Machine Learning for OR & FE

Deterministic Inference

Martin Haugh

Department of Industrial Engineering and Operations Research
Columbia University
Email: `martin.b.haugh@gmail.com`

Additional References: Christopher Bishop's *PRML* and David Barber's *BRML*

Outline

Why Deterministic Inference?

Modal Methods
- Normal And Mixture Approximations
- Approximations Based on Marginal Posterior Modes

Properties of KL Divergence

Variational Bayes
- Examples
- Application: Control by Inference

Expectation Propagation

Why Deterministic Inference?

Many inference problems in probabilistic modeling amount to evaluating posterior distributions of the form $p(z \mid x)$

– arises in Bayesian modeling and other domains, e.g. graphical models.

Can evaluate the posterior by simulating samples using MCMC methods

– can work very well in practice but can be very time-consuming.

An alternative approach is to use deterministic schemes to approximate the posterior

- results in analytic approximations to the posterior
- can often be found very quickly in comparison to MCMC, e.g. seconds or minutes v. possibly hours or days.
- depending on the inference goals, these approximations may be more than sufficient.

Deterministic inference is less well-known than sampling-based inference, i.e. MCMC

- but it has become very popular in recent years
- has its origins in analysis of large-scale physical systems.

Methods of Deterministic Inference

There are many approaches to the deterministic inference problem of evaluating $p(z \mid x)$ where x is the observed data. They include:

1. Normal and mixture approximations
   - first need methods for finding the posterior modes.
2. Methods based on finding marginal posterior modes
   - sometimes makes sense to approximate lower-dimensional marginals rather than full joint distributions.
3. Variational Bayes
   - based on minimizing $\mathsf{KL}(q(z) \,||\, p(z \mid x))$ with respect to a family of distributions $q(Z)$.
4. Expectation Propagation
   - based on minimizing $\mathsf{KL}(p(z \mid x) \,||\, q(z))$ with respect to a (parametric) family of distributions, $q(Z)$.

Note the difference between variational Bayes and expectation propagation.

See Chapter 13 of 3^{rd} edition of *Bayesian Data Analysis* by Gelman et al. for details on all of these methods.

Chapter 10 of Bishop's *PR and ML* and Chapter 28 of Barber's *BR and ML* have more detailed introductions to variational Bayes and expectation propagation.

Finding Posterior Modes

Standard numerical methods exist for finding posterior modes including:

1. Conditional maximization or coordinate ascent
    - will converge to a local maximum if posterior is bounded.
2. Newton's method – based on quadratic Taylor series approximation to $p(z \mid x)$
    - not guaranteed to converge from all starting points z_0
    - but convergence is extremely fast once we are close to a solution
    - can be computationally expensive since inverse of matrix of second derivatives, i.e. the Hessian, is required at each step.
3. Quasi-Newton methods, e.g. BFGS, which iteratively approximates the inverse Hessian.

Note these algorithms only require the unnormalized posterior.

Derivatives may be computed analytically or numerically.

Use multiple starting points if distribution is suspected to be multi-modal and we want to find all modes associated with non-negligible probability mass.

If posterior mode on boundary of parameter space then it may not be suitable as a point summary of posterior distribution

- may even want to use a prior that moves posterior to mode to the interior.

Normal And Mixture Approximations

When $p(z \mid x)$ is believed to be unimodal can use the approximation

$$p(z \mid x) \approx p_{\text{norm-approx}}(z \mid x) := \mathsf{N}\left(\hat{z}, V_z\right)$$

where $\hat{z}$ is the mode of $p(z \mid x)$ and

$$V_z := \left[- \left.\frac{d^2 \log p(z \mid x)}{dz^2}\right|_{z=\hat{z}}\right]^{-1}$$

is the inverse of the curvature of the $\log$ posterior evaluated at $\hat{z}$

- can be calculated analytically or numerically.

Also straightforward to approximate $p(z \mid x)$ with a t-distribution if preferred.

If there are multiple modes then straightforward to approximate $p(z \mid x)$ with a normal mixture or t mixture

- straightforward to simulate directly from any of these approximations.

Sometimes we're more interested in computing $\mathsf{E}\left[h(z) \mid x\right] = \int_z h(z)p(z \mid x)\, dz$

- Laplace's method uses the approximation $h(z)p(z \mid x) \approx \mathsf{N}\left(z_0, V_0\right)$ for suitable z_0 and V_0 – can also apply to mixture distributions.

Approximations Based on Marginal Posterior Modes

Normal approximations to the posterior often not suitable

- e.g. hierarchical models where # parameters / latent variables grows with # observations.

If $\boldsymbol{\theta} = (\boldsymbol{\gamma}, \boldsymbol{\phi})$ denotes set of parameters / latent variables then often convenient to work with posterior in the form

$$p(\boldsymbol{\gamma}, \boldsymbol{\phi} \mid x) = p(\boldsymbol{\phi} \mid x)\, p(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x).$$

Can approximate $p(\boldsymbol{\phi} \mid x)$ with a normal or t approximation

- may be suitable if dim($\boldsymbol{\phi}$) small and does not grow with # of observations.

Must also approximate $p(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)$ – possibly using normal or t (mixtures) again

- although with parameters now depending on $\boldsymbol{\phi}$.

Can use modal methods to approximate $p(\boldsymbol{\phi} \mid x)$

- e.g. use EM algorithm to find mode with $\boldsymbol{\gamma}$ playing role of "missing data"
- also possible to approximate $p(\boldsymbol{\phi} \mid x)$ with

$$p(\boldsymbol{\phi} \mid x) = \frac{p(\boldsymbol{\gamma}, \boldsymbol{\phi} \mid x)}{p(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)} \approx \frac{p(\hat{\boldsymbol{\gamma}}, \boldsymbol{\phi} \mid x)}{p_{\text{approx}}(\hat{\boldsymbol{\gamma}} \mid \boldsymbol{\phi}, x)}$$

if $p_{\text{approx}}(\boldsymbol{\gamma} \mid \boldsymbol{\phi}, x)$ has recognizable form so normalizing factor (which depends on $\boldsymbol{\phi}$) is available analytically.

Review: Kullback-Leibler Divergence

Recall the Kullback-Leibler (KL) divergence or relative entropy of Q from P is defined to be

$$\mathsf{KL}(P\,||\,Q) \;=\; \int_{\mathbf{x}} P(\mathbf{x}) \ln\left(\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right) \tag{1}$$

with the understanding that $0 \log 0 = 0$.

The KL divergence is a fundamental concept in information theory & statistics.

Can imagine P representing some true but unknown distribution, e.g. $p(z \mid x)$, that we approximate with Q

– $\mathsf{KL}(P\,||\,Q)$ then measures the “distance” between P and Q.

This interpretation is valid because $\mathsf{KL}(P\,||\,Q) \geq 0$

– with equality if and only if $P = Q$

The KL divergence is not a true measure of distance since it is:

1. Asymmetric in that $\mathsf{KL}(P\,||\,Q) \neq \mathsf{KL}(Q\,||\,P)$
2. And does not satisfy the triangle inequality.

The KL-Divergence Lower Bound

Recall again that we want to compute (or approximate) $p(\mathbf{z} \mid \mathbf{x})$

- so the data $\mathbf{x}$ has been observed but the parameters / latent variables $\mathbf{z}$ are unobserved.

For any probability distribution $q(\mathbf{z})$ we have

$$\begin{aligned}
\ln p(\mathbf{x}) &= \sum_{\mathbf{z}} q(\mathbf{z}) \ln(p(\mathbf{x})) \\
&= \sum_{\mathbf{z}} q(\mathbf{z}) \ln\left(\frac{p(\mathbf{x},\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})}\right) \\
&= \sum_{\mathbf{z}} q(\mathbf{z}) \ln\left(\frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z})} \cdot \frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})}\right) \\
&= \underbrace{\sum_{\mathbf{z}} q(\mathbf{z}) \ln\left(\frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right)}_{=:\mathcal{L}(q)} + \underbrace{\sum_{\mathbf{z}} q(\mathbf{z}) \ln\left(\frac{q(\mathbf{z})}{p(\mathbf{z} \mid \mathbf{x})}\right)}_{\equiv \mathsf{KL}(q \,||\, p(\cdot|\mathbf{x}))} \qquad (2)
\end{aligned}$$

(For continuous random variables we can replace sums with integrals.)

The KL-Divergence Lower Bound

It follows from (2) that

$$\begin{aligned} \max_q \{\mathcal{L}(q)\} &= \max_q \left\{ \ln p(\mathbf{x}) - \mathsf{KL}(q \,||\, p(\cdot \mid \mathbf{x})) \right\} \\ &= \ln p(\mathbf{x}) - \min_q \{\mathsf{KL}(q \,||\, p(\cdot \mid \mathbf{x}))\} \end{aligned}$$

But $\operatorname{argmin}_q \{\mathsf{KL}(p(\cdot \mid \mathbf{x}) \,||\, q)\} = \{p(\cdot \mid \mathbf{x})\}$ (a singleton set), it follows that

$$\operatorname*{argmax}_q \{\mathcal{L}(q)\} = \{p(\cdot \mid \mathbf{x})\}.$$

– posterior is therefore the solution of an optimization problem!

Variational Bayes approximates $p(\mathbf{z} \mid \mathbf{x})$ with $q(\mathbf{z})$ where q obtained by minimizing $\mathsf{KL}(q \,||\, p(\cdot \mid \mathbf{x}))$ over a family of tractable distributions.

In contrast, expectation propagation approximates $p(\mathbf{z} \mid \mathbf{x})$ with $q(\mathbf{z})$ where q obtained by minimizing $\mathsf{KL}(p(\cdot \mid \mathbf{x}) \,||\, q)$ over a family of tractable distributions.

Before describing these algorithms will first consider difference between minimizing $\mathsf{KL}(p \,||\, q)$ and minimizing $\mathsf{KL}(q \,||\, p)$.

Minimizing $\mathsf{KL}(p \,||\, q)$ or $\mathsf{KL}(q \,||\, p)$?

As before, we want to approximate $p(\mathbf{z})$ with another distribution $q(\mathbf{z})$.

Since $\mathsf{KL}(q \,||\, p) \geq 0$ for all q with equality only if $q = p$, this suggests two reasonable approaches:

1. Solve

$$\min_q \mathsf{KL}(q \,||\, p) = \sum_{\mathbf{z}} q(\mathbf{z}) \ln \left(\frac{q(\mathbf{z})}{p(\mathbf{z})} \right)$$

   Here the expectation is taken wrt the approximation, q, so that:

   - regions in the support of p can be ignored if q places zero mass on them
   - will also want to choose a q that avoids regions where p is very small.

   And so often obtain a more **local** approximation to p as a result

   - therefore inference methods based on this approach, i.e. variational Bayes, can sometimes overfit and underestimate posterior variances.

Minimizing $\mathsf{KL}(p\,||\,q)$ or $\mathsf{KL}(q\,||\,p)$?

The other approach is:

2. Solve

$$\min_q \mathsf{KL}(p\,||\,q) = \sum_{\mathbf{z}} p(\mathbf{z}) \ln\left(\frac{p(\mathbf{z})}{q(\mathbf{z})}\right)$$

   Here the expectation is taken wrt the target distribution, p, which is fixed, so that:

   - q must be non-negligible in regions where p is non-negligible.

   This results in a more **global** approximation where q tries to approximates p across entire support of p.

Exercise: Suppose we minimize $\mathsf{KL}(p\,||\,q)$ over the distributions q under which components of $\mathbf{z}$ are independent, i.e. we solve

$$\min_{q\,:\,q(\mathbf{z})=\prod q_i(z_i)} \mathsf{KL}(p\,||\,q).$$

Show the optimal q satisfies $q(\mathbf{z}) = \prod p_i(z_i)$.

Approximating a Bimodal Distribution

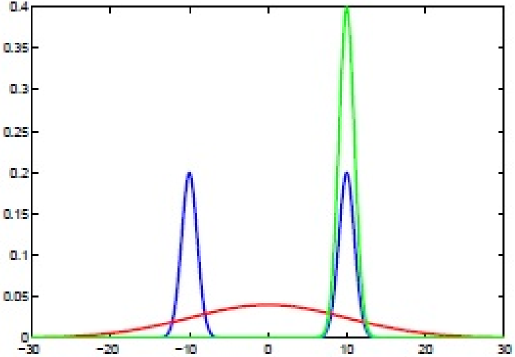


Figure 28.1 from Barber: Fitting a mixture of Gaussians $p(x)$ (blue) with a single Gaussian. The green curve minimises KL$(q||p)$ corresponding to fitting a local model. The red curve minimises KL$(p||q)$ corresponding to moment matching.

Approximating a Correlated Bivariate Gaussian

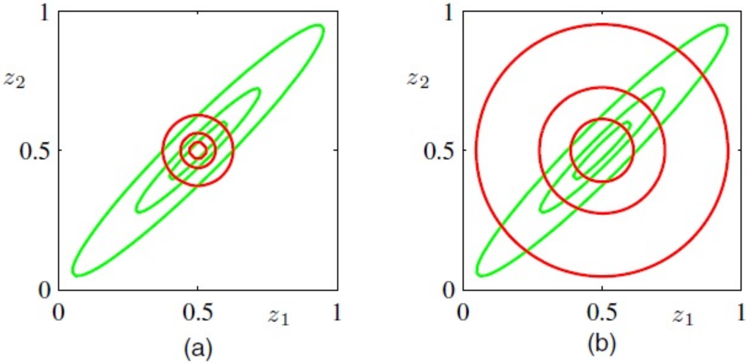


Figure 10.2 from Bishop: Comparison of the two alternative forms for the Kullback-Leibler divergence. The green contours corresponding to 1, 2, and 3 standard deviations for a correlated Gaussian distribution $p(\mathbf{z})$ over two variables z_1 and z_2, and the red contours represent the corresponding levels for an approximating distribution $q(\mathbf{z})$ over the same variables given by the product of two independent univariate Gaussian distributions whose parameters are obtained by minimization of (a) the Kullback-Leibler divergence $\mathsf{KL}(q||p)$, and (b) the reverse Kullback-Leibler divergence $\mathsf{KL}(p||q)$.

Approximating a Correlated Bivariate Gaussian Mixture

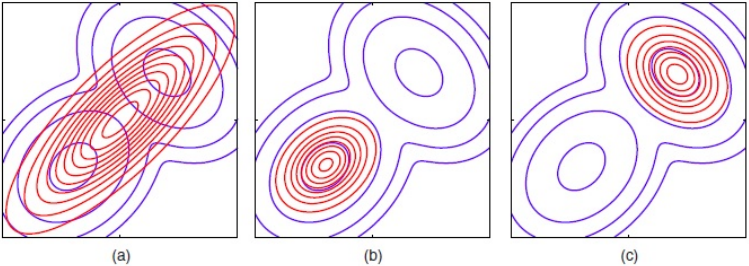


Figure 10.3 from Bishop: Another comparison of the two alternative forms for the Kullback-Leibler divergence. (a) The blue contours show a bimodal distribution $p(\mathbf{Z})$ given by a mixture of two Gaussians, and the red contours correspond to the single Gaussian distribution $q(\mathbf{Z})$ that best approximates $p(\mathbf{Z})$ in the sense of minimizing the Kullback- Leibler divergence KL$(p||q)$. (b) As in (a) but now the red contours correspond to a Gaussian distribution $q(\mathbf{Z})$ found by numerical minimization of the Kullback-Leibler divergence KL$(q||p)$. (c) As in (b) but showing a different local minimum of the Kullback-Leibler divergence.

Variational Bayes

We write $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_m)$ where each $\mathbf{z}_i$ is a sub-vector and then define $\mathcal{Q}$ to be the set of product distributions over these sub-vectors so that

$$\mathcal{Q} := \left\{ q \,:\, q(\mathbf{z}) = \prod_{i=1}^{m} q_i(\mathbf{z}_i) \right\}.$$

The variational Bayes approach approximates $p_x := p(\mathbf{z} \mid \mathbf{x})$ with the solution to

$$\min_{q \in \mathcal{Q}} \mathsf{KL}(q \,||\, p_x).$$

This is a computationally hard optimization problem but we can use coordinate descent to find a local minimum.

Let $\bar{q}_j := \prod_{i \neq j} q_i$ and $\bar{\mathbf{z}}_j := (z_1, \ldots, z_{j-1}, z_{j+1}, \ldots, z_m)$.

Coordinate descent works by minimizing $\mathsf{KL}(q \,||\, p_x)$ over q_j keeping $\bar{q}_j$ fixed for $j = 1, \ldots, m$

- and then iterating until convergence.

Variational Bayes Via Coordinate Descent

So let $q = q_j \bar{q}_j \in \mathcal{Q}$ and suppose we want to minimize over q_j in the current step. We have

$$\begin{aligned}
\mathsf{KL}(q \,||\, p_x) &= \sum_{\mathbf{z}} q(\mathbf{z}) \ln \left(\frac{q(\mathbf{z})}{p(\mathbf{x}, \mathbf{z})} \right) + C_1 \\
&= \sum_{\mathbf{z}} q(\mathbf{z}) \ln q(\mathbf{z}) - \sum_{\mathbf{z}} q(\mathbf{z}) \ln p(\mathbf{x}, \mathbf{z}) + C_1 \\
&= \sum_{\bar{\mathbf{z}}_j} \bar{q}_j(\bar{\mathbf{z}}_j) \ln \bar{q}_j(\bar{\mathbf{z}}_j) + \sum_{z_j} q_j(z_j) \ln q_j(z_j) \\
&\qquad\qquad - \sum_{z_j} q_j(z_j) \sum_{\bar{\mathbf{z}}_j} \bar{q}_j(\bar{\mathbf{z}}_j) \ln p(\mathbf{x}, \mathbf{z}) + C_1 \\
&= C_2 + \sum_{z_j} q_j(z_j) \ln q_j(z_j) \; - \; \sum_{z_j} q_j(z_j) \mathsf{E}_{\bar{q}_j} \left[\ln p(\mathbf{X}, \mathbf{Z}) \right] \\
&= C_2 + \sum_{z_j} q_j(z_j) \ln q_j(z_j) \; - \; \sum_{z_j} q_j(z_j) \ln \exp \left(\mathsf{E}_{\bar{q}_j} \left[\ln p(\mathbf{X}, \mathbf{Z}) \right] \right) \\
&= C_3 + \mathsf{KL}(q_j \,||\, C_4 \exp \left(\mathsf{E}_{\bar{q}_j} \left[\ln p(\mathbf{X}, \mathbf{Z}) \right] \right)). \qquad (3)
\end{aligned}$$

Mean-Field Equations

Can view C_1 to C_3 as constants since they do not depend on q_j.

C_4 is a normalization constant ensuring that the distribution integrates to 1.

From (3) it follows that minimizing $\mathsf{KL}(q \,||\, p_x)$ over q_j has an optimal solution (why?)

$$q_j^*(Z_j) \propto \exp\left(\mathsf{E}_{\bar{q}_j}\left[\ln p(\mathbf{X}, \mathbf{Z})\right]\right)$$

or equivalently

$$\ln q_j^*(Z_j) = \mathsf{E}_{\bar{q}_j}\left[\ln p(\mathbf{X}, \mathbf{Z})\right] + \text{constant}. \tag{4}$$

The equations in (4) for $j = 1, \ldots, m$ are known as the mean-field equations

- and they are iterated (when tractable) until convergence
- and convergence is guaranteed; see Section 13.7 of Gelman et al. or Section 28.4.3 of Barber.

The constant term in (4) is a normalization constant that can be found by simply recognizing the distribution q_j^*.

Variational Bayes for Mean and Precision of a Gaussian

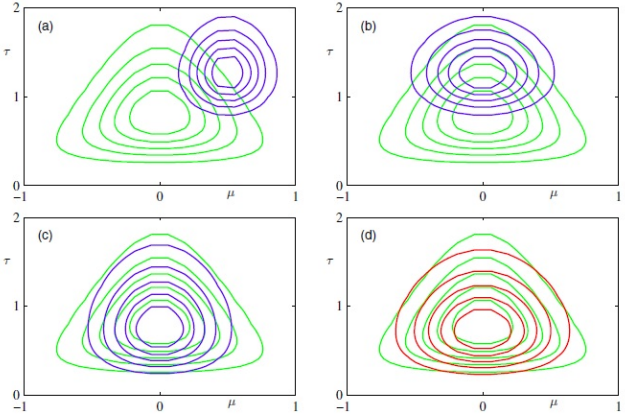


Figure 10.4 from Bishop: Illustration of variational inference for the mean μ and precision τ of a univariate Gaussian distribution. Contours of the true posterior distribution $p(\mu, \tau|D)$ are shown in green. (a) Contours of the initial factorized approximation $q_\mu(\mu)q_\tau(\tau)$ are shown in blue. (b) After re-estimating the factor $q_\mu(\mu)$. (c) After re-estimating the factor $q_\tau(\tau)$. (d) Contours of the optimal factorized approximation, to which the iterative scheme converges, are shown in red.

A Simple Example

Consider a bivariate Gaussian distribution with:

$$\text{mean } \boldsymbol{\mu} = \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix} \qquad \text{and precision matrix } \boldsymbol{\Lambda} = \begin{bmatrix}\Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22}\end{bmatrix}$$

so the joint distribution is

$$p(\mathbf{z}) = \frac{|\Lambda|^{\frac{1}{2}}}{(2\pi)} e^{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^\top \boldsymbol{\Lambda}(\mathbf{z}-\boldsymbol{\mu})}$$

There is no data **x** here but that's ok: can still use variational Bayes to approximate $p(\mathbf{z}\,;\,\boldsymbol{\mu},\boldsymbol{\Lambda})$ with a product distribution $q(\mathbf{z}) = q_1(z_1)q_2(z_2)$.

Mean-field equation for q_1 is:

$$\begin{aligned}
\ln q_1^*(z_1) &= \mathsf{E}_{\bar{q}_1}\left[\ln p(\mathbf{z})\right] + \text{constant} \\
&= \mathsf{E}_{\bar{q}_1}\left[-\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 - (z_1-\mu_1)\Lambda_{12}(z_2-\mu_2)\right] + \text{constant} \\
&= -\frac{1}{2}\Lambda_{11}z_1^2 + z_1\Big(\Lambda_{11}\mu_1 - \Lambda_{12}\left(\mathsf{E}_{\bar{q}_1}[z_2]-\mu_2\right)\Big) + \text{constant} \\
&= -\frac{1}{2}\Lambda_{11}\left(z_1 - \left(\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathsf{E}_{\bar{q}_1}[z_2]-\mu_2)\right)\right)^2 + \text{constant} \qquad (5)
\end{aligned}$$

A Simple Example

We recognize the form of (5) so that

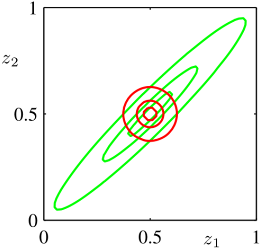
$$q_1^*(z_1) = \mathsf{N}\Big(\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathsf{E}_{\bar{q}_1}[z_2] - \mu_2),\ \Lambda_{11}\Big) \tag{6}$$

By symmetry can also have

$$q_2^*(z_2) = \mathsf{N}\Big(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathsf{E}_{\bar{q}_2}[z_1] - \mu_1),\ \Lambda_{22}\Big) \tag{7}$$

Starting with some initial q_2 we can iterate (6) and (7) until convergence.

What does the converged product distribution $q_1^*(z_1)q_2^*(z_2)$ miss?



- it captures the mean correctly.
- but the variance is underestimated.
- and the directionality is completely missed.

Example: Variational Linear Regression

Consider the following Bayesian linear regression model:

- Distribution of data: $p(\mathbf{y} \mid \mathbf{w}) = \prod_{i=1}^{N} \mathsf{N}(y_i \mid \boldsymbol{\phi}(\mathbf{x}_i)^\top \mathbf{w},\ \beta^{-1})$
- Distribution of weights: $p(\mathbf{w} \mid \alpha) = \mathsf{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I})$
- Distribution of weight precision: $p(\alpha) = \mathsf{Gamma}(\alpha \mid \nu_0, b_0) \propto \alpha^{\nu_0 - 1} e^{-b_0 \alpha}$

Want to compute posterior $p(\alpha, \mathbf{w} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \mathbf{w}) p(\mathbf{w} \mid \alpha) p(\alpha)$

- but instead could use variational Bayes to approximate it with $q(\alpha, \mathbf{w}) := q_\alpha(\alpha) q_{\mathbf{w}}(\mathbf{w})$.

The mean-field equation for $q_\alpha^*(\alpha)$ is:

$$\begin{aligned}
\ln q_\alpha^*(\alpha) &= \ln p(\alpha) + \mathsf{E}_{q_{\mathbf{w}}}[\ln p(\mathbf{w} \mid \alpha)] + \text{constant} \\
&= (\nu_0 - 1) \ln \alpha - \alpha b_0 + \frac{d}{2} \ln \alpha - \frac{\alpha}{2} \mathsf{E}_{q_{\mathbf{w}}}[\mathbf{w}^\top \mathbf{w}] + \text{constant} \\
&= \mathsf{Gamma}(\nu_N, b_N)
\end{aligned}$$

where $\nu_N = \nu_0 + \frac{d}{2}$ and $b_N = b_0 + \frac{1}{2} \mathsf{E}_{q_{\mathbf{w}}}[\mathbf{w}^\top \mathbf{w}]$.

Do not yet know how to compute $\mathsf{E}_{q_{\mathbf{w}}}[\mathbf{w}^\top \mathbf{w}]$.

Variational Linear Regression (continued)

The mean-field equation for $q^*_{\mathbf{w}}(\mathbf{w})$ is:

$$\begin{aligned}\ln q^*_{\mathbf{w}}(\mathbf{w}) &= \ln p(\mathbf{y} \mid \mathbf{w}) + \mathsf{E}_{q_\alpha}[\ln p(\mathbf{w} \mid \alpha)] + \text{constant} \\ &= -\frac{\beta}{2}\sum_{i=1}^{N}(\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_i) - y_i)^2 - \frac{1}{2}\mathsf{E}_{q_\alpha}[\alpha]\mathbf{w}^\top\mathbf{w} + \text{const} \\ &= -\frac{1}{2}\mathbf{w}^\top\left(\mathsf{E}_{q_\alpha}[\alpha]\mathbf{I} + \beta\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\right)\mathbf{w} + \beta\mathbf{w}^\top\boldsymbol{\Phi}\mathbf{y} + \text{constant}\end{aligned}$$

Therefore $q^*_{\mathbf{w}}(\mathbf{w}) \equiv \mathsf{N}(\mathbf{w} \mid \mathbf{m}_N, \boldsymbol{\Lambda}_N^{-1})$ where

$$\boldsymbol{\Lambda}_N := \mathsf{E}_{q_\alpha}[\alpha]\mathbf{I} + \beta\boldsymbol{\Phi}\boldsymbol{\Phi}^\top \qquad \text{and} \qquad \mathbf{m}_N := \beta\boldsymbol{\Lambda}_N^{-1}\boldsymbol{\Phi}\mathbf{y}.$$

Moments easily calculated as

- Moment of the weights: $\mathsf{E}_{q_{\mathbf{w}}}[\mathbf{w}^\top\mathbf{w}] = \mathbf{m}_N^\top\mathbf{m}_N + \mathbf{Tr}(\boldsymbol{\Lambda}_N^{-1})$
- Moment of the precision α: $\mathsf{E}_{q_\alpha}[\alpha] = \alpha_N b_N$
  - so can now iterate mean-field equations until convergence.

Computing an Approximate Posterior Predictive Distribution

Can now use (the converged) $q(\alpha, \mathbf{w}) := q_\alpha(\alpha) q_{\mathbf{w}}(\mathbf{w})$ to do approximate posterior inference.

e.g. Suppose we want to predict the outcome y for a new datapoint x. If we let $\mathcal{D}$ denote the training data then we need $p(y \mid \mathbf{x}, \mathcal{D})$ to do predictions.

Can approximate this posterior predictive distribution as

$$
\begin{aligned}
p(y \mid \mathbf{x}, \mathcal{D}) &= \int p(y \mid \mathbf{w}, \mathbf{x}) p(\mathbf{w} \mid \mathcal{D}) \, d\mathbf{w} \\
&\approx \int p(y \mid \mathbf{w}, \mathbf{x}) q_{\mathbf{w}}(\mathbf{w}) \, d\mathbf{w} \\
&= \int \mathsf{N}(y \mid \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}) \, \mathsf{N}(\mathbf{w} \mid \mathbf{m}_N, \boldsymbol{\Lambda}_N^{-1}) \, d\mathbf{w} \\
&= \mathsf{N}\big(y \mid \mathbf{m}_N^\top \boldsymbol{\phi}(\mathbf{x}), \sigma^2(\mathbf{x})\big)
\end{aligned}
$$

where $\sigma^2(\mathbf{x}) := \beta^{-1} + \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\Lambda}_N^{-1} \boldsymbol{\phi}(\mathbf{x})$.

Can use this approximation to make predict y for new datapoint $\mathbf{x}$.

Variational Bayes and the Exponential Family

Variational Bayes is particularly easy to implement with the exponential family of distributions and a conjugate prior.

- a very rich class of distributions.

In particular, suppose $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is the observed data and let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ be corresponding hidden or latent data.

If the observations are IID and the complete data $(\mathbf{x}, \mathbf{z})$ has an exponential family distribution then

$$p(\mathbf{x}, \mathbf{z}; \boldsymbol{\eta}) = \prod_{i=1}^{n} h(\mathbf{x}_i, \mathbf{z}_i) g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{x}_i, \mathbf{z}_i)\right).$$

If we assume a conjugate prior for $\boldsymbol{\eta}$ so that

$$p(\boldsymbol{\eta}; \nu_0, \boldsymbol{\chi}_0) = f(\nu_0, \boldsymbol{\chi}_0) g(\boldsymbol{\eta})^{\nu_0} \exp\left(\nu_0 \boldsymbol{\eta}^\top \boldsymbol{\chi}_0\right)$$

then the posterior is $p(\mathbf{z}; \boldsymbol{\eta} \mid \mathbf{x}) \propto p(\boldsymbol{\eta}; \nu_0, \boldsymbol{\chi}_0) p(\mathbf{x}, \mathbf{z}; \boldsymbol{\eta})$.

Variational Bayes with $q(\mathbf{z}; \boldsymbol{\eta}) = q_{\mathbf{z}}(\mathbf{z}) q_{\boldsymbol{\eta}}(\boldsymbol{\eta})$ is then straightforward to implement

- see Section 10.4 of Bishop for further deails.

Variational Bayes for a Gaussian Mixture Model

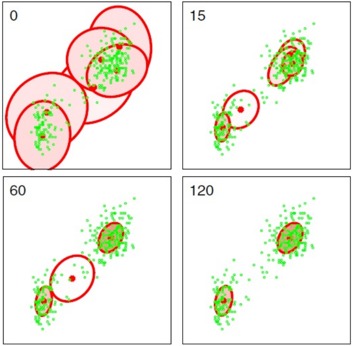


Figure 10.6 from Bishop: Variational Bayesian mixture of $K = 6$ Gaussians applied to the Old Faithful data set, in which the ellipses denote the one standard-deviation density contours for each of the components, and the density of red ink inside each ellipse corresponds to the mean value of the mixing coefficient for each component. The number in the top left of each diagram shows the number of iterations of variational inference. Components whose expected mixing coefficient are numerically indistinguishable from zero are not plotted.

See Section 10.2 of Bishop for details.

Application: Control by Inference

Example 28.2 from Barber:

1. Let $\mathbf{v}_t = (x_t, y_t)^\top$ be the time t position of an n-link robot arm in the plane.

2. Each link $i \in \{1, \ldots, n\}$ in the arm has unit length and angle, $h_{i,t}$, so that

$$x_t = \sum_{i=1}^{n} \cos h_{i,t}, \qquad y_t = \sum_{i=1}^{n} \sin h_{i,t}$$

3. **Goal:** choose the $h_{i,t}$'s so that the arm will track a given sequence $\mathbf{v}_{1:T}$ such that the joint angles, $\mathbf{h}_t$, do not change much from one time to the next.
4. This is a classical control problem which may be formulated as an inference problem using the model

$$\mathsf{P}(\mathbf{v}_{1:T}, \mathbf{h}_{1:T}) \; = \; \mathsf{P}(\mathbf{v}_1 \,|\, \mathbf{h}_1)\mathsf{P}(\mathbf{h}_1) \prod_{t=2}^{T} \mathsf{P}(\mathbf{v}_t \,|\, \mathbf{h}_t)\mathsf{P}(\mathbf{h}_t \,|\, \mathbf{h}_{t-1}) \tag{8}$$

   where we assume

$$\mathsf{P}(\mathbf{v}_t \,|\, \mathbf{h}_t) \sim \mathsf{N}\left(\left(\sum_{i=1}^{n} \cos h_{i,t}, \sum_{i=1}^{n} \sin h_{i,t}\right)^\top, \sigma^2 \mathbf{I}\right), \qquad \mathsf{P}(\mathbf{h}_t \,|\, \mathbf{h}_{t-1}) \sim \mathsf{N}\left(\mathbf{h}_{t-1}, \nu^2 \mathbf{I}\right). \tag{9}$$

Control by Inference

One approach is to solve for the most likely posterior sequence $\text{argmax}_{\mathbf{h}_{1:T}} \ \mathsf{P}(\mathbf{h}_{1:T} \,|\, \mathbf{v}_{1:T})$.

Question: How does this model formulation ensure that the joint angles, $\mathbf{h}_t$, do not change much from one time to the next?

We could restrict ourselves to finding the most likely marginal state at each time $\text{argmax}_{\mathbf{h}_t} \ \mathsf{P}(\mathbf{h}_t \,|\, \mathbf{v}_{1:T})$

- but cannot compute the marginals exactly due to the non-linear observations.

Instead we will use variational Bayes to find an approximation $q(\mathbf{h}_{1:T}) \approx \mathsf{P}(\mathbf{h}_{1:T} \,|\, \mathbf{v}_{1:T})$ where we assume a fully factorized form for $q(\cdot)$ so that

$$q(\mathbf{h}_{1:T}) \ = \ \prod_{t=1}^{T} \prod_{i=1}^{n} q(h_{i,t})$$

Variational Bayes and the Mean-Field Equations

Recall the mean-field equations imply that the optimal choice for $q(h_{i,t})$ satisfies

$$q(h_{i,t}) \;\propto\; \exp\left(\mathsf{E}_{(i',t')\neq(i,t)}\left[\ln \mathsf{P}(\mathbf{h}_{1:T}, \mathbf{v}_{1:T})\right]\right) \tag{10}$$

where $\mathsf{E}_{(i',t')\neq(i,t)}\left[\cdot\right]$ means the expectation is calculated with respect to $\prod_{(i',t')\neq(i,t)} q(h_{i',t'})$.

Question: In what sense is $q(h_{i,t})$ in (10) optimal?

Using (8) and (9), it is straightforward to see that (10) yields for $1 < t < T$

$$\begin{aligned} -2\log q(h_{i,t}) \quad &= \quad \frac{1}{\nu^2}\left(h_{i,t} - \bar{h}_{i,t-1}\right)^2 + \frac{1}{\nu^2}\left(h_{i,t} - \bar{h}_{i,t+1}\right)^2 \\ &\quad + \frac{1}{\sigma^2}\left(\cos h_{i,t} - \alpha_{i,t}\right)^2 + \frac{1}{\sigma^2}\left(\sin h_{i,t} - \beta_{i,t}\right)^2 + \text{const} \end{aligned} \tag{11}$$

where

$$\bar{h}_{i,t+1} \quad := \quad \mathsf{E}_q\left[h_{i,t+1}\right] \tag{12}$$

$$\alpha_{i,t} \quad := \quad x_t - \sum_{j\neq i} \mathsf{E}_q\left[\cos h_{j,t}\right] \tag{13}$$

$$\beta_{i,t} \quad := \quad y_t - \sum_{j\neq i} \mathsf{E}_q\left[\sin h_{j,t}\right]. \tag{14}$$

Variational Bayes and the Mean-Field Equations

The marginal distributions are clearly non-Gaussian because of the $\cos(\cdot)$ and $\sin(\cdot)$ terms.

But because these distributions are one-dimensional the expectations in (12) to (14) are easy to calculate numerically

– and so the mean-field equations can be iterated to convergence.

Question: How would you use this approach to move the arm smoothly from $\mathbf{v}_1 = (x_1, y_1)$ to $\mathbf{v}_T = (x_T, y_T)$ where the intermediate locations are not specified?

A particular application is displayed in Figure 28.6 from Barber

– note how well the approach works.

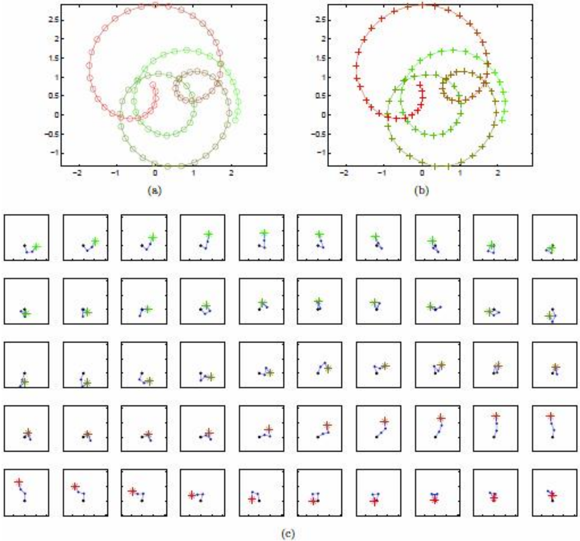


Figure 28.6 from Barber: (a): The desired trajectory of the end point of a three link robot arm. Green denotes time 1 and red time 100. (b): The learned trajectory based on a fully factorised KL variational approximation. (c): The robot arm sections every 2^{nd} time-step, from time 1 (top left) to time 100 (bottom right). The control problem of matching the trajectory using a smoothly changing angle set is solved using this simple approximation.

Expectation Propagation: Motivation

Reverse Kullback-Leibler minimization solves

$$\min_{q} \mathsf{KL}(p \,||\, q) = \min_{q} \mathsf{E}_p \left[\log \left(\frac{p}{q} \right) \right]$$

Suppose now we insist that q be a member of the exponential family so that

$$q(\mathbf{z}) = h(\mathbf{z}) g(\boldsymbol{\eta}) e^{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{z})} \tag{15}$$

– where we've suppressed any dependence on observed data $\mathcal{D}$.

Then

$$\mathsf{KL}(p \,||\, q) = -\ln g(\boldsymbol{\eta}) - \boldsymbol{\eta}^\top \mathsf{E}_p[u(\mathbf{z})] + \text{constant}$$

so the optimal estimate $\boldsymbol{\eta}^*$ satisfies

$$-\boldsymbol{\nabla}_{\boldsymbol{\eta}} \ln g(\boldsymbol{\eta}^*) = \mathsf{E}_p[\mathbf{u}(\mathbf{z})].$$

But also known that $-\boldsymbol{\nabla}_{\boldsymbol{\eta}} \ln g(\boldsymbol{\eta}^*) = \mathsf{E}_{q(\boldsymbol{\eta}^*)}[\mathbf{u}(\mathbf{z})]$

- follows by differentiating $\int h(\mathbf{z}) g(\boldsymbol{\eta}) e^{\boldsymbol{\eta}^\top \mathbf{u}(\mathbf{z})} \, d\mathbf{z} = 1$ wrt $\boldsymbol{\eta}$.

Therefore obtain the following moment matching result:

$$\mathsf{E}_{q(\boldsymbol{\eta}^*)}[\mathbf{u}(\mathbf{z})] = \mathsf{E}_p[\mathbf{u}(\mathbf{z})] \tag{16}$$

Expectation Propagation

Can use this observation to create an algorithm – expectation propagation (EP) – for approximate inference.

In many applications the joint distribution of the hidden variables (including unknown parameters) $\mathbf{z}$ has the form

$$p(\mathcal{D}, \mathbf{z}) = \prod_i f_i(\mathbf{z}) \tag{17}$$

Note that (17) arises in many applications including:

1. A model for IID data where there is one factor $f_i(\mathbf{z}) := p(\mathbf{x}_i \mid \mathbf{z})$ for each datapoint $\mathbf{x}_i$ and a factor $f_0(\mathbf{z}) = p(\mathbf{z})$ for the prior.
2. Models defined by directed acyclic graphs (DAGs) or undirected graphs.

The posterior is then given by

$$p(\mathbf{z} \mid \mathcal{D}) = \frac{1}{Z_p} \prod_i f_i(\mathbf{z}).$$

EP approximates this posterior with a distribution q that has the same factorization so that

$$q(\mathbf{z}) = \frac{1}{Z_q} \prod_i \tilde{f}_i(\mathbf{z}).$$

Expectation Propagation

Will assume that the $\tilde{f}_i(\mathbf{z})$'s come from the exponential family. Would then like to solve

$$\min_{q} \mathsf{KL}(p \,||\, q) = \min_{\text{all } \tilde{f}_i} \mathsf{KL}\left(\frac{1}{Z_p} \prod_i f_i(\mathbf{z}) \,\Big\|\, \frac{1}{Z_q} \prod_i \tilde{f}_i(\mathbf{z}) \right)$$

- intractable in general since expectation is wrt unknown true distribution p.

Could try to solve $\min_{\tilde{f}_i} \mathsf{KL}(f_i \,||\, \tilde{f}_i)$ for all i

- much easier but tends to result in poor approximation to joint distribution.

Instead EP approximates each factor f_i in turn **in context** of all remaining factors.

In particular, suppose we have initialized all factors $\tilde{f}_i$ so that our initial approximation is

$$q(\mathbf{z}) \propto \prod_i \tilde{f}_i(\mathbf{z}).$$

Expectation Propagation

We then iterate the following steps until **convergence**:

1. Pick a factor $\tilde{f}_j(\mathbf{z})$
2. Define $\bar{q}_j(\mathbf{z}) := q(\mathbf{z})/\tilde{f}_j(\mathbf{z})$
3. Evaluate new posterior approximation by setting

$$\tilde{f}_j^*(\mathbf{z}) \;=\; \underset{\tilde{f}_j}{\operatorname{argmin}}\, \mathsf{KL}\left(\frac{f_j(\mathbf{z})\bar{q}_j(\mathbf{z})}{Z_j} \,\middle\|\, \frac{\tilde{f}_j(\mathbf{z})\bar{q}_j(\mathbf{z})}{\tilde{Z}_j}\right) \tag{18}$$

   where Z_j and $\tilde{Z}_j$ are just normalization factors.

   Can solve (18) using result in (16) since $\tilde{f}_j(\mathbf{z})\bar{q}_j(\mathbf{z})$ is from exponential family

   - assuming expectation in (16) with $p \propto f_j(\mathbf{z})\bar{q}_j(\mathbf{z})$ can be calculated.
4. Set $\tilde{f}_j(\mathbf{z}) = \tilde{f}_j^*(\mathbf{z})$.

Note that **convergence** is **not guaranteed** in general

- see Section 10.7 of Bishop for examples and further details.