Binary probit regression with I-priors

Haziq Jamil
Supervisors: Dr. Wicher Bergsma & Prof. Irini Moustaki

Social Statistics (Year 3)
London School of Economics & Political Science

8-9 May 2017

PhD Presentation Event

http://phd3.haziqj.ml

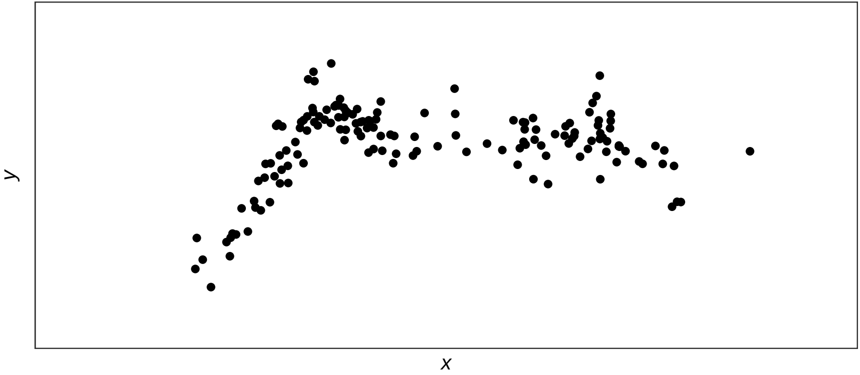
Outline

The regression model

- For $i = 1, \dots, n$, consider the regression model

$$y_i = f(x_i) + \epsilon_i$$
$$(\epsilon_1, \dots, \epsilon_n) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi}^{-1})$$

where $f \in \mathcal{F}$, $y_i \in \mathbb{R}$, and $x_i = (x_{i1}, \dots, x_{ip}) \in \mathcal{X}$.



I-priors

- Let $\mathcal{F}$ be a reproducing kernel Hilbert space (RKHS) with reproducing kernel $h_\lambda : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. An I-prior on f is

$$\big(f(x_1), \dots, f(x_n)\big)^\top \sim \mathrm{N}\left(\mathbf{f}_0, \mathcal{I}(f)\right)$$

with $\mathbf{f}_0$ a prior mean, and $\mathcal{I}$ the Fisher information for f, given by

$$\mathcal{I}\big(f(x), f(x')\big) = \sum_{k=1}^n \sum_{l=1}^n \psi_{kl} h_\lambda(x, x_k) h_\lambda(x', x_l).$$

- The I-prior regression model for $i = 1, \dots, n$ becomes

$$y_i = f_0(x_i) + \sum_{k=1}^n h_\lambda(x_i, x_k) w_k + \epsilon_i$$

$$(w_1, \dots, w_n) \sim \mathrm{N}(\mathbf{0}, \mathbf{\Psi})$$

$$(\epsilon_1, \dots, \epsilon_n) \sim \mathrm{N}(\mathbf{0}, \mathbf{\Psi}^{-1})$$

W. Bergsma (2017). "Regression with I-priors". *Manuscript in preparation*

I-priors (cont.)

- Of interest is the posterior of the function

$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}},$$

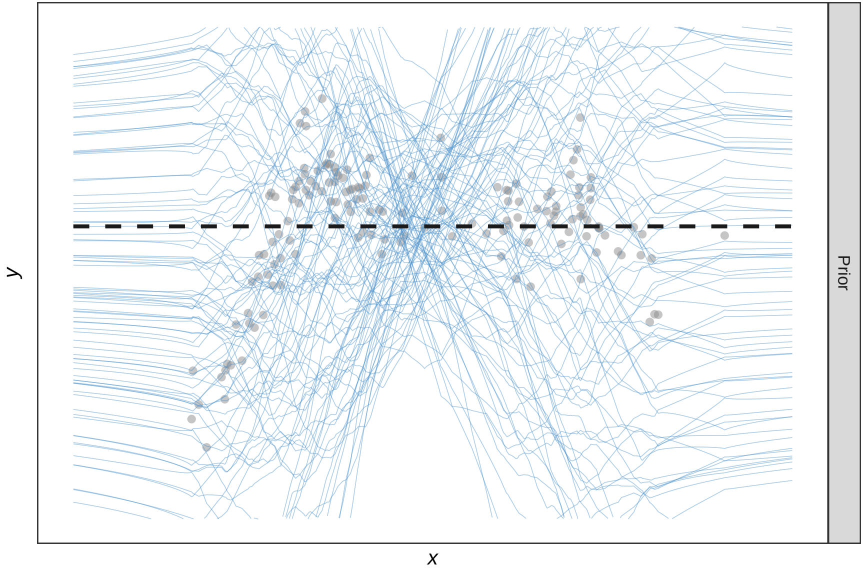
  and also the posterior predictive distribution

$$p(\mathbf{y}^*|\mathbf{y}) = \int p(\mathbf{y}^*|\mathbf{y},\mathbf{f})p(\mathbf{f}|\mathbf{y})\,\mathrm{d}\mathbf{f}.$$

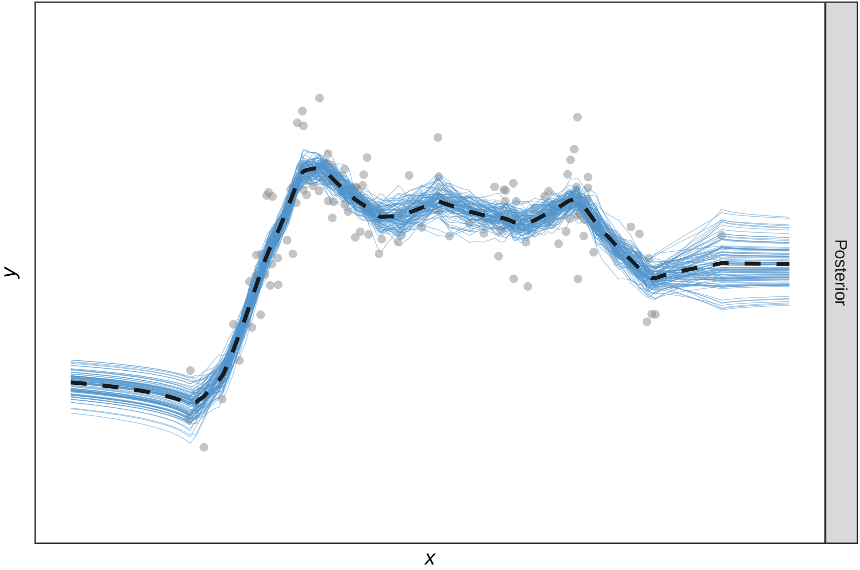
- Estimation using EM algorithm or direct maximisation of the marginal likelihood $\log p(y)$.
- Fully Bayesian estimation also possible.

HJ (2017). *iprior: Linear Regression using I-Priors*. R Package version 0.6.4: CRAN/GitHub

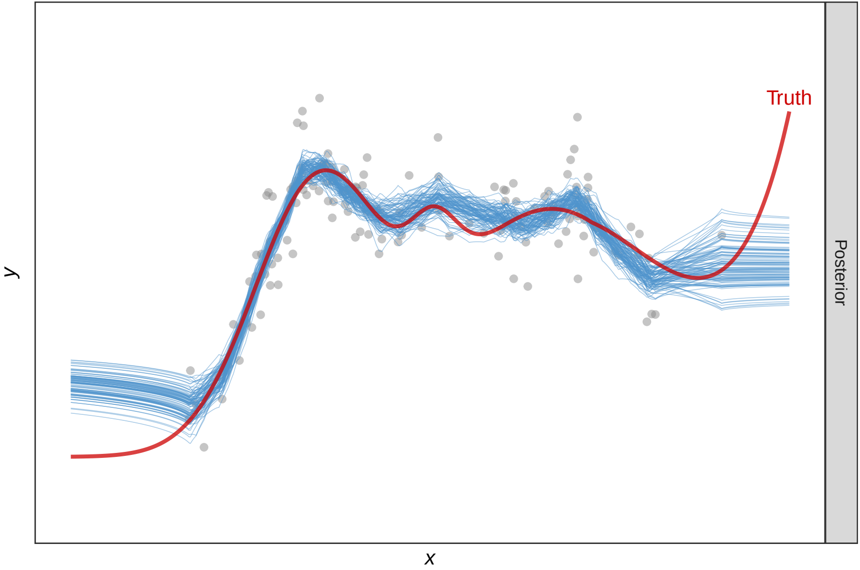
Fractional Brownian motion (FBM) RKHS



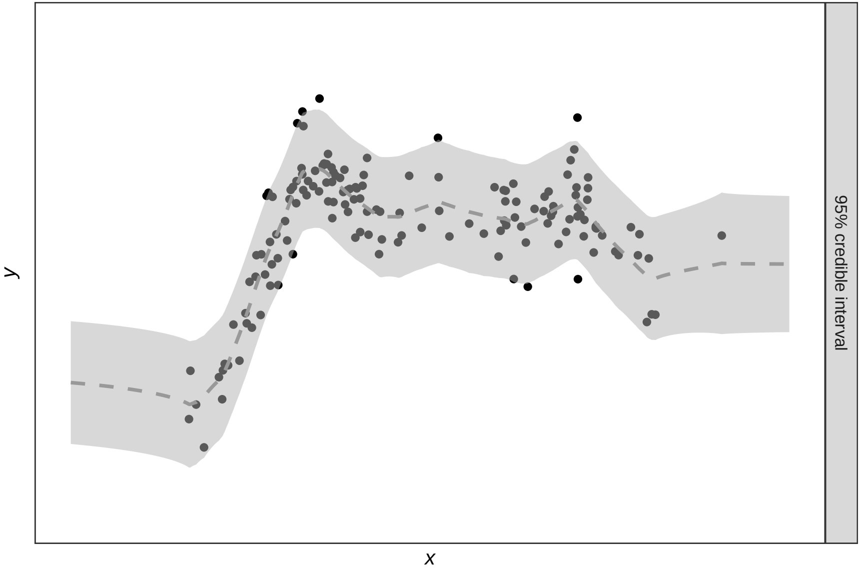
Fractional Brownian motion (FBM) RKHS



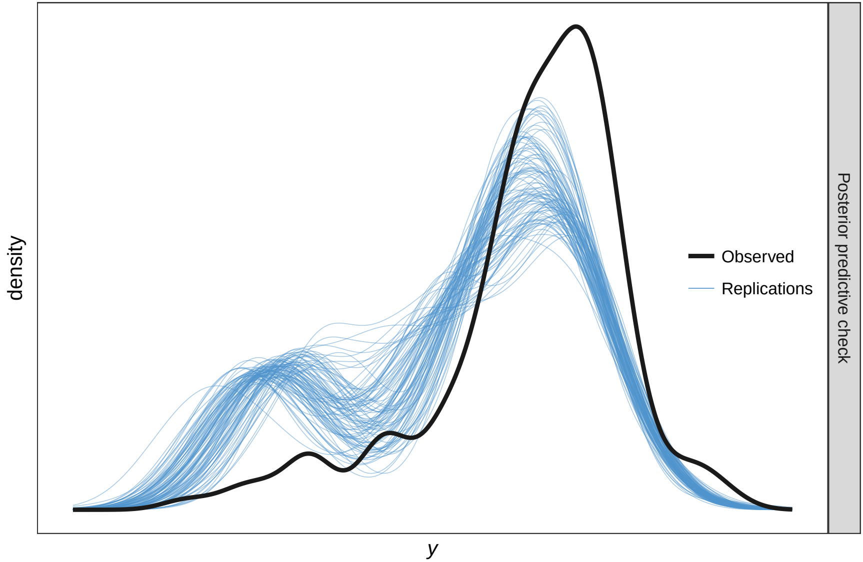
Fractional Brownian motion (FBM) RKHS



Posterior predictive distribution



Posterior predictive distribution



Roadmap

The latent variable motivation

- Consider binary responses $y_1, \dots, y_n$ together with their corresponding covariates $x_1, \dots, x_n$.
- For $i = 1, \dots, n$, model the responses as

$$y_i \sim \text{Bern}(p_i).$$

- Assume that there exists continuous, underlying latent variables $y_1^*, \dots, y_n^*$, such that

$$y_i = \begin{cases} 1 & \text{if } y_i^* \geq 0 \\ 0 & \text{if } y_i^* < 0. \end{cases}$$

- Model these continuous latent variables according to

$$y_i^* = f(x_i) + \epsilon_i$$

where $(\epsilon_1, \dots, \epsilon_n) \sim \text{N}(\mathbf{0}, \mathbf{\Psi}^{-1})$ and $f \in \mathcal{F}$ (some RKHS).

Using I-priors

- Assume an I-prior on f. Then,

$$f(x_i) = \alpha + \sum_{k=1}^{n} h_\lambda(x_i, x_k) w_k$$
$$(w_1, \dots, w_n) \sim \mathrm{N}(\mathbf{0}, \mathbf{\Psi})$$

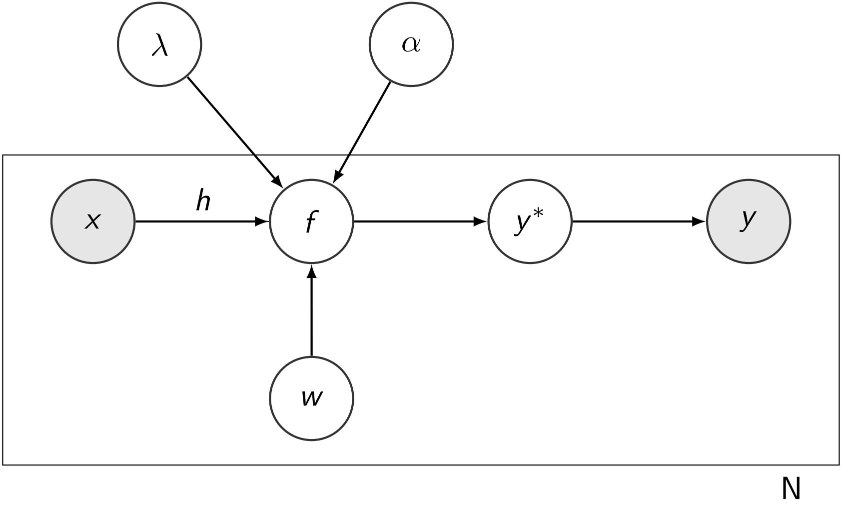
- For now, consider iid errors $\mathbf{\Psi} = \psi \mathbf{I}_n$. In this case,

$$\begin{aligned} p_i = \mathrm{P}[y_i = 1] &= \mathrm{P}[y_i^* \geq 0] \\ &= \mathrm{P}[\epsilon_i \leq f(x_i)] \\ &= \Phi\Big(\psi^{1/2}(\alpha + \textstyle\sum_{k=1}^n h_\lambda(x_i, x_k) w_k)\Big) \end{aligned}$$

  where Φ is the CDF of a standard normal.

- No loss of generality compared with using an arbitrary threshold τ or error precision ψ. Thus, set $\psi = 1$.

The probit I-prior model



ADD JOINT DISTRIBUTIONS AND COLOUR CODE

Estimation

- Denote $f_i = f(x_i)$ for short.
- The marginal density

$$
\begin{aligned}
p(\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f} \\
&= \int \prod_{i=1}^{n} \Big[\Phi(f_i)^{y_i} \big(1 - \Phi(f_i)\big)^{1-y_i} \Big] \cdot \mathrm{N}(\alpha\mathbf{1}_n, \mathbf{H}_\lambda^2)\,\mathrm{d}\mathbf{f}
\end{aligned}
$$

for which $p(\mathbf{f}|\mathbf{y})$ depends, cannot be evaluated analytically.

Estimation

- Denote $f_i = f(x_i)$ for short.
- The marginal density

$$\begin{aligned} p(\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f} \\ &= \int \prod_{i=1}^{n} \Big[\Phi(f_i)^{y_i} \big(1 - \Phi(f_i)\big)^{1-y_i} \Big] \cdot \mathrm{N}(\alpha \mathbf{1}_n, \mathbf{H}_\lambda^2)\,\mathrm{d}\mathbf{f} \end{aligned}$$

for which $p(\mathbf{f}|\mathbf{y})$ depends, cannot be evaluated analytically.

- Some strategies:
  - ✗ Naive Monte-Carlo integral

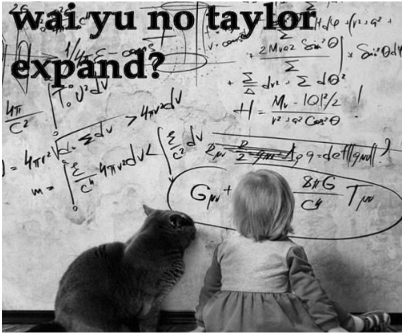
Estimation

- Denote $f_i = f(x_i)$ for short.
- The marginal density

$$
\begin{aligned}
p(\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f} \\
&= \int \prod_{i=1}^{n} \Big[\Phi(f_i)^{y_i} \big(1 - \Phi(f_i)\big)^{1-y_i} \Big] \cdot \mathrm{N}(\alpha \mathbf{1}_n, \mathbf{H}_\lambda^2)\,\mathrm{d}\mathbf{f}
\end{aligned}
$$

  for which $p(\mathbf{f}|\mathbf{y})$ depends, cannot be evaluated analytically.
- Some strategies:
  - ✗ Naive Monte-Carlo integral
  - ✗ EM algorithm with a MCMC E-step

Estimation



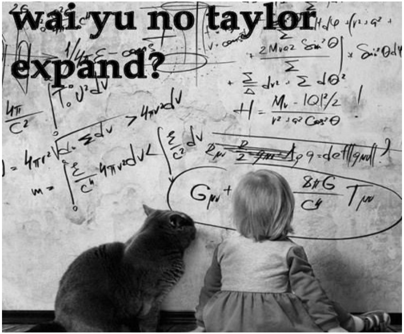
- Denote $f_i = f(x_i)$ for short.
- The marginal density

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}$$
$$= \int \prod_{i=1}^{n} \left[\Phi(f_i)^{y_i} \big(1 - \Phi(f_i)\big)^{1-y_i} \right] \cdot \mathrm{N}(\alpha\mathbf{1}_n, \mathbf{H}_\lambda^2)\,\mathrm{d}\mathbf{f}$$

for which $p(\mathbf{f}|\mathbf{y})$ depends, cannot be evaluated analytically.

- Some strategies:
  - ✗ Naive Monte-Carlo integral
  - ✗ EM algorithm with a MCMC E-step
  - ✓ Laplace approximation

Estimation



- Denote $f_i = f(x_i)$ for short.
- The marginal density

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}$$
$$= \int \prod_{i=1}^{n} \left[\Phi(f_i)^{y_i} \big(1 - \Phi(f_i)\big)^{1-y_i} \right] \cdot \mathrm{N}(\alpha\mathbf{1}_n, \mathbf{H}_\lambda^2)\,\mathrm{d}\mathbf{f}$$

for which $p(\mathbf{f}|\mathbf{y})$ depends, cannot be evaluated analytically.

- Some strategies:
  - ✗ Naive Monte-Carlo integral
  - ✗ EM algorithm with a MCMC E-step
  - ✓ Laplace approximation
  - ✓ MCMC sampling

Laplace's method

- Interested in $p(\mathbf{f}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{f})p(\mathbf{f}) =: e^{Q(\mathbf{f})}$, with normalising constant $p(\mathbf{y}) = \int e^{Q(\mathbf{f})} \, \mathrm{d}\mathbf{f}$. The Taylor expansion of Q about its mode $\tilde{\mathbf{f}}$

$$Q(\mathbf{f}) \approx Q(\tilde{\mathbf{f}}) - \frac{1}{2}(\mathbf{f} - \tilde{\mathbf{f}})^\top \mathbf{A}(\mathbf{f} - \tilde{\mathbf{f}})$$

  is recognised as the logarithm of an unnormalised Gaussian density, with $\mathbf{A} = -\mathrm{D}^2 Q(\mathbf{f})$ being the negative Hessian of Q evaluated at $\tilde{\mathbf{f}}$.

- The posterior $p(\mathbf{f}|\mathbf{y})$ is approximated by $\mathrm{N}(\tilde{\mathbf{f}}, \mathbf{A}^{-1})$, and the marginal by

$$p(\mathbf{y}) \approx (2\pi)^{n/2} |\mathbf{A}|^{-1/2} p(\mathbf{y}|\tilde{\mathbf{f}}) p(\tilde{\mathbf{f}})$$

- Won't scale with large n; difficult to find modes in high dimensions.

R. Kass and A. Raftery (1995). "Bayes Factors". *Journal of the American Statistical Association* 90.430, §4.1, pp. 777-778.

Full Bayesian analysis using MCMC

- Assign hyperpriors on parameters of the I-prior, e.g.
  - $\lambda^2 \sim \Gamma^{-1}(a, b)$
  - $\alpha \sim \mathrm{N}(c, d^2)$

  for a hierarchical model to be estimated fully Bayes.
- No closed-form posteriors - need to resort to MCMC sampling.
- Computationally slow, and sampling difficulty results in unreliable posterior samples.

DENSITY PLOTS OF LAMBDA HERE

Variational inference

- Name derived from calculus of variations which deals with maximising or minimising functionals.

$$\begin{array}{lll} \text{Functions} & p : \theta \mapsto \mathbb{R} & \text{(standard calculus)} \\ \text{Functionals} & \mathcal{H} : p \mapsto \mathbb{R} & \text{(variational calculus)} \end{array}$$

- Using standard calculus, we can solve

$$\arg\max_{\theta} p(\theta) =: \hat{\theta}$$

  e.g. p is a likelihood function, and $\hat{\theta}$ is the ML estimate.

- Using variational calculus, we can solve

$$\arg\max_{p} \mathcal{H}(p) =: \tilde{p}$$

  e.g. $\mathcal{H}$ is the entropy $\mathcal{H} = -\int p(x) \log p(x) \, dx$, and $\tilde{p}$ is the entropy maximising distribution.

Variational inference (cont.)

- Consider a statistical model where we have observations $(y_1, \dots, y_n)$ and also some latent variables $(z_1, \dots, z_n)$.
- The z_i could be random effects or some auxiliary latent variables.
- In a Bayesian setting, this could also include the parameters to be estimated.
- **GOAL**: Find approximations for
  - The posterior distribution $p(\mathbf{z}|\mathbf{y})$; and
  - The marginal likelihood (or model evidence) $p(\mathbf{y})$.
- Variational inference is a deterministic approach, unlike MCMC.

C. M. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer, Ch. 10

K. P. Murphy (1991). *Machine Learning: A Probabilistic Perspective*. The MIT Press. DOI: 10.1007/SpringerReference_35834, Ch. 21

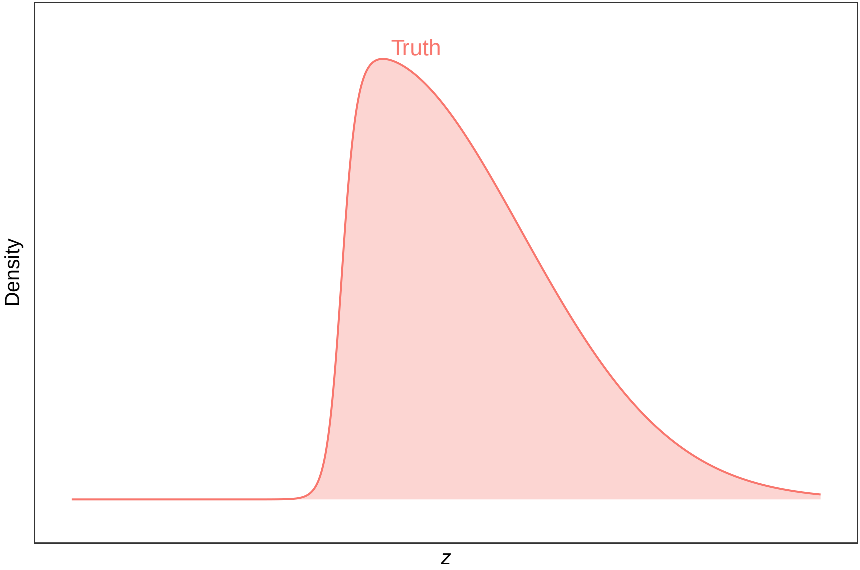
Decomposition of the log marginal

- Let $q(\mathbf{z})$ be some density function to approximate $p(\mathbf{z}|\mathbf{y})$. Then the log-marginal density can be decomposed into

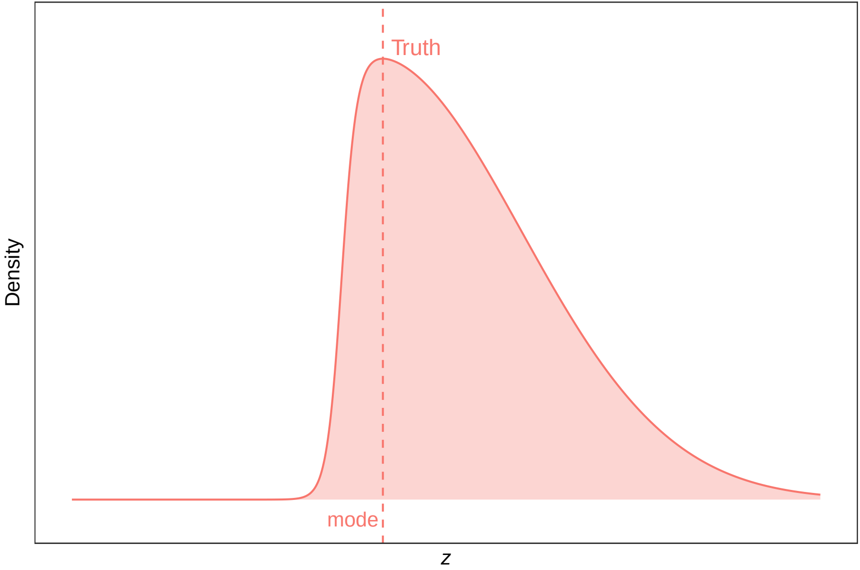
$$\begin{aligned}\log p(\mathbf{y}) &= \log p(\mathbf{y},\mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \\ &= \int \left\{ \log \frac{p(\mathbf{y},\mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z}) \, d\mathbf{z} \\ &= \mathcal{L}(q) + \mathrm{KL}(q\|p) \\ &\geq \mathcal{L}(q)\end{aligned}$$

- $\mathcal{L}$ is referred to as the "lower-bound", and it serves as a surrogate function to the marginal.
- Maximising the $\mathcal{L}(q)$ is equivalent to minimising $\mathrm{KL}(q\|p)$.
- Although $\mathrm{KL}(q\|p)$ is minimised at $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y})$ (c.f. EM algorithm), we are unable to work with $p(\mathbf{z}|\mathbf{y})$.

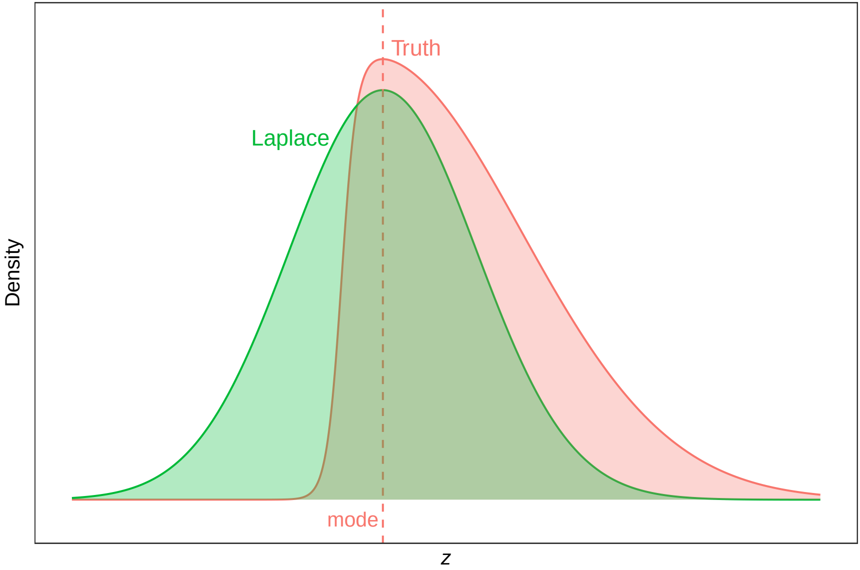
Comparison of approximations (density)



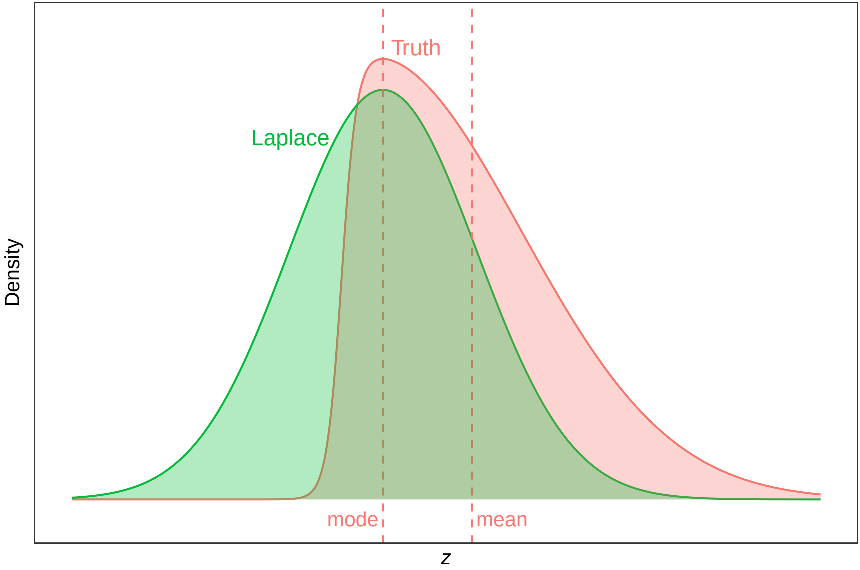
Comparison of approximations (density)



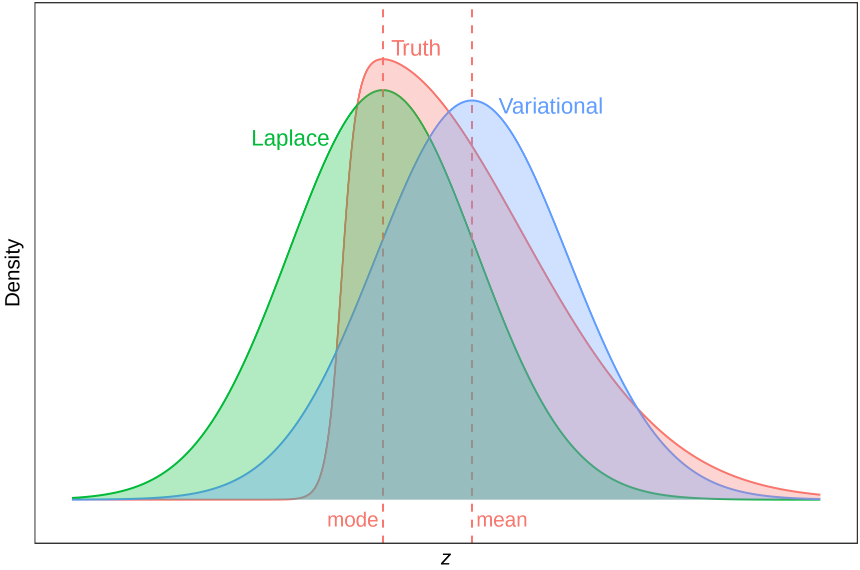
Comparison of approximations (density)



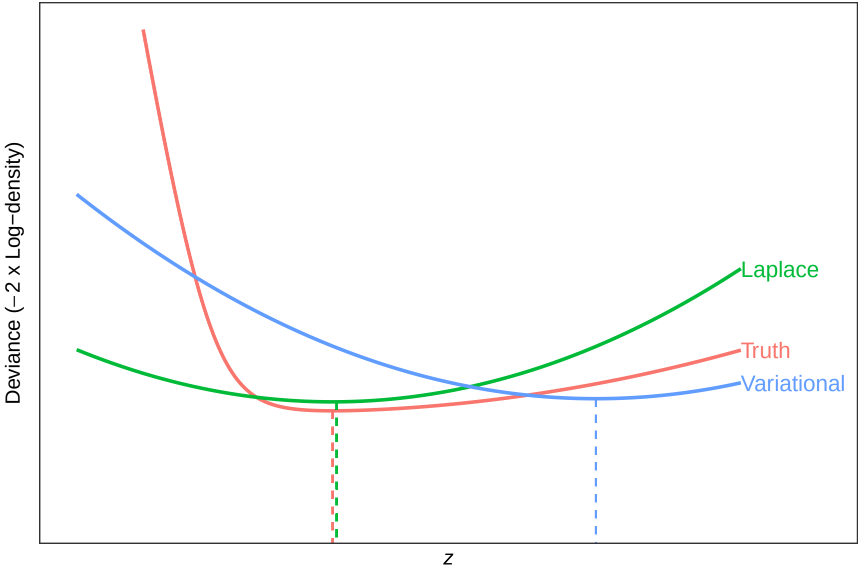
Comparison of approximations (density)



Comparison of approximations (density)



Comparison of approximations (deviance)



Factorised distributions (Mean-field theory)

- Maximising $\mathcal{L}$ over all possible q not feasible. Need some restrictions, but only to achieve tractability.
- Suppose we partition elements of $\mathbf{z}$ into m disjoint groups $\mathbf{z} = (\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)})$, and assume

$$q(\mathbf{z}) = \prod_{j=1}^{m} q_j(\mathbf{z}^{(j)})$$

- Under this restriction, the solution to $\arg\max_q \mathcal{L}(q)$ is

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp\big(\mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\big) \tag{1}$$

  for $j \in \{1, \dots, m\}$.
- In practice, these unnormalised densities are of recognisable form (especially if conjugate priors are used).

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe (2016). "Variational Inference: A Review for Statisticians". arXiv: 1601.00670

Coordinate ascent mean-field variational inference (CAVI)

- The optimal distributions are coupled with another, i.e. each $\tilde{q}_j(\mathbf{z}^{(j)})$ depends on the optimal moments of $\mathbf{z}^{(k)}$, $k \in \{1, \dots, m : k \neq j\}$.
- One way around this to employ an iterative procedure.
- Assess convergence by monitoring the lower bound

$$\mathcal{L}(q) = \mathrm{E}_q[\log p(\mathbf{y}, \mathbf{z})] - \mathrm{E}_q[\log q(\mathbf{z})].$$

Algorithm 1 CAVI

initialise Variational factors $q_j(\mathbf{z}^{(j)})$
while $\mathcal{L}(q)$ not converged **do**
  for $j = 1, \dots, m$ **do**
    $\log q_j(\mathbf{z}^{(j)}) \leftarrow \mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})] + \text{const.}$
  end for
  $\mathcal{L}(q) \leftarrow \mathrm{E}_q[\log p(\mathbf{y}, \mathbf{z})] - \mathrm{E}_q[\log q(\mathbf{z})]$
end while
return $\tilde{q}(\mathbf{z}) = \prod_{j=1}^m \tilde{q}_j(\mathbf{z}^{(j)})$

Estimation of a 1-dim Gaussian mean and variance

- GOAL: Bayesian inference of mean μ and variance ψ^{-1}

$$y_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \psi^{-1}) \qquad \text{Data}$$

$$\mu|\psi \sim \mathrm{N}\left(\mu_0, (\kappa_0\psi)^{-1}\right)$$

$$\psi \sim \Gamma(a_0, b_0) \qquad \text{Priors}$$

$$i = 1, \dots, n$$

Estimation of a 1-dim Gaussian mean and variance

- GOAL: Bayesian inference of mean μ and variance ψ^{-1}

$$\begin{aligned} y_i &\overset{\text{iid}}{\sim} \mathrm{N}(\mu, \psi^{-1}) && \text{Data} \\ \mu|\psi &\sim \mathrm{N}\left(\mu_0, (\kappa_0\psi)^{-1}\right) && \text{Priors} \\ \psi &\sim \Gamma(a_0, b_0) && \\ i &= 1, \dots, n \end{aligned}$$

- Substitute $p(\mu, \psi|\mathbf{y})$ with the mean-field approximation

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

Estimation of a 1-dim Gaussian mean and variance

- GOAL: Bayesian inference of mean μ and variance ψ^{-1}

$$
\begin{aligned}
y_i &\overset{\text{iid}}{\sim} \mathrm{N}(\mu, \psi^{-1}) && \text{Data} \\
\mu|\psi &\sim \mathrm{N}\left(\mu_0, (\kappa_0\psi)^{-1}\right) && \multirow{2}{*}{Priors} \\
\psi &\sim \Gamma(a_0, b_0) \\
i &= 1, \dots, n
\end{aligned}
$$

- Substitute $p(\mu, \psi|\mathbf{y})$ with the mean-field approximation

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

- From (1), we can work out the solutions

Estimation of a 1-dim Gaussian mean and variance

- GOAL: Bayesian inference of mean μ and variance ψ^{-1}

- Under the mean-field restriction, the solution to $\arg\max_q \mathcal{L}(q)$ is

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp\left(\mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\right) \tag{1}$$

for $j \in \{1, \dots, m\}$.

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

- From (1), we can work out the solutions

Estimation of a 1-dim Gaussian mean and variance

- GOAL: Bayesian inference of mean μ and variance ψ^{-1}

- Under the mean-field restriction, the solution to $\arg\max_q \mathcal{L}(q)$ is

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp\left(\mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\right) \tag{1}$$

for $j \in \{1, \dots, m\}$.

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

- From (1), we can work out the solutions

$$\begin{aligned}
\log \tilde{q}_\mu(\mu) &= \mathrm{E}_\psi[\log p(\mathbf{y}|\mu,\psi)] + \mathrm{E}_\psi[\log p(\mu|\psi)] + \text{const.} \\
\log \tilde{q}_\psi(\psi) &= \mathrm{E}_\mu[\log p(\mathbf{y}|\mu,\psi)] + \mathrm{E}_\mu[\log p(\mu|\psi)] + \log p(\psi) \\
&\quad + \text{const.}
\end{aligned}$$

Estimation of a 1-dim Gaussian mean and variance

- GOAL: Bayesian inference of mean μ and variance ψ^{-1}

- Under the mean-field restriction, the solution to $\arg\max_q \mathcal{L}(q)$ is

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp\big(\,\mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\big) \qquad (1)$$

for $j \in \{1, \dots, m\}$.

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

- From (1), we can work out the solutions

$$\tilde{q}_\mu(\mu) \equiv \mathrm{N}\left(\frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n}, \frac{1}{(\kappa_0 + n)\,\mathrm{E}_q[\psi]}\right)$$

Estimation of a 1-dim Gaussian mean and variance

- GOAL: Bayesian inference of mean μ and variance ψ^{-1}

> - Under the mean-field restriction, the solution to $\arg\max_q \mathcal{L}(q)$ is
>
> $$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp\left(\mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\right) \tag{1}$$
>
> for $j \in \{1, \dots, m\}$.

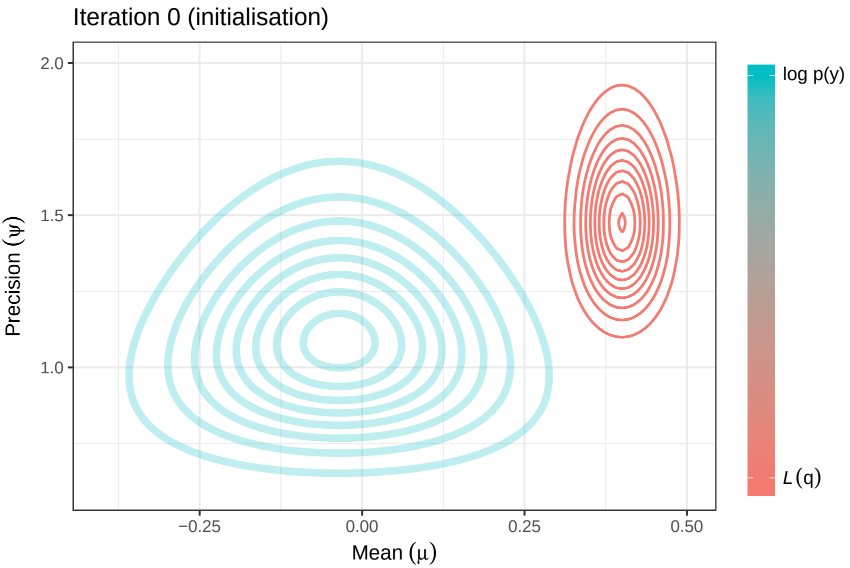
$$q(\mu, \psi) = q_\mu(\mu) q_\psi(\psi)$$

- From (1), we can work out the solutions

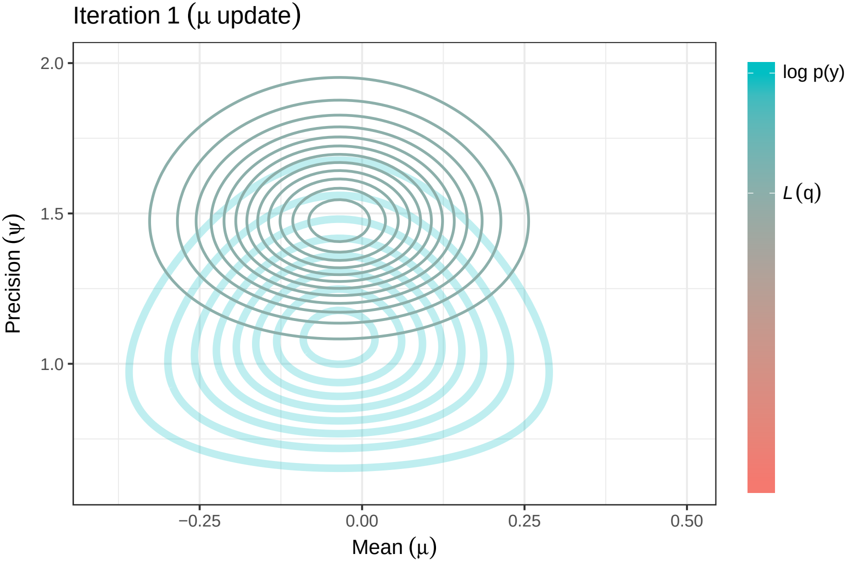
$$\tilde{q}_\mu(\mu) \equiv \mathrm{N}\left(\frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n}, \frac{1}{(\kappa_0 + n)\,\mathrm{E}_q[\psi]}\right) \quad \text{and} \quad \tilde{q}_\psi(\psi) \equiv \Gamma(\tilde{a}, \tilde{b})$$

$$\tilde{a} = a_0 + \frac{n}{2} \qquad \tilde{b} = b_0 + \frac{1}{2}\mathrm{E}_q\left[\textstyle\sum_{i=1}^n (y_i - \mu)^2 + \kappa_0(\mu - \mu_0)^2\right]$$

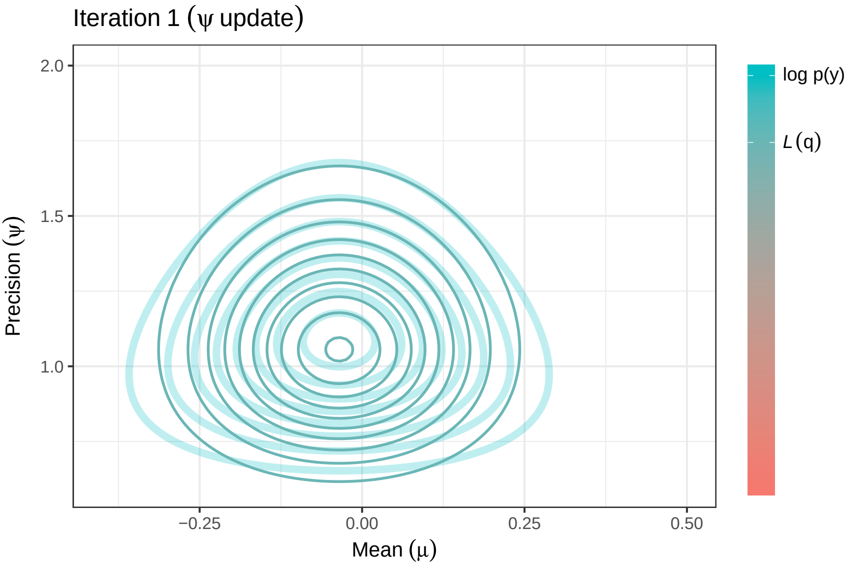
Estimation of a 1-dim Gaussian mean and variance (cont.)



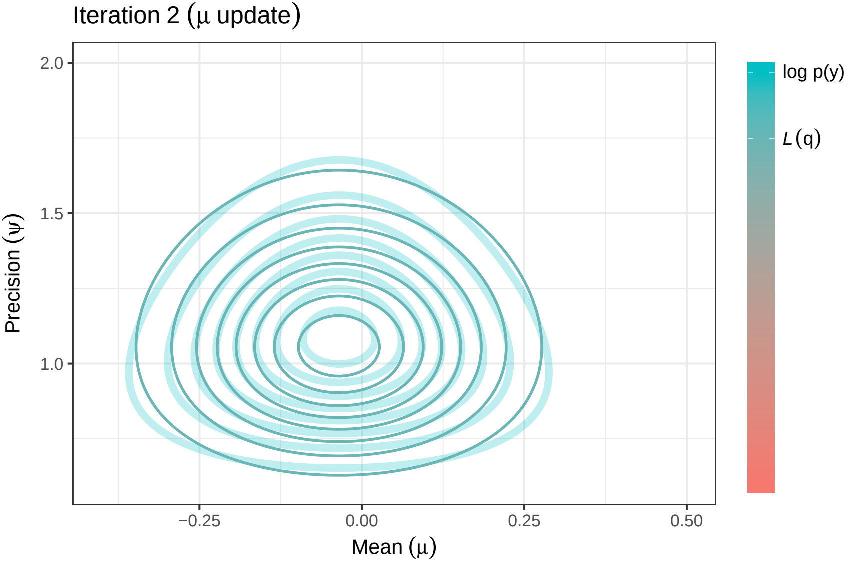
Estimation of a 1-dim Gaussian mean and variance (cont.)



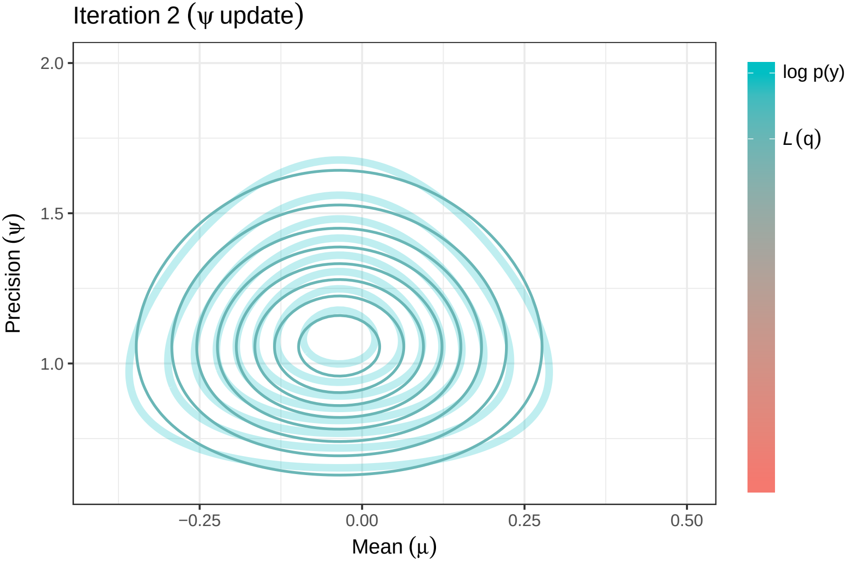
Estimation of a 1-dim Gaussian mean and variance (cont.)



Estimation of a 1-dim Gaussian mean and variance (cont.)



Estimation of a 1-dim Gaussian mean and variance (cont.)



Simulated data

R code

Timings, parameter estimates, training error rate, test error rate

Diagnostics

Monitor the lower bound

Cardiac arrhythmia data set

Multilevel example

Longitudinal example

Summary

Way forward

End

Thank you!

References I

Bergsma, W. (2017). "Regression with I-priors". *Manuscript in preparation.*

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2016). "Variational Inference: A Review for Statisticians". arXiv: 1601.00670.

HJ (2017). *iprior: Linear Regression using I-Priors*. R Package version 0.6.4: CRAN/GitHub.

Kass, R. and A. Raftery (1995). "Bayes Factors". *Journal of the American Statistical Association* 90.430.

Murphy, K. P. (1991). *Machine Learning: A Probabilistic Perspective*. The MIT Press. DOI: 10.1007/SpringerReference_35834.