Variational Bayesian inference

Kay H. Brodersen
Translational Neuromodeling Unit (TNU)
Institute of Biomedical Engineering
University of Zurich & ETH Zurich

Illustration adapted from: Mike West
“An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.”

John W. Tukey, 1915 – 2000
Bayesian inference formalizes *model inversion*, the process of passing from a prior to a posterior in light of data.

\[
p(\theta | y) = \frac{p(y | \theta) \cdot p(\theta)}{\int p(y, \theta) d\theta}
\]

In practice, evaluating the posterior is usually difficult because we cannot easily evaluate \( p(y) \), especially when:

- analytical solutions are not available
- numerical integration is too expensive
There are two approaches to approximate inference. They have complementary strengths and weaknesses.
Approximate Bayesian inference

There are two approaches to approximate inference. They have complementary strengths and weaknesses.

Stochastic approximate inference
in particular sampling

1. design an algorithm that draws samples $\theta^{(1)}, \ldots, \theta^{(m)}$ from $p(\theta|y)$
2. inspect sample statistics (e.g., histogram, sample quantiles, ...)

☑ asymptotically exact
☒ computationally expensive
☒ tricky engineering concerns

Structural approximate inference
in particular variational Bayes

1. find an analytical proxy $q(\theta)$ that is maximally similar to $p(\theta|y)$
2. inspect distribution statistics of $q(\theta)$ (e.g., mean, quantiles, intervals, ...)

☑ often insightful – and lightning-fast!
☒ often hard work to derive
☒ requires validation via sampling
Overview

1. The Laplace approximation
2. Variational Bayes
3. Variational density estimation
4. Variational linear regression
5. Variational clustering
1  The Laplace approximation
2  Variational Bayes
3  Variational density estimation
4  Variational linear regression
5  Variational clustering
The Laplace approximation provides a way of approximating a density whose normalization constant we cannot evaluate, by fitting a Gaussian distribution to its mode.

\[ p(z) = \frac{1}{Z} \times f(z) \]

- **Normalization constant** (unknown)
- **Main part of the density** (easy to evaluate)

This is exactly the situation we face in Bayesian inference:

\[ p(\theta | y) = \frac{1}{p(y)} \times p(y, \theta) \]

- **Model evidence** (unknown)
- **Joint density** (easy to evaluate)
The Taylor approximation

The evaluation of any function \( f(x) \) can be approximated by a series:

\[
f(x) \approx f(x^*) + f'(x^*)(x - x^*) + \frac{1}{2!} f''(x^*)(x - x^*)^2 + \frac{1}{3!} f'''(x^*)(x - x^*)^3 + \cdots
\]

Brook Taylor
(1685 – 1731)
English mathematician, introduced Taylor series
Deriving the Laplace approximation

We begin by expressing the log-joint density $\mathcal{L}(\theta) \equiv \ln p(y, \theta)$ in terms of a second-order Taylor approximation around the mode $\theta^*$:

$$\mathcal{L}(\theta) \approx \mathcal{L}(\theta^*) + \mathcal{L}'(\theta^*)(\theta - \theta^*) + \frac{1}{2} \mathcal{L}''(\theta^*)(\theta - \theta^*)^2$$

$$= \mathcal{L}(\theta^*) + \frac{1}{2} \mathcal{L}''(\theta^*) (\theta - \theta^*)^2$$

This already has the same form as a Gaussian density:

$$\ln \mathcal{N}(\theta | \mu, \eta^{-1}) = \frac{1}{2} \ln \eta - \frac{1}{2} \ln 2\pi - \frac{\eta}{2} (\theta - \mu)^2$$

$$= \frac{1}{2} \ln \frac{\eta}{2\pi} + \frac{1}{2} (-\eta) (\theta - \mu)^2$$

And so we have an approximate posterior:

$$q(\theta) = \mathcal{N}(\theta | \mu, \eta^{-1}) \text{ with } \mu = \theta^* \quad \text{(mode of the log-posterior)}$$

$$\eta = -\mathcal{L}''(\theta^*) \quad \text{(negative curvature at the mode)}$$
Applying the Laplace approximation

Given a model with parameters $\theta = (\theta_1, \ldots, \theta_p)$, the Laplace approximation reduces to a simple three-step procedure:

1. Find the mode of the log-joint:
   \[
   \theta^* = \arg \max_{\theta} \ln p(y, \theta)
   \]

2. Evaluate the curvature of the log-joint at the mode:
   \[
   \nabla^2 \ln p(y, \theta^*)
   \]

3. We obtain a Gaussian approximation:
   \[
   \mathcal{N}(\theta | \mu, \Lambda^{-1}) \quad \text{with} \quad \mu = \theta^*, \quad \Lambda = -\nabla^2 \ln p(y, \theta^*)
   \]
The Laplace approximation: demo

~kbroders/teaching/vb_gui.m
The Laplace approximation is often too strong a simplification.

- Ignores global properties of the posterior when the posterior is multimodal.
- Becomes brittle when the posterior is multimodal.
- Only directly applicable to real-valued parameters.
Overview

1. The Laplace approximation
2. Variational Bayes
3. Variational density estimation
4. Variational linear regression
5. Variational clustering
Variational Bayesian (VB) inference generalizes the idea behind the Laplace approximation. In VB, we wish to find an approximate density that is maximally similar to the true posterior.
Variational Bayesian inference is based on variational calculus.

**Standard calculus**

Newton, Leibniz, and others

- functions  \( f: x \mapsto f(x) \)
- derivatives  \( \frac{df}{dx} \)

**Example:** maximize the likelihood expression  \( p(y|\theta) \) w.r.t.  \( \theta \)

**Variational calculus**

Euler, Lagrange, and others

- functionals  \( F: f \mapsto F(f) \)
- derivatives  \( \frac{dF}{df} \)

**Example:** maximize the entropy  \( H[p] \) w.r.t. a probability distribution  \( p(x) \)

Leonhard Euler (1707 – 1783)
Swiss mathematician, ‘Elementa Calculi Variationum’
Variational calculus and the free energy

Variational calculus lends itself nicely to approximate Bayesian inference.

\[
\ln p(y) = \ln \frac{p(y, \theta)}{p(\theta | y)} = \int q(\theta) \ln \frac{p(y, \theta)}{p(\theta | y)} \, d\theta = \int q(\theta) \ln \frac{p(y, \theta)}{q(\theta)} \frac{q(\theta)}{p(\theta | y)} \, d\theta = \int q(\theta) \left( \ln \frac{q(\theta)}{p(\theta | y)} + \ln \frac{p(y, \theta)}{q(\theta)} \right) \, d\theta = \int q(\theta) \ln \frac{q(\theta)}{p(\theta | y)} \, d\theta + \int q(\theta) \ln \frac{p(y, \theta)}{q(\theta)} \, d\theta
\]

- \( \text{KL}[q || p] \) divergence between \( q(\theta) \) and \( p(\theta | y) \)
- \( F(q, y) \) free energy
Variational calculus and the free energy

In summary, the log model evidence can be expressed as:

\[ \ln p(y) = \text{KL}[q||p] + F(q, y) \]

- divergence \( \geq 0 \) (unknown)
- free energy (easy to evaluate for a given \( q \))

Maximizing \( F(q, y) \) is equivalent to:

- minimizing \( \text{KL}[q||p] \)
- tightening \( F(q, y) \) as a lower bound to the log model evidence

\* In this illustrative example, the log model evidence and the free energy are positive; but the above equivalences hold just as well when the log model evidence is negative.
Computing the free energy

We can decompose the free energy $F(q, y)$ as follows:

$$F(q, y) = \int q(\theta) \ln \frac{p(y, \theta)}{q(\theta)} \, d\theta$$

$$= \int q(\theta) \ln p(y, \theta) \, d\theta - \int q(\theta) \ln q(\theta) \, d\theta$$

$$= \langle \ln p(y, \theta) \rangle_q + H[q]$$

- **expected log-joint**
- **Shannon entropy**
The Laplace approximation: demo

~kbroders/teaching/vb_gui.m
When inverting models with several parameters, a common way of restricting the class of approximate posteriors $q(\theta)$ is to consider those posteriors that factorize into independent partitions,

$$q(\theta) = \prod_i q_i(\theta_i),$$

where $q_i(\theta_i)$ is the approximate posterior for the $i^{th}$ subset of parameters.
Typical strategies in variational inference

| No mean-field assumption | No parametric assumptions | Parametric assumptions $q(\theta) = F(\theta|\delta)$ |
|--------------------------|---------------------------|-----------------------------------------------------|
|                          | (variational inference = exact inference) | fixed-form optimization of moments |
| Mean-field assumption $q(\theta) = \prod q(\theta_i)$ | Iterative free-form variational optimization | Iterative fixed-form variational optimization |
Variational inference under the mean-field assumption

\[
F(q, y) = \int q(\theta) \ln \frac{p(y, \theta)}{q(\theta)} d\theta
\]

\[
= \int \prod_i q_i \times \left( \ln p(y, \theta) - \sum_i \ln q_i \right) d\theta
\]

\[
= \int q_j \prod_{\backslash j} q_i \left( \ln p(y, \theta) - \ln q_j \right) d\theta - \int q_j \prod_{\backslash j} q_i \sum_{\backslash j} \ln q_i d\theta
\]

\[
= \int q_j \left( \int \prod_{\backslash j} q_i \ln p(y, \theta) d\theta_{\backslash j} - \ln q_j \right) d\theta_j - \int q_j \int \prod_{\backslash j} q_i \ln \prod_{\backslash j} q_i d\theta_{\backslash j} d\theta_j
\]

\[
= \int q_j \ln \frac{\exp \left( \langle \ln p(y, \theta) \rangle_{q_{\backslash j}} \right)}{q_j} d\theta_j + c
\]

\[
= -\text{KL} \left[ q_j || \exp \left( \langle \ln p(y, \theta) \rangle_{q_{\backslash j}} \right) \right] + c
\]
In summary:
\[
F(q, y) = -\text{KL} \left[ q_j \| \exp \left( \langle \ln p(y, \theta) \rangle_{q_{\theta \setminus j}} \right) \right] + c
\]

Suppose the densities \( q_{\theta \setminus j} \equiv q(\theta_{\setminus j}) \) are kept fixed. Then the approximate posterior \( q(\theta_j) \) that maximizes \( F(q, y) \) is given by:
\[
q_j^* = \arg \max_{q_j} F(q, y) \\
= \frac{1}{Z} \exp \left( \langle \ln p(y, \theta) \rangle_{q_{\theta \setminus j}} \right)
\]

Therefore:
\[
\ln q_j^* = \langle \ln p(y, \theta) \rangle_{q_{\theta \setminus j}} - \ln Z \\
\quad =: I(\theta_j)
\]

This implies a straightforward algorithm for variational inference:

1. Initialize all approximate posteriors \( q(\theta_i) \), e.g., by setting them to their priors.
2. Cycle over the parameters, revising each given the current estimates of the others.
3. Loop until convergence.
Frameworks for approximate inference

**Variational Bayes**
minimize $\text{KL}[q(\theta)||p(\theta|y)]$

$q(\theta)$ will tend to be zero where $p(\theta|y)$ is zero

may lead to a local minimum

**Expectation propagation**
minimize $\text{KL}[p(\theta|y)||q(\theta)]$

$q(\theta)$ will tend to be nonzero where $p(\theta|y)$ is nonzero

averaging across modes may lead to poor predictive performance

Bishop (2005) PRML, pp. 468 – 469
Overview

1. The Laplace approximation
2. Variational Bayes
3. Variational density estimation
4. Variational linear regression
5. Variational clustering
We are given a univariate dataset \( \{y_1, \ldots, y_n\} \) which we model by a simple univariate Gaussian distribution. We wish to infer on its mean and precision:

\[
p(\mu, \tau | y)
\]

Although in this case a closed-form solution exists*, we shall pretend it does not. Instead, we consider approximations that satisfy the mean-field assumption:

\[
q(\mu, \tau) = q_\mu(\mu) \ q_\tau(\tau)
\]

Application 1: variational density estimation

\[
p(\mu | \tau) = \mathcal{N}(\mu | \mu_0, (\lambda_0 \tau)^{-1})
\]

\[
p(\tau) = \text{Ga}(\tau | a_0, b_0)
\]

\[
p(y_i | \mu, \tau) = \mathcal{N}(y_i | \mu, \tau^{-1})
\]

\[i = 1 \ldots n\]

* Exercise 2.44; Bishop (2005) PRML
Recurring expressions in Bayesian inference

**Univariate normal distribution**

\[
\ln \mathcal{N}(x|\mu, \lambda^{-1}) = \frac{1}{2} \ln \lambda - \frac{1}{2} \ln \pi - \frac{\lambda}{2} (x - \mu)^2
\]

\[= -\frac{1}{2} \lambda x^2 + \lambda \mu x + c\]

**Multivariate normal distribution**

\[
\ln \mathcal{N}_d(x|\mu, \Lambda^{-1}) = -\frac{1}{2} \ln |\Lambda^{-1}| - \frac{d}{2} \ln 2\pi - \frac{1}{2} (x - \mu)^T \Lambda (x - \mu)
\]

\[= -\frac{1}{2} x^T \Lambda x + x^T \Lambda \mu + c\]

**Gamma distribution**

\[
\ln \text{Ga}(x|a, b) = a \ln b - \ln \Gamma(a) + (a - 1) \ln x - b x
\]

\[= (a - 1) \ln x - b x + c\]
Variational density estimation: mean $\mu$

$$\ln q^*(\mu) = \langle \ln p(y, \mu, \tau) \rangle_{q(\tau)} + c$$

$$= \left( \ln \prod_{i} p(y_i | \mu, \tau) \right)_{q(\tau)} + \langle \ln p(\mu | \tau) \rangle_{q(\tau)} + \langle \ln p(\tau) \rangle_{q(\tau)} + c$$

$$= \langle \ln \prod \mathcal{N}(y_i | \mu, \tau^{-1}) \rangle_{q(\tau)} + \langle \ln \mathcal{N}(\mu | \mu_0, (\lambda_0 \tau)^{-1}) \rangle_{q(\tau)} + \langle \ln \text{Ga}(\tau | a_0, b_0) \rangle_{q(\tau)} + c$$

$$= \sum \left( -\frac{\tau}{2} (y_i - \mu)^2 \right)_{q(\tau)} + \left( -\frac{\lambda_0 \tau}{2} (\mu - \mu_0)^2 \right)_{q(\tau)} + c$$

$$= \sum -\frac{\langle \tau \rangle_{q(\tau)}}{2} y_i^2 + \langle \tau \rangle_{q(\tau)} n \bar{y} \mu - n \frac{\langle \tau \rangle_{q(\tau)}}{2} \mu^2 - \frac{\lambda_0 \langle \tau \rangle_{q(\tau)}}{2} \mu^2 + \lambda_0 \mu \mu_0 \langle \tau \rangle_{q(\tau)} - \frac{\lambda_0}{2} \mu_0^2 + c$$

$$= -\frac{1}{2} \left\{ n \langle \tau \rangle_{q(\tau)} + \lambda_0 \langle \tau \rangle_{q(\tau)} \right\} \mu^2 + \left\{ n \bar{y} \langle \tau \rangle_{q(\tau)} + \lambda_0 \mu_0 \langle \tau \rangle_{q(\tau)} \right\} \mu + c$$

$$\Rightarrow q^*(\mu) = \mathcal{N}(\mu | \mu_n, \lambda_n^{-1}) \quad \text{with} \quad \lambda_n = (\lambda_0 + n) \langle \tau \rangle_{q(\tau)}$$

$$\mu_n = \frac{n \bar{y} \langle \tau \rangle_{q(\tau)} + \lambda_0 \mu_0 \langle \tau \rangle_{q(\tau)}}{\lambda_n} = \frac{\lambda_0 \mu_0 + n \bar{y}}{\lambda_0 + n}$$
\[\ln q^*(\tau) = \langle \ln p(y, \mu, \tau) \rangle_{q(\mu)} + c\]
\[= \left( \ln \prod_{i=1}^{n} \mathcal{N}(y_i | \mu, \tau^{-1}) \right)_{q(\mu)} + \langle \ln \mathcal{N}(\mu | \mu_0, (\lambda_0 \tau)^{-1}) \rangle_{q(\mu)} + \langle \ln \text{Ga}(\tau | a_0, b_0) \rangle_{q(\mu)} + c\]
\[= \sum_{i=1}^{n} \left( \frac{1}{2} \ln \tau - \frac{\tau}{2} (y_i - \mu)^2 \right)_{q(\mu)} + \left( \frac{1}{2} \ln \lambda_0 \tau - \frac{\lambda_0 \tau}{2} (\mu - \mu_0)^2 \right)_{q(\mu)} + \langle (a_0 - 1) \ln \tau - b_0 \tau \rangle_{q(\mu)} + c\]
\[= \frac{n}{2} \ln \tau - \frac{\tau}{2} \langle \Sigma (y_i - \mu)^2 \rangle_{q(\mu)} + \frac{1}{2} \ln \lambda_0 + \frac{1}{2} \ln \tau - \frac{\lambda_0 \tau}{2} \langle (\mu - \mu_0)^2 \rangle_{q(\mu)} + (a_0 - 1) \ln \tau - b_0 \tau + c\]
\[= \left\{ \frac{n}{2} + \frac{1}{2} + (a_0 - 1) \right\} \ln \tau - \left\{ \frac{1}{2} \langle \Sigma (y_i - \mu)^2 \rangle_{q(\mu)} + \frac{\lambda_0}{2} \langle (\mu - \mu_0)^2 \rangle_{q(\mu)} + b_0 \right\} \tau + c\]
\[\Rightarrow q^*(\tau) = \text{Ga}(\tau | a_n, b_n) \quad \text{with} \quad a_n = a_0 + \frac{n + 1}{2}\]
\[b_n = b_0 + \frac{\lambda_0}{2} \langle (\mu - \mu_0)^2 \rangle_{q(\mu)} + \frac{1}{2} \langle \Sigma (y_i - \mu)^2 \rangle_{q(\mu)}\]
Variational density estimation: illustration

\[ p(\theta | y) \quad q(\theta) \]

\[ q^*(\theta) \]

Bishop (2005) PRML, p. 472
We consider a multiple linear regression model with a shrinkage prior on the regression coefficients.

We wish to infer on the coefficients $\beta$, their precision $\alpha$, and the noise precision $\lambda$. There is no analytical posterior

$$p(\beta, \alpha, \lambda|y).$$

We therefore seek a variational approximation:

$$q(\beta, \alpha, \lambda) = q_\beta(\beta) q_\alpha(\alpha) q_\lambda(\lambda).$$
Variational linear regression: coefficients precision $\alpha$

\[
\ln q^*(\alpha) = \langle \ln p(y, \beta, \alpha, \lambda) \rangle_{q(\beta, \lambda)} + c
\]
\[
= \left( \ln \prod \mathcal{N}(y_i | \beta^T x_i, \lambda^{-1}) \right)_{q(\beta)q(\lambda)} + \ln \mathcal{N}_d(\beta | 0, \alpha^{-1} I)_{q(\beta)q(\lambda)} + \langle \ln \Gamma(\alpha | a_0, b_0) \rangle_{q(\beta)q(\lambda)} + c
\]
\[
= \left( \frac{1}{2} \ln |\alpha^{-1} I| - \frac{d}{2} \ln 2\pi - \frac{1}{2} (\beta - 0)^T \alpha I (\beta - 0) \right)_{q(\beta)} + \langle a_0 \ln b_0 - \ln \Gamma(a_0) + (a_0 - 1) \ln \alpha - b_0 \alpha \rangle_{q(\beta)} + c
\]
\[
= \frac{d}{2} \ln \alpha - \frac{\alpha}{2} \langle \beta^T \beta \rangle_{q(\beta)} + (a_0 - 1) \ln \alpha - b_0 \alpha + c
\]
\[
= \left( \frac{d}{2} + a_0 - 1 \right) \ln \alpha - \left( \frac{1}{2} \langle \beta^T \beta \rangle_{q(\beta)} + b_0 \right) \alpha + c
\]

$\implies$ $q^*(\alpha) = \Gamma(\alpha | a_n, b_n)$ with $a_n = a_0 + \frac{d}{2}$

\[
b_n = b_0 + \frac{1}{2} \langle \beta^T \beta \rangle_{q(\beta)}
\]
Variational linear regression: coefficients $\beta$

\[
\ln q^*(\beta) = \langle \ln p(y, \beta, \alpha, \lambda) \rangle_{q(\alpha, \lambda)} + c
\]

\[
= \langle \ln \prod \mathcal{N}(y_i | \beta^T x_i, \lambda^{-1}) \rangle_{q(\alpha)q(\lambda)} + \langle \ln \mathcal{N}_d(\beta | 0, \alpha^{-1} I) \rangle_{q(\alpha)q(\lambda)} + \langle \ln \mathcal{Ga}(\alpha | a_0, b_0) \rangle_{q(\alpha)q(\lambda)} + c
\]

\[
= \sum_{i} \left( \frac{1}{2} \ln \lambda - \frac{1}{2} \ln 2\pi - \frac{\lambda}{2} (y_i - \beta^T x_i)^2 \right)_{q(\alpha)q(\lambda)} + \left( -\frac{1}{2} \ln |\alpha^{-1} I| - \frac{d}{2} \ln 2\pi - \frac{1}{2} \beta^T \alpha d\beta \right)_{q(\alpha)} + c
\]

\[
= -\frac{\langle \lambda \rangle_{q(\lambda)}}{2} \sum_{i} (y_i - \beta^T x_i)^2 - \frac{1}{2} \langle \alpha \rangle_{q(\alpha)} \beta^T \beta + c
\]

\[
= -\frac{\langle \lambda \rangle_{q(\lambda)}}{2} y^T y + \langle \lambda \rangle_{q(\lambda)} \beta^T X^T y - \frac{\langle \lambda \rangle_{q(\lambda)}}{2} \beta^T X^T X \beta - \frac{1}{2} \beta^T \langle \alpha \rangle_{q(\alpha)} I \beta + c
\]

\[
= -\frac{1}{2} \beta^T \left\{ \langle \lambda \rangle_{q(\lambda)} X^T X + \langle \alpha \rangle_{q(\alpha)} I \right\} \beta + \beta^T \langle \lambda \rangle_{q(\lambda)} X^T y + c
\]

\[
\implies q^*(\beta) = \mathcal{N}_d(\beta | \mu_n, \Lambda_n^{-1}) \quad \text{with} \quad \Lambda_n = \langle \alpha \rangle_{q(\alpha)} I + \langle \lambda \rangle_{q(\lambda)} X^T X, \quad \mu_n = \Lambda_n^{-1} \langle \lambda \rangle_{q(\lambda)} X^T y
\]
Variational linear regression: noise precision $\lambda$

$$\ln q^*(\lambda) = \langle \ln p(y, \beta, \alpha, \lambda) \rangle_{q(\beta, \alpha)} + c$$

$$= \left( \sum_i^n \frac{1}{2} \ln \lambda - \frac{1}{2} \ln 2\pi - \frac{\lambda}{2} (y_i - \beta^T x_i)^2 \right)_{q(\beta)q(\alpha)}$$

$$+ \left( c_0 \ln d_0 - \ln \Gamma(c_0) + (c_0 - 1) \ln \lambda - d_0 \lambda \right)_{q(\beta)q(\alpha)} + c$$

$$= \frac{n}{2} \ln \lambda - \frac{\lambda}{2} y^T y + \lambda \langle \beta \rangle^T_{q(\beta)} X^T y - \frac{\lambda}{2} \langle \beta \rangle^T_{q(\beta)} X^T X \langle \beta \rangle_{q(\beta)} + (c_0 - 1) \ln \lambda - d_0 \lambda + c$$

$$= \left\{ c_0 + \frac{n}{2} - 1 \right\} \ln \lambda - \left\{ \frac{1}{2} y^T y - \langle \beta \rangle^T_{q(\beta)} X^T y + \frac{1}{2} \langle \beta \rangle^T_{q(\beta)} X^T X \langle \beta \rangle_{q(\beta)} + d_0 \right\} \lambda + c$$

$$\Rightarrow q^*(\lambda) = \text{Ga}(\lambda|c_n, d_n), \quad c_n = c_0 + \frac{n}{2}$$

$$d_n = d_0 + \frac{1}{2} y^T y - \langle \beta \rangle^T_{q(\beta)} X^T y + \frac{1}{2} \langle \beta \rangle^T_{q(\beta)} X^T X \langle \beta \rangle_{q(\beta)}$$
Variational linear regression: example

Data $y^T$

Design matrix $X^T$

- regressor 1 (sinusoid)
- regressor 2 (linear slope)
- regressor 3 (constant)
Variational linear regression: example

Iteration 0

F = -14771.09

coefficients precision $\alpha$

noise precision $\lambda$

$\beta_1$

$\beta_2$

$\beta_3$

correlation matrix of $\beta$
Variational linear regression: example

Iteration 1

F = 95.95

coefficients precision $\alpha$

noise precision $\lambda$

Correlation matrix of $\beta$
Variational linear regression: example

 Iteration 2 (convergence)

\[ F = 96.27 \]

\[ \alpha, \lambda \]

\[ \beta_1, \beta_2, \beta_3 \]

Correlation matrix of \( \beta \)
Variational linear regression: example

Variational inference
(Bayesian model comparison)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Variational BF</th>
<th>Frequentist p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>$\ln BF = 51.5$</td>
<td>$\beta_1: p = 0.0000$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$\ln BF = 293.9$</td>
<td>$\beta_2: p = 0.0000$</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>$\ln BF = 3.6$</td>
<td>$\beta_3: p = 0.0003$</td>
</tr>
<tr>
<td>$\phi$</td>
<td>$\ln BF = 320.7$</td>
<td>$\phi: p = 0.0000$</td>
</tr>
</tbody>
</table>

Frequentist inference
(classical t- and F-test)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Frequentist p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>$p = 0.0000$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$p = 0.0000$</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>$p = 0.0003$</td>
</tr>
<tr>
<td>$\phi$</td>
<td>$p = 0.0000$</td>
</tr>
</tbody>
</table>

Correlation matrix of $\beta$
Variational linear regression: free energy

\[ F(q, y) = \langle \ln p(y, \beta, \alpha, \lambda) \rangle_q(\beta, \alpha, \lambda) + H[q] \]

**Expected log-joint**

\[ = \langle \ln \prod \mathcal{N}(y_i | \beta^T x_i, \lambda^{-1}) \rangle_q + \langle \ln \mathcal{N}_d(\beta | 0, \alpha^{-1} I) \rangle_q + \langle \ln \mathcal{G}_a(\alpha | a_0, b_0) \rangle_q + \langle \ln \mathcal{G}_c(\lambda | c_0, d_0) \rangle_q \]

\[ + H[\mathcal{N}_d(\beta | \mu_n, \Lambda_n^{-1})] + H[\mathcal{G}_a(\alpha | a_n, b_n)] + H[\mathcal{G}_c(\lambda | c_n, d_n)] \]

\[ = \frac{n}{2} (\psi(c_n) - \ln d_n) - \frac{n}{2} \ln 2\pi - \frac{c_n}{2d_n} y^T y + \frac{c_n}{d_n} \mu_n X^T y - \frac{c_n}{2d_n} \text{Tr}[X^T X (\mu_n \mu_n^T + \Lambda_n^{-1})] \]

\[ - \frac{d}{2} \ln 2\pi + \frac{n}{2} (\psi(a_n) - \ln b_n) - \frac{a_n}{2b_n} (\mu_n^T \mu_n + \text{Tr}[\Lambda_n^{-1}]) \]

\[ + a_0 \ln b_0 - \ln \Gamma(a_0) + (a_0 - 1)(\psi(a_n) - \ln b_n) - \frac{b_0 a_n}{b_n} \]

\[ + c_0 \ln d_0 - \ln \Gamma(c_0) + (c_0 - 1)(\psi(c_n) - \ln d_n) - \frac{d_0 c_n}{d_n} \]

\[ + \frac{d}{2} (1 + \ln 2\pi) + \frac{1}{2} \ln |\Lambda_n^{-1}| \]

\[ + a_n - \ln b_n + \ln \Gamma(a_n) + (1 - a_n) \psi(a_n) \]

\[ + c_n - \ln d_n + \ln \Gamma(c_n) + (1 - c_n) \psi(c_n) \]
Variational linear regression: predictive density

\[
p(y_{n+1}|x_{n+1}, X, y) = \int p(y_{n+1}|x_{n+1}, \beta, \lambda) \ p(\beta, \lambda|X, y) \ d\beta \ d\lambda \\
\approx \int p(y_{n+1}|x_{n+1}, \beta, \lambda) \ q(\beta) \ q(\lambda) \ d\beta \ d\lambda
\]

-6 -4 -2 0 2 4 6 8 10 12

-10 -5 0 5

\( \text{data (n = 200)} \)

\( \text{posterior predictive mean +/- 2 SD} \)
Variational linear regression: predictive density

\[ p(y_{n+1}|x_{n+1}, X, y) = \int p(y_{n+1}|x_{n+1}, \beta, \lambda) p(\beta, \lambda|X, y) \, d\beta \, d\lambda \]
\[ \approx \int p(y_{n+1}|x_{n+1}, \beta, \lambda) q(\beta) \, q(\lambda) \, d\beta \, d\lambda \]
% Variational Bayesian multiple linear regression.
%
% Usage:
%    q = vblm(y, X)
%    [q, stats, q_trace] = vblm(y, X, a_0, b_0, c_0, d_0)
%
% Args:
%    y:   <n x 1> vector of observations (response variable)
%    X:   <n x d> design matrix (regressors)
%    a_0: shape parameter of the prior precision of coefficients
%    b_0: rate parameter of the prior precision of coefficients
%    c_0: shape parameter of the prior noise precision
%    d_0: rate parameter of the prior noise precision
%
% Returns:
%    q:   moments of the variational posterior
%    q.F: free energy of the model given the data
%
% See also:
%    vblm_predict

% Kay H. Brodersen, TNU, University of Zurich & ETH Zurich
% $Id: vblm.m 19126 2013-03-18 18:33:05Z bkay $
Frequentist linear regression

\[ \hat{\beta} = (X^T X)^{-1} X^T y \]
\[ \text{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} \]
\[ p = P(t \geq t^* | H_0) \]

Variational Bayesian linear regression

\[ \langle \beta | X, y \rangle \approx \langle \beta \rangle_{q(\beta)} \]
\[ = (\langle \alpha \rangle_{q(\alpha)} + \langle \lambda \rangle_{q(\lambda)} X^T X)^{-1} \langle \lambda \rangle_{q(\lambda)} X^T y \]
\[ \text{Cov}(\beta | X, y) = (\langle \alpha \rangle_{q(\alpha)} I + \langle \lambda \rangle_{q(\lambda)} X^T X)^{-1} \]
\[ \ln BF = \ln F_1 - \ln F_2 \]
Overview

1. The Laplace approximation
2. Variational Bayes
3. Variational density estimation
4. Variational linear regression
5. Variational clustering
Extending the univariate model to a mixture model yields a variational clustering algorithm.

The only assumption required to obtain a tractable solution is:

\[ q(Z, \pi, \mu, \Lambda) = q(Z) q(\pi, \mu, \Lambda) \]

Iterating between these two densities gives the variational equivalent of an EM algorithm.

\[
p(\pi) = \text{Dir}(\pi | \alpha_0) \\
p(\Lambda) = \prod_{k=1}^{K} \mathcal{W}(\Lambda_k | W_0, \nu_0) \\
p(\mu | \Lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k | m_0, (\beta_0 \Lambda_k)^{-1}) \\
p(z_i | \pi) = \prod_{k=1}^{K} \pi_k^{z_{ik}} \\
p(x_i | Z, \mu, \Lambda) = \prod_{k=1}^{K} \mathcal{N}(x_i | \mu_k, \Lambda_k^{-1})^{z_{ik}}
\]
Variational clustering

**Variational E-step**

\[
\ln q^*(\pi, \mu, \Lambda) = \langle \ln p(X, Z, \pi, \mu, \Lambda) \rangle_{q(Z)} \implies q(\pi, \mu, \Lambda) = q(\pi) \prod_{k=1}^{K} q(\mu_k, \Lambda_k)
\]

\[
q^*(\pi) = \text{Dir}(\pi|\alpha)
\]

where \( \alpha = (\alpha_k)_{k=1,...,K}, \quad \alpha_k = \alpha_0 + n_k \)

\[
n_k := \sum_{i=1}^{n} r_{i,k}
\]

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

where \( m_k = \frac{1}{\beta_k} (\beta_0 m_0 + n_k \bar{x}_k) \)

\[
\beta_k = \beta_0 + n_k
\]

\[
W_k^{-1} = W_0^{-1} + n_k S_k + \frac{\beta_0 n_k}{\beta_0 + n_k} (\bar{x}_k - m_0)(\bar{x}_k - m_0)^T
\]

\[
\nu_k = \nu_0 + n_k + 1
\]

\[
\bar{x}_k := \frac{1}{n_k} \sum_{i=1}^{n} r_{i,k} x_i
\]

\[
S_k := \frac{1}{n_k} \sum_{i=1}^{n} r_{i,k} (x_i - \bar{x}_k)(x_i - \bar{x}_k)^T
\]
Variational clustering

**Variational M-step**

\[
\ln q^*(Z) = \langle \ln p(X, Z, \pi, \mu, \Lambda) \rangle_{q(\pi, \mu, \Lambda)} \Rightarrow q^*(Z) = \prod_{i=1}^{n} \prod_{k=1}^{K} r_{i,k}^{z_{i,k}}
\]

where \( r_{i,k} := \frac{\rho_{i,k}}{\sum_{j=1}^{K} \rho_{i,j}} \)

where \( \rho_{i,j} := \exp \left( -\frac{1}{2} \left( d \beta_k^{-1} + v_k (x_i - m_k)^T W_k (x_i - m_k) \right) \right) \)
Advantages of variational clustering over the maximum-likelihood approach:

• no singularity issues (components that collapse onto a single data point)
• no overfitting (even with many components)
• number of clusters determined by model selection

Bishop (2005) PRML, p. 480
Two approaches to approximate inference
• stochastic inference (sampling)
• deterministic inference (variational Bayes)

The Laplace approximation
• simple local approximation
• often used in conjunction with VB

Variational inference under the mean-field assumption
• to maximize $F$ means to minimize $KL[q||p]$
• variational algorithm
Summary (2)

**Variational univariate density estimation**
- exact solution available

**Variational multiple linear regression**
- `vblm.m`

**Variational clustering using a Gaussian mixture model**
- `spm_mix.m`