Probabilistic & Unsupervised Learning

Factored Variational Approximations and Variational Bayes

Maneesh Sahani
maneesh@gatsby.ucl.ac.uk

Gatsby Computational Neuroscience Unit, and
MSc ML/CSML, Dept Computer Science
University College London

Term 1, Autumn 2017

Expectations in Statistical Modelling

- **Parameter estimation**

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \int d\mathcal{Y}\; P(\mathcal{Y}|\theta) P(\mathcal{X}|\mathcal{Y},\theta)$$

(or, using EM)

$$\theta^{\text{new}} = \operatorname*{argmax}_{\theta} \int d\mathcal{Y}\; P(\mathcal{Y}|\mathcal{X},\theta^{\text{old}}) \log P(\mathcal{X},\mathcal{Y}|\theta)$$

- **Prediction**

$$p(x|\mathcal{D},m) = \int d\theta\; p(\theta|\mathcal{D},m) p(x|\theta,\mathcal{D},m)$$

- **Model selection or weighting** (by marginal likelihood)

$$p(\mathcal{D}|m) = \int d\theta\; p(\theta|m) p(\mathcal{D}|\theta,m)$$

These integrals are often intractable:

- **Analytic intractability**: integrals may not have closed form in non-linear, non-Gaussian models $\Rightarrow$ numerical integration.
- **Computational intractability**: Numerical integral (or sum if $\mathcal{Y}$ or θ are discrete) may be exponential in data or model size.

Examples of Intractability

- Marginal likelihood/model evidence for Mixture of Gaussians: exact computations are exponential in number of data points

$$p(\mathbf{x}_1,\ldots,\mathbf{x}_N) = \int d\theta\; p(\theta) \prod_{i=1}^{N} \sum_{s_i} p(\mathbf{x}_i|s_i,\theta) p(s_i|\theta)$$

$$= \sum_{s_1}\sum_{s_2}\cdots\sum_{s_N} \int d\theta\; p(\theta) \prod_{i=1}^{N} p(\mathbf{x}_i|s_i,\theta) p(s_i|\theta)$$

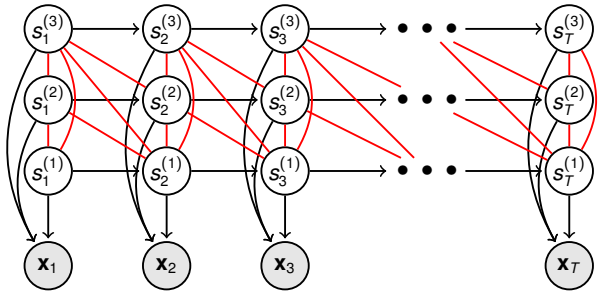
- Computing the conditional probabilities in a very large multiply-connected DAG:

$$p(x_i|X_j=a) = \sum_{\text{all settings of } \mathbf{y}\setminus\{i,j\}} p(x_i,\mathbf{y},X_j=a)/p(X_j=a)$$

- Computing the hidden state distribution in a general nonlinear dynamical system

$$p(\mathbf{y}_t|\mathbf{x}_1,\ldots,\mathbf{x}_t) \propto \int d\mathbf{y}_{t-1} p(\mathbf{y}_t|f(\mathbf{y}_{t-1})) p(\mathbf{x}_t|g(\mathbf{y}_t)) p(\mathbf{y}_{t-1}|\mathbf{x}_1,\ldots,\mathbf{x}_{t-1})$$

Distributed models



Consider an FHMM with M state variables taking on K values each.

- Moralisation puts simultaneous states $(s_t^{(1)}, s_t^{(2)}, \ldots, s_t^{(M)})$ into a single clique
- Triangulation extends cliques to size $M+1$
- Each state takes K values $\Rightarrow$ sums over K^{M+1} terms.
- Factorial *prior* $\not\Rightarrow$ Factorial *posterior* (explaining away).

Variational methods approximate the posterior, often in a factored form.
To see how they work, we need to review the free-energy interpretation of EM.

The Free Energy for a Latent Variable Model

Observed data $\mathcal{X} = \{\mathbf{x}_i\}$; Latent variables $\mathcal{Y} = \{\mathbf{y}_i\}$; Parameters θ.

Goal: Maximize the log likelihood wrt θ (i.e. ML learning):

$$\ell(\theta) = \log P(\mathcal{X}|\theta) = \log \int P(\mathcal{Y}, \mathcal{X}|\theta) d\mathcal{Y}$$

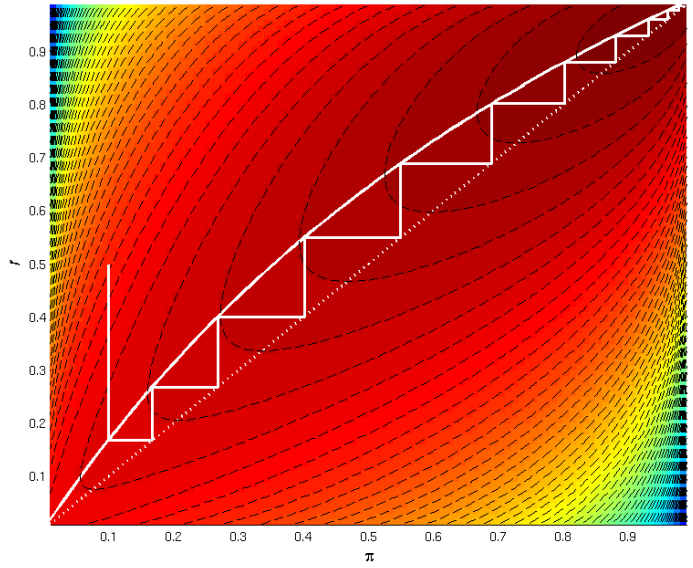
Any distribution, $q(\mathcal{Y})$, over the hidden variables can be used to obtain a lower bound on the log likelihood using Jensen's inequality:

$$\ell(\theta) = \log \int q(\mathcal{Y}) \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \, d\mathcal{Y} \geq \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \, d\mathcal{Y} \stackrel{\text{def}}{=} \mathcal{F}(q, \theta)$$

$$\begin{aligned} \int q(\mathcal{Y}) \log \frac{P(\mathcal{Y}, \mathcal{X}|\theta)}{q(\mathcal{Y})} \, d\mathcal{Y} &= \int q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) \, d\mathcal{Y} - \int q(\mathcal{Y}) \log q(\mathcal{Y}) \, d\mathcal{Y} \\ &= \int q(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}|\theta) \, d\mathcal{Y} + \mathbf{H}[q], \end{aligned}$$

where $\mathbf{H}[q]$ is the entropy of $q(\mathcal{Y})$.
So: $\mathcal{F}(q, \theta) = \langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{q(\mathcal{Y})} + \mathbf{H}[q]$

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



The E and M steps of EM

The log likelihood is bounded below by:

$$\mathcal{F}(q, \theta) = \langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{q(\mathcal{Y})} + \mathbf{H}[q] = \ell(\theta) - \mathbf{KL}[q(\mathcal{Y}) \| P(\mathcal{Y}|\mathcal{X}, \theta)]$$

EM alternates between:

E step: optimise $\mathcal{F}(q, \theta)$ wrt distribution over hidden variables holding parameters fixed:

$$q^{(k)}(\mathcal{Y}) := \underset{q(\mathcal{Y})}{\operatorname{argmax}} \; \mathcal{F}(q(\mathcal{Y}), \theta^{(k-1)}) = P(\mathcal{Y}|\mathcal{X}, \theta^{(k-1)})$$

M step: maximise $\mathcal{F}(q, \theta)$ wrt parameters holding hidden distribution fixed:

$$\theta^{(k)} := \underset{\theta}{\operatorname{argmax}} \; \mathcal{F}(q^{(k)}(\mathcal{Y}), \theta) = \underset{\theta}{\operatorname{argmax}} \; \langle \log P(\mathcal{Y}, \mathcal{X}|\theta) \rangle_{q^{(k)}(\mathcal{Y})}$$

EM Never Decreases the Likelihood

The E and M steps together never decrease the log likelihood:

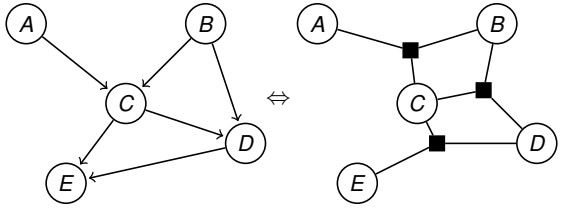
$$\ell(\theta^{(k-1)}) \underset{\text{E step}}{=} \mathcal{F}(q^{(k)}, \theta^{(k-1)}) \underset{\text{M step}}{\leq} \mathcal{F}(q^{(k)}, \theta^{(k)}) \underset{\text{Jensen}}{\leq} \ell(\theta^{(k)}),$$

- The E step brings the free energy to the likelihood.
- The M-step maximises the free energy wrt θ.
- $\mathcal{F} \leq \ell$ by Jensen – or, equivalently, from the non-negativity of KL

If the M-step is executed so that $\theta^{(k)} \neq \theta^{(k-1)}$ iff $\mathcal{F}$ increases, then the overall EM iteration will step to a new value of θ iff the likelihood increases.

Intractability

The M-step for a graphical model is usually (relatively) easy.



$$P(A,B,C,D,E) = \underbrace{P(A)P(B)P(C|A,B)}_{f_1(A,B,C)}\underbrace{P(D|B,C)}_{f_2(B,C,D)}\underbrace{P(E|C,D)}_{f_3(C,D,E)}$$

- Need expected sufficient stats from marginal posteriors on each factor group.
- Then (at least for a DAG) can optimise each factor parameter vector separately.
- Intractability in EM comes from the difficulty of computing marginal posteriors in graphs with large tree-width or non-linear/non-conjugate conditionals.
- [For non-DAG models, partition function (normalising constant) may also be intractable.]

What do we lose?

What does restricting q to $\mathcal{Q}$ cost us?

- Recall that the free-energy is bounded above by Jensen:

  $$\mathcal{F}(q,\theta) \leq \ell(\theta^{\mathrm{ML}})$$

  Thus, as long as every step increases $\mathcal{F}$, convergence is still guaranteed.

- But, since $P(\mathcal{Y}|\mathcal{X},\theta^{(k)})$ may not lie in $\mathcal{Q}$, we no longer saturate the bound after the E-step. Thus, the likelihood may not increase on each full EM step.

  $$\ell(\theta^{(k-1)}) \underset{\text{E step}}{\not\geq} \mathcal{F}(q^{(k)},\theta^{(k-1)}) \underset{\text{M step}}{\leq} \mathcal{F}(q^{(k)},\theta^{(k)}) \underset{\text{Jensen}}{\leq} \ell(\theta^{(k)}),$$

- This means we may not converge to a maximum of ℓ.

The hope is that by *increasing a lower bound* on ℓ we will find a decent solution.
[Note that if $P(\mathcal{Y}|\mathcal{X},\theta^{\mathrm{ML}}) \in \mathcal{Q}$, then θ^{ML} is a fixed point of the variational algorithm.]

Free-energy-based variational approximation

What if finding expected sufficient stats under $P(\mathcal{Y}|\mathcal{X},\theta)$ is computationally intractable?

For the **generalised EM** algorithm, we argued that intractable maximisations could be replaced by gradient M-steps.

- Each step increases the likelihood.
- A fixed point of the gradient M-step must be at a mode of the expected log-joint.

For the E-step we could:

- Parameterise $q = q_\rho(\mathcal{Y})$ and take a gradient step in ρ.
- Assume some simplified form for q, usually factored: $q = \prod_i q_i(\mathcal{Y}_i)$ where $\mathcal{Y}_i$ partition $\mathcal{Y}$, and maximise within this form.

In either case, we choose q from within a limited set $\mathcal{Q}$:

VE step: maximise $\mathcal{F}(q,\theta)$ wrt constrained latent distribution given parameters:

$$q^{(k)}(\mathcal{Y}) := \underset{q(\mathcal{Y})\in\mathcal{Q}\leftarrow\textbf{Constraint}}{\operatorname{argmax}} \mathcal{F}(q(\mathcal{Y}),\theta^{(k-1)}).$$

M step: unchanged

$$\theta^{(k)} := \underset{\theta}{\operatorname{argmax}}\ \mathcal{F}(q^{(k)}(\mathcal{Y}),\theta) = \underset{\theta}{\operatorname{argmax}} \int q^{(k)}(\mathcal{Y}) \log p(\mathcal{Y},\mathcal{X}|\theta) d\mathcal{Y},$$

Unlike in GEM, the fixed point may not be at an unconstrained optimum of $\mathcal{F}$.

KL divergence

Recall that

$$\begin{aligned}\mathcal{F}(q,\theta) &= \langle \log P(\mathcal{X},\mathcal{Y}|\theta)\rangle_{q(\mathcal{Y})} + \mathbf{H}[q] \\ &= \langle \log P(\mathcal{X}|\theta) + \log P(\mathcal{Y}|\mathcal{X},\theta)\rangle_{q(\mathcal{Y})} - \langle \log q(\mathcal{Y})\rangle_{q(\mathcal{Y})} \\ &= \langle \log P(\mathcal{X}|\theta)\rangle_{q(\mathcal{Y})} - \mathbf{KL}[q\|P(\mathcal{Y}|\mathcal{X},\theta)].\end{aligned}$$

Thus,

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

$$q^{(k)}(\mathcal{Y}) := \underset{q(\mathcal{Y})\in\mathcal{Q}}{\operatorname{argmax}} \mathcal{F}(q(\mathcal{Y}),\theta^{(k-1)}).$$

is equivalent to:

E step minimise $\mathbf{KL}[q\|p(\mathcal{Y}|\mathcal{X},\theta)]$ wrt distribution over latents, given parameters:

$$q^{(k)}(\mathcal{Y}) := \underset{q(\mathcal{Y})\in\mathcal{Q}}{\operatorname{argmin}} \int q(\mathcal{Y}) \log \frac{q(\mathcal{Y})}{p(\mathcal{Y}|\mathcal{X},\theta^{(k-1)})} d\mathcal{Y}$$

So, in each E step, the algorithm is trying to find the best approximation to $P(\mathcal{Y}|\mathcal{X})$ in $\mathcal{Q}$ in a KL sense. This is related to ideas in *information geometry*. It also suggests generalisations to other distance measures.

Factored Variational E-step

The most common form of variational approximation partitions $\mathcal{Y}$ into disjoint sets $\mathcal{Y}_i$ with

$$\mathcal{Q} = \{q \mid q(\mathcal{Y}) = \prod_i q_i(\mathcal{Y}_i)\}.$$

In this case the E-step is itself iterative:

(Factored VE step)$_i$: maximise $\mathcal{F}(q, \theta)$ wrt $q_i(\mathcal{Y}_i)$ given other q_j and parameters:

$$q_i^{(k)}(\mathcal{Y}_i) := \operatorname*{argmax}_{q_i(\mathcal{Y}_i)} \mathcal{F}\Big(q_i(\mathcal{Y}_i)\prod_{j\neq i} q_j(\mathcal{Y}_j), \theta^{(k-1)}\Big).$$

- q_i updates iterated to convergence to "complete" VE-step.
- In fact, every (VE)$_i$-step separately increases $\mathcal{F}$, so any schedule of (VE)$_i$- and M-steps will converge. Choice can be dictated by practical issues (rarely efficient to fully converge E-step before updating parameters).

Factored Variational E-step

The Factored Variational E-step has a general form.

The free energy is:

$$\begin{aligned}\mathcal{F}\Big(\prod_j q_j(\mathcal{Y}_j), \theta^{(k-1)}\Big) &= \Big\langle \log P(\mathcal{X}, \mathcal{Y}|\theta^{(k-1)})\Big\rangle_{\prod_j q_j(\mathcal{Y}_j)} + \mathbf{H}\Big[\prod_j q_j(\mathcal{Y}_j)\Big] \\ &= \int d\mathcal{Y}_i \; q_i(\mathcal{Y}_i)\Big\langle \log P(\mathcal{X}, \mathcal{Y}|\theta^{(k-1)})\Big\rangle_{\prod_{j\neq i} q_j(\mathcal{Y}_j)} + \mathbf{H}[q_i] + \sum_{j\neq i}\mathbf{H}[q_j]\end{aligned}$$

Now, taking the variational derivative of the Lagrangian (enforcing normalisation of q_i):

$$\frac{\delta}{\delta q_i}\left(\mathcal{F} + \lambda\left(\int q_i - 1\right)\right) = \Big\langle \log P(\mathcal{X}, \mathcal{Y}|\theta^{(k-1)})\Big\rangle_{\prod_{j\neq i} q_j(\mathcal{Y}_j)} - \log q_i(\mathcal{Y}_i) - \frac{q_i(\mathcal{Y}_i)}{q_i(\mathcal{Y}_i)} + \lambda$$

$$(=0) \quad \Rightarrow \quad q_i(\mathcal{Y}_i) \propto \exp\Big\langle \log P(\mathcal{X}, \mathcal{Y}|\theta^{(k-1)})\Big\rangle_{\prod_{j\neq i} q_j(\mathcal{Y}_j)}$$

In general, this depends only on the expected sufficient statistics under q_j. Thus, again, we don't actually need the *entire* distributions, just the relevant expectations (now for approximate inference as well as learning).

Mean-field approximations

If $\mathcal{Y}_i = y_i$ (*i.e.*, q is factored over all variables) then the variational technique is often called a "mean field" approximation.

- Suppose $P(\mathcal{X}, \mathcal{Y})$ has sufficient statistics that are separable in the latent variables: *e.g.* the Boltzmann machine

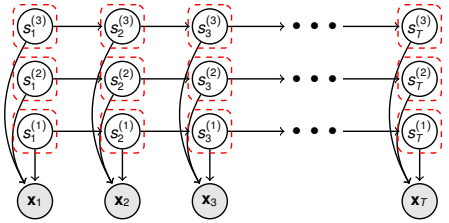
$$P(\mathcal{X}, \mathcal{Y}) = \frac{1}{Z}\exp\Big(\sum_{ij} W_{ij}s_i s_j + \sum_i b_i s_i\Big)$$

  with some $s_i \in \mathcal{Y}$ and others observed.
- Expectations wrt a fully-factored q distribute over all $s_i \in \mathcal{Y}$

$$\langle \log P(\mathcal{X}, \mathcal{Y})\rangle_{\prod q_i} = \sum_{ij} W_{ij}\langle s_i\rangle_{q_i}\langle s_j\rangle_{q_j} + \sum_i b_i\langle s_i\rangle_{q_i}$$

  (where q_i for $s_i \in \mathcal{X}$ is a delta function on the observed value).
- Thus, we can update each q_i in turn given the means (or, in general, mean sufficient statistics) of the others.
- Each variable sees the mean field imposed by its neighbours, and we update these fields until they all agree.

Mean-field FHMM



$$q(s_{1:T}^{1:M}) = \prod_{m,t} q_t^m(s_t^m)$$

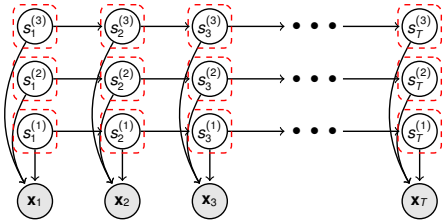
$$q_t^m(s_t^m) \propto \exp\Big\langle \log P(\mathbf{s}_{1:T}^{1:M}, \mathbf{x}_{1:T})\Big\rangle_{\prod_{\neg(m,t)} q_{t'}^{m'}(s_{t'}^{m'})}$$

$$= \exp\left\langle \sum_\mu\sum_\tau \log P(s_\tau^\mu|s_{\tau-1}^\mu) + \sum_\tau \log P(\mathbf{x}_\tau|s_\tau^{1:M})\right\rangle_{\prod_{\neg(m,t)} q_{t'}^{m'}}$$

$$\propto \exp\Big[\underbrace{\big\langle \log P(s_t^m|s_{t-1}^m)\big\rangle_{q_{t-1}^m} + \big\langle \log P(\mathbf{x}_t|s_t^{1:M})\big\rangle_{\prod_{\neg m} q_t^{m'}}}_{\alpha_t^m(i) \propto e^{\sum_j \log \Phi_{ji}^m q_{t-1}^m(j)} \cdot e^{\langle \log A_i(\mathbf{x}_t)\rangle_{q_t^{\neg m}}}} + \underbrace{\big\langle \log P(s_{t+1}^m|s_t^m)\big\rangle_{q_{t+1}^m}}_{\beta_t^m(i) \propto e^{\sum_j \log \Phi_{ij}^m q_{t+1}^m(j)}}\Big]$$

Cf. forward-backward: $\alpha_t(i) \propto \sum_j \alpha_{t-1}(j)\Phi_{ji}\cdot A_i(\mathbf{x}_t)$ $\qquad$ $\beta_t(i) \propto \sum_j \Phi_{ij}A_j(\mathbf{x}_{t+1})\beta_{t+1}(j)$

Mean-field FHMM



$$q(s_{1:T}^{1:M}) = \prod_{m,t} q_t^m(s_t^m)$$

$$q_t^m(s_t^m) \propto \exp\Big[\underbrace{\big\langle \log P(s_t^m|s_{t-1}^m)\big\rangle_{q_{t-1}^m} + \big\langle \log P(\mathbf{x}_t|s_t^{1:M})\big\rangle_{\prod_{\neg m} q_t^{m'}}}_{\alpha_t^m(i) \propto e^{\sum_j \log \Phi_{ji}^m q_{t-1}^m(j)} \cdot e^{\langle \log A_i(\mathbf{x}_t)\rangle_{q_t^{\neg m}}}} + \underbrace{\big\langle \log P(s_{t+1}^m|s_t^m)\big\rangle_{q_{t+1}^m}}_{\beta_t^m(i) \propto e^{\sum_j \log \Phi_{ij}^m q_{t+1}^m(j)}}\Big]$$

Cf. forward-backward: $\alpha_t(i) \propto \sum_j \alpha_{t-1}(j)\Phi_{ji}\cdot A_i(\mathbf{x}_t)$ $\quad$ $\beta_t(i) \propto \sum_j \Phi_{ij} A_j(\mathbf{x}_{t+1})\beta_{t+1}(j)$

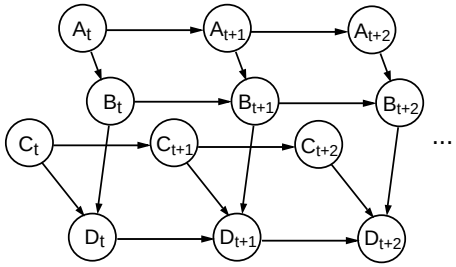
- Yields a message-passing algorithm like forward-backward
- Updates depend only on immediate neighbours in chain
- Chains couple only through joint output
- Multiple passes; messages depend on (approximate) marginals
- Evidence does not appear explicitly in backward message (cf Kalman smoothing)

Structured variational approximation

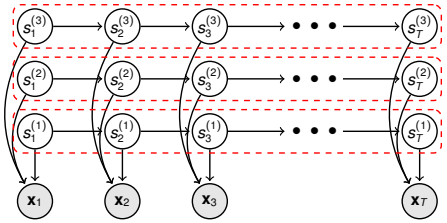
- $q(\mathcal{Y})$ need not be completely factorized.
- For example, suppose $\mathcal{Y}$ can be partitioned into sets $\mathcal{Y}_1$ and $\mathcal{Y}_2$ such that computing the expected sufficient statistics under $P(\mathcal{Y}_1|\mathcal{Y}_2,\mathcal{X})$ and $P(\mathcal{Y}_2|\mathcal{Y}_1,\mathcal{X})$ would be tractable.

$\Rightarrow$ Then the factored approximation $q(\mathcal{Y}) = q(\mathcal{Y}_1)q(\mathcal{Y}_2)$ is tractable.

- In particular, any factorisation of $q(\mathcal{Y})$ into a product of distributions on trees, yields a tractable approximation.



Stuctured FHMM



For the FHMM we can factor the chains:

$$q(s_{1:T}^{1:M}) = \prod_m q^m(s_{1:T}^m)$$

$$q^m(s_{1:T}^m) \propto \exp\Big\langle \log P(\mathbf{s}_{1:T}^{1:M}, \mathbf{x}_{1:T})\Big\rangle_{\prod_{\neg m} q^{m'}(s_{1:T}^{m'})}$$

$$= \exp\Big\langle \sum_\mu \sum_t \log P(s_t^\mu|s_{t-1}^\mu) + \sum_t \log P(\mathbf{x}_t|s_t^{1:M})\Big\rangle_{\prod_{\neg m} q^{m'}}$$

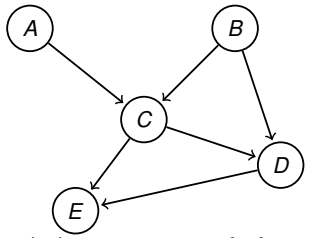
$$\propto \exp\Big[\sum_t \log P(s_t^m|s_{t-1}^m) + \sum_t \Big\langle \log P(\mathbf{x}_t|s_t^{1:M})\Big\rangle_{\prod_{\neg m} q^{m'}(s_t^{m'})}\Big]$$

$$= \prod_t P(s_t^m|s_{t-1}^m) \prod_t e^{\big\langle \log P(\mathbf{x}_t|s_t^{1:M})\big\rangle_{\prod_{\neg m} q^{m'}(s_t^{m'})}}$$

This looks like a standard HMM joint, with a modified likelihood term $\Rightarrow$ cycle through multiple forward-backward passes, updating likelihood terms each time.

Messages on an arbitrary graph

Consider a DAG:



$$P(\mathcal{X},\mathcal{Y}) = \prod_k P(Z_k|\,\mathrm{pa}(Z_k))$$

and let $q(\mathcal{Y}) = \prod_i q_i(\mathcal{Y}_i)$ for disjoint sets $\{\mathcal{Y}_i\}$.

We have that the VE update for q_i is given by $q_i^*(\mathcal{Y}_i) \propto \exp\langle \log p(\mathcal{Y},\mathcal{X})\rangle_{q_{\neg i}(\mathcal{Y})}$ where $\langle\cdot\rangle_{q_{\neg i}(\mathcal{Y})}$ denotes averaging with respect to $q_j(\mathcal{Y}_j)$ for all $j \neq i$

Then:

$$\log q_i^*(\mathcal{Y}_i) = \Big\langle \sum_k \log P(Z_k|\,\mathrm{pa}(Z_k))\Big\rangle_{q_{\neg i}(\mathcal{Y})} + \text{const}$$

$$= \sum_{j\in\mathcal{Y}_i} \langle \log P(Y_j|\,\mathrm{pa}(Y_j))\rangle_{q_{\neg i}(\mathcal{Y})} + \sum_{j\in \mathrm{ch}(\mathcal{Y}_i)} \langle \log P(Z_j|\,\mathrm{pa}(Z_j))\rangle_{q_{\neg i}(\mathcal{Y})} + \text{const}$$

This defines messages that are passed between nodes in the graph. Each node receives messages from its Markov boundary: parents, children and parents of children (all neighbours in the corresponding factor graph).

Non-factored variational methods

The term variational approximation is used whenever a bound on the likelihood (or on another estimation cost function) is optimised, but does not necessarily become tight.

Many further variational approximations have been developed, including:

- parametric forms (e.g. Gaussian) for non-linear models
- non-free-energy-based bounds (both upper and lower) on the likelihood.

We can also see MAP- or zero-temperature EM and recognition models as parametric forms of variational inference.

Variational methods can also be used to find an approximate posterior on the parameters.

Variational Bayes

So far, we have applied Jensen's bound and factorisations to help with integrals over latent variables.

We can do the same for integrals over parameters in order to bound the log marginal likelihood or evidence.

$$\begin{aligned}\log P(\mathcal{X}|\mathcal{M}) &= \log \iint d\mathcal{Y}\, d\boldsymbol{\theta}\;\; P(\mathcal{X},\mathcal{Y}|\boldsymbol{\theta},\mathcal{M})P(\boldsymbol{\theta}|\mathcal{M}) \\ &= \max_{Q} \iint d\mathcal{Y}\, d\boldsymbol{\theta}\;\; Q(\mathcal{Y},\boldsymbol{\theta}) \log \frac{P(\mathcal{X},\mathcal{Y},\boldsymbol{\theta}|\mathcal{M})}{Q(\mathcal{Y},\boldsymbol{\theta})} \\ &\geq \max_{Q_{\mathcal{Y}},Q_{\boldsymbol{\theta}}} \iint d\mathcal{Y}\, d\boldsymbol{\theta}\;\; Q_{\mathcal{Y}}(\mathcal{Y})Q_{\theta}(\boldsymbol{\theta}) \log \frac{P(\mathcal{X},\mathcal{Y},\boldsymbol{\theta}|\mathcal{M})}{Q_{\mathcal{Y}}(\mathcal{Y})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}\end{aligned}$$

The constraint that the distribution Q must factor into the product $Q_y(\mathcal{Y})Q_\theta(\boldsymbol{\theta})$ leads to the **variational Bayesian EM algorithm** or just **"Variational Bayes"**.

Some call this the "Evidence Lower Bound" (ELBO). I'm not fond of that term.

Variational Bayesian EM . . .

Coordinate maximization of the VB free-energy lower bound

$$\mathcal{F}(Q_{\mathcal{Y}}, Q_{\boldsymbol{\theta}}) = \iint d\mathcal{Y}\, d\boldsymbol{\theta}\;\; Q_{\mathcal{Y}}(\mathcal{Y})Q_{\theta}(\boldsymbol{\theta}) \log \frac{p(\mathcal{X},\mathcal{Y},\boldsymbol{\theta}|\mathcal{M})}{Q_{\mathcal{Y}}(\mathcal{Y})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$$

leads to **EM-like** updates:

$$Q^*_{\mathcal{Y}}(\mathcal{Y}) \propto \exp \langle \log P(\mathcal{Y},\mathcal{X}|\boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \qquad \textit{E-like step}$$

$$Q^*_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto P(\boldsymbol{\theta}) \exp \langle \log P(\mathcal{Y},\mathcal{X}|\boldsymbol{\theta}) \rangle_{Q_{\mathcal{Y}}(\mathcal{Y})} \qquad \textit{M-like step}$$

Maximizing $\mathcal{F}$ is equivalent to minimizing KL-divergence between the *approximate posterior*, $Q(\boldsymbol{\theta})Q(\mathcal{Y})$ and the *true posterior*, $P(\boldsymbol{\theta},\mathcal{Y}|\mathcal{X})$.

$$\begin{aligned}\log P(\mathcal{X}) - \mathcal{F}(Q_{\mathcal{Y}}, Q_{\boldsymbol{\theta}}) &= \log P(\mathcal{X}) - \iint d\mathcal{Y}\, d\boldsymbol{\theta}\;\; Q_{\mathcal{Y}}(\mathcal{Y})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{P(\mathcal{X},\mathcal{Y},\boldsymbol{\theta})}{Q_{\mathcal{Y}}(\mathcal{Y})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \\ &= \iint d\mathcal{Y}\, d\boldsymbol{\theta}\;\; Q_{\mathcal{Y}}(\mathcal{Y})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \log \frac{Q_{\mathcal{Y}}(\mathcal{Y})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{P(\mathcal{Y},\boldsymbol{\theta}|\mathcal{X})} = KL(Q||P)\end{aligned}$$

Conjugate-Exponential models

Let's focus on *conjugate-exponential* (**CE**) latent-variable models:

- **Condition (1)**. The joint probability over *variables* is in the exponential family:

  $$P(\mathcal{Y},\mathcal{X}|\boldsymbol{\theta}) = f(\mathcal{Y},\mathcal{X})\; g(\boldsymbol{\theta}) \exp\left\{\boldsymbol{\phi}(\boldsymbol{\theta})^{\mathsf{T}} \mathsf{T}(\mathcal{Y},\mathcal{X})\right\}$$

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

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

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

  where ν and τ are hyperparameters of the prior.

Conjugate priors are computationally convenient and have an intuitive interpretation:

- ν: number of pseudo-observations
- τ: 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 combinations of Gaussian, Gamma, Poisson, Dirichlet, Wishart, Multinomial and others.

Not in the **CE** family:

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

Note: one can often approximate such models with a suitable choice from the **CE** family.

Conjugate-exponential VB

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

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

$$\begin{aligned} Q_\theta(\boldsymbol{\theta}) &\propto P(\boldsymbol{\theta}) \quad \exp\left\langle \sum_i \log P(\mathbf{y}_i, \mathbf{x}_i|\boldsymbol{\theta}) \right\rangle_{Q_\mathcal{Y}} \\ &= h(\nu, \boldsymbol{\tau}) g(\boldsymbol{\theta})^\nu e^{\phi(\boldsymbol{\theta})^\mathsf{T} \boldsymbol{\tau}} \quad g(\boldsymbol{\theta})^n e^{\left\langle \log f(\mathcal{Y}, \mathcal{X}) \right\rangle_{Q_\mathcal{Y}}} e^{\phi(\boldsymbol{\theta})^\mathsf{T} \left\langle \sum_i \mathsf{T}(\mathbf{y}_i, \mathbf{x}_i) \right\rangle_{Q_\mathcal{Y}}} \\ &\propto h(\tilde{\nu}, \tilde{\boldsymbol{\tau}}) g(\boldsymbol{\theta})^{\tilde{\nu}} e^{\phi(\boldsymbol{\theta})^\mathsf{T} \tilde{\boldsymbol{\tau}}} \end{aligned}$$

  with $\tilde{\nu} = \nu + n$ and $\tilde{\boldsymbol{\tau}} = \tau + \sum_i \langle \mathsf{T}(\mathbf{y}_i, \mathbf{x}_i) \rangle_{Q_\mathcal{Y}}$ $\Rightarrow$ only need to track $\tilde{\nu}, \tilde{\boldsymbol{\tau}}$.

- $Q_\mathcal{Y}(\mathcal{Y}) = \prod_{i=1}^n Q_{\mathbf{y}_i}(\mathbf{y}_i)$ takes the same form as in the E-step of regular EM

$$\begin{aligned} Q_{\mathbf{y}_i}(\mathbf{y}_i) &\propto \exp \langle \log P(\mathbf{y}_i, \mathbf{x}_i|\boldsymbol{\theta}) \rangle_{Q_\theta} \\ &\propto f(\mathbf{y}_i, \mathbf{x}_i) e^{\langle \phi(\boldsymbol{\theta}) \rangle_{Q_\theta}^\mathsf{T} \mathsf{T}(\mathbf{y}_i, \mathbf{x}_i)} = P(\mathbf{y}_i|\mathbf{x}_i, \overline{\phi}(\boldsymbol{\theta})) \end{aligned}$$

  with natural parameters $\overline{\phi}(\boldsymbol{\theta}) = \langle \phi(\boldsymbol{\theta}) \rangle_{Q_\theta}$ $\Rightarrow$ inference unchanged from regular EM.

The Variational Bayesian EM algorithm

EM for MAP estimation

Goal: maximize $P(\boldsymbol{\theta}|\mathcal{X}, m)$ wrt $\boldsymbol{\theta}$

E Step: compute

$Q_\mathcal{Y}(\mathcal{Y}) \leftarrow p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})$

M Step:

$\boldsymbol{\theta} \leftarrow \underset{\boldsymbol{\theta}}{\text{argmax}} \int d\mathcal{Y}\, Q_\mathcal{Y}(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta})$

Variational Bayesian EM

Goal: maximise bound on $P(\mathcal{X}|m)$ wrt Q_θ

VB-E Step: compute

$Q_\mathcal{Y}(\mathcal{Y}) \leftarrow p(\mathcal{Y}|\mathcal{X}, \bar{\phi})$

VB-M Step:

$Q_\theta(\boldsymbol{\theta}) \leftarrow \exp \int d\mathcal{Y}\, Q_\mathcal{Y}(\mathcal{Y}) \log P(\mathcal{Y}, \mathcal{X}, \boldsymbol{\theta})$

Properties:

- Reduces to the EM algorithm if $Q_\theta(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$.
- $\mathcal{F}_m$ 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**, $\bar{\phi}$.

VB and model selection

- Variational Bayesian EM yields an approximate posterior Q_θ over model parameters.
- It also yields an optimised lower bound on the model evidence

$$\max \mathcal{F}_\mathcal{M}(Q_\mathcal{Y}, Q_\theta) \leq P(\mathcal{D}|\mathcal{M})$$

- These lower bounds can be compared amongst models to learn the right (structure, connectivity ... of the) model
- If a continuous domain of models is specified by a hyperparameter $\boldsymbol{\eta}$, then the VB free energy depends on that parameter:

$$\mathcal{F}(Q_\mathcal{Y}, Q_\theta, \boldsymbol{\eta}) = \iint d\mathcal{Y}\, d\boldsymbol{\theta}\;\; Q_\mathcal{Y}(\mathcal{Y}) Q_\theta(\boldsymbol{\theta}) \log \frac{P(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta}|\boldsymbol{\eta})}{Q_\mathcal{Y}(\mathcal{Y}) Q_\theta(\boldsymbol{\theta})} \leq P(\mathcal{X}|\eta)$$

  A hyper-M step maximises the current bound wrt η:

$$\boldsymbol{\eta} \leftarrow \underset{\boldsymbol{\eta}}{\text{argmax}} \iint d\mathcal{Y}\, d\boldsymbol{\theta}\;\; Q_\mathcal{Y}(\mathcal{Y}) Q_\theta(\boldsymbol{\theta}) \log P(\mathcal{X}, \mathcal{Y}, \boldsymbol{\theta}|\boldsymbol{\eta})$$

ARD for unsupervised learning

Recall that ARD (automatic relevance determination) was a hyperparameter method to select relevant or useful inputs in regression.

- A similar idea used with variational Bayesian methods can learn a latent dimensionality.
- Consider factor analysis:

$$\mathbf{x} \sim \mathcal{N}(\Lambda\mathbf{y}, \Psi) \qquad \mathbf{y} \sim \mathcal{N}(0, I) \qquad \text{with a column-wise prior} \quad \Lambda_{:i} \sim \mathcal{N}\left(0, \alpha_i^{-1} I\right)$$

- The VB free energy is

$$\mathcal{F}(Q_{\mathcal{Y}}(\mathcal{Y}), Q_\Lambda(\Lambda), \Psi, \boldsymbol{\alpha}) = \left\langle \log P(\mathcal{X}, \mathcal{Y}|\Lambda, \Psi) + \log P(\Lambda|\boldsymbol{\alpha}) + \log P(\Psi) \right\rangle_{Q_{\mathcal{Y}} Q_\Lambda} + \ldots$$

  and so hyperparameter optimisation requires

$$\boldsymbol{\alpha} \leftarrow \text{argmax} \left\langle \log P(\Lambda|\boldsymbol{\alpha}) \right\rangle_{Q_\Lambda}$$

- Now Q_Λ is Gaussian, with the same form as in linear regression, but with expected moments of $\mathbf{y}$ appearing in place of the inputs.
- Optimisation wrt the distributions, Ψ and $\boldsymbol{\alpha}$ in turn causes some α_i to diverge as in regression ARD.
- In this case, these parameters select "relevant" latent dimensions, effectively learning the dimensionality of $\mathbf{y}$.

Augmented Variational Methods

In our examples so far, the approximate variational distribution has been over the "natural" latent variables (and parameters) of the generative model.

Sometimes it may be useful to introduce additional latent variables, solely to achieve computational tractability.

Two examples are GP regression and the GPLVM.

Sparse GP approximations

GP predictions:

$$y'|X, Y, \mathbf{x}' \sim \mathcal{N}\left(K_{\mathbf{x}'X}(K_{XX} + \sigma^2 I)^{-1} Y, K_{\mathbf{x}'\mathbf{x}'} - K_{\mathbf{x}'X}(K_{XX+\sigma^2 I)}^{-1} K_{X\mathbf{x}'} + \sigma^2\right)$$

Evidence (for learning kernel hyperparameters):

$$\log P(Y|X) = -\frac{1}{2}\log|2\pi(K_{XX} + \sigma^2 I)| - \frac{1}{2} Y (K_{XX} + \sigma^2 I)^{-1} Y^{\mathsf{T}}$$

Computing either form requires inverting the $N \times N$ matrix K_{XX}, in $\mathcal{O}(N^3)$ time.

One proposal to make this more efficient is to find (or select) a smaller set of possibly fictitious measurements U at inputs Z such that

$$P(y'|Z, U, \mathbf{x}') \approx P(y'|X, Y, \mathbf{x}').$$

What values should U and Z take?

Variational Sparse GP approximations

Write F for the (smooth) GP function values that underlie Y (so $Y \sim \mathcal{N}\left(F, \sigma^2 I\right)$).

Introduce latent measurements U at inputs Z (and integrate over U).

The likelihood can be written

$$P(Y|X) = \iint dF\, dU\, P(Y, F, U|X, Z) = \iint dF\, dU\, P(Y|F) P(F|U, X, Z) P(U|Z)$$

Now, both U and F are latent, so we introduce a variational distribution $q(F, U)$ to form a free-energy.

$$\mathcal{F}(q(F, U), \theta) = \left\langle \log \frac{P(Y|F) P(F|U, X, Z) P(U|Z)}{q(F, U)} \right\rangle_{q(F,U)}$$

Now, choose the variational form $q(F, U) = P(F|U, X, Z) q(U)$. That is, fix $F|U$ without reference to Y – so information about Y will need to be "compressed" into $q(U)$.
Then

$$\begin{aligned}\mathcal{F}(q(F, U), \theta, Z) &= \left\langle \log \frac{P(Y|F) \cancel{P(F|U, X, Z)} P(U|Z)}{\cancel{P(F|U, X, Z)} q(U)} \right\rangle_{P(F|U)q(U)} \\ &= \left\langle \langle \log P(Y|F) \rangle_{P(F|U)} + \log P(U|Z) - \log q(U) \right\rangle_{q(U)}\end{aligned}$$

Variational Sparse GP approximations

$$\mathcal{F}(q(U),\theta,Z) = \left\langle \langle \log P(Y|F)\rangle_{P(F|U)} + \log P(U|Z) - \log q(U) \right\rangle_{q(U)}$$

Now $P(F|U)$ is fixed by the generative model (rather than being subject to free optimisation). So we can evaluate that expectation:

$$\begin{aligned}
&\langle \log P(Y|F)\rangle_{P(F|U)} \\
&= \left\langle -\frac{1}{2}\log\left|2\pi\sigma^2 I\right| - \frac{1}{2\sigma^2}\mathrm{Tr}\left[(Y-F)(Y-F)^\mathsf{T}\right] \right\rangle_{P(F|U)} \\
&= -\frac{1}{2}\log\left|2\pi\sigma^2 I\right| - \frac{1}{2\sigma^2}\mathrm{Tr}\left[(Y-\langle F\rangle_{P(F|U)})(Y-\langle F\rangle_{P(F|U)})^\mathsf{T}\right] - \frac{1}{2\sigma^2}\mathrm{Tr}\left[\Sigma_{F|U}\right] \\
&= \log\mathcal{N}\left(Y|K_{XZ}K_{ZZ}^{-1}U,\sigma^2 I\right) - \frac{1}{2\sigma^2}\mathrm{Tr}\left[K_{XX}-K_{XZ}K_{ZZ}^{-1}K_{ZX}\right]
\end{aligned}$$

So,

$$\mathcal{F}(q(U),\theta,Z) = \left\langle \log\mathcal{N}\left(Y|K_{XZ}K_{ZZ}^{-1}U,\sigma^2 I\right) + \log P(U|Z) - \log q(U)\right\rangle_{q(U)} - \frac{1}{2\sigma^2}\mathrm{Tr}\left[K_{XX}-K_{XZ}K_{ZZ}^{-1}K_{ZX}\right] .$$

A few references

- Jordan, Ghahramani, Jaakkola, Saul, 1999. An introduction to variational methods for graphical models. *Machine Learning* **37**:183–233.
- Attias, 2000. A variational Bayesian framework for graphical models. *NIPS 12*. `http://www.gatsby.ucl.ac.uk/publications/papers/03-2000.ps`
- Beal, 2003. Variational algorithms for approximate Bayesian inference. *PhD thesis*, Gatsby Unit, UCL. `http://www.cse.buffalo.edu/faculty/mbeal/thesis/`
- Winn, 2003. Variational message passing and its applications. *PhD thesis*, Cambridge. `http://johnwinn.org/Publications/Thesis.html`; also `VIBES` software for conjugate-exponential graphs.

Some complexities:

- MacKay, 2001. A problem with variational free energy minimization. `http://www.inference.phy.cam.ac.uk/mackay/minima.pdf`
- Turner, MS, 2011. Two problems with variational expectation maximisation for time-series models. In Barber, Cemgil, Chiappa, eds., *Bayesian Time Series Models*. `http://www.gatsby.ucl.ac.uk/~maneesh/papers/turner-sahani-2010-ildn.pdf`
- Berkes, Turner, MS, 2008. On sparsity and overcompleteness in image models. *NIPS 20*. `http://www.gatsby.ucl.ac.uk/~maneesh/papers/berkes-etal-2008-nips.pdf`
- Giordano, R, Broderick, T, and Jordan, MI, 2015. Linear response methods for accurate covariance estimates from mean field variational Bayes. *NIPS*

Variational Sparse GP approximations

$$\mathcal{F}(q(U),\theta,Z) = \left\langle \log\frac{\mathcal{N}\left(Y|K_{XZ}K_{ZZ}^{-1}U,\sigma^2 I\right)P(U|Z)}{q(U)}\right\rangle_{q(U)} - \frac{1}{2\sigma^2}\mathrm{Tr}\left[K_{XX}-K_{XZ}K_{ZZ}^{-1}K_{ZX}\right] .$$

The expectation is the free energy of a PPCA-like model with normal prior $U\sim\mathcal{N}(0,K_{ZZ})$ and loading matrix $K_{XZ}K_{ZZ}^{-1}$. The maximum of this free energy is the log-likelihood (achieved with q equal to the posterior under the PPCA-like model).
This gives

$$\mathcal{F}(q^*(U),\theta,Z) = \log\mathcal{N}\left(Y|0,K_{XZ}\cancel{K_{ZZ}^{-1}K_{ZZ}}K_{ZZ}^{-1}K_{ZX}+\sigma^2 I\right) - \frac{1}{2\sigma^2}\mathrm{Tr}\left[K_{XX}-K_{XZ}K_{ZZ}^{-1}K_{ZX}\right] .$$

Note that we have eliminated all terms in K_{XX}^{-1}.

We can optimise the free energy numerically with respect to Z and θ to adjust the GP prior and quality of variational approximation.

A similar approach can be used to learn X if they are unobserved (*i.e.* in the GPLVM). Assume $q(X,F,U) = q(X)P(F|X,U)q(U)$. Then $\mathcal{F} = \langle \log P(Y,F,U|X)\log P(X)\rangle_{q(U)q(X)}$ which simplifies into tractable components in much the same way as above.