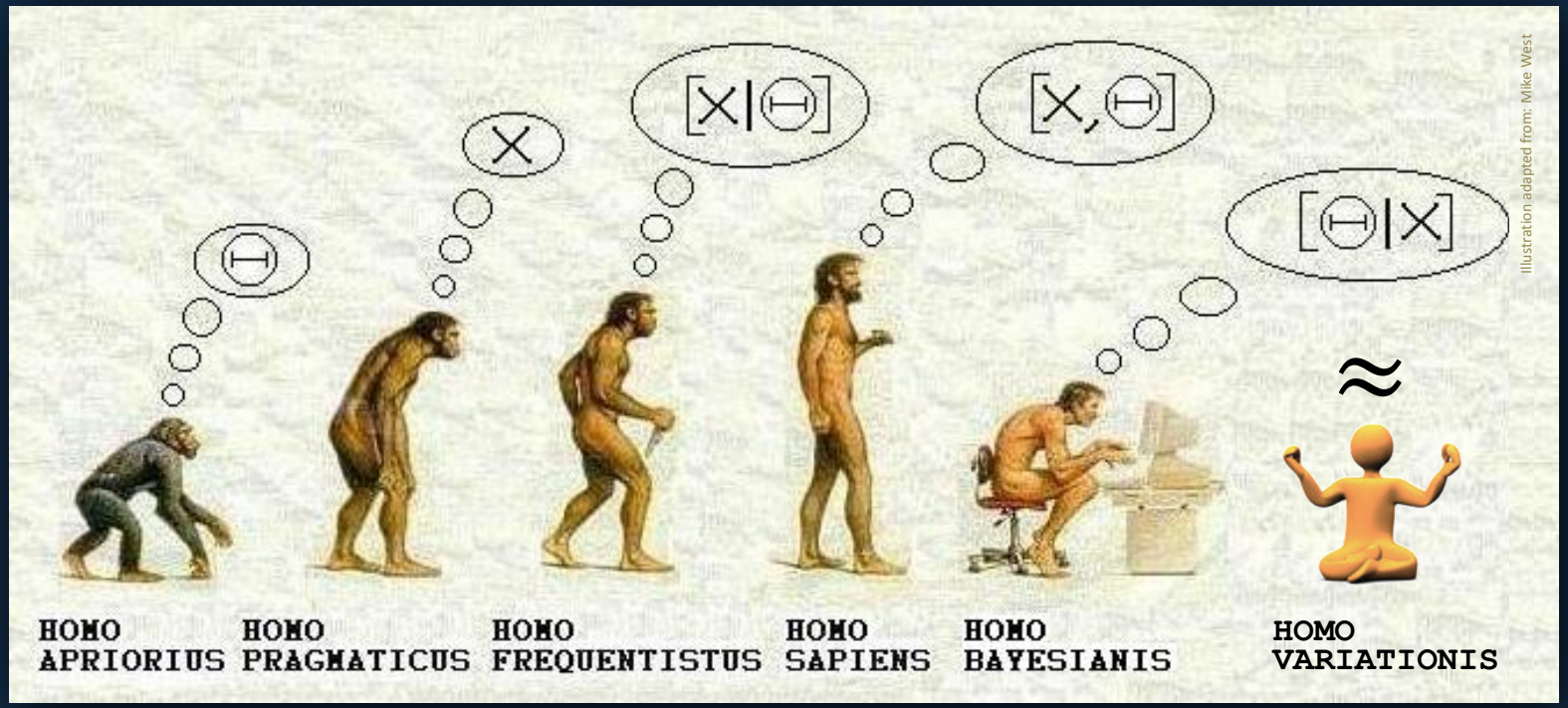
Variational Bayesian inference

Kay H. Brodersen

Translational Neuromodeling Unit (TNU)
Institute of Biomedical Engineering
University of Zurich & ETH Zurich



Variational Bayesian inference

"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

Approximate Bayesian inference

Bayesian inference formalizes *model inversion*, the process of passing from a prior to a posterior in light of data.

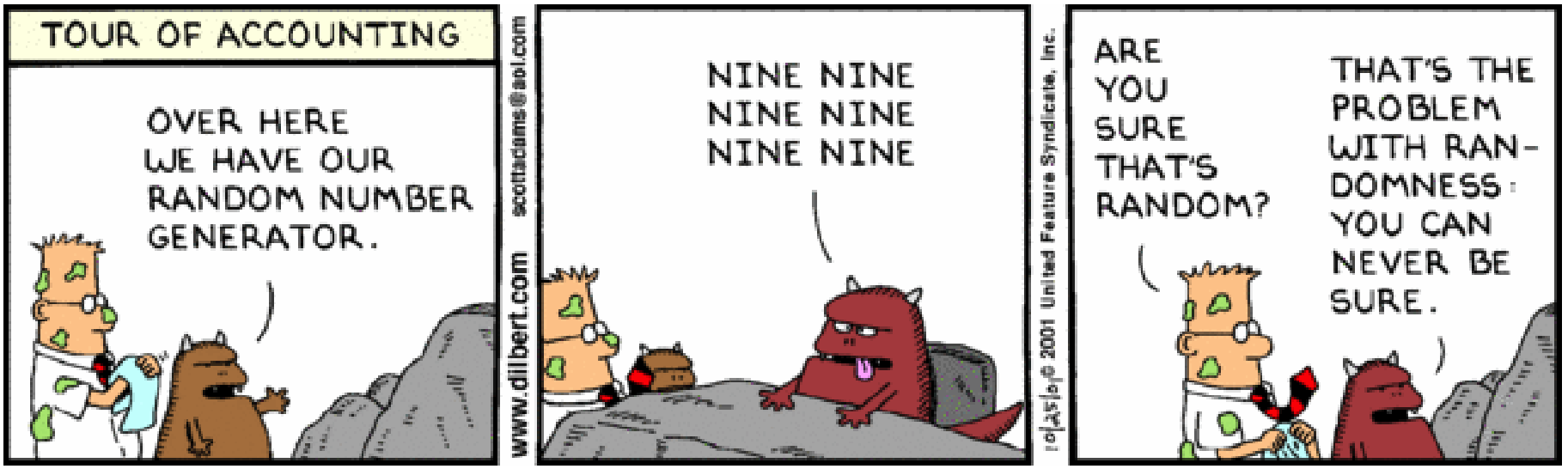
$$\underbrace{p(\theta|y)}_{\text{posterior}} = \frac{\overbrace{p(y|\theta)}^{\text{likelihood}}\ \overbrace{p(\theta)}^{\text{prior}}}{\underbrace{\int p(y,\theta)\mathrm{d}\theta}_{\text{marginal likelihood } p(y) \text{ (model evidence)}}}$$

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

Approximate Bayesian inference

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

❶ design an algorithm that draws samples $\theta^{(1)}, \ldots, \theta^{(m)}$ from $p(\theta|y)$

❷ inspect sample statistics (e.g., histogram, sample quantiles, …)

☑ asymptotically exact
☒ computationally expensive
☒ tricky engineering concerns

Structural approximate inference

in particular variational Bayes

❶ find an analytical proxy $q(\theta)$ that is maximally similar to $p(\theta|y)$

❷ 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

Overview

1 The Laplace approximation

2 Variational Bayes

3 Variational density estimation

4 Variational linear regression

5 Variational clustering

The Laplace approximation

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)

Pierre-Simon Laplace
(1749 – 1827)

French mathematician and astronomer

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)

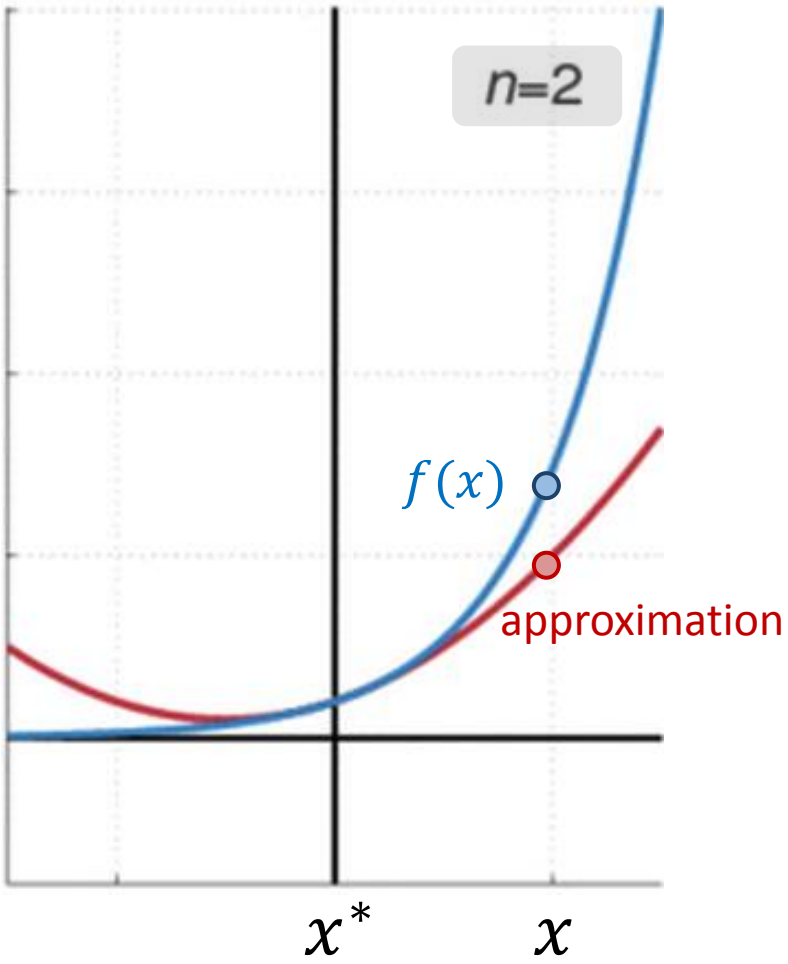
The Taylor approximation

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

$$
\begin{aligned}
f(x) \approx\ & f(x^*) \\
& + f'(x^*)(x - x^*) \\
& + \frac{1}{2!} f''(x^*)(x - x^*)^2 \\
& + \frac{1}{3!} f'''(x^*)(x - x^*)^3 \\
& + \cdots
\end{aligned}
$$

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 θ^*:

$$\begin{aligned}\mathcal{L}(\theta) &\approx \mathcal{L}(\theta^*) + \underbrace{\mathcal{L}'(\theta^*)}_{0}(\theta - \theta^*) + \frac{1}{2}\mathcal{L}''(\theta^*)(\theta - \theta^*)^2 \\ &= \mathcal{L}(\theta^*) + \frac{1}{2}\,\mathcal{L}''(\theta^*)\,(\theta - \theta^*)^2\end{aligned}$$

This already has the same form as a Gaussian density:

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

And so we have an approximate posterior:

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

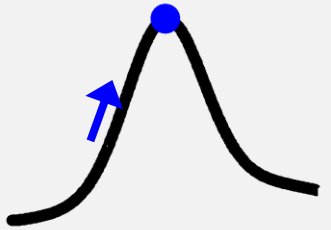
$\eta = -\mathcal{L}''(\theta^*)$ (negative curvature at the mode)

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:

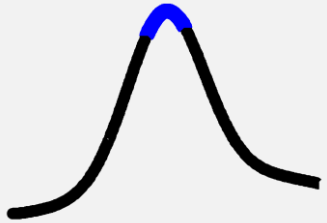
1. Find the mode of the log-joint:

$$\theta^* = \arg\max_{\theta} \ln p(y, \theta)$$



2. Evaluate the curvature of the log-joint at the mode:

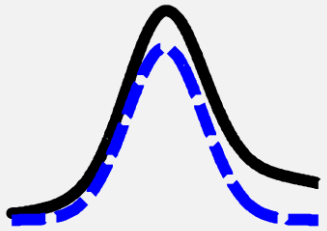
$$\nabla\nabla \ln p(y, \theta^*)$$



3. We obtain a Gaussian approximation:

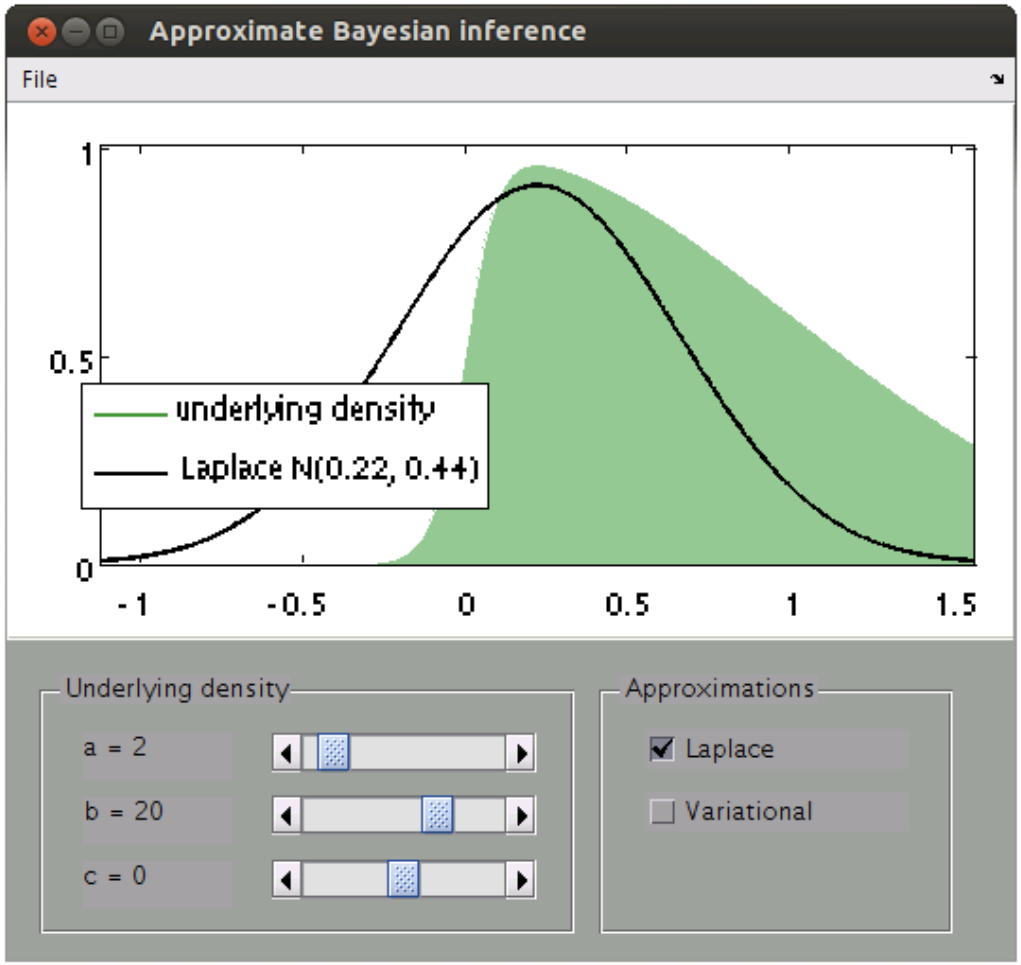
$$\mathcal{N}(\theta|\mu, \Lambda^{-1}) \qquad \text{with } \mu = \theta^*$$

$$\Lambda = -\nabla\nabla \ln p(y, \theta^*)$$



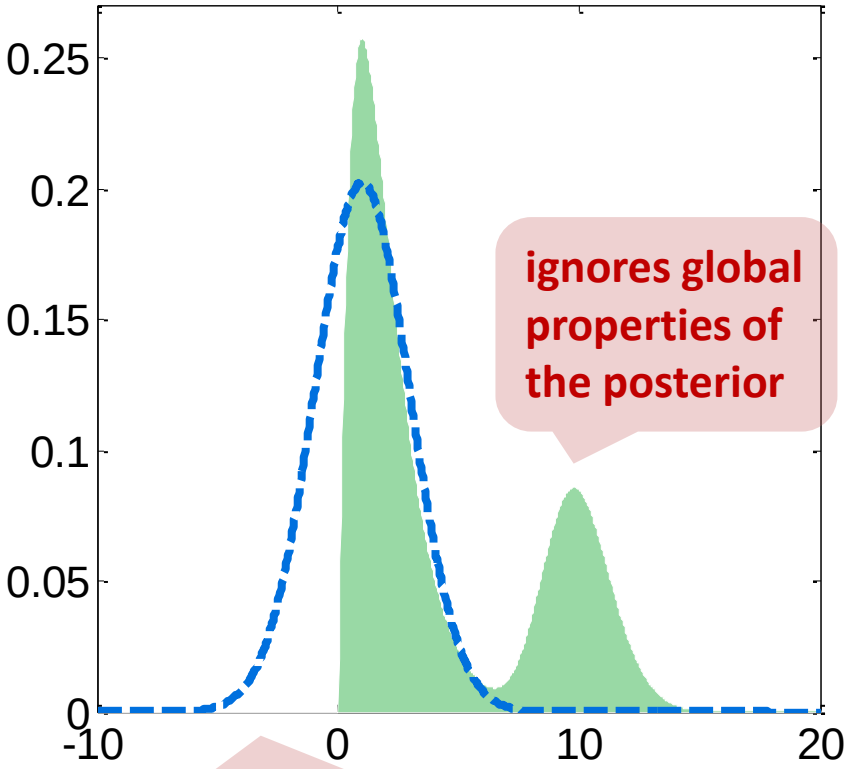
The Laplace approximation: demo

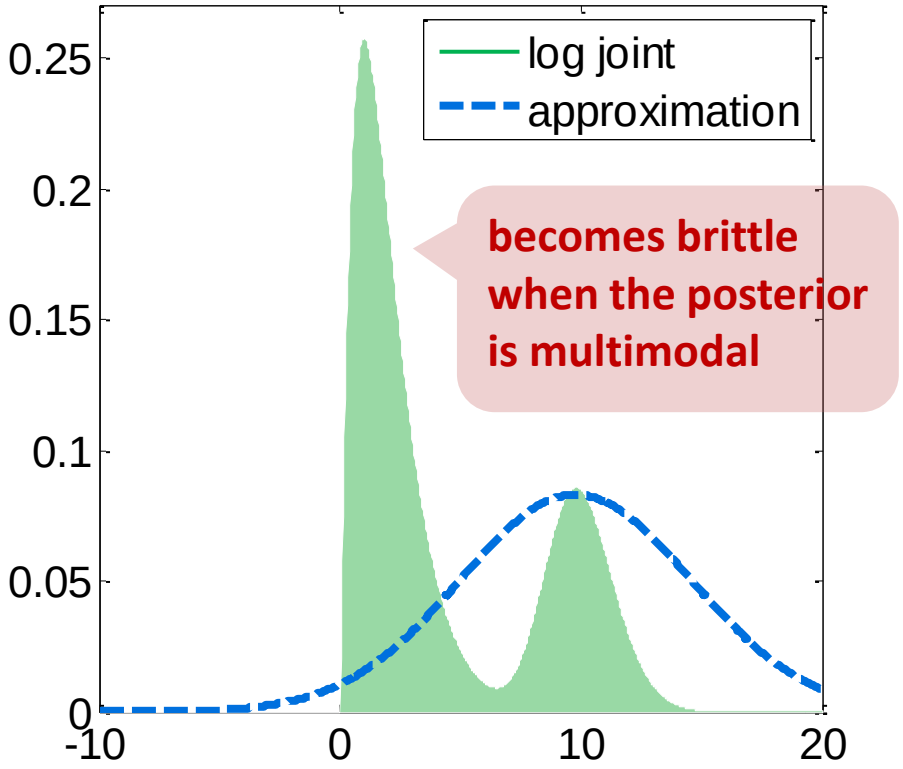
~kbroders/teaching/vb_gui.m



Limitations of the Laplace approximation

The Laplace approximation is often too strong a simplification.





Overview

1 The Laplace approximation

2 Variational Bayes

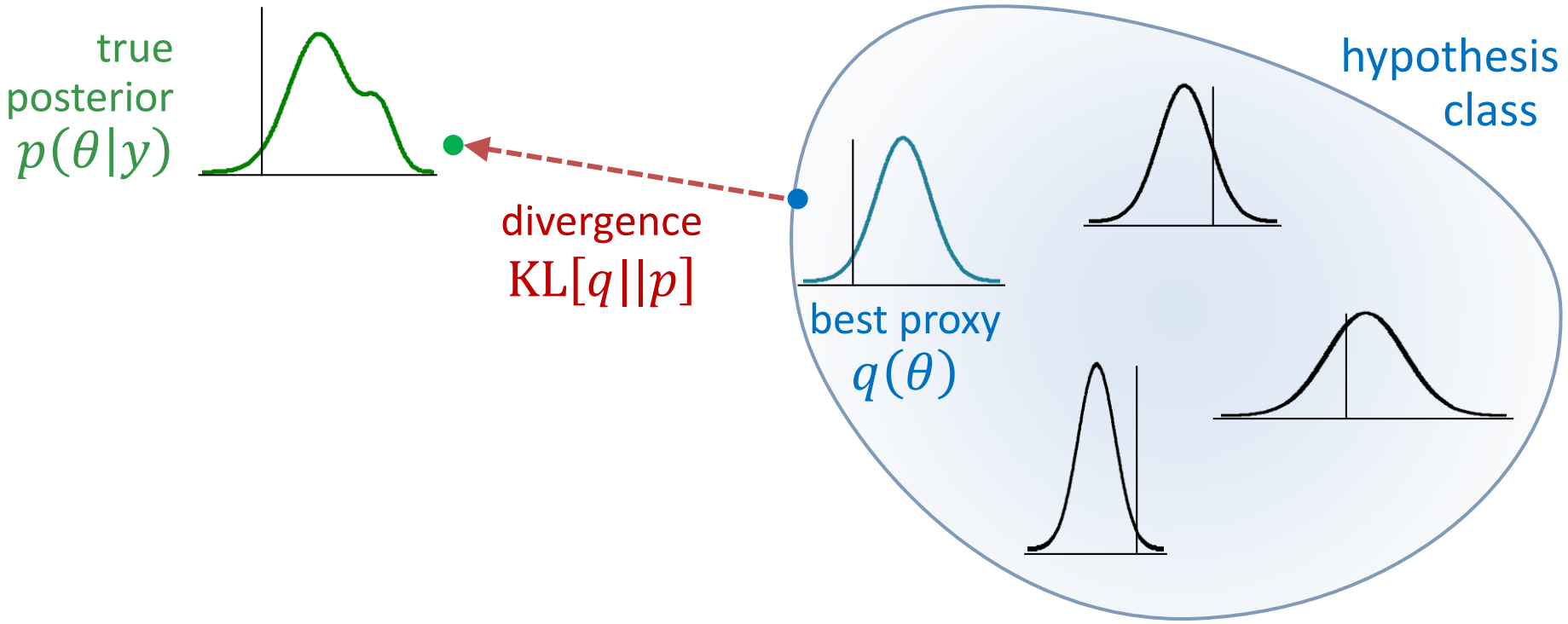
3 Variational density estimation

4 Variational linear regression

5 Variational clustering

Variational Bayesian inference

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 calculus

Variational Bayesian inference is based on variational calculus.

Standard calculus

Newton, Leibniz, and others

- functions $f: x \mapsto f(x)$
- derivatives $\frac{\mathrm{d}f}{\mathrm{d}x}$

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

Variational calculus

Euler, Lagrange, and others

- functionals $F: f \mapsto F(f)$
- derivatives $\frac{\mathrm{d}F}{\mathrm{d}f}$

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.

$$
\begin{aligned}
\ln p(y) &= \ln \frac{p(y,\theta)}{p(\theta|y)} \\
&= \int q(\theta) \ln \frac{p(y,\theta)}{p(\theta|y)} \, \mathrm{d}\theta \\
&= \int q(\theta) \ln \frac{p(y,\theta)}{p(\theta|y)} \frac{q(\theta)}{q(\theta)} \, \mathrm{d}\theta \\
&= \int q(\theta) \left(\ln \frac{q(\theta)}{p(\theta|y)} + \ln \frac{p(y,\theta)}{q(\theta)} \right) \mathrm{d}\theta \\
&= \underbrace{\int q(\theta) \ln \frac{q(\theta)}{p(\theta|y)} \, \mathrm{d}\theta}_{\mathrm{KL}[q||p]} + \underbrace{\int q(\theta) \ln \frac{p(y,\theta)}{q(\theta)} \, \mathrm{d}\theta}_{F(q,y)}
\end{aligned}
$$

$\mathrm{KL}[q||p]$
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:

$$\ln p(y) = \underbrace{\mathrm{KL}[q||p]}_{\text{divergence}} + \underbrace{F(q,y)}_{\text{free energy}}$$

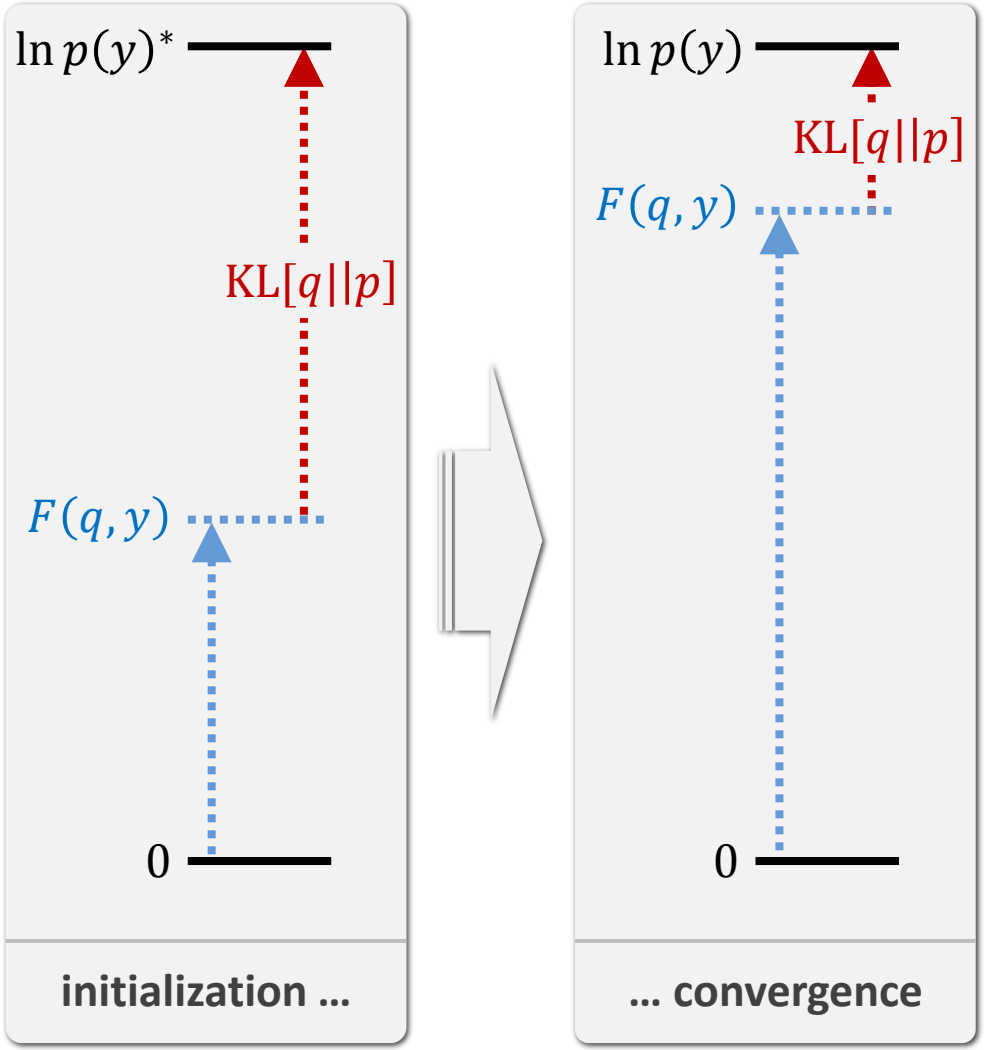
divergence
≥ 0
(unknown)

free energy
(easy to evaluate for a given q)

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

- minimizing $\mathrm{KL}[q||p]$
- 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:

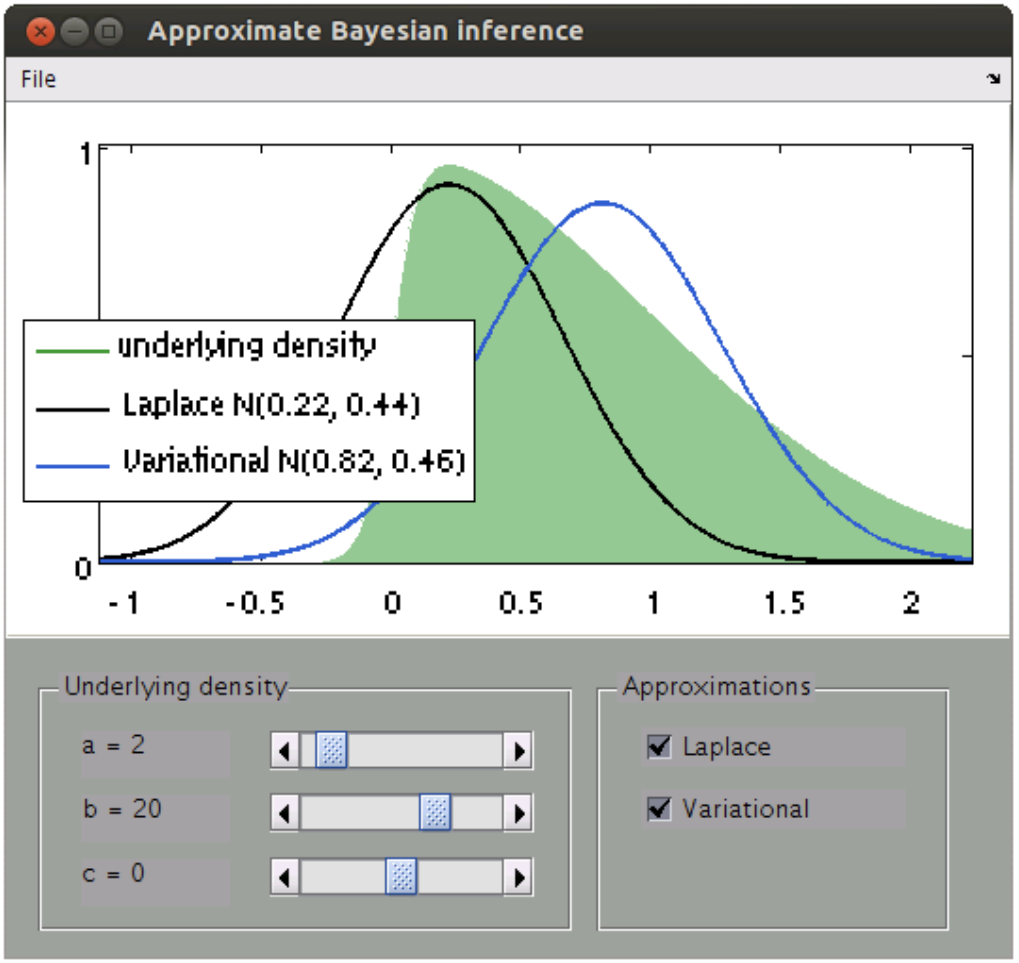
$$F(q,y) = \int q(\theta) \ln \frac{p(y,\theta)}{q(\theta)} \mathrm{d}\theta$$

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

$$= \underbrace{\langle \ln p(y,\theta) \rangle_q}_{\text{expected log-joint}} + \underbrace{H[q]}_{\text{Shannon entropy}}$$

The Laplace approximation: demo

~kbroders/teaching/vb_gui.m

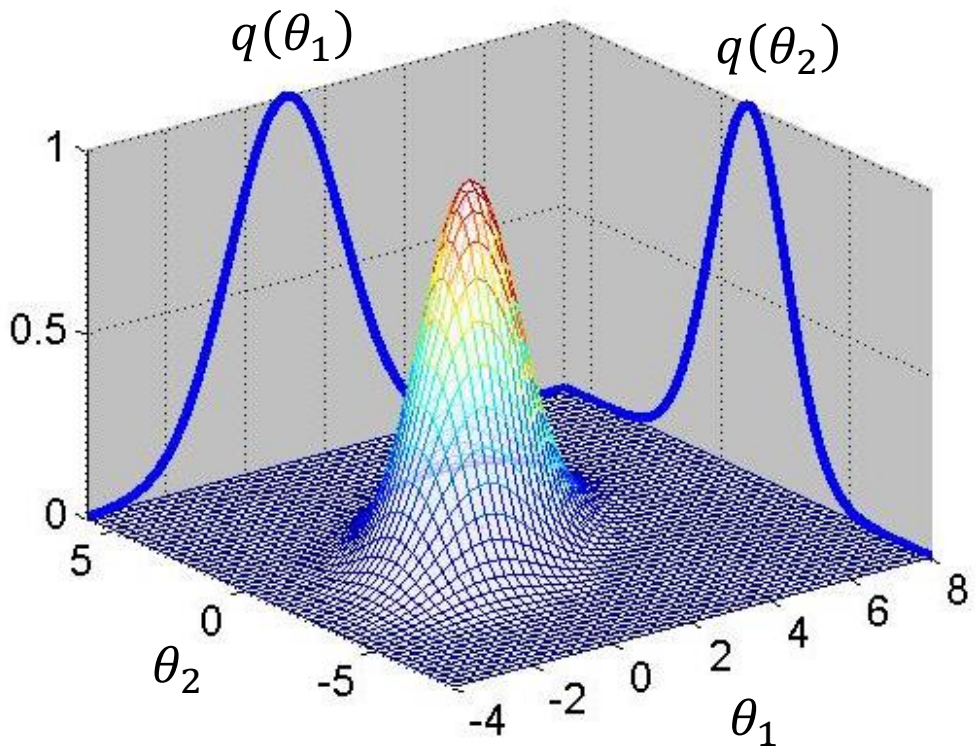


The mean-field assumption

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.



Jean Daunizeau, www.fil.ion.ucl.ac.uk/~jdaunize/presentations/Bayes2.pdf

Typical strategies in variational inference

	no parametric assumptions	**parametric assumptions** $q(\theta) = F(\theta\|\delta)$
no mean-field assumption	(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

$$
\begin{aligned}
F(q,y) &= \int q(\theta) \ln \frac{p(y,\theta)}{q(\theta)} \mathrm{d}\theta \\
&= \int \prod_i q_i \times \left(\ln p(y,\theta) - \sum_i \ln q_i \right) \mathrm{d}\theta \\
&= \int q_j \prod_{\setminus j} q_i \left(\ln p(y,\theta) - \ln q_j \right) \mathrm{d}\theta - \int q_j \prod_{\setminus j} q_i \sum_{\setminus j} \ln q_i \, \mathrm{d}\theta \\
&= \int q_j \left(\underbrace{\int \prod_{\setminus j} q_i \ln p(y,\theta) \, \mathrm{d}\theta_{\setminus j}}_{\langle \ln p(y,\theta) \rangle_{q_{\setminus j}}} - \ln q_j \right) \mathrm{d}\theta_j - \int q_j \int \prod_{\setminus j} q_i \ln \prod_{\setminus j} q_i \, \mathrm{d}\theta_{\setminus j} \, \mathrm{d}\theta_j \\
&= \int q_j \ln \frac{\exp\left(\langle \ln p(y,\theta) \rangle_{q_{\setminus j}} \right)}{q_j} \mathrm{d}\theta_j \quad + \quad c \\
&= -\mathrm{KL}\left[q_j \,||\, \exp\left(\langle \ln p(y,\theta) \rangle_{q_{\setminus j}} \right) \right] \quad + \quad c
\end{aligned}
$$

mean-field assumption:
$q(\theta) = \prod_i q_i(\theta_i)$

Variational algorithm under the mean-field assumption

In summary:

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

Suppose the densities $q_{\setminus j} \equiv q(\theta_{\setminus j})$ are kept fixed. Then the approximate posterior $q(\theta_j)$ that maximizes $F(q,y)$ is given by:

$$\begin{aligned} q_j^* &= \arg\max_{q_j} F(q,y) \\ &= \frac{1}{Z}\exp\left(\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}\right) \end{aligned}$$

Therefore:

$$\ln q_j^* = \underbrace{\langle \ln p(y,\theta)\rangle_{q_{\setminus j}}}_{=:I(\theta_j)} - \ln Z$$

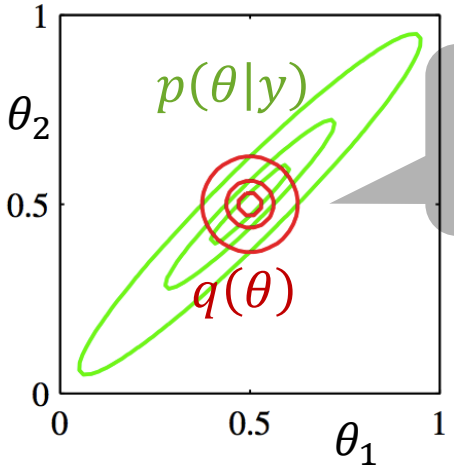
This implies a straightforward algorithm for variational inference:

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.

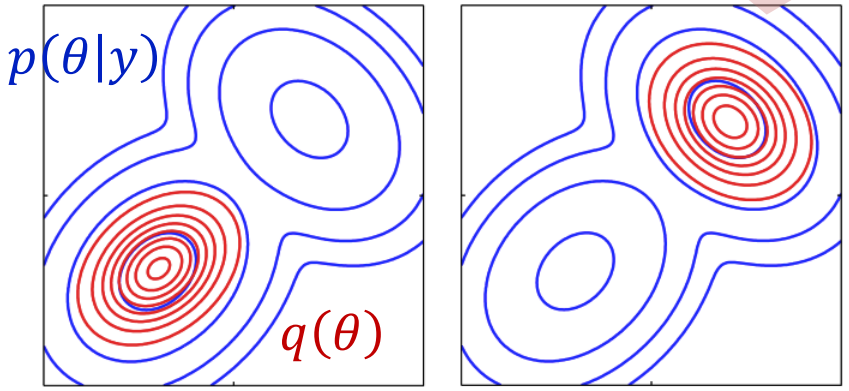
Frameworks for approximate inference

Variational Bayes

minimize $\mathrm{KL}[q(\theta)||p(\theta|y)]$



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

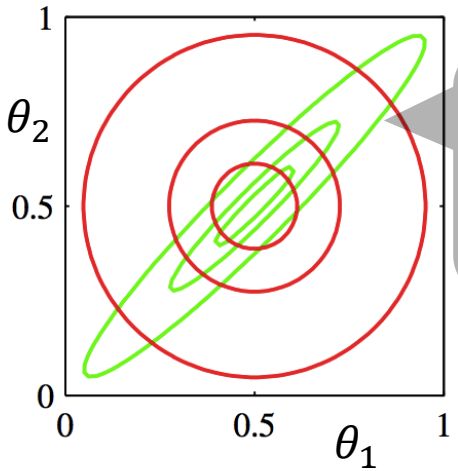


may lead to a local minimum

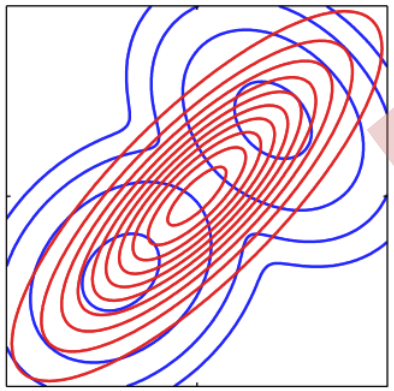
Bishop (2005) PRML, pp. 468 – 469

Expectation propagation

minimize $\mathrm{KL}[p(\theta|y)||q(\theta)]$



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



averaging across modes may lead to poor predictive performance

Overview

1 The Laplace approximation

2 Variational Bayes

3 Variational density estimation

4 Variational linear regression

5 Variational clustering

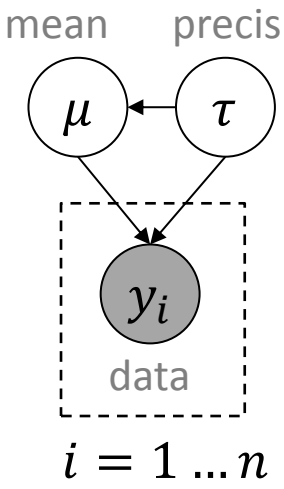
Application 1: variational density estimation

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:

$$q(\mu, \tau) = q_\mu(\mu)\, q_\tau(\tau)$$



$$p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1})$$
$$p(\tau) = \mathrm{Ga}(\tau|a_0, b_0)$$
$$p(y_i|\mu, \tau) = \mathcal{N}(y_i|\mu, \tau^{-1})$$

* Exercise 2.44; Bishop (2005) PRML

Recurring expressions in Bayesian inference

Univariate normal distribution

$$\begin{aligned}\ln \mathcal{N}(x|\mu,\lambda^{-1}) &= \frac{1}{2}\ln\lambda - \frac{1}{2}\ln\pi - \frac{\lambda}{2}(x-\mu)^2 \\ &= -\frac{1}{2}\lambda x^2 + \lambda\mu x + c\end{aligned}$$

Multivariate normal distribution

$$\begin{aligned}\ln \mathcal{N}_d(x|\mu,\Lambda^{-1}) &= -\frac{1}{2}\ln|\Lambda^{-1}| - \frac{d}{2}\ln 2\pi - \frac{1}{2}(x-\mu)^T\Lambda(x-\mu) \\ &= -\frac{1}{2}x^T\Lambda x + x^T\Lambda\mu + c\end{aligned}$$

Gamma distribution

$$\begin{aligned}\ln \mathrm{Ga}(x|a,b) &= a\ln b - \ln\Gamma(a) + (a-1)\ln x - b\,x \\ &= (a-1)\ln x - b\,x + c\end{aligned}$$

Variational density estimation: mean μ

$$
\begin{aligned}
\ln q^*(\mu) &= \langle \ln p(y, \mu, \tau) \rangle_{q(\tau)} + c \\
&= \left\langle \ln \prod_i^n p(y_i|\mu, \tau) \right\rangle_{q(\tau)} + \langle \ln p(\mu|\tau) \rangle_{q(\tau)} + \langle \ln p(\tau) \rangle_{q(\tau)} + c \\
&= \langle \ln \prod \mathcal{N}(y_i|\mu, \tau^{-1}) \rangle_{q(\tau)} + \langle \ln \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}) \rangle_{q(\tau)} + \langle \ln \mathrm{Ga}(\tau|a_0, b_0) \rangle_{q(\tau)} + c \\
&= \sum \left\langle -\frac{\tau}{2}(y_i - \mu)^2 \right\rangle_{q(\tau)} + \left\langle -\frac{\lambda_0\tau}{2}(\mu - \mu_0)^2 \right\rangle_{q(\tau)} + c \\
&= \sum -\frac{\langle\tau\rangle_{q(\tau)}}{2} y_i^2 + \langle\tau\rangle_{q(\tau)} n\bar{y}\mu - n\frac{\langle\tau\rangle_{q(\tau)}}{2}\mu^2 - \frac{\lambda_0\langle\tau\rangle_{q(\tau)}}{2}\mu^2 + \lambda_0\mu\mu_0\langle\tau\rangle_{q(\tau)} - \frac{\lambda_0}{2}\mu_0^2 + c \\
&= -\frac{1}{2}\left\{n\langle\tau\rangle_{q(\tau)} + \lambda_0\langle\tau\rangle_{q(\tau)}\right\}\mu^2 + \left\{n\bar{y}\langle\tau\rangle_{q(\tau)} + \lambda_0\mu_0\langle\tau\rangle_{q(\tau)}\right\}\mu + c
\end{aligned}
$$

reinstation by inspection

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

$$
\mu_n = \frac{n\bar{y}\langle\tau\rangle_{q(\tau)} + \lambda_0\mu_0\langle\tau\rangle_{q(\tau)}}{\lambda_n} = \frac{\lambda_0\mu_0 + n\bar{y}}{\lambda_0 + n}
$$

Variational density estimation: precision τ

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

$$= \left\langle \ln \prod_{i=1}^{n} \mathcal{N}(y_i | \mu, \tau^{-1}) \right\rangle_{q(\mu)} + \langle \ln \mathcal{N}(\mu | \mu_0, (\lambda_0 \tau)^{-1}) \rangle_{q(\mu)} + \langle \ln \mathrm{Ga}(\tau | a_0, b_0) \rangle_{q(\mu)} + c$$

$$= \sum_{i=1}^{n} \left\langle \frac{1}{2} \ln \tau - \frac{\tau}{2} (y_i - \mu)^2 \right\rangle_{q(\mu)} + \left\langle \frac{1}{2} \ln(\lambda_0 \tau) - \frac{\lambda_0 \tau}{2} (\mu - \mu_0)^2 \right\rangle_{q(\mu)}$$
$$+ \langle (a_0 - 1) \ln \tau - b_0 \tau \rangle_{q(\mu)} + c$$

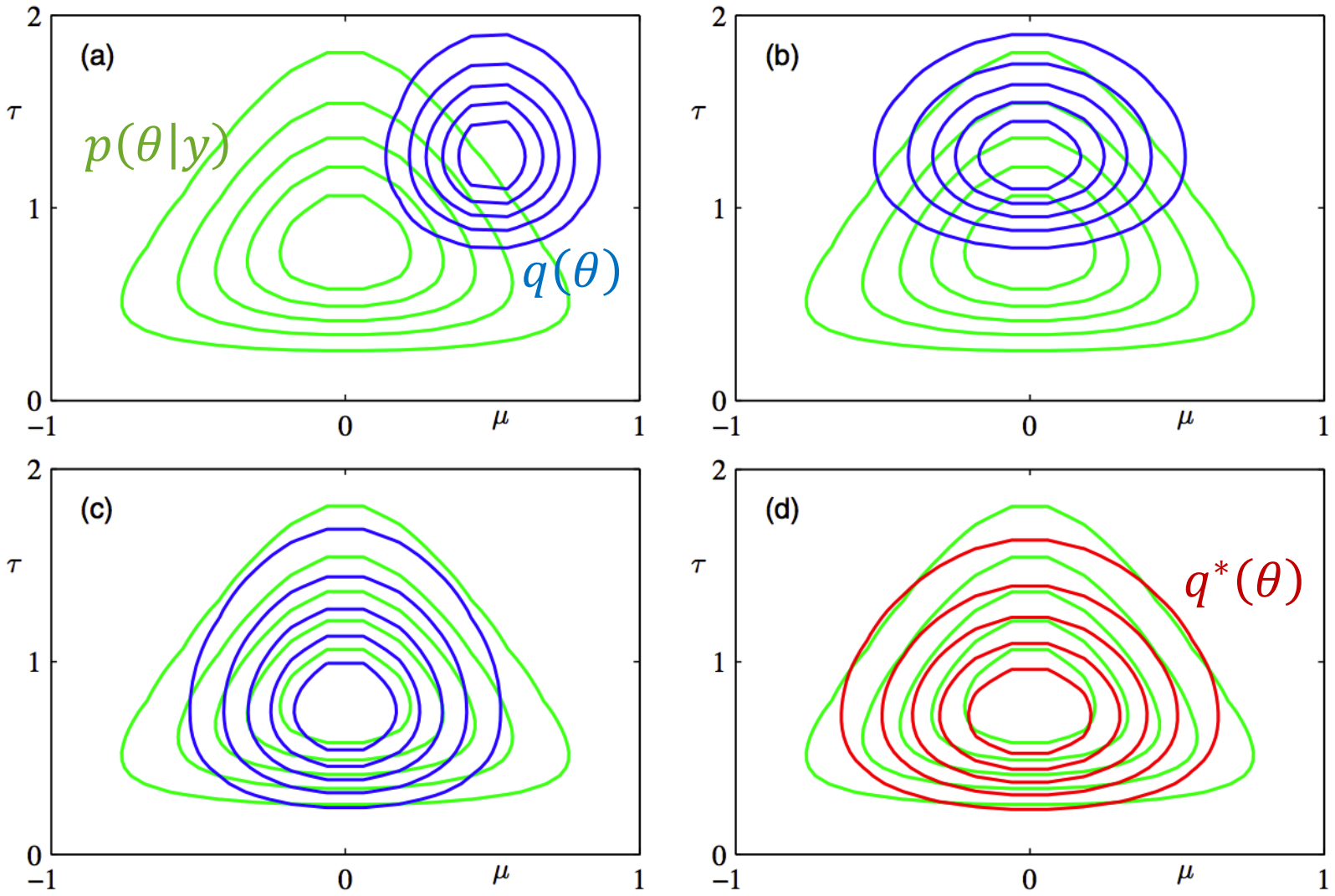
$$= \frac{n}{2} \ln \tau - \frac{\tau}{2} \langle \sum (y_i - \mu)^2 \rangle_{q(\mu)} + \frac{1}{2} \ln \lambda_0 + \frac{1}{2} \ln \tau - \frac{\lambda_0 \tau}{2} \langle (\mu - \mu_0)^2 \rangle_{q(\mu)} + (a_0 - 1) \ln \tau - b_0 \tau + c$$

$$= \left\{ \frac{n}{2} + \frac{1}{2} + (a_0 - 1) \right\} \ln \tau - \left\{ \frac{1}{2} \langle \sum (y_i - \mu)^2 \rangle_{q(\mu)} + \frac{\lambda_0}{2} \langle (\mu - \mu_0)^2 \rangle_{q(\mu)} + b_0 \right\} \tau + c$$

$$\Longrightarrow q^*(\tau) = \mathrm{Ga}(\tau | a_n, b_n) \quad \text{with} \quad a_n = a_0 + \frac{n+1}{2}$$

$$b_n = b_0 + \frac{\lambda_0}{2} \langle (\mu - \mu_0)^2 \rangle_{q(\mu)} + \frac{1}{2} \langle \sum (y_i - \mu)^2 \rangle_{q(\mu)}$$

Variational density estimation: illustration



Bishop (2005) PRML, p. 472

Overview

Application 2: variational linear regression

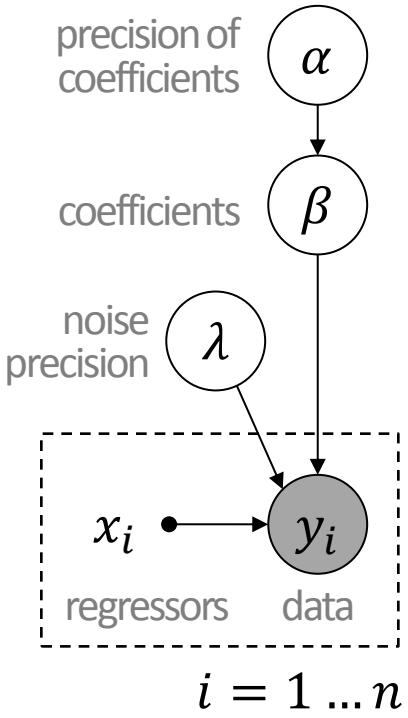
We consider a multiple linear regression model with a shrinkage prior on the regression coefficients.

We wish to infer on the coefficients β, their precision α, and the noise precision λ. There is no analytical posterior

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

We therefore seek a variational approximation:

$$q(\beta, \alpha, \lambda) = q_\beta(\beta)\, q_\alpha(\alpha)\, q_\lambda(\lambda).$$



$$p(\alpha) = \mathrm{Ga}(\alpha | a_0, b_0)$$

$$p(\beta | \alpha) = \mathcal{N}_d(\beta | 0, \alpha^{-1} I_{d \times d})$$

$$p(\lambda) = \mathrm{Ga}(\lambda | c_0, d_0)$$

$$p(y_i | \beta) = \mathcal{N}(y_i | \beta^T x_i, \lambda^{-1})$$

Variational linear regression: coefficients precision α

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

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

$$= \left\langle -\frac{1}{2} \ln \underbrace{|\alpha^{-1} I|}_{\alpha^{-d}} \underbrace{-\frac{d}{2} \ln 2\pi}_{c} - \frac{1}{2} (\beta - 0)^T \alpha I (\beta - 0) \right\rangle_{q(\beta)}$$

$$+ \langle a_0 \ln b_0 - \ln \Gamma(a_0) + (a_0 - 1) \ln \alpha - b_0 \alpha \rangle_{q(\beta)} + c$$

$$= \frac{d}{2} \ln \alpha - \frac{\alpha}{2} \langle \beta^T \beta \rangle_{q(\beta)} + (a_0 - 1) \ln \alpha - b_0 \alpha + c$$

$$= \left(\frac{d}{2} + a_0 - 1 \right) \ln \alpha - \left(\frac{1}{2} \langle \beta^T \beta \rangle_{q(\beta)} + b_0 \right) \alpha + c$$

$$\Longrightarrow q^*(\alpha) = \mathrm{Ga}(\alpha | a_n, b_n) \quad \text{with} \quad a_n = a_0 + \frac{d}{2}$$

$$b_n = b_0 + \frac{1}{2} \langle \beta^T \beta \rangle_{q(\beta)}$$

Variational linear regression: cofficients β

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

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

$$= \sum_i^n \left\langle \underbrace{\frac{1}{2}\ln\lambda}_{c} - \underbrace{\frac{1}{2}\ln 2\pi}_{c} - \frac{\lambda}{2}(y_i - \beta^T x_i)^2 \right\rangle_{q(\alpha)q(\lambda)} + \left\langle \underbrace{-\frac{1}{2}\ln|\alpha^{-1} I|}_{c} \underbrace{-\frac{d}{2}\ln 2\pi}_{c} - \frac{1}{2}\beta^T \alpha I \beta \right\rangle_{q(\alpha)} + c$$

$$= -\frac{\langle\lambda\rangle_{q(\lambda)}}{2} \sum_i^n (y_i - \beta^T x_i)^2 - \frac{1}{2}\langle\alpha\rangle_{q(\alpha)} \beta^T \beta + c$$

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

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

$$\Longrightarrow q^*(\beta) = \mathcal{N}_d(\beta | \mu_n, \Lambda_n^{-1}) \quad \text{with} \quad \Lambda_n = \langle\alpha\rangle_{q(\alpha)} I + \langle\lambda\rangle_{q(\lambda)} X^T X, \quad \mu_n = \Lambda_n^{-1} \langle\lambda\rangle_{q(\lambda)} X^T y$$

Variational linear regression: noise precision λ

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

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

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

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

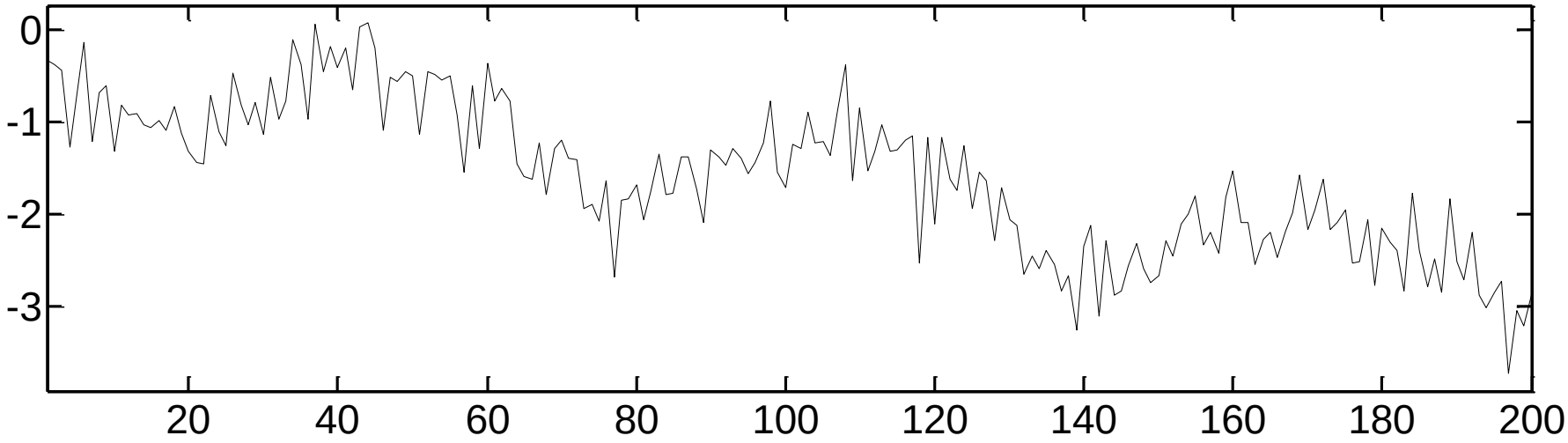
$$= \left\{ c_0 + \frac{n}{2} - 1 \right\} \ln \lambda - \left\{ \frac{1}{2} y^T y - \langle \beta \rangle_{q(\beta)}^T X^T y + \frac{1}{2} \langle \beta \rangle_{q(\beta)}^T X^T X \langle \beta \rangle_{q(\beta)} + d_0 \right\} \lambda + c$$

$$\Longrightarrow q^*(\lambda) = \mathrm{Ga}(\lambda | c_n, d_n), \qquad c_n = c_0 + \frac{n}{2}$$

$$d_n = d_0 + \frac{1}{2} y^T y - \langle \beta \rangle_{q(\beta)}^T X^T y + \frac{1}{2} \langle \beta \rangle_{q(\beta)}^T X^T X \langle \beta \rangle_{q(\beta)}$$

Variational linear regression: example

Data y^T

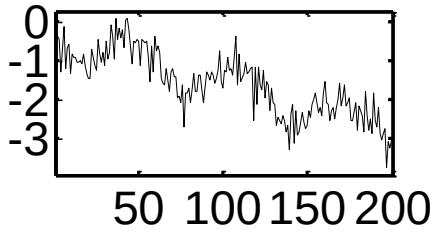


Design matrix X^T

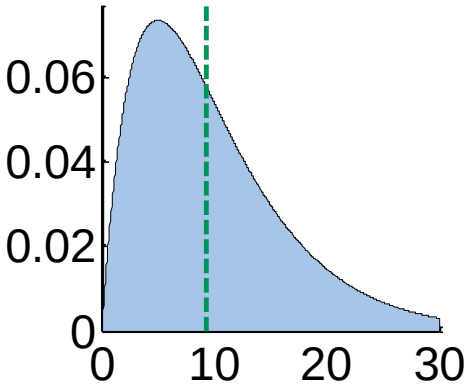
Variational linear regression: example

Iteration 0

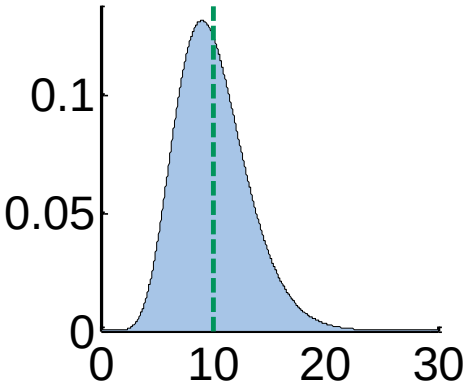
0:00:00'000



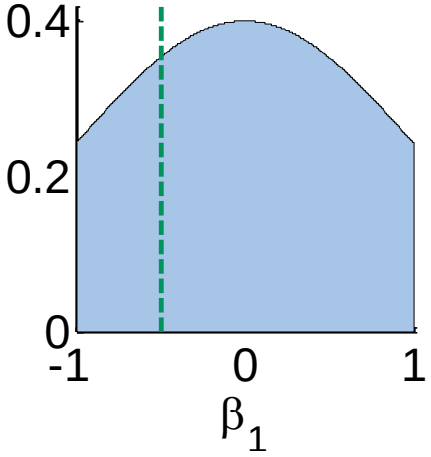
F = -14771.09



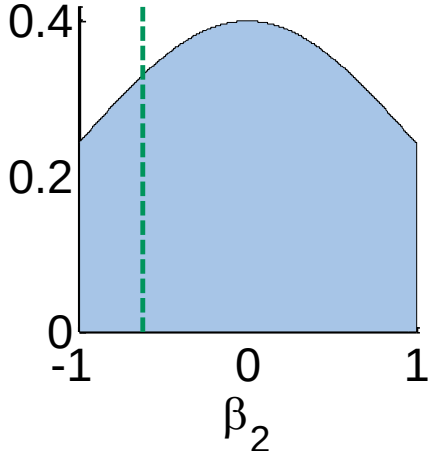
coefficients precision α



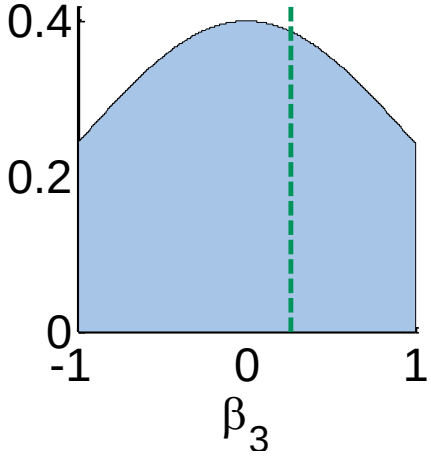
noise precision λ



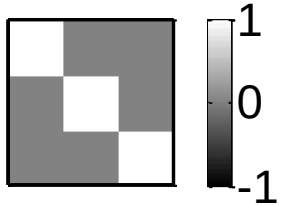
β_1



β_2



β_3

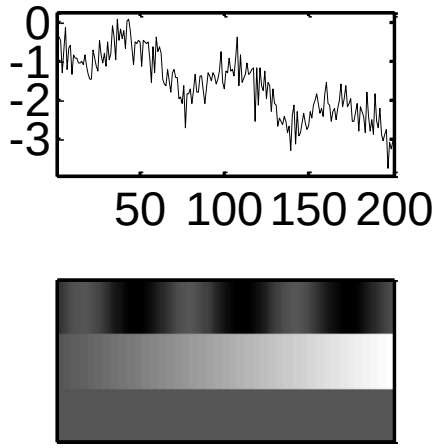


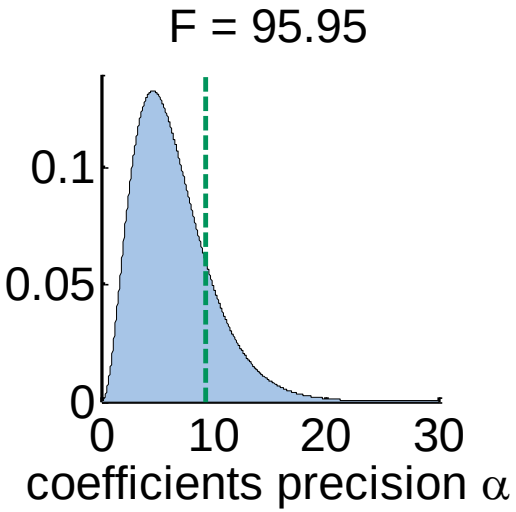
correlation matrix of β

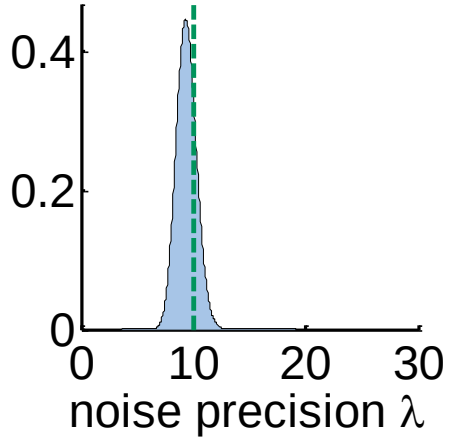
Variational linear regression: example

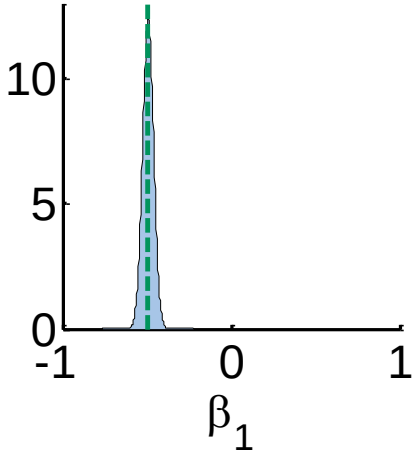
Iteration 1

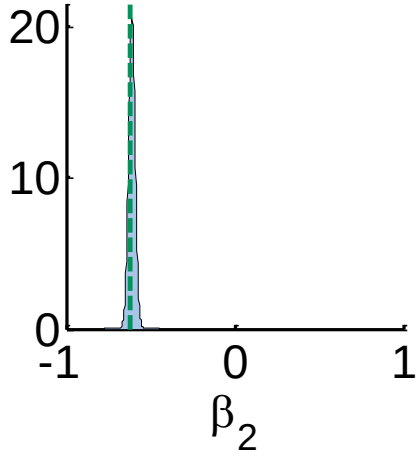
0:00:00’002

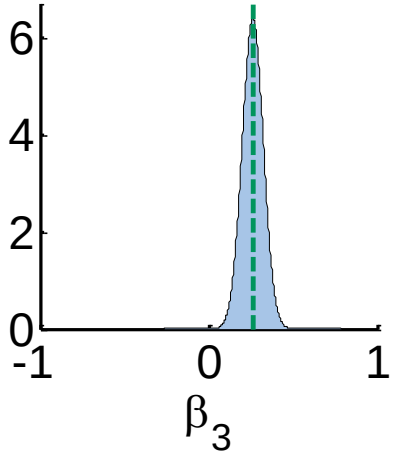


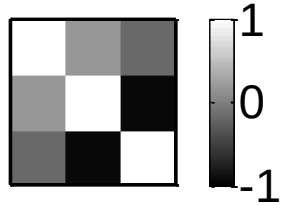










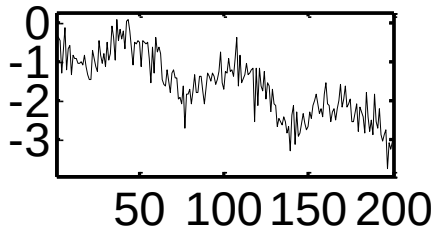


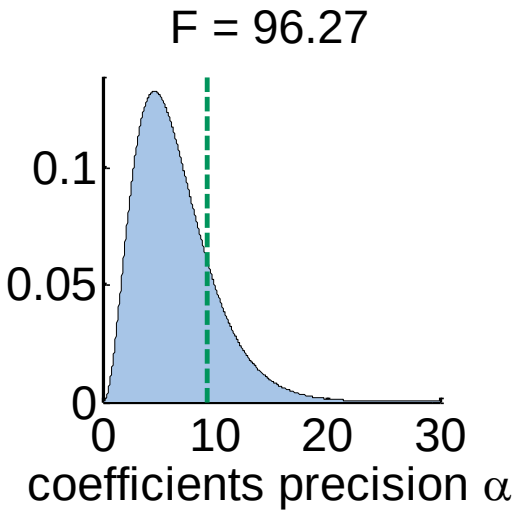
correlation matrix of β

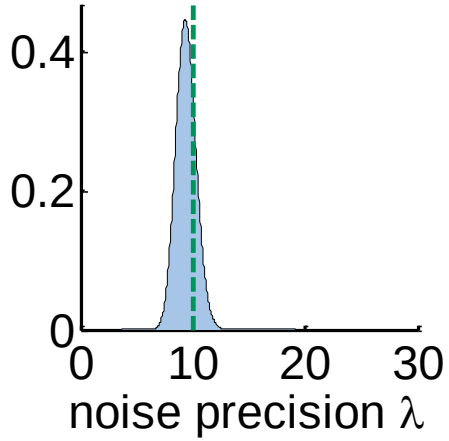
Variational linear regression: example

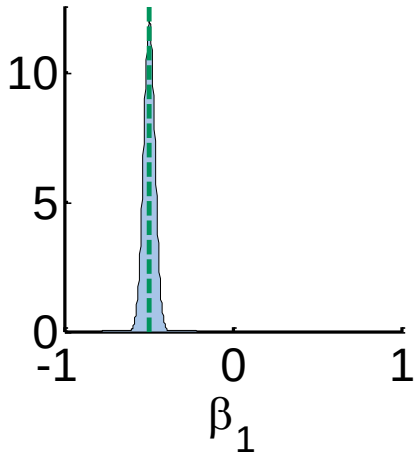
Iteration 2 (convergence)

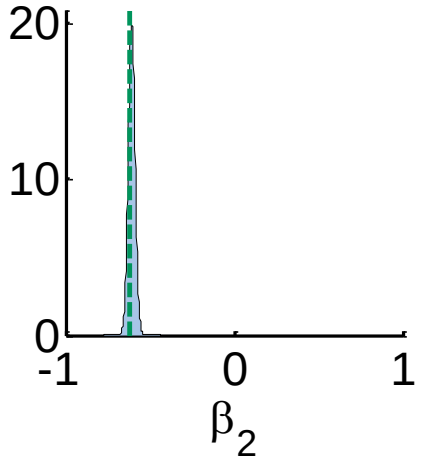
0:00:00’003

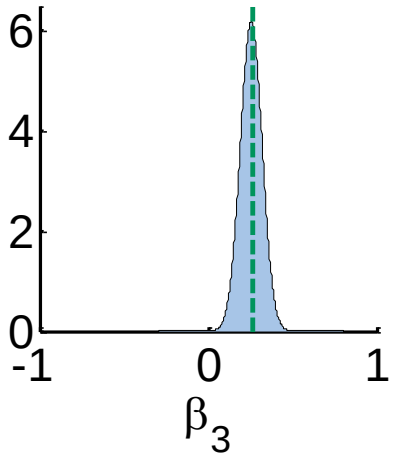


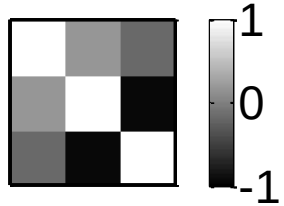








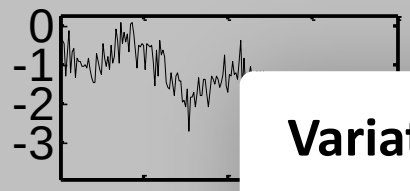




correlation matrix of β

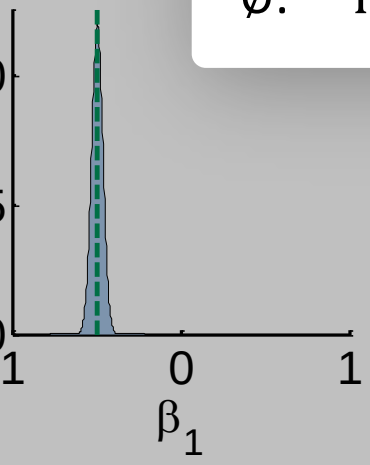
Variational linear regression: example

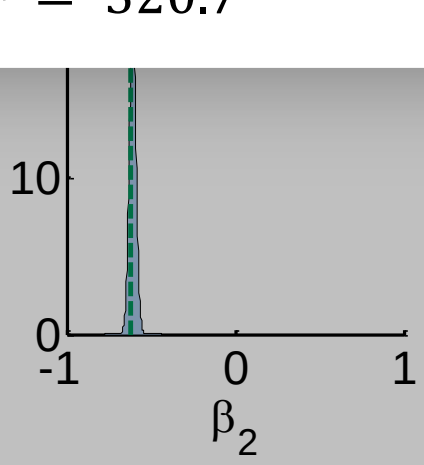
Iteration 2 (convergence) **0:00:00'003**

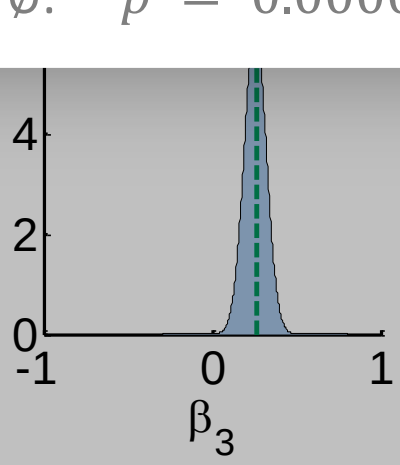


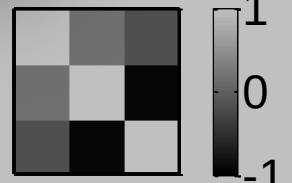
F = 96.27

Variational inference (Bayesian model comparsion)	**Frequentist inference** (classical *t*- and *F*-test)
β_1: $\ln BF = 51.5$	β_1: $p = 0.0000$
β_2: $\ln BF = 293.9$	β_2: $p = 0.0000$
β_3: $\ln BF = 3.6$	β_3: $p = 0.0003$
$\emptyset$: $\ln BF = 320.7$	$\emptyset$: $p = 0.0000$









correlation matrix of β

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 — Shannon entropy

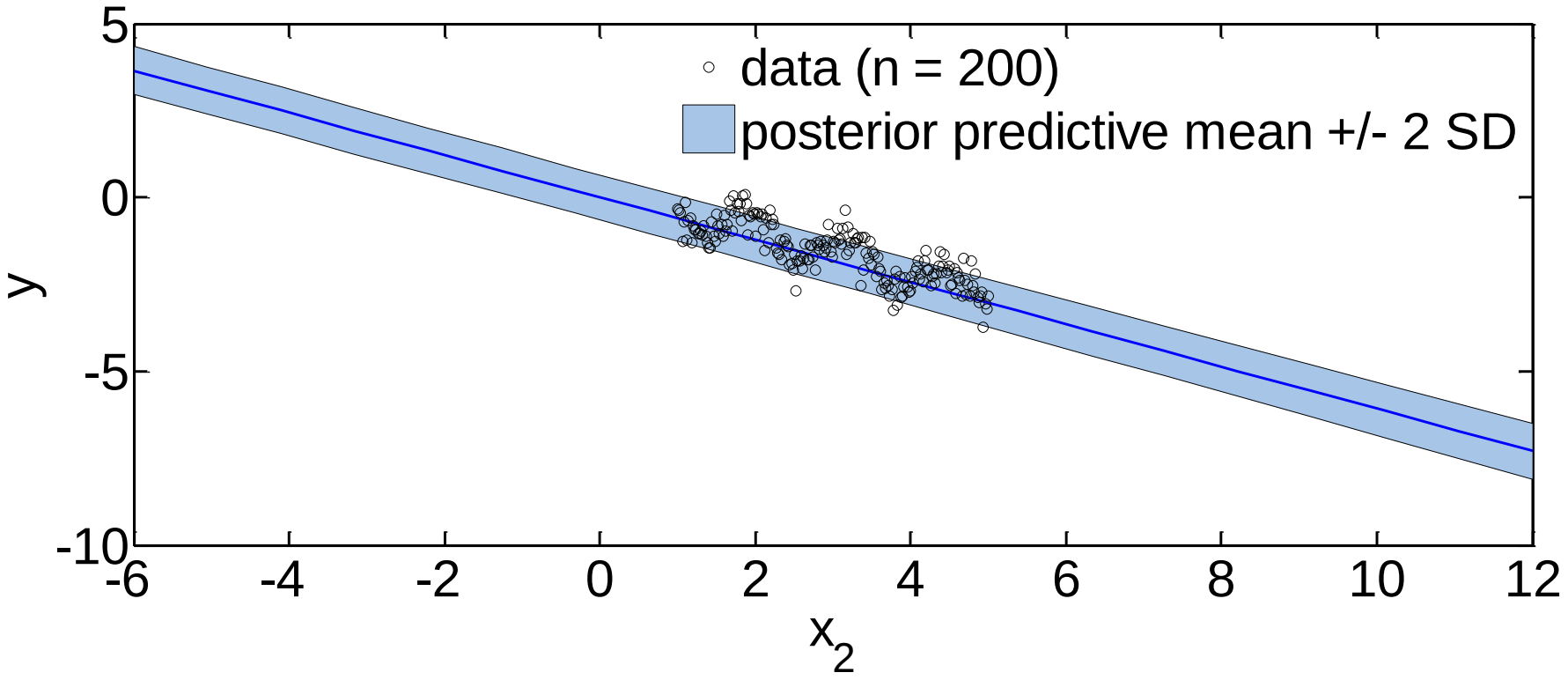
$$= \langle \ln \prod \mathcal{N}(y_i|\beta^T x_i, \lambda^{-1}) \rangle_q + \langle \ln \mathcal{N}_d(\beta|0,\alpha^{-1}I) \rangle_q + \langle \ln \mathrm{Ga}(\alpha|a_0,b_0) \rangle_q + \langle \ln \mathrm{Ga}(\lambda|c_0,d_0) \rangle_q$$
$$+H[\mathcal{N}_d(\beta|\mu_n,\Lambda_n^{-1})] + H[\mathrm{Ga}(\alpha|a_n,b_n)] + H[\mathrm{Ga}(\lambda|c_n,d_n)]$$

$$= \frac{n}{2}(\psi(c_n) - \ln d_n) - \frac{n}{2}\ln 2\pi - \frac{c_n}{2d_n} y^T y + \frac{c_n}{d_n} \mu_n^T X^T y - \frac{c_n}{2d_n} \mathrm{Tr}[X^T X(\mu_n \mu_n^T + \Lambda_n^{-1})]$$
$$-\frac{d}{2}\ln 2\pi + \frac{n}{2}(\psi(a_n) - \ln b_n) - \frac{a_n}{2b_n}(\mu_n^T \mu_n + \mathrm{Tr}[\Lambda_n^{-1}])$$
$$+a_0 \ln b_0 - \ln \Gamma(a_0) + (a_0 - 1)(\psi(a_n) - \ln b_n) - \frac{b_0 a_n}{b_n}$$
$$+c_0 \ln d_0 - \ln \Gamma(c_0) + (c_0 - 1)(\psi(c_n) - \ln d_n) - \frac{d_0 c_n}{d_n}$$

$$+\frac{d}{2}(1 + \ln 2\pi) + \frac{1}{2}\ln|\Lambda_n^{-1}|$$
$$+a_n - \ln b_n + \ln \Gamma(a_n) + (1 - a_n)\psi(a_n)$$
$$+c_n - \ln d_n + \ln \Gamma(c_n) + (1 - c_n)\psi(c_n)$$

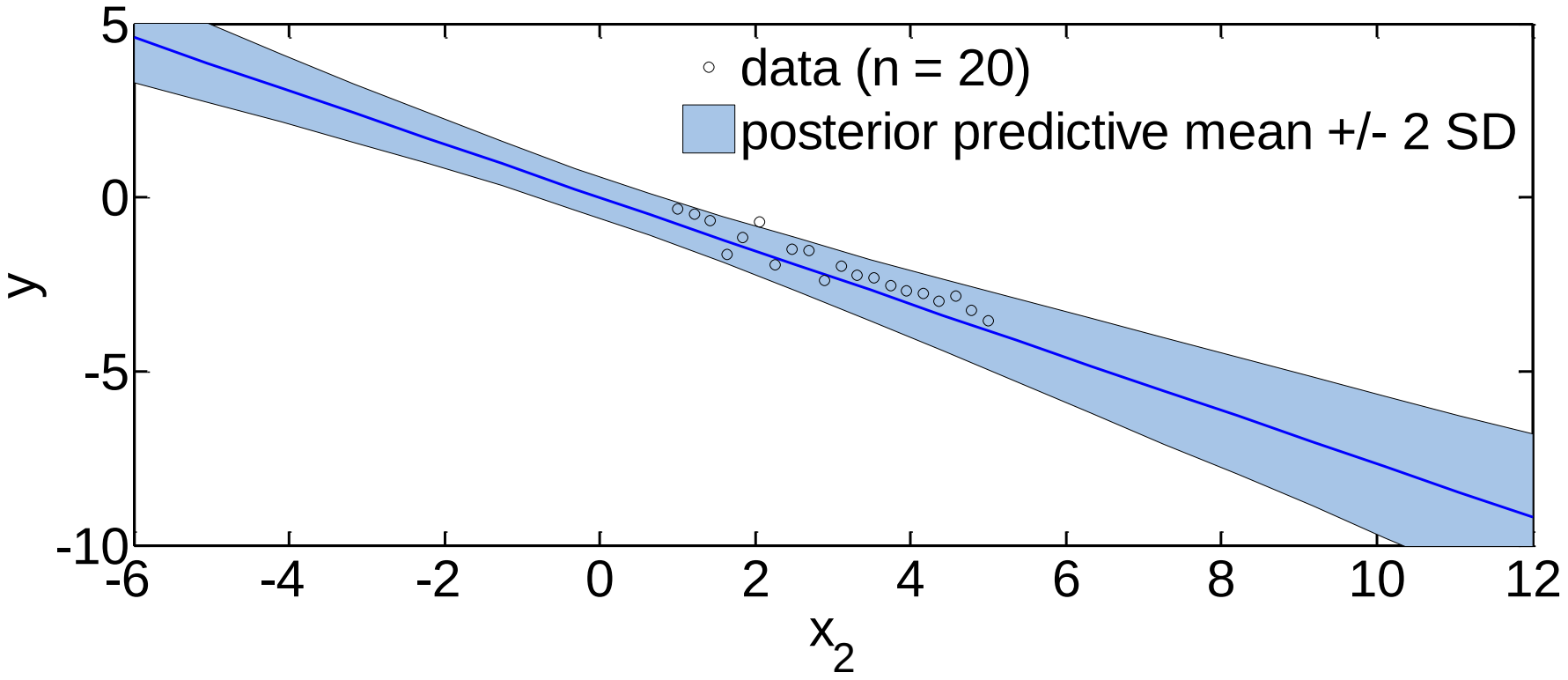
Variational linear regression: predictive density

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



Variational linear regression: predictive density

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



MATLAB implementation

vblm.m

```
% 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 vs. variational inference

Frequentist linear regression

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

$$p = P(t \geq t^* | H_0)$$

Variational Bayesian linear regression

$$\langle \beta | X, y \rangle \approx \langle \beta \rangle_{q(\beta)}$$

$$= \left(\langle \alpha \rangle_{q(\alpha)} + \langle \lambda \rangle_{q(\lambda)} X^T X \right)^{-1} \langle \lambda \rangle_{q(\lambda)} \, X^T y$$

$$\mathrm{Cov}(\beta | X, y) = \left(\langle \alpha \rangle_{q(\alpha)} I + \langle \lambda \rangle_{q(\lambda)} X^T X \right)^{-1}$$

$$\ln BF = \ln F_1 - \ln F_2$$

Overview

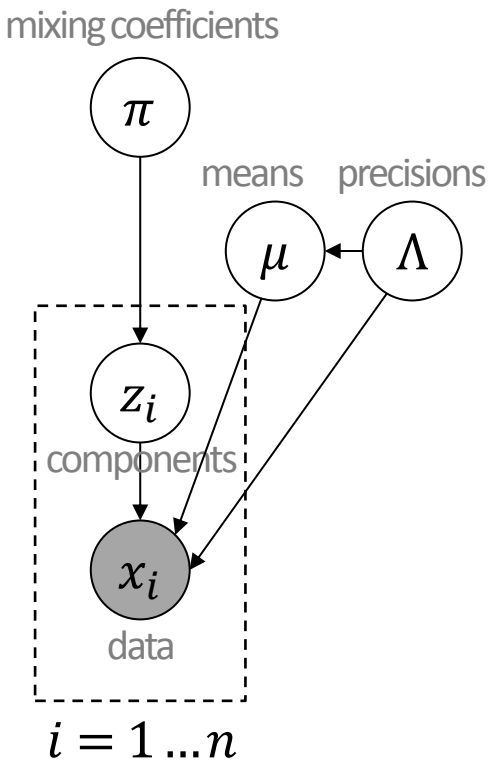
Application 3: variational clustering using a Gaussian mixture model

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.



$$p(\pi) = \mathrm{Dir}(\pi|\alpha_0)$$

$$p(\Lambda) = \prod_{k=1}^{K} \mathcal{W}(\Lambda_k|W_0, \nu_0)$$

$$p(\mu|\Lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k|m_0, (\beta_0\Lambda_k)^{-1})$$

$$p(z_i|\pi) = \prod_{k=1}^{K} \pi_k^{z_{i,k}}$$

$$p(x_i|Z, \mu, \Lambda) = \prod_{k=1}^{K} \mathcal{N}\left(x_i|\mu_k, \Lambda_k^{-1}\right)^{z_{i,k}}$$

Variational clustering

Variational E-step

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

$$q^*(\pi) = \mathrm{Dir}(\pi|\alpha)$$

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

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

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

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

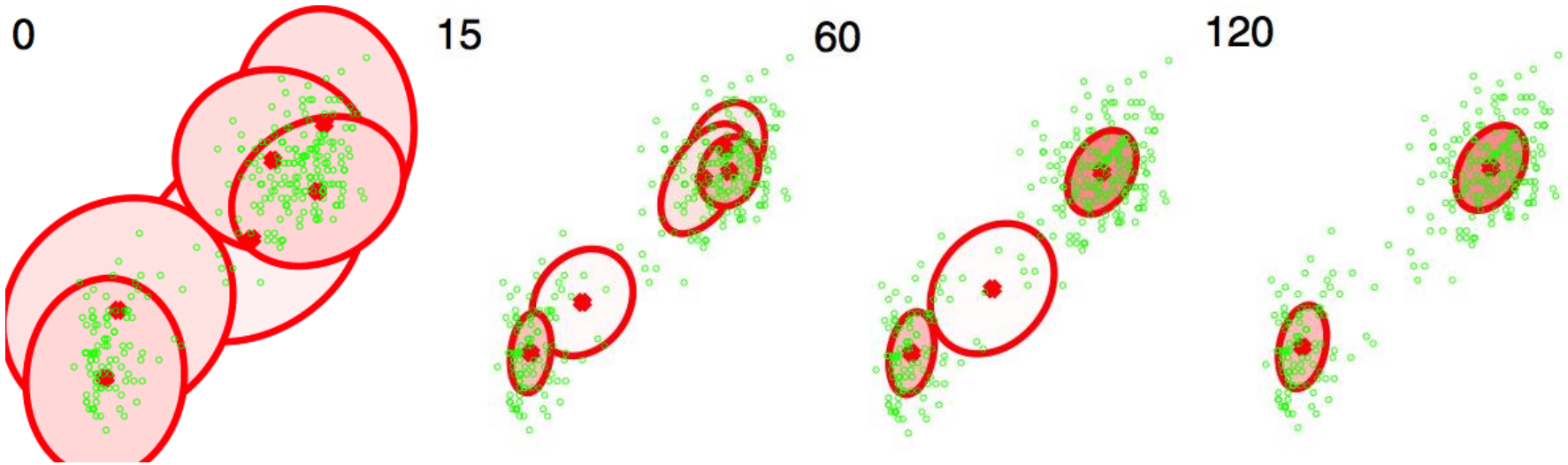
Variational M-step

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

where $r_{i,k} \coloneqq \frac{\rho_{i,k}}{\sum_{j=1}^{K} \rho_{i,j}}$

where $\rho_{i,j} \coloneqq \exp\left(-\frac{1}{2}\left(d\beta_k^{-1} + \nu_k (x_i - m_k)^T W_k (x_i - m_k)\right)\right)$

Variational clustering: example

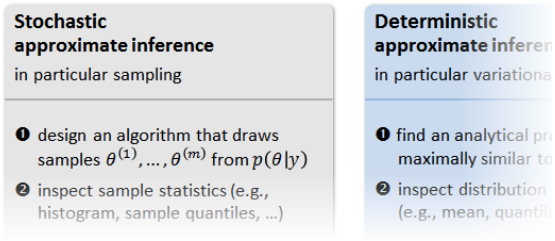


Advantages of variational clustering over the maximum-likelihood approach:

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

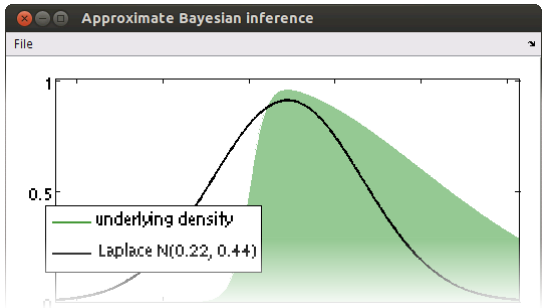
Bishop (2005) PRML, p. 480

Summary (1)



Two approaches to approximate inference

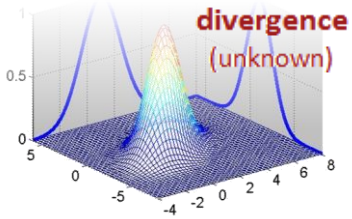
- stochastic inference (sampling)
- deterministic inference (variational Bayes)



The Laplace approximation

- simple local approximation
- often used in conjunction with VB

$$\ln p(y) = \underbrace{\mathrm{KL}[q||p]}_{\text{divergence}} + \underbrace{F(q,y)}_{\text{free energy}}$$



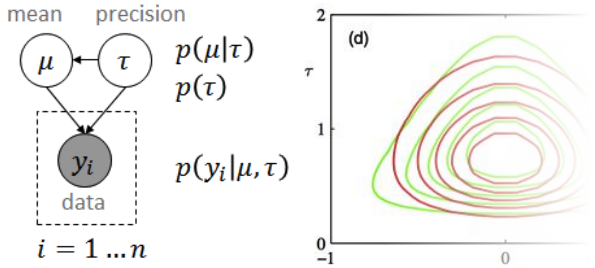
divergence (unknown)

free energy (easy to evaluate for a given q)

Variational inference under the mean-field assumption

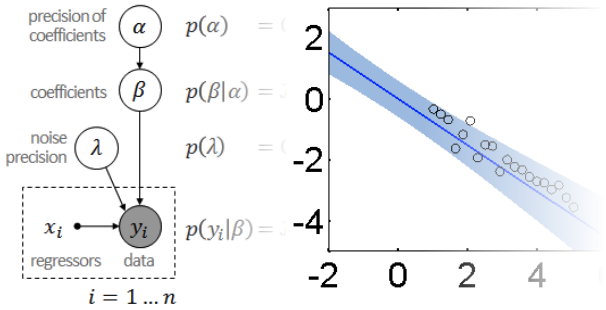
- to maximize F means to minimize $KL[q||p]$
- variational algorithm

Summary (2)



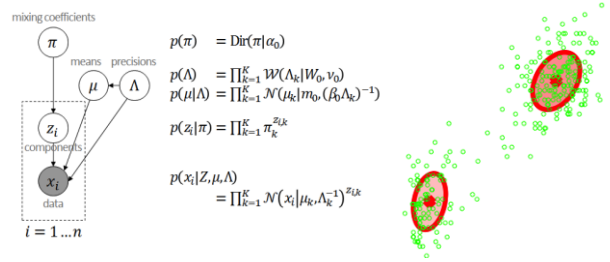
Variational univariate density estimation

- exact solution available



Variational multiple linear regression

- `vblm.m`



Variational clustering using a Gaussian mixture model

- `spm_mix.m`