what $\perp\!\!\!\perp$ relations hold)?



- **Feature Selection**
  Is some input relevant to predicting some output ?

- **Cardinality of Discrete Latent Variables**

  How many clusters in the data?



  How many states in a hidden Markov model?

SVYDAAAQLTADVKKDLRDSWKVIGSDKKGNGVALMTTY

- **Dimensionality of Real Valued Latent Vectors**

  What choice of dimensionality in a PCA/FA model of the data?

  How many state variables in a linear-Gaussian state-space model?

Using Bayesian Occam's Razor to Learn Model Structure

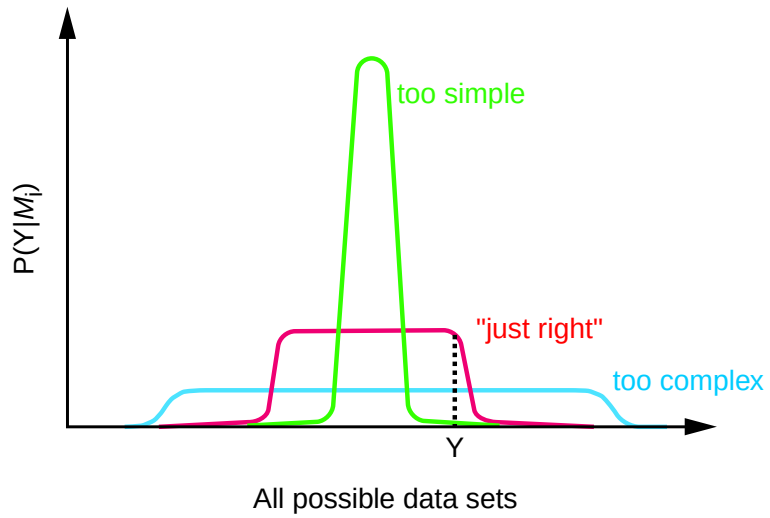
Select the model class, m, with the highest probability given the data, $\mathbf{y}$:

$$p(m|\mathbf{y}) = \frac{p(\mathbf{y}|m)p(m)}{p(\mathbf{y})}, \qquad p(\mathbf{y}|m) = \int p(\mathbf{y}|\boldsymbol{\theta}, m)\, p(\boldsymbol{\theta}|m)\, d\boldsymbol{\theta}$$

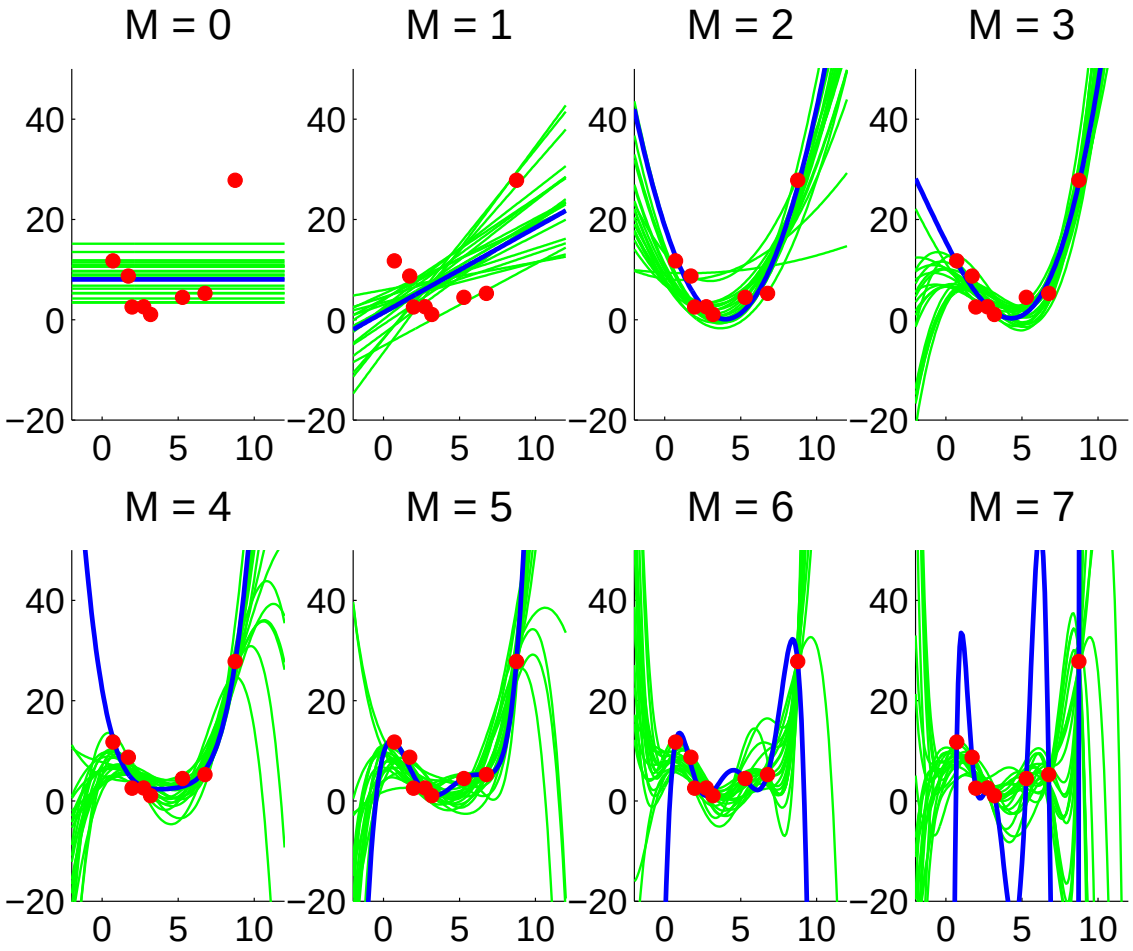
Interpretation of the marginal likelihood ("evidence"): The probability that *randomly selected* parameters from the prior would generate $\mathbf{y}$.

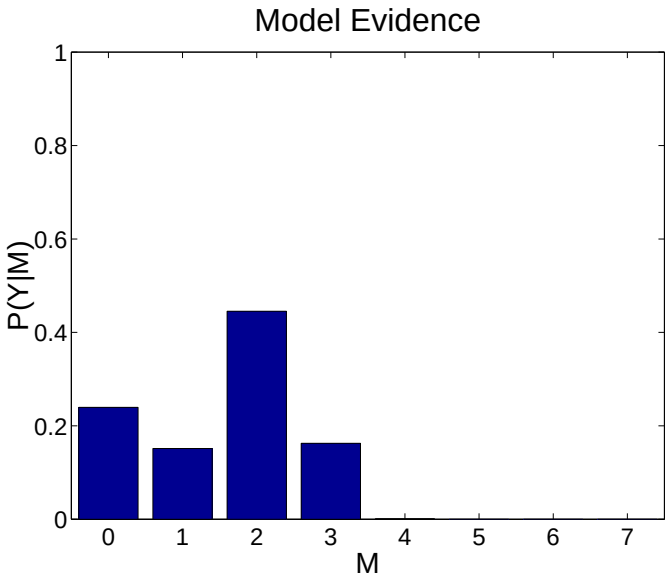
Model classes that are too simple are unlikely to generate the data set.

Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.



Bayesian Model Selection: Occam's Razor at Work





`demo: polybayes`

Subtleties of Occam's Hill

Computing Marginal Likelihoods can be Computationally Intractable

$$p(\mathbf{y}|m) = \int p(\mathbf{y}|\boldsymbol{\theta}, m)\, p(\boldsymbol{\theta}|m)\, d\boldsymbol{\theta}$$

- This can be a very **high dimensional integral**.

- The presence of **latent variables** results in additional dimensions that need to be marginalized out.

$$p(\mathbf{y}|m) = \int \int p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}, m)\, p(\boldsymbol{\theta}|m)\, d\mathbf{x}\, d\boldsymbol{\theta}$$

- The likelihood term can be **complicated**.

Practical Bayesian approaches

- Laplace approximations:
  - Appeals to asymptotic normality to make a Gaussian approximation about the posterior mode of the parameters.
- Large sample approximations (e.g. BIC).

- Markov chain Monte Carlo methods (MCMC):
  - converge to the desired distribution in the limit, but:
  - many samples are required to ensure accuracy.
  - sometimes hard to assess convergence and reliably compute marginal likelihood.
- Variational approximations...

Note: other deterministic approximations are also available now: e.g. Bethe/Kikuchi approximations, Expectation Propagation, Tree-based reparameterizations.

Lower Bounding the Marginal Likelihood

Variational Bayesian Learning

Let the latent variables be $\mathbf{x}$, data $\mathbf{y}$ and the parameters $\boldsymbol{\theta}$.
We can lower bound the marginal likelihood (by Jensen's inequality):

$$\begin{aligned}
\ln p(\mathbf{y}|m) &= \ln \int p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}|m)\, d\mathbf{x}\, d\boldsymbol{\theta} \\
&= \ln \int q(\mathbf{x}, \boldsymbol{\theta}) \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}|m)}{q(\mathbf{x}, \boldsymbol{\theta})}\, d\mathbf{x}\, d\boldsymbol{\theta} \\
&\geq \int q(\mathbf{x}, \boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}|m)}{q(\mathbf{x}, \boldsymbol{\theta})}\, d\mathbf{x}\, d\boldsymbol{\theta}.
\end{aligned}$$

Use a simpler, factorised approximation to $q(\mathbf{x}, \boldsymbol{\theta}) \approx q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$:

$$\begin{aligned}
\ln p(\mathbf{y}|m) &\geq \int q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}|m)}{q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}\, d\mathbf{x}\, d\boldsymbol{\theta} \\
&= \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y}).
\end{aligned}$$

Variational Bayesian Learning . . .

Maximizing this lower bound, $\mathcal{F}_m$, leads to **EM-like** iterative updates:

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \quad \propto \quad \exp\left[\int \ln p(\mathbf{x},\mathbf{y}|\boldsymbol{\theta},m)\, q_{\boldsymbol{\theta}}^{(t)}(\boldsymbol{\theta})\, d\boldsymbol{\theta}\right] \qquad \text{E-like step}$$

$$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) \quad \propto \quad p(\boldsymbol{\theta}|m)\ \exp\left[\int \ln p(\mathbf{x},\mathbf{y}|\boldsymbol{\theta},m)\, q_{\mathbf{x}}^{(t+1)}(\mathbf{x})\, d\mathbf{x}\right] \qquad \text{M-like step}$$

Maximizing $\mathcal{F}_m$ is equivalent to minimizing KL-divergence between the *approximate posterior*, $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})\ q_{\mathbf{x}}(\mathbf{x})$ and the *true posterior*, $p(\boldsymbol{\theta},\mathbf{x}|\mathbf{y},m)$:

$$\ln p(\mathbf{y}|m) - \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y}) = \int q_{\mathbf{x}}(\mathbf{x})\ q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{q_{\mathbf{x}}(\mathbf{x})\ q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{p(\boldsymbol{\theta},\mathbf{x}|\mathbf{y},m)}\, d\mathbf{x}\, d\boldsymbol{\theta} = \mathbf{KL}(q\|p)$$

In the limit as $n \to \infty$, for identifiable models, the variational lower bound approaches Schwartz's (1978) BIC criterion.

Conjugate-Exponential models

Let's focus on *conjugate-exponential* (**CE**) models, which satisfy **(1)** and **(2)**:

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 $\boldsymbol{\phi}(\boldsymbol{\theta})$ is the vector of *natural parameters*, $\mathbf{u}$ are *sufficient statistics*

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

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

where η and ν are hyperparameters of the prior.

Conjugate priors are computationally convenient and have an intuitive interpretation:

- η: number of pseudo-observations
- ν: values of pseudo-observations

Conjugate-Exponential examples

In the **CE** family:

- Gaussian mixtures
- factor analysis, probabilistic PCA
- hidden Markov models and factorial HMMs
- linear dynamical systems and switching 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, MRFs (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.

The Variational Bayesian EM algorithm

EM for MAP estimation

Goal: maximize $p(\boldsymbol{\theta}|\mathbf{y}, m)$ w.r.t. $\boldsymbol{\theta}$

E Step: compute

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}^{(t)})$$

M Step:

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \int q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) \, d\mathbf{x}$$

Variational Bayesian EM

Goal: lower bound $p(\mathbf{y}|m)$

VB-E Step: compute

$$q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}, \bar{\boldsymbol{\phi}}^{(t)})$$

VB-M Step:

$$q_{\boldsymbol{\theta}}^{(t+1)}(\boldsymbol{\theta}) \propto \exp\left[\int q_{\mathbf{x}}^{(t+1)}(\mathbf{x}) \ln p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) \, d\mathbf{x}\right]$$

Properties:

- Reduces to the EM algorithm if $q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$.
- $\mathcal{F}_m$ increases monotonically, and incorporates the model complexity penalty.
- Analytical parameter distributions (but not constrained to be Gaussian).
- VB-E step has same complexity as corresponding E step.
- We can use the junction tree, belief propagation, Kalman filter, etc, algorithms in the VB-E step of VB-EM, but **using expected natural parameters**, $\bar{\boldsymbol{\phi}}$.

Variational Bayesian EM

The Variational Bayesian EM algorithm has been used to approximate Bayesian learning in a wide range of models such as:

- probabilistic PCA and factor analysis
- mixtures of Gaussians and mixtures of factor analysers
- hidden Markov models
- state-space models (linear dynamical systems)
- independent components analysis (ICA) and mixtures
- discrete graphical models...

The main advantage is that it can be used to **automatically do model selection** and does not suffer from overfitting to the same extent as ML methods do.

Also it is about as computationally demanding as the usual EM algorithm.

See: `www.variational-bayes.org`

demos: `mixture of Gaussians`, `hidden Markov models`

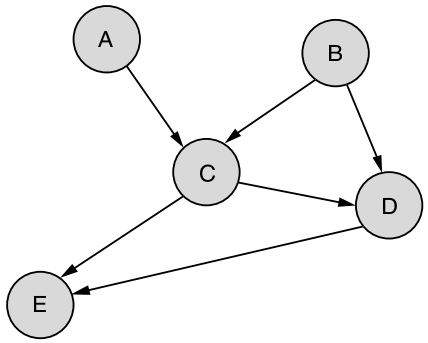
Model Selection Task

Which of the following graphical models is the data generating process?

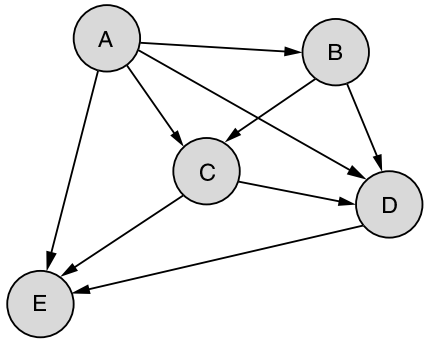
Discrete directed acyclic graphical models: data $\mathbf{y} = (A, B, C, D, E)^n$

ALL OBSERVED

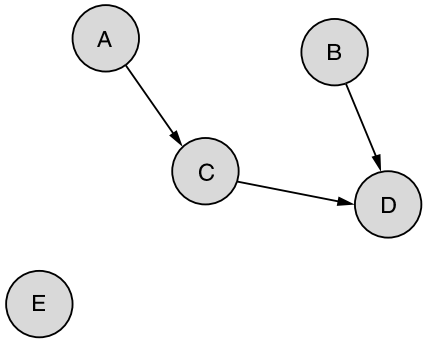
true



complex



simple

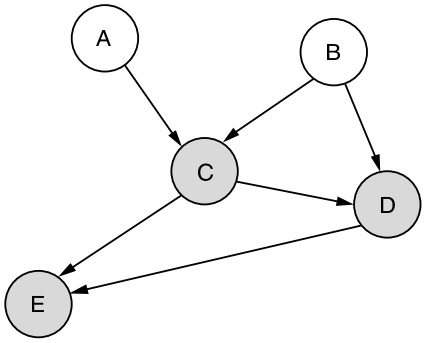


marginal likelihood tractable

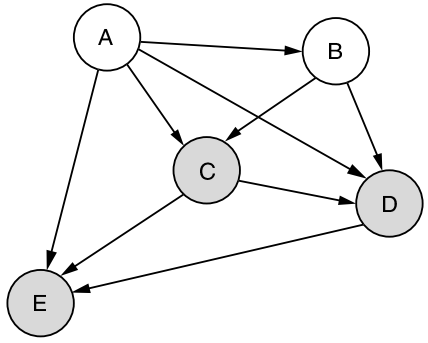
If the data are just $\mathbf{y} = (C, D, E)^n$, and (A, B) are **hidden** variables... ?

OBS.+HIDDEN

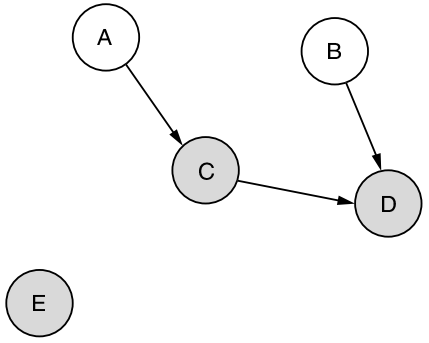
true



complex



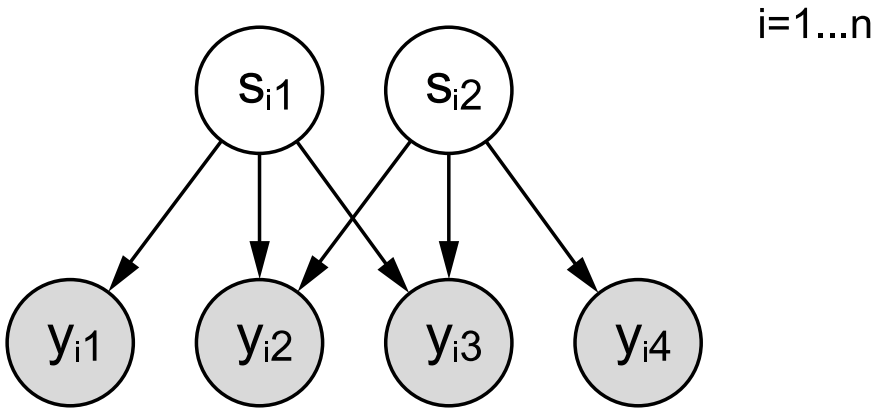
simple



marginal likelihood intractable

A thorough Case Study

- Bipartite structure: only hidden variables can be parents of observed variables.
- **Two** binary hidden variables, and **four** five-valued discrete observed variables.



- Conjugate prior is Dirichlet, Conjugate-Exponential model, so VB-EM algorithm is a straightforward modification of EM.

- **Experiment:** There are 136 distinct structures (out of 256) with 2 latent variables as potential parents of 4 conditionally independent observed vars.

- **Score** each structure for twenty varying size data sets:

$n \in \{10, 20, 40, 80, 110, 160, 230, 320, 400, 430, 480, 560, 640, 800, 960, 1120, 1280, 2560, 5120, 10240\}$

using 3 methods: BIC, VB, and the gold standard AIS

- 2720 graph scores computed, times for each: BIC (1.5s), VB (4s), AIS (400s).

Annealed Importance Sampling (AIS)

AIS is a state-of-the-art method for estimating marginal likelihoods, by breaking a difficult integral into a series of easier ones.

Combines ideas from importance sampling, Markov chain Monte Carlo, & annealing.

$$\text{Define} \qquad \mathcal{Z}_k = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}\,|\,m) p(\mathbf{y}\,|\,\boldsymbol{\theta}, m)^{\tau(k)} = \int d\boldsymbol{\theta}\; f_k(\boldsymbol{\theta})$$

with $\tau(0) = 0 \implies \mathcal{Z}_0 = \int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}\,|\,m) = 1$ ← normalisation of prior,

and $\tau(K) = 1 \implies \mathcal{Z}_K = p(\mathbf{y}\,|\,m)$ ← marginal likelihood.

$$\frac{\mathcal{Z}_K}{\mathcal{Z}_0} \equiv \frac{\mathcal{Z}_1}{\mathcal{Z}_0}\,\frac{\mathcal{Z}_2}{\mathcal{Z}_1}\,\cdots\,\frac{\mathcal{Z}_K}{\mathcal{Z}_{K-1}}$$

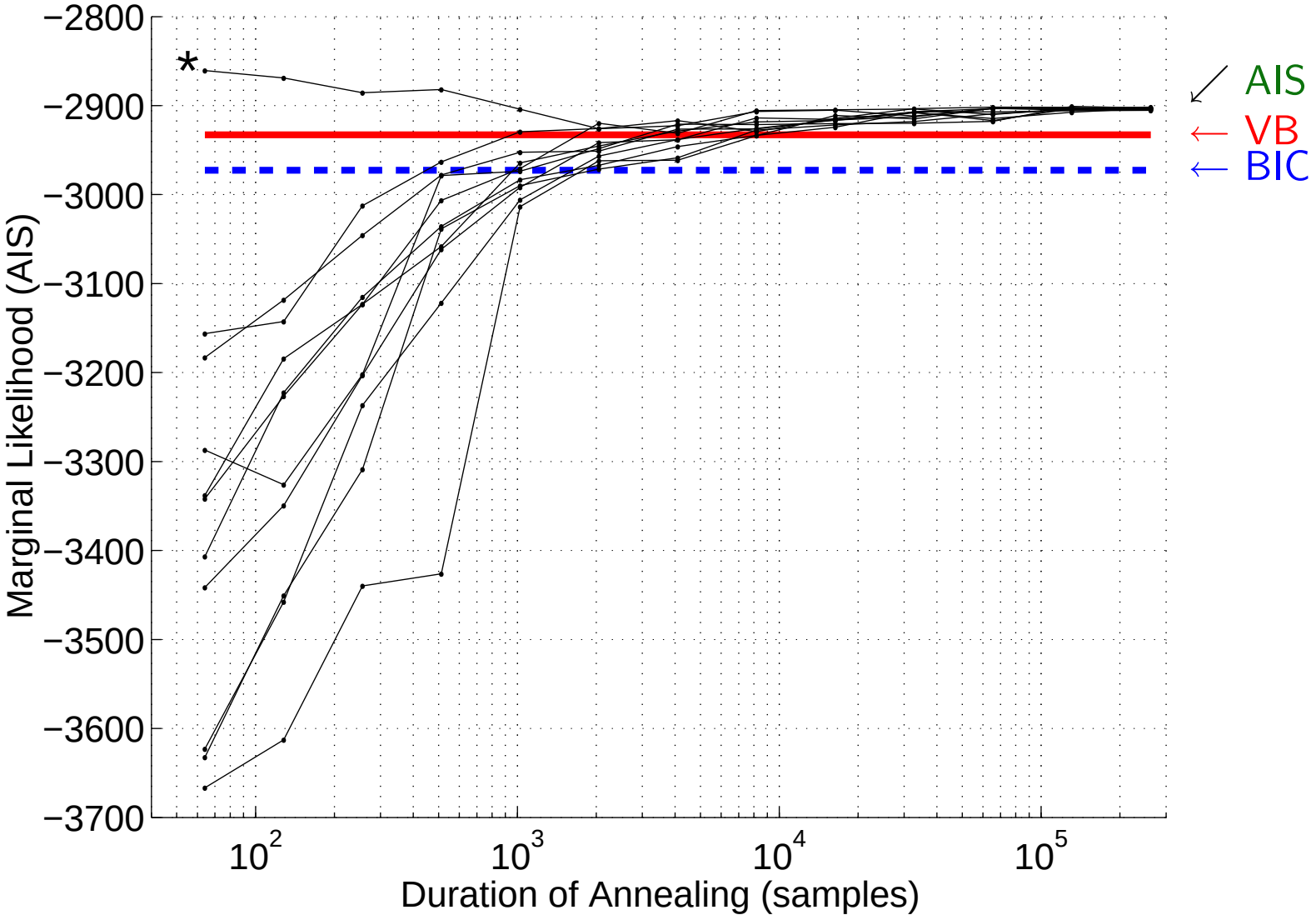
Importance sample from $f_{k-1}(\boldsymbol{\theta})$ as follows: with $\boldsymbol{\theta}^{(r)} \sim f_{k-1}(\boldsymbol{\theta})$,

$$\frac{\mathcal{Z}_k}{\mathcal{Z}_{k-1}} \equiv \int d\boldsymbol{\theta}\; \frac{f_k(\boldsymbol{\theta})}{f_{k-1}(\boldsymbol{\theta})}\frac{f_{k-1}(\boldsymbol{\theta})}{\mathcal{Z}_{k-1}} \approx \frac{1}{R}\sum_{r=1}^{R} \frac{f_k(\boldsymbol{\theta}^{(r)})}{f_{k-1}(\boldsymbol{\theta}^{(r)})} = \frac{1}{R}\sum_{r=1}^{R} p(\mathbf{y}\,|\,\boldsymbol{\theta}^{(r)}, m)^{\tau(k)-\tau(k-1)}$$

- How tight are the variational bounds? Now we have a Gold Standard.

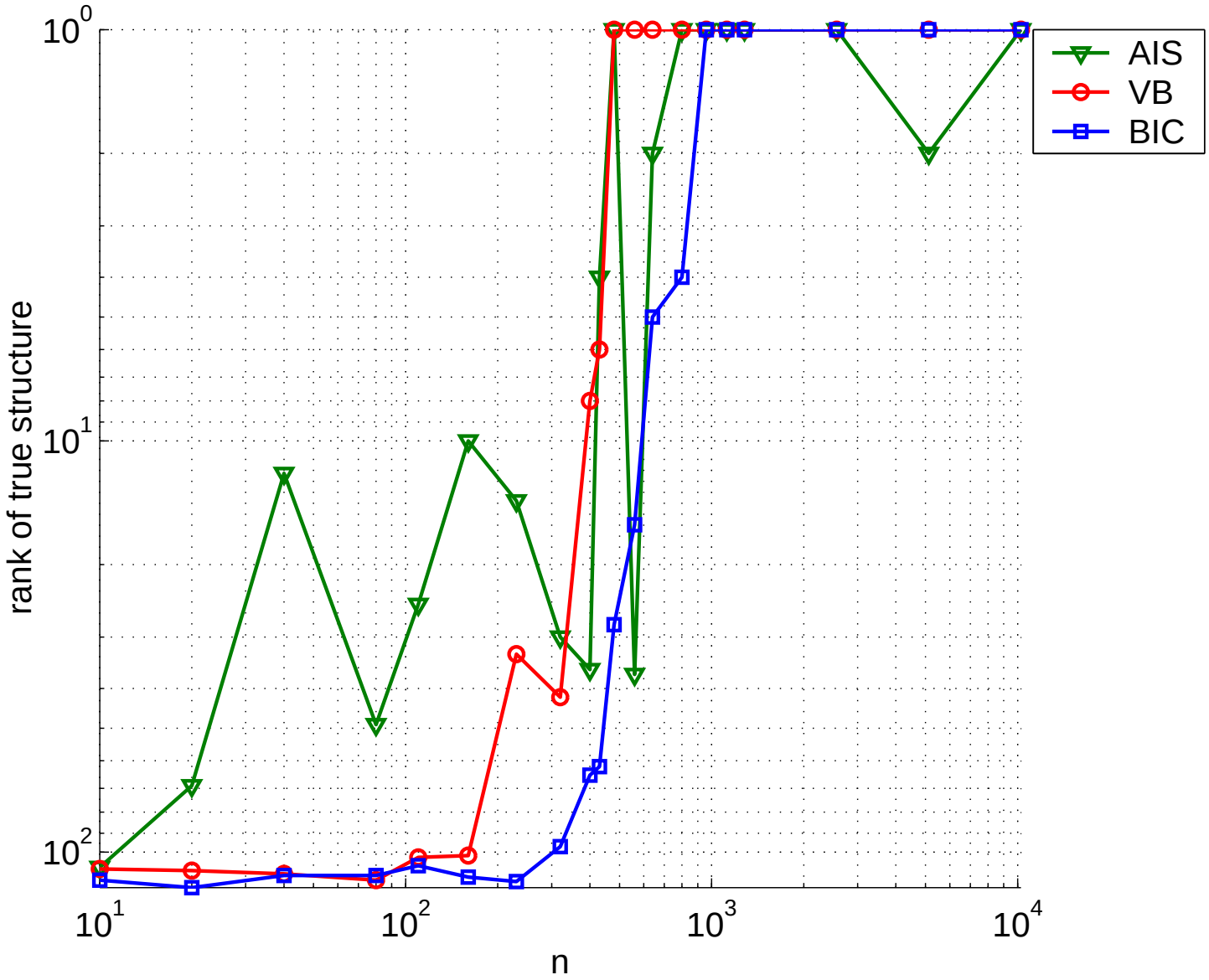
How reliable is the AIS Gold Standard?

Varying the annealing schedule with random initialisation. $n = 480, R = 2^6 \ldots 2^{18}$

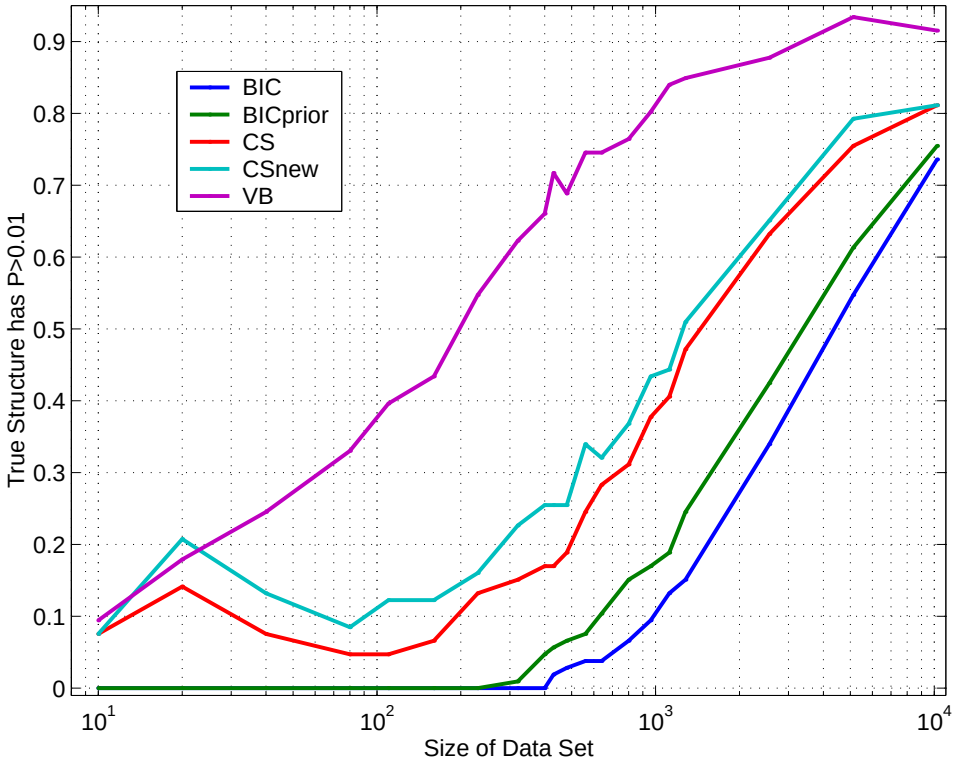


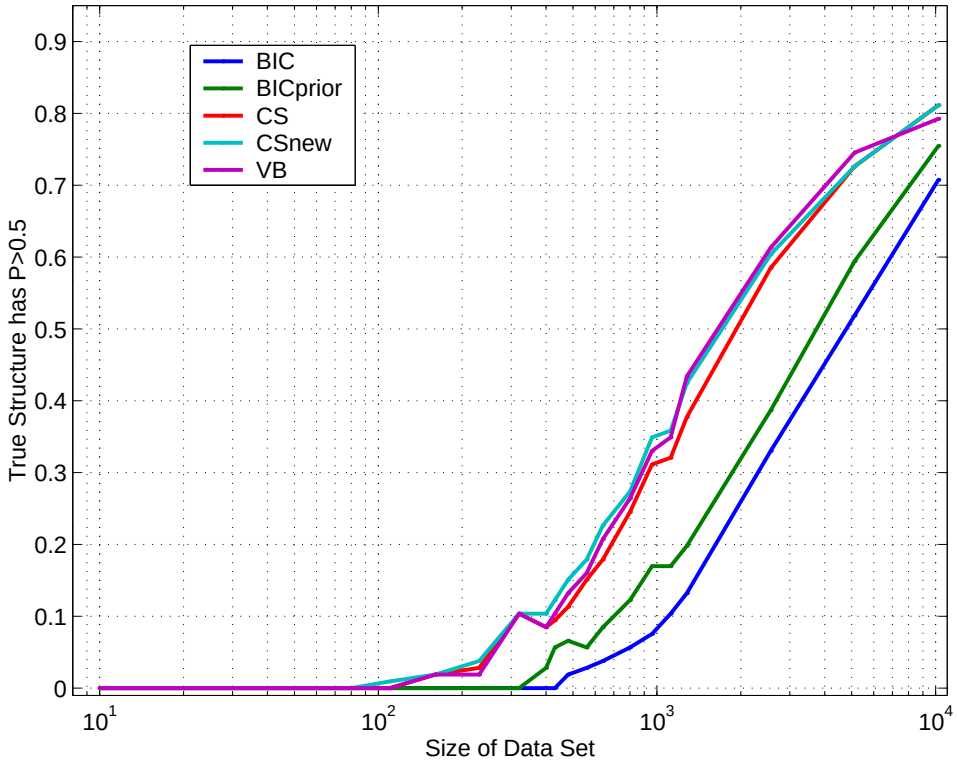
Ranking the true structure

VB score finds correct structure earlier, and more reliably



Comparison to Cheeseman-Stutz and BIC





- Averaging over about 100 samples.
- CS is much better than BIC, under some measures as good as VB.

Note: BIC and CS require estimates of the effective number of parameters. This can be difficult to compute. We estimate the effective number of parameters using a variant of the procedure described in Geiger, Heckerman and Meek (1996).

Summary and Conclusions

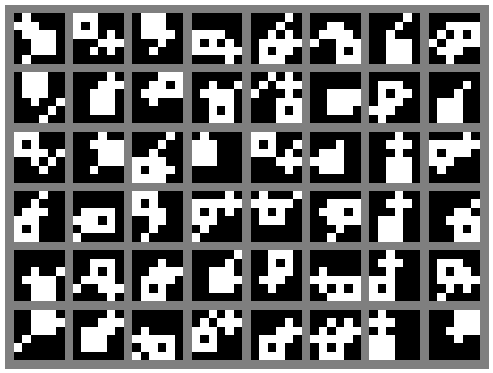
- EM can be interpreted as a lower bound maximization algorithm.
- For many models of interest the E step is intractable.
- For such models an approximate E step can be used in a variational lower bound optimization algorithm.
- This lower bound idea can also be used to do variational Bayesian learning.
- Bayesian learning embodies automatic Occam's razor via the marginal likelihood.
- This makes it possible to avoid overfitting and select models.

Appendix

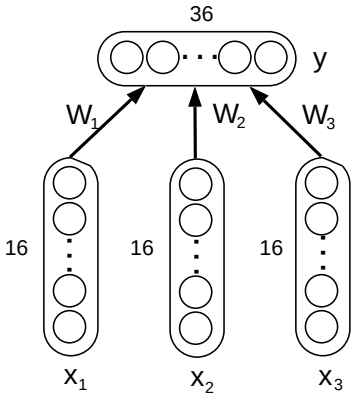
Appendix

Example: A Multiple Cause Model

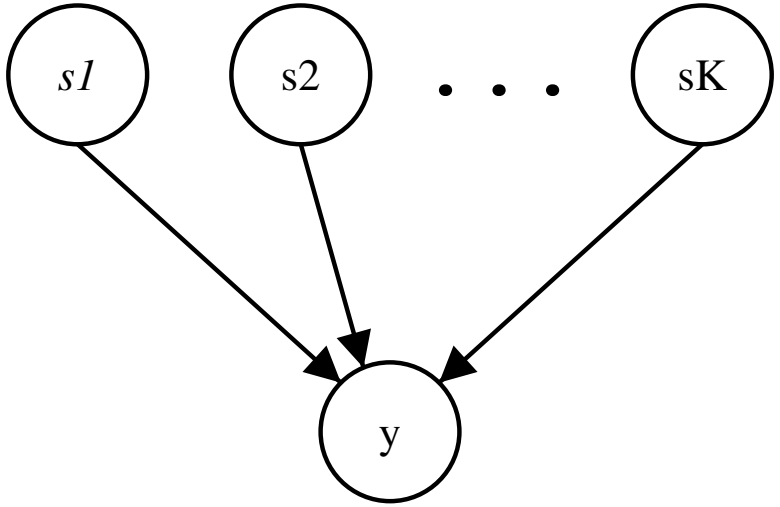
Shapes Problem

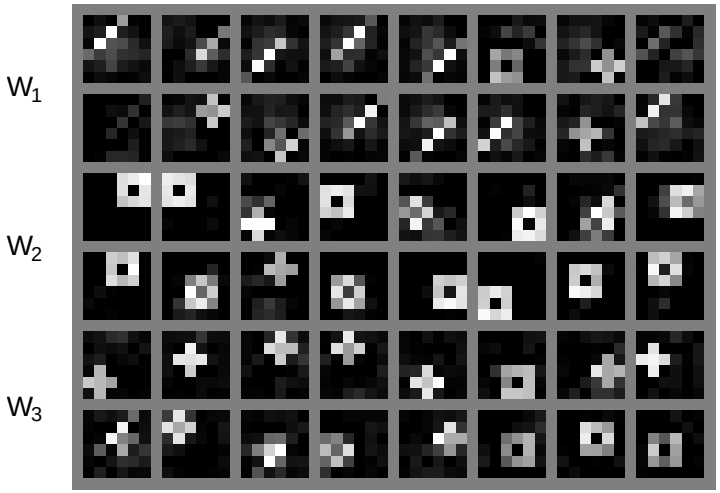


Training Data



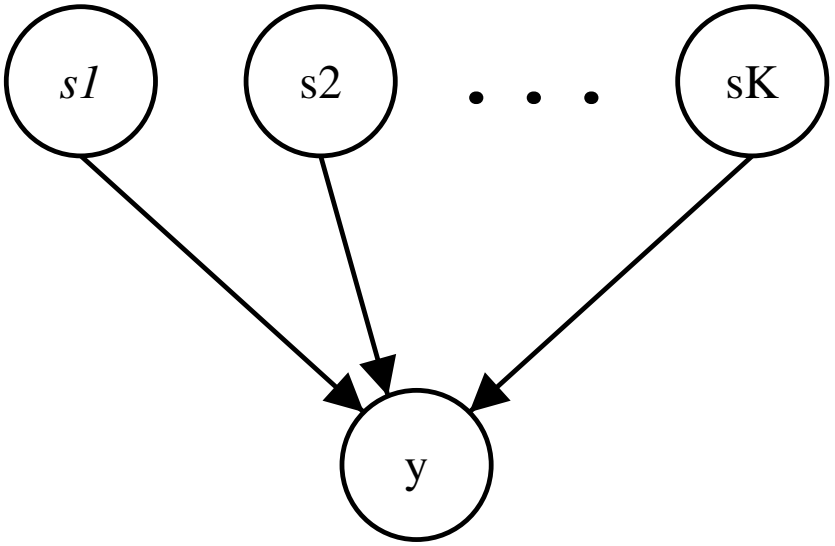
Architecture





Output Weight Matrix

Example: A Multiple Cause Model



Model with binary latent variables $s_i \in \{0, 1\}$, real-valued observed vector $\mathbf{y}$ and parameters $\boldsymbol{\theta} = \{\{\mu_i, \pi_i\}_{i=1}^K, \sigma^2\}$

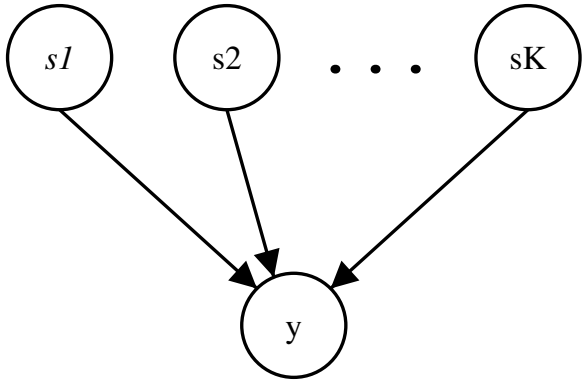
$$p(s_1, \ldots, s_K|\boldsymbol{\pi}) = \prod_{i=1}^K p(s_i|\pi_i) = \prod_{i=1}^K \pi_i^{s_i}(1-\pi_i)^{(1-s_i)}$$

$$p(\mathbf{y}|s_1, \ldots, s_K, \mu, \sigma^2) = \mathcal{N}(\sum_i s_i\mu_i, \sigma^2 I)$$

EM optimizes lower bound on likelihood: $\mathcal{F}(q, \boldsymbol{\theta}) = \langle \log p(\mathbf{s}, \mathbf{y}|\boldsymbol{\theta}) \rangle_{q(\mathbf{s})} - \langle \log q(\mathbf{s}) \rangle_{q(\mathbf{s})}$ where $\langle \rangle_q$ is expectation under q.

Optimum E step: $q(\mathbf{s}) = p(\mathbf{s}|\mathbf{y}, \boldsymbol{\theta})$ is **exponential** in K.

Example: A Multiple Cause Model (cont)



$$\mathcal{F}(q, \boldsymbol{\theta}) = \langle \log p(\mathbf{s}, \mathbf{y}|\boldsymbol{\theta}) \rangle_{q(\mathbf{s})} - \langle \log q(\mathbf{s}) \rangle_{q(\mathbf{s})} \quad (1)$$

$$\begin{aligned}
\log \quad & p(\mathbf{s}, \mathbf{y}|\boldsymbol{\theta}) + c \\
= \quad & \textstyle\sum_{i=1}^{K} s_i \log \pi_i \quad + (1 - s_i) \log(1 - \pi_i) - D \log \sigma - \frac{1}{2\sigma^2} (\mathbf{y} - \sum_i s_i \mu_i)^\top (\mathbf{y} - \sum_i s_i \mu_i) \\
= \quad & \textstyle\sum_{i=1}^{K} s_i \log \pi_i \quad + (1 - s_i) \log(1 - \pi_i) - D \log \sigma \\
& \qquad - \frac{1}{2\sigma^2} \left(\mathbf{y}^\top \mathbf{y} - 2 \sum_i s_i {\mu_i}^\top \mathbf{y} + \sum_i \sum_j s_i s_j {\mu_i}^\top \mu_j \right)
\end{aligned}$$

we therefore need $\langle s_i \rangle$ and $\langle s_i s_j \rangle$ to compute $\mathcal{F}$.

These are the expected *sufficient statistics* of the hidden variables.

Example: A Multiple Cause Model (cont)

Variational approximation:

$$q(\mathbf{s}) = \prod_i q_i(s_i) = \prod_{i=1}^{K} \lambda_i^{s_i}(1-\lambda_i)^{(1-s_i)} \tag{2}$$

Under this approximation we know $\langle s_i \rangle = \lambda_i$ and $\langle s_i s_j \rangle = \lambda_i \lambda_j + \delta_{ij}(\lambda_i - \lambda_i^2)$.

$$\begin{aligned} \mathcal{F}(\lambda, \boldsymbol{\theta}) = & \sum_i \lambda_i \log \frac{\pi_i}{\lambda_i} + (1-\lambda_i) \log \frac{(1-\pi_i)}{(1-\lambda_i)} \\ & - D \log \sigma - \frac{1}{2\sigma^2} (\mathbf{y} - \sum_i \lambda_i \mu_i)^\top (\mathbf{y} - \sum_i \lambda_i \mu_i) + C(\lambda, \mu) + c \end{aligned} \tag{3}$$

where $C(\lambda, \mu) = -\frac{1}{2\sigma^2} \sum_i (\lambda_i - \lambda_i^2) \mu_i{}^\top \mu_i$, and $c = -\frac{D}{2} \log(2\pi)$ is a constant.

Fixed point equations for multiple cause model

Taking derivatives w.r.t. λ_i:

$$\frac{\partial \mathcal{F}}{\partial \lambda_i} = \log \frac{\pi_i}{1-\pi_i} - \log \frac{\lambda_i}{1-\lambda_i} + \frac{1}{\sigma^2}(\mathbf{y} - \sum_{j \neq i} \lambda_j \mu_j)^\top \mu_i - \frac{1}{2\sigma^2} {\mu_i}^\top \mu_i \tag{4}$$

Setting to zero we get fixed point equations:

$$\lambda_i = f\left(\log \frac{\pi_i}{1-\pi_i} + \frac{1}{\sigma^2}(\mathbf{y} - \sum_{j \neq i} \lambda_j \mu_j)^\top \mu_i - \frac{1}{2\sigma^2} {\mu_i}^\top \mu_i \right) \tag{5}$$

where $f(x) = 1/(1+\exp(-x))$ is the logistic (sigmoid) function.

Learning algorithm:

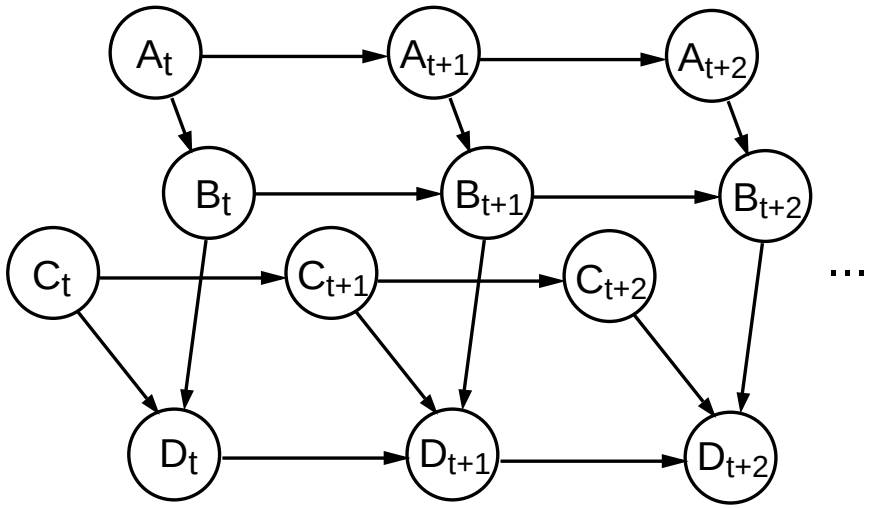
E step: run fixed point equations until convergence of λ for each data point.
M step: re-estimate $\boldsymbol{\theta}$ given λs.

Structured Variational Approximations

$q(\mathbf{s})$ need not be completely factorized.

For example, suppose you can partition $\mathbf{s}$ into sets $\mathbf{s}_1$ and $\mathbf{s}_2$ such that computing the expected sufficient statistics under $q(\mathbf{s}_1)$ and $q(\mathbf{s}_2)$ is tractable.
Then $q(\mathbf{s}) = q(\mathbf{s}_1)q(\mathbf{s}_2)$ is tractable.

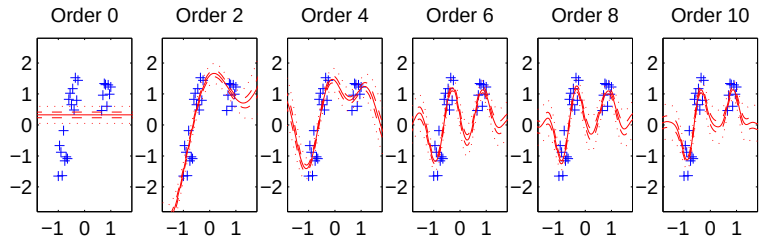
If you have a graphical model, you may want to factorize $q(\mathbf{s})$ into aproduct of trees, which are tractable distributions.

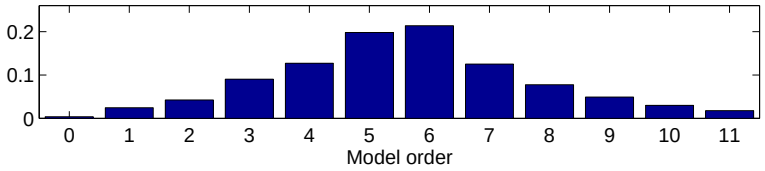


Scaling the parameter priors

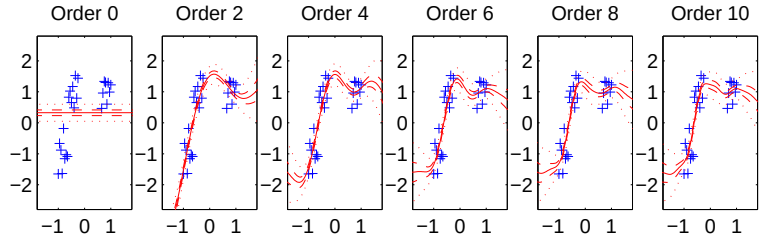
How the parameter priors are scaled determines whether an Occam's hill is present or not.

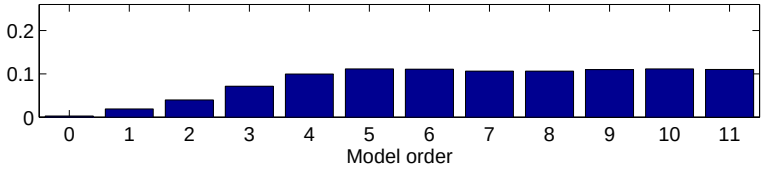
Unscaled models:





Scaled models:





(Rasmussen & Ghahramani, 2000)

back

The Cheeseman-Stutz (CS) Approximation

The Cheeseman-Stutz approximation is based on:

$$p(\mathbf{y}|m) = p(\mathbf{z}|m)\frac{p(\mathbf{y}|m)}{p(\mathbf{z}|m)} = p(\mathbf{z}|m)\frac{\int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}|m)p(\mathbf{y}|\boldsymbol{\theta},m)}{\int d\boldsymbol{\theta}\; p(\boldsymbol{\theta}'|m)p(\mathbf{z}|\boldsymbol{\theta}',m)}$$

which is true for any completion of the data: $\mathbf{z} = \{\hat{\mathbf{s}}, \mathbf{y}\}$.

We use the BIC approximation for both top and bottom integrals:

$$\begin{aligned}\ln p(\mathbf{y}|m) \approx \ln p(\hat{\mathbf{s}}, \mathbf{y}|m) + \ln p(\hat{\boldsymbol{\theta}}|m) + \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}) - \frac{d}{2}\ln n \\ - \ln p(\hat{\boldsymbol{\theta}}'|m) - \ln p(\hat{\mathbf{s}}, \mathbf{y}|\hat{\boldsymbol{\theta}}) + \frac{d'}{2}\ln n \\ = \ln p(\hat{\mathbf{s}}, \mathbf{y}|m) + \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}) - \ln p(\hat{\mathbf{s}}, \mathbf{y}|\hat{\boldsymbol{\theta}})\ ,\end{aligned}$$

This can be corrected for $d \neq d'$.

Cheeseman-Stutz: Run MAP to get $\hat{\boldsymbol{\theta}}$, complete data with expectations under $\hat{\boldsymbol{\theta}}$, compute CS approximation as above.