Variational Algorithms for Approximate Bayesian Inference: Linear Regression

Nicholas Zabaras
Warwick Centre of Predictive Modelling (WCPM)
University of Warwick

February 7, 2015
Contents

- Variational Linear Regression, Predictive Distribution, Lower Bound, Selection of the order of the polynomial
- Variational Linear Regression with Gam(\beta|c_0, d_0)
- Exponential Family Distributions, Mixture of Gaussians
- Variational Message Passing

Following:
- Pattern Recognition and Machine Learning, Christopher M. Bishop, Chapter 10
- Machine Learning: A Probabilistic Perspective, Kevin Murphy, Chapter 21.
Variational Linear Regression

We can return to the Bayesian linear regression model considered earlier.

Recall that the likelihood function is given by

\[ p(t|w) = \prod_{n=1}^{N} \mathcal{N}(t_n|w^T \phi_n, \beta^{-1}) \]

where \( \phi_n = \phi(x_n) \).

We use the following conjugate prior distributions (i.e. Gaussian-Gamma):

\[ p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I) \]
\[ p(\alpha) = \text{Gam}(\alpha|a_0, b_0) \]

To simplify discussion the noise precision parameter \( \beta \) is fixed and assumed to be known. The framework can be extended with this assumption being relaxed.

Variational Linear Regression

Collectively we have the joint distribution

\[ p(t, w, \alpha) = p(t|w)p(w|\alpha)p(\alpha) \]

which can be represented by the following graphical model:

We assume a variational posterior of the following form to obtain the unknown hyperparameters

\[ q(w, \alpha) = q(w)q(\alpha) \]
Variational Linear Regression

Using the same framework as previously discussed, we obtain for $q(\alpha)$:

$$
\ln q^*(\alpha) = \ln p(\alpha) + \mathbb{E}_w[\ln p(w|\alpha)] + \text{const}
$$

$$
= (a_0 - 1) \ln \alpha - b_0 \alpha + \frac{M}{2} \ln \alpha - \frac{\alpha}{2} \mathbb{E}[w^T w] + \text{const}
$$

As expected this gives a Gamma distribution

$$
q^*(\alpha) = \text{Gam}(\alpha|a_N, b_N)
$$

with

$$
a_N = a_0 + \frac{M}{2}
$$

$$
b_N = b_0 + \frac{1}{2} \mathbb{E}[w^T w]$$
Variational Linear Regression

Similarly we obtain for \( q(\mathbf{w}) \):

\[
\ln q^*(\mathbf{w}) = \ln p(\mathbf{t}|\mathbf{w}) + \mathbb{E}_\alpha [\ln p(\mathbf{w}|\alpha)] + \text{const}
\]

\[
= -\frac{\beta}{2} \sum_{n=1}^{N} \{\mathbf{w}^T \phi_n - t_n\}^2 - \frac{1}{2} \mathbb{E}[\alpha] \mathbf{w}^T \mathbf{w} + \text{const}
\]

\[
= -\frac{1}{2} \mathbf{w}^T (\mathbb{E}[\alpha] \mathbf{I} + \beta \Phi^T \Phi) \mathbf{w} + \beta \mathbf{w}^T \Phi^T \mathbf{t} + \text{const}
\]

As expected this gives a normal distribution

\[
q^*(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)
\]

with

\[
\mathbf{m}_N = \beta \mathbf{S}_N \Phi^T \mathbf{t}
\]

\[
\mathbf{S}_N = (\mathbb{E}[\alpha] \mathbf{I} + \beta \Phi^T \Phi)^{-1}
\]

Recall that \( \Phi^T \Phi = (\phi_1 \ldots \phi_N) \begin{pmatrix} \phi_1^T \\ \vdots \\ \phi_N^T \end{pmatrix} = \sum_{n=1}^{N} \phi_n \phi_n^T \)

These are the same formulas we obtained earlier when \( \alpha \) was treated as constant. In the VI approach, \( \alpha \) is replaced by \( \mathbb{E}[\alpha] \).
Variational Linear Regression

By standard properties for Gaussian and Gamma distributions we finally have

\[
\mathbb{E}[\alpha] = \frac{a_N}{b_N}
\]
\[
\mathbb{E}[ww^T] = m_N m_N^T + S_N
\]

which collectively give

\[
\mathbb{E}[\alpha] = \frac{a_N}{b_N} = \frac{a_0 + M/2}{b_0 + \mathbb{E}[w^T w]/2}
\]

For the case of \(a_0=b_0=0\) (infinitely broad prior over \(\alpha\)):

\[
\mathbb{E}[\alpha] = \frac{M}{m_N^T m_N + \text{tr}(S_N)}.
\]

This is similar to the result obtained earlier using the model evidence approximation.

It then remains a task to cycle between computing \(a_N, b_N, m_N\) and \(S_N\) until some convergence criterion is met.

These results are consistent with those obtained by maximizing the evidence using EM (except that the point estimate of \(\alpha\) is now replaced by \(\mathbb{E}[\alpha]\)) and give identical results in the case of an infinitely broad prior for which

\[
\mathbb{E}[\alpha] = \frac{M}{m_N^T m_N + \text{tr}(S_N)}.
\]
Predictive Distribution

The predictive distribution of $t$ for a new input $x$ is given as:

$$p(t|x, t) = \int p(t|x, w)p(w|t)dw \approx \int p(t|x, w)q(w)dw$$

$$= \int \mathcal{N}(t|w^T \phi(x), \beta^{-1})\mathcal{N}(w|m_N, S_N) \, dw = \mathcal{N}(t|m_N^T \phi(x), \sigma^2(x))$$

Here we used an earlier result for linear Gaussian models. Given a marginal Gaussian distribution for $x$ and a conditional Gaussian distribution for $y$ given $x$ in the form

$$p(x) = \mathcal{M}(x|\mu, \Lambda^{-1})$$

$$p(y|x) = \mathcal{N}(y|Ax + b, L^{-1})$$

the marginal distribution of $y$ is given by

$$p(y) = \mathcal{N}(y|A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$$

Using the above result, the input dependent variance is given as:

$$\sigma^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x), \text{ with } S_N = (\mathbb{E}[\alpha]I + \beta \Phi^T \Phi)^{-1}$$

This is the same form we obtained earlier when $\alpha$ was treated as constant but now $\alpha$ is replaced by $\mathbb{E}[\alpha]$. 
But how can we select M (degree of polynomial)? We can compute the lower bound

\[ \mathcal{L}(q) = \mathbb{E}[\ln p(w, \alpha, t)] - \mathbb{E}[\ln q(w, \alpha)] \]

\[ = \mathbb{E}_w[\ln p(t|w)] + \mathbb{E}_{w,\alpha}[\ln p(w|\alpha)] + \mathbb{E}_\alpha[\ln p(\alpha)] - \mathbb{E}_w[\ln q(w)] - \mathbb{E}_\alpha[\ln q(\alpha)] \]

where

\[ \mathbb{E}_w[\ln p(t|w)] = -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln(\beta) - \frac{\beta}{2} \mathbb{E} \left[ \sum_{n=1}^{N} (w^T \phi_n - t_n)^2 \right] \]

\[ = -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln(\beta) - \frac{\beta}{2} \left\{ t^T t - 2\mathbb{E}[w^T \Phi^T t] + Tr(\mathbb{E}[ww^T] \Phi^T \Phi) \right\} \]

\[ = \frac{N}{2} \ln \frac{\beta}{2\pi} - \frac{\beta}{2} t^T t + \beta m_N^T \Phi^T t - \frac{\beta}{2} Tr[\Phi^T \Phi (m_N m_N^T + S_N)] \]

\[ \mathbb{E}_{w,\alpha}[\ln p(w|\alpha)] = -\frac{M}{2}\ln(2\pi) + \frac{M}{2}\mathbb{E}[\ln \alpha] - \frac{\mathbb{E}[\alpha]}{2} \mathbb{E}[w^T w] \]

\[ = -\frac{M}{2}\ln(2\pi) + \frac{M}{2}(\psi(a_N) - \ln b_N) - \frac{a_N}{2b_N} [m_N^T m_N + Tr(S_N)] \]

\[ \mathbb{E}_\alpha[\ln p(\alpha)] = a_0 \ln b_0 + (a_0 - 1)\mathbb{E}[\ln \alpha] - b_0 \mathbb{E}[\alpha] - \ln \Gamma(a_0) \]

\[ = a_0 \ln b_0 + (a_0 - 1)[\psi(a_N) - \ln b_N] - \frac{b_0 a_N}{b_N} - \ln \Gamma(a_0) \]
Lower Bound

The final two terms in $\mathcal{L}(q)$ represent the entropies of the Gaussian and Gamma distributions:

$$-\mathbb{E}_w[\ln q(w)] = \frac{1}{2} \ln |\mathbf{S}_N| + \frac{M}{2} [1 + \ln 2\pi]$$

$$-\mathbb{E}_\alpha[\ln q(\alpha)] = \ln \Gamma(a_N) - (a_N - 1) \psi(a_N) - \ln b_N + a_N$$

We substitute in the Eqs above the following expressions for the moments:

$$\mathbb{E}[\mathbf{w}] = \mathbf{m}_N$$
$$\mathbb{E}[\mathbf{w}\mathbf{w}^T] = \mathbf{m}_N \mathbf{m}_N^T + \mathbf{S}_N$$
$$\mathbb{E}[\alpha] = \frac{a_N}{b_N}$$
$$\mathbb{E}[\ln \alpha] = \psi(a_N) - \ln b_N$$

to obtain the final expression for $\mathcal{L}(q)$:

$$\mathcal{L}(q) = \mathbb{E}_w[\ln p(t|\mathbf{w})] + \mathbb{E}_{w,\alpha}[\ln p(\mathbf{w}|\alpha)] + \mathbb{E}_\alpha[\ln p(\alpha)] - \mathbb{E}_w[\ln q(\mathbf{w})] - \mathbb{E}[\ln q(\alpha)]$$
10 data points have been generated from a polynomial of degree 3 over \([-5, 5]\) with additive Gaussian noise. The lower bound is maximized for \(M=3\) corresponding to the true model form which the data was generated.

\(\mathcal{L}\) represents lower bound on the log marginal likelihood \(\ln p(t|M)\) for the model.

If we assign equal prior probabilities \(p(M)\) to the different values of \(M\), then we can interpret \(\mathcal{L}\) as an approximation to \(p(M|t)\).

Thus the variational framework assigns the highest probability to the model with \(M = 3\).

This should be contrasted with the MLE result, which assigns ever smaller residual error to models of increasing complexity until the residual error is driven to zero, causing MLE to over-fitted models.

Prior parameters \(a_0 = b_0 = 0\), corresponding to the noninformative prior \(p(\alpha) \propto 1/\alpha\), which is uniform over \(\ln \alpha\).

Python Code
Lower Bound Vs the Order of the Polynomial

Run `linregEbModelSelVsN` in the PMTK3 toolbox
Variational Linear Regression with \( \text{Gam}(\beta|c_0, d_0) \)

We now extend the variational treatment of Bayesian linear regression to include a gamma hyperprior \( \text{Gam}(\beta|c_0, d_0) \) over \( \beta \) and solve variationally, by assuming a factorized variational distribution of the form \( q(w)q(\alpha)q(\beta) \).

We modify the joint distribution of all variables as:

\[
p(t,w,\alpha, \beta) = (t|w, \beta)p(w|\alpha)p(\alpha)p(\beta)
\]

The formulae for \( p(\alpha) \) remain the same:

\[
q^*(\alpha) = \text{Gam}(\alpha|a_N, b_N), a_N = a_0 + \frac{M}{2}, = b_0 + \frac{1}{2} \mathbb{E}[w^T w]
\]

For \( q^*(w) \) we have:

\[
\ln q^*(w) = \ln p(t|w, \beta) + \mathbb{E}_\alpha [\ln p(w|\alpha)] + \text{const}
= - \frac{\mathbb{E}[\beta]}{2} \sum_{n=1}^{N} \{w^T \phi_n - t_n\}^2 - \frac{1}{2} \mathbb{E}[\alpha]w^T w + \text{const}
= - \frac{1}{2} w^T (\mathbb{E}[\alpha] I + \mathbb{E}[\beta] \Phi^T \Phi) w + \mathbb{E}[\beta] w^T \Phi^T t + \text{const}
\]

Thus \( q^*(w) = \mathcal{N}(w|m_N, S_N) \) with

\[
m_N = \mathbb{E}[\beta] S_N \Phi^T t
\]
\[
S_N = (\mathbb{E}[\alpha] I + \mathbb{E}[\beta] \Phi^T \Phi)^{-1}
\]
Variational Linear Regression with \( \text{Gam}(\beta | c_0, d_0) \)

For \( q^*(\beta) \)

\[
\ln q^*(\beta) = \mathbb{E} [\ln p(t|w, \beta)] + \ln p(\beta) + \text{const} \\
= \frac{N}{2} \ln \beta - \frac{\beta}{2} \mathbb{E}_w \left[ \sum_{n=1}^{N} \{w^T \phi_n - t_n\}^2 \right] + (c_0 - 1)\ln \beta - d_0 \beta + \text{const}
\]

We recognize the log of a Gamma distribution with:

\[
q^*(\beta) = \text{Gam}(\beta | c_N, d_N),
\]

\[
d_N = d_0 + \frac{1}{2} \mathbb{E} [\sum_{n=1}^{N} \{w^T \phi_n - t_n\}^2] = d_0 + \frac{1}{2} (\text{Tr}(\Phi^T \Phi) \mathbb{E}[ww^T] + t^T t - t^T \Phi \mathbb{E}[w]) = d_0 + \frac{1}{2} (\|t - \Phi m_N\|^2 + \text{Tr}(\Phi^T \Phi S_N))
\]

Where we used: \( \mathbb{E}[ww^T] = m_N m_N^T + S_N \) and \( \mathbb{E}[w] = m_N \), with \( m_N = \beta S_N \Phi^T t, S_N = (\mathbb{E}[\alpha] I + \beta \Phi^T \Phi)^{-1} \)

Thus \( \mathbb{E}[\beta] = \frac{c_N}{d_N} \)
### Variational Linear Regression with \( \text{Gam}(\beta|c_0, d_0) \)

The lower bound also needs to be modified. Starting with the modified log-likelihood and using 
\[
\mathbb{E}[\mathbf{w}] = \mathbf{m}_N, \quad \mathbb{E}[\beta] = \frac{c_N}{d_N}, \quad \mathbb{E}[\mathbf{w}^T \mathbf{w}] = \mathbf{m}_N \mathbf{m}_N^T + \mathbf{S}_N \quad \text{and} \quad \mathbb{E}[\beta] = \psi(c_N) - \ln d_N:
\]

\[
\mathbb{E}_\beta \left[ \mathbb{E}_\mathbf{w} \left[ \ln p(t|\mathbf{w}) \right] \right] = \frac{N}{2} \left( \mathbb{E}[\beta] - \ln(2\pi) \right) - \frac{\mathbb{E}[\beta]}{2} - \mathbb{E}[||t - \Phi \mathbf{w}||^2]
\]
\[
= \frac{N}{2} \left( \psi(c_N) - \ln d_N - \ln(2\pi) \right) - \frac{c_N}{2d_N} \left( ||t - \Phi \mathbf{w}||^2 + Tr(\Phi^T \Phi S_N) \right)
\]

Next using \( \mathbb{E}[\beta] = \frac{c_N}{d_N} \) and \( \mathbb{E}[\beta] = \psi(c_N) - \ln d_N \), we consider the term corresponding to log prior over \( \beta \):

\[
\mathbb{E}[\ln p(\beta)] = (c_0 - 1) \mathbb{E} [\ln \beta] - d_0 \mathbb{E} [\beta] + c_0 \ln d_0 - \ln \Gamma(c_0)
\]
\[
= (c_0 - 1) \left( \psi(c_N) - \ln d_N \right) - \frac{d_0 c_N}{d_N} + c_0 \ln d_0 - \ln \Gamma(c_0)
\]

Finally we compute the negative entropy of the posterior over \( \beta \):

\[
-\mathbb{E}[\ln q^*(\beta)] = (c_N - 1) \psi(c_N) + \ln d_N - c_N - \ln \Gamma(c_N)
\]

Finally the predictive distribution is given as:

\[
p(t|x, t) \approx \mathcal{N}(t | \mathbf{m}_N^T \Phi(x), \sigma^2(x)), \quad \sigma^2(x) = \frac{1}{\mathbb{E}[\beta]} + \Phi(x)^T \mathbf{S}_N \Phi(x)
\]
Exponential Family Distributions

For many models the complete-data likelihood is drawn from the exponential family. However, in general this will not be the case for the marginal likelihood function for the observed data.

E.g. in a mixture of Gaussians, the joint distribution of observations \( x_n \) and hidden variables \( z_n \) is a member of the exponential family but the marginal of \( x_n \) is a Gaussian mixture.

Here we make a distinction between latent variables, \( Z \), and parameters, \( \theta \), where parameters are intensive (fixed in number independent of the size of the data set), whereas latent variables are extensive (scale in number with the size of the data set).

E.g., in a Gaussian mixture model, the indicator variables \( z_{kn} \) (which specify which component \( k \) is responsible for generating data point \( x_n \)) represent the latent variables, whereas the means \( \mu_k \), precisions \( \Lambda_k \) and mixing proportions \( \pi_k \) represent the parameters.

Consider i.i.d. data \( X = \{x_n\}, n = 1, \ldots N \), with latent variables \( Z = \{z_n\} \). Let the joint distribution of observed and latent variables be a member of the exponential family, parameterized by natural parameters \( \eta \). Using a conjugate prior \( p(\eta | \xi_0, v_0) \), we write:

\[
p(X, Z | \eta) = \prod_{n=1}^{N} h(x_n, z_n) g(\eta) \exp{\eta^T u(x_n, z_n)}
\]

\[
p(\eta | \xi_0, v_0) = f(\xi_0, v_0) g(\eta)^{v_0} \exp{v_0 \eta^T \xi_0}
\]
Exponential Family Distributions

Consider a variational distribution \( q(Z, \eta) = q(Z)q(\eta) \). We solve for the two factors as follows

\[
\ln q^*(Z) = \mathbb{E}_\eta[\ln p(X, Z|\eta)] + \text{const} = \sum_{n=1}^{N} \{ \ln h(x_n, z_n) + \mathbb{E}[\eta^T]u(x_n, z_n) \} + \text{const}
\]

So it factorizes as: \( q^*(Z) = \prod_{n=1}^{N} q^*(z_n) \). This is an induced factorization leading (after normalization) to:

\[
q^*(z_n) = h(x_n, z_n)g(\mathbb{E}[\eta]) \exp\{\mathbb{E}[\eta^T]u(x_n, z_n)\}
\]

Similarly:

\[
\ln q^*(\eta) = \ln p(\eta|X_0, v_0) + \mathbb{E}_Z[\ln p(X, Z|\eta)] + \text{const} = v_0 \ln g(\eta) + v_0 \eta^T x_0 + \sum_{n=1}^{N} \{ \ln g(\eta) + \eta^T \mathbb{E}_Z[u(x_n, z_n)] \} + \text{const}
\]

From this after normalization we compute:

\[
q^*(\eta) = f(x_N, v_N)g(\eta)^v_N \exp\{v_N \eta^T x_N\}
\]

\[
v_N = v_0 + N, \quad v_N x_N = v_0 x_0 + \sum_{n=1}^{N} \mathbb{E}_Z[u(x_n, z_n)]
\]

\( q^*(z_n) \) and \( q^*(\eta) \) are coupled and we solve them iteratively.

- **E step:** compute \( \mathbb{E}_Z[u(x_n, z_n)] \) using \( q(z_n) \) and use this to compute a revised \( q(\eta) \).
- **M step:** use \( q(\eta) \) to find \( \mathbb{E}[\eta^T] \), which gives rise to a revised \( q^*(z_n) \).
Rewrite the model for the Bayesian mixture of Gaussians as a conjugate model from the exponential family.

Hence use the general results

\[
\ln q^*(Z) = \mathbb{E}_\eta[\ln p(X, Z|\eta)] + \text{const} = \sum_{n=1}^{N} \{\ln h(x_n, z_n) + \mathbb{E}[\eta^T]u(x_n, z_n)\} + \text{const}
\]

\[
q^*(\eta) = f(\chi_N, \nu_N)g(\eta)^{\nu_N} \exp\{\nu_N\eta^T\chi_N\}, \nu_N = \nu_0 + N, \nu_N \chi_N = \nu_0\chi_0 + \sum_{n=1}^{N} \mathbb{E}_{z_n}[u(x_n, z_n)]
\]

to derive the specific results

\[
q^*(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} \tau_{nk}^{z_{nk}}
\]

\[
q^*(\pi) = \text{Dir}(\pi|\alpha) \text{ where } \alpha_k = \alpha_0 + N_k
\]

\[
q^*(\mu_k, \Lambda_k) = N(\mu_k|m_k, (\beta_k\Lambda_k)^{-1})W(\Lambda_k|W_k, \nu_k)
\]

where we have defined:

\[
\beta_k = \beta_0 + N_k, m_k = \frac{1}{\beta_k}(\beta_0 m_0 + N_k \bar{x}_k), W_k^{-1} = W_0^{-1} + N_k S_k + \frac{\beta_0 N_k}{\beta_0 + N_k} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^T, \nu_k = \nu_0 + N_k
\]
Exponential Family and the Mixture of Gaussians

We start with the complete data log-likelihood using $p(X|Z, \mu, \Lambda) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{nk}}$ and $p(Z|\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}}$:

$$p(X, Z|\pi, \mu, \Lambda) = p(X|Z, \mu, \Lambda)p(Z|\pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left( \pi_k^{z_{nk}} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1}) \right)^{z_{nk}}$$

$$= \prod_{n=1}^{N} \exp \left\{ \sum_{k=1}^{K} z_{nk} \left( \ln \pi_k + \frac{1}{2} \ln |\Lambda_k| - \frac{D}{2} \ln (2\pi) - \frac{1}{2} (x_n - \mu_k)^T \Lambda_k (x_n - \mu_k) \right) \right\}$$

Exploring the similarities with $p(X, Z|\eta) = \prod_{n=1}^{N} h(x_n, z_n) g(\eta) \exp\{\eta^T u(x_n, z_n)\}$, we can write by inspection:

$$\eta = \begin{pmatrix} \Lambda_k & \mu_k \\ \mu_k^T \Lambda_k & \mu_k \\ \ln |\Lambda_k| \\ \ln \pi_k \end{pmatrix}_{k=1,\ldots,K}, u(x_n, z_n) = \begin{pmatrix} x_n \\ \frac{1}{2} x_n x_n^T \\ -\frac{1}{2} \\ \frac{1}{2} \\ 1 \end{pmatrix}_{k=1,\ldots,K}$$

$$, h(x_n, z_n) = \prod_{k=1}^{K} \left( (2\pi)^{-D/2} \right)^{z_{nk}}, g(\eta) = 1$$

Arrows above matrices return a vector formed by stacking the columns of the matrix on top of each other.

Also $(v_k)_{k=1,\ldots,K} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_K \end{pmatrix}$
Exponential Family and the Mixture of Gaussians

Similarly, the prior over the parameters can be written as:

\[
p(\pi, \mu, \Lambda) = \text{Dir}(\pi | a_0) \prod_{k=1}^{K} \mathcal{N}\left( \mu_k | m_0, (\beta_0 \Lambda_k)^{-1} \right) W(\Lambda_k | W_0, \nu_0)
\]

\[
= C(a_0) \left( \frac{\beta_0}{2\pi} \right)^{K D / 2} B(W_0, \nu_0)^K \exp \left\{ \sum_{k=1}^{K} (\alpha_0 - 1) \ln \pi_k + \frac{\nu_0 - D}{2} \ln |\Lambda_k| - \frac{1}{2} Tr(\Lambda_k [\beta_0 \mu_k]ight\}
\]
Exponential Family and the Mixture of Gaussians

By exponentiating both sides of

$$\ln q^*(Z) = \sum_{n=1}^{N} \{ \ln h(x_n, z_n) + \mathbb{E}[\eta^T u(x_n, z_n)] \} + \text{const},$$

we obtain

$$q^*(Z) = \prod_{n=1}^{N} \prod_{k=1}^{K} \rho_{nk}^{z_{nk}} \quad \text{with} \quad \ln \rho_{nk} = \mathbb{E}[\ln \pi_k] + \frac{1}{2} \mathbb{E}[\ln |\Lambda_k|] - \frac{D}{2} \ln(2\pi) - \frac{1}{2} \mathbb{E}_{\mu_k, \Lambda_k}(x_n - \mu_k)^T \Lambda_k^{-1} (x_n - \mu_k)$$
Exponential Family and the Mixture of Gaussians

Next we can use $\mathbb{E}[z_{nk}] = r_{nk}$ to take the expectation wrt $Z$ in $v_N x_N = v_0 x_0 + \Sigma_{n=1}^N \mathbb{E}_{z_n}[u(x_n, z_n)]$, substituting $r_{nk}$ for $\mathbb{E}[z_{nk}]$ in

$$u(x_n, z_n) = \begin{pmatrix} x_n \\ \frac{1}{2} x_n x_n^T \\ -\frac{1}{2} \\ 1 \\ 2 \\ 1 \end{pmatrix}$$

$k=1,...,K$

Combining this with $\chi_0 = \left(\begin{array}{c} \beta_0 m_0 \\ -\frac{1}{2} \left(\beta_0 m_0 m_0^T + W_0^{-1}\right) \\ -\frac{\beta_0}{2} \\ \frac{\nu_0 - D}{2} \\ \alpha_0 - 1 \end{array}\right)$, Eqs. $v_N = v_0 + N and v_N x_N = v_0 x_0 + \Sigma_{n=1}^N \mathbb{E}_{z_n}[u(x_n, z_n)]$ become as follows:
Exponential Family and the Mixture of Gaussians

\[ v_N = v_0 + N = 1 + N \] and,

\[ v_N \mathbf{x}_N = \left( \begin{array}{c}
\frac{\beta_0 \mathbf{m}_0}{2} \\
- \frac{1}{2} \left( \frac{\beta_0 \mathbf{m}_0 \mathbf{m}_0^T}{2} + \mathbf{W}_0^{-1} \right) \\
- \frac{\beta_0}{2} \\
\nu_0 - D \\
\alpha_0 - 1
\end{array} \right) + \sum_{n=1}^{N} \left( \begin{array}{c}
\frac{x_n}{2} \\
- \frac{1}{2} x_n x_n^T \\
\frac{1}{2} \\
1
\end{array} \right)_{k=1, \ldots, K}

From the bottom row of \( \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n) \) and the Eq. above, we see that the inner product of \( \eta \) and \( v_N \mathbf{x}_N \) gives us the r.h.s. of \( \ln q^*(\mathbf{\pi}) = (\alpha_0 - 1) \sum_{k=1}^{K} \ln \pi_k + \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} \ln \pi_k + \text{const} \) from which \( \ln q^*(\mathbf{\pi}) = (\alpha_0 - 1) \sum_{k=1}^{K} \ln \pi_k + N_k \sum_{k=1}^{K} \ln \pi_k + \text{const} \). e. \( q^*(\mathbf{\pi}) = \text{Dir}(\mathbf{\pi}|\alpha) \) where \( \alpha_k = \alpha_0 + N_k \)
Exponential Family and the Mixture of Gaussians

The remaining terms of the inner product are:

\[
\sum_{k=1}^{K} \left\{ \mu_k^T A_k \left( \beta_0 m_0 + N_k \bar{x}_k \right) - \frac{1}{2} Tr \left( A_k \left[ \beta_0 m_0 m_0^T + W_0^{-1} + N_k \left( S_k + \bar{x}_k \bar{x}_k^T \right) \right] \right) \right\}
\]
Exponential Family and the Mixture of Gaussians

The remaining terms of the inner product are:

\[
\sum_{k=1}^{K} \left\{ \mu_k^T \Lambda_k \left( \beta_0 m_0 + N_k \bar{x}_k \right) - \frac{1}{2} Tr \left( \Lambda_k \left[ \beta_0 m_0 m_0^T + W_0^{-1} + N_k \left( S_k + \bar{x}_k \bar{x}_k^T \right) \right] \right\}
\]
Variational Message Passing

Here we consider more generally the use of variational methods for models described by directed graphs and derive a number of widely applicable results.

The joint distribution corresponding to a directed graph can be written using the decomposition

\[ p(x) = \prod_i p(x_i | pa_i) \]

where \( x_i \) denotes the variable(s) associated with node \( i \), and \( pa_i \) denotes the parent set corresponding to node \( i \). Note that \( x_i \) may be a latent variable or it may belong to the set of observed variables.

Now consider a variational approximation in which the distribution \( q(x) \) is assumed to factorize with respect to the \( x_i \) so that

\[ q(x) = \prod_i q_i(x_i) \]

Note that for observed nodes, there is no factor \( q_i(x_i) \) in the variational distribution.

Using our general result to give

\[ \ln q_j^*(x_j) = \mathbb{E}_{i \neq j} \left[ \sum_i \ln p(x_i | pa_i) \right] + \text{const}. \]
Variational Message Passing

\[ \ln q^*_j(x_j) = \mathbb{E}_{i \neq j} \left[ \sum_i \ln p(x_i | p_a_i) \right] + \text{const}. \]

The only terms that do depend on \( x_j \) are \( p(x_j | p_a_j) \) together with the conditional distributions corresponding to the children of node \( j \), and they therefore also depend on the co-parents of the child nodes, i.e., the other parents of the child nodes besides node \( x_j \) itself.

The set of all nodes on which \( q(x_j) \) depends correspond to the Markov blanket of node \( x_j \).

Thus the update of the factors in the variational posterior distribution represents a local calculation on the graph.

This makes possible the construction of general purpose software for variational inference in which the form of the model does not need to be specified in advance (Bishop et al., 2003).

If we now specialize to the case of a model in which all of the conditional distributions have a conjugate-exponential structure, then the variational update procedure can be cast in terms of a local message passing algorithm (Winn and Bishop, 2005).

Variational Message Passing

\[ \ln q_j^*(x_j) = \mathbb{E}_{i \neq j} \left[ \sum_i \ln p(x_i | p_{a_i}) \right] + \text{const}. \]

The distribution associated with a particular node can be updated once that node has received messages from all of its parents and all of its children. This in turn requires that the children have already received messages from their coparents.

The evaluation of the lower bound can also be simplified because many of the required quantities are already evaluated as part of the message passing scheme.

This distributed message passing formulation has good scaling properties and is well suited to large networks.