Binary probit regression with I-priors

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

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

8 May 2017

PhD Presentation Event

http://phd3.haziqj.ml

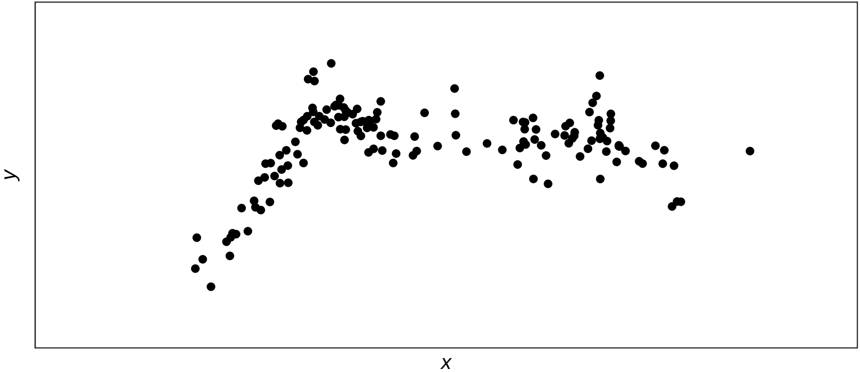
Outline

The regression model

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

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

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

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

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 regression function characterised by the distribution

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

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

I-priors (cont.)

- Of interest is the posterior regression function characterised by the distribution

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

and also the posterior predictive distribution for new data points x_{new}

$$p(y_{\text{new}}|\mathbf{y}) = \int p(y_{\text{new}}|\mathbf{y}, f_{\text{new}})p(f_{\text{new}}|\mathbf{y})\,\mathrm{d}f_{\text{new}}$$

with $f_{\text{new}} = f(x_{\text{new}})$.

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

I-priors (cont.)

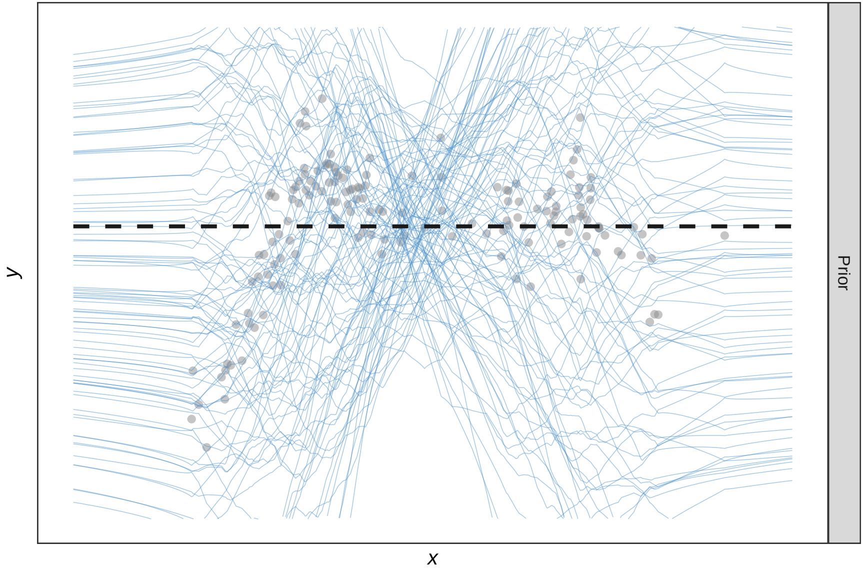
- Of interest is the posterior regression function characterised by the distribution
$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}},$$
and also the posterior predictive distribution for new data points x_{new}
$$p(y_{\text{new}}|\mathbf{y}) = \int p(y_{\text{new}}|\mathbf{y}, f_{\text{new}})p(f_{\text{new}}|\mathbf{y})\,\mathrm{d}f_{\text{new}}$$
with $f_{\text{new}} = f(x_{\text{new}})$.

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

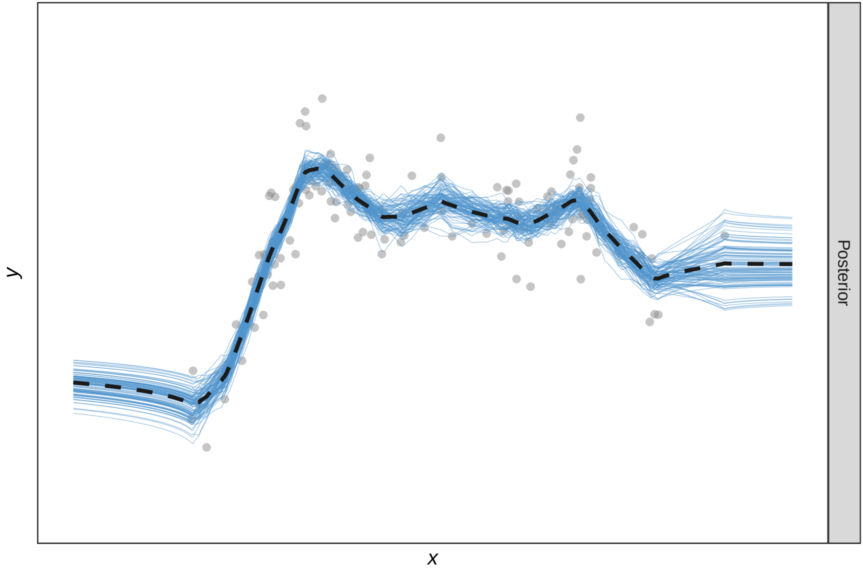
- Complete Bayesian estimation also possible.

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

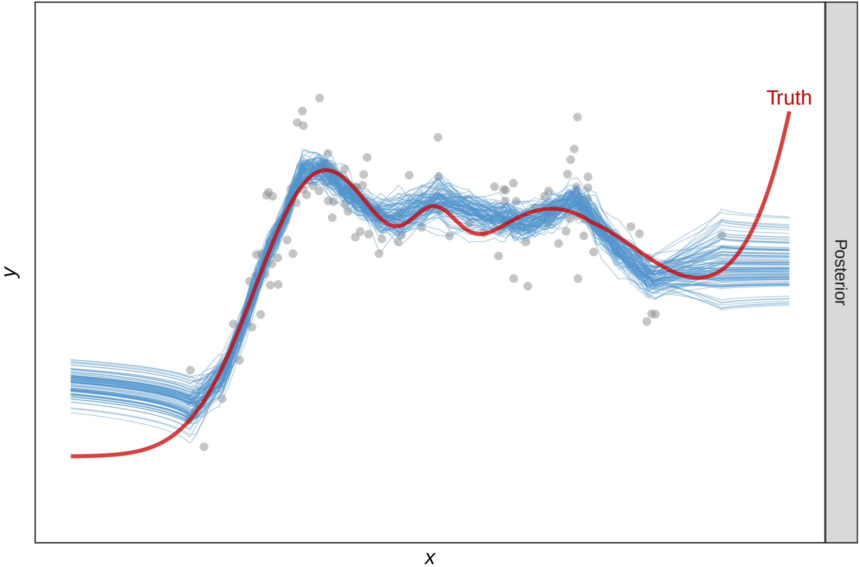
Fractional Brownian motion (FBM) RKHS



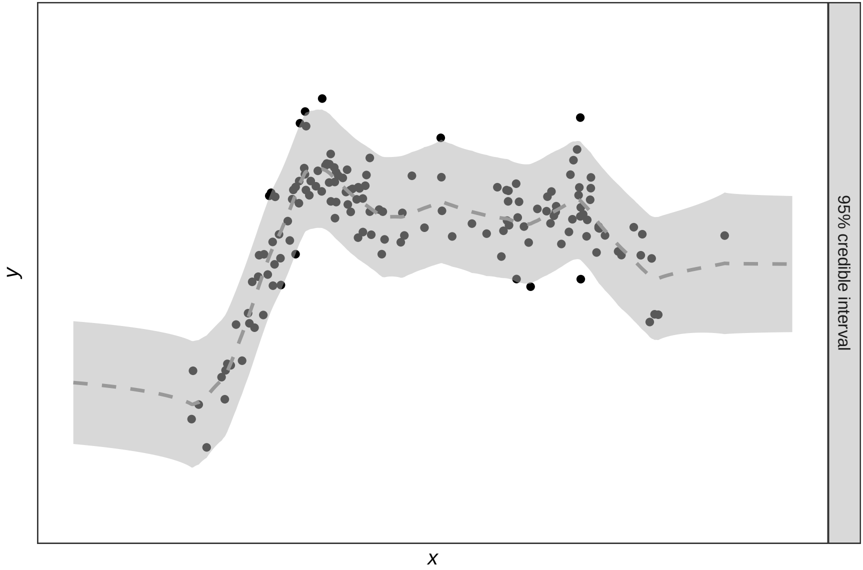
Fractional Brownian motion (FBM) RKHS



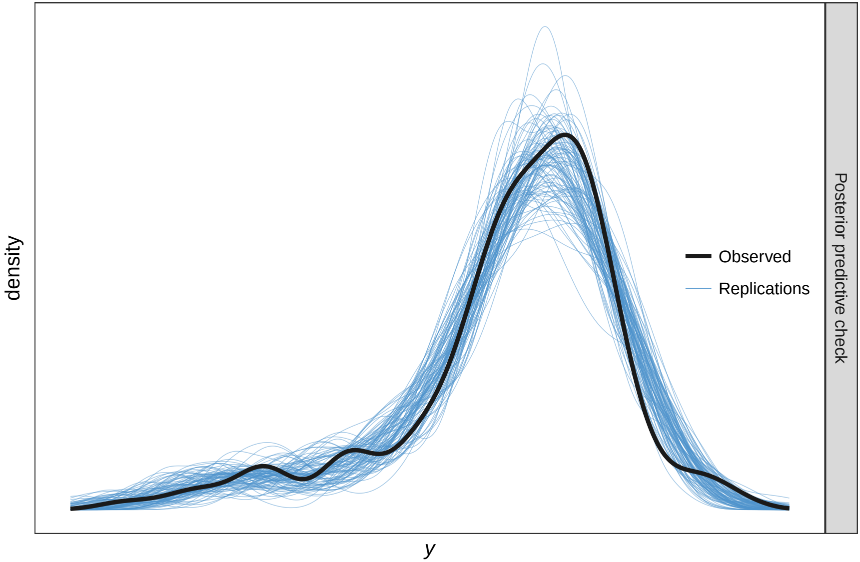
Fractional Brownian motion (FBM) RKHS



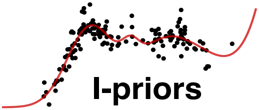
Posterior predictive distribution

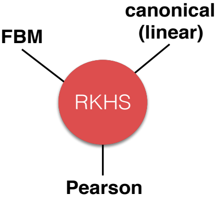


Posterior predictive distribution



PhD Roadmap

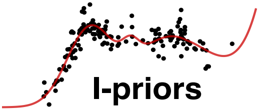


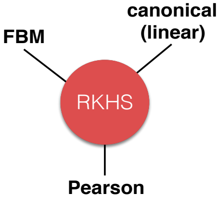


Unified methodology for
- additive models
- multilevel models
- models with functional covariates

<u>Advantages</u>
- Minimal assumptions
- Straightforward inference
- Performance competetive

PhD Roadmap





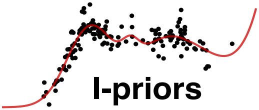
Unified methodology for
- additive models
- multilevel models
- models with functional covariates

Advantages
- Minimal assumptions
- Straightforward inference
- Performance competetive

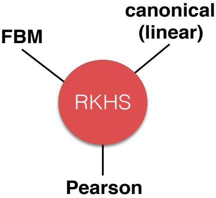
R/iprior

Estimation:
- Direct maximisation
- **EM algorithm**
- MCMC (Gibbs/HMC)

PhD Roadmap



Unified methodology for
- additive models
- multilevel models
- models with functional covariates



<u>Advantages</u>
- Minimal assumptions
- Straightforward inference
- Performance competetive

Bayesian Variable Selection
(using I-priors in the canonical RKHS)

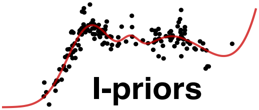
✓	✓	✗	✗	✓
X_1	X_2	X_3	X_4	X_5

Good performance in cases with multicollinearity

`R/iprior`

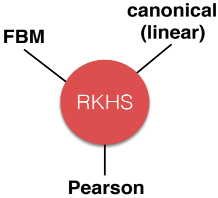
Estimation:
- Direct maximisation
- **<u>EM algorithm</u>**
- MCMC (Gibbs/HMC)

PhD Roadmap



I-priors

Unified methodology for
- additive models
- multilevel models
- models with functional covariates



Bayesian Variable Selection
(using I-priors in the canonical RKHS)

✓ X_1 ✓ X_2 ✗ X_3 ✗ X_4 ✓ X_5

Good performance in cases with multicollinearity

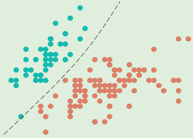
<u>Advantages</u>
- Minimal assumptions
- Straightforward inference
- Performance competetive

R/iprior

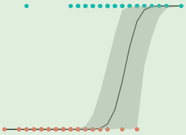
Estimation:
- Direct maximisation
- <u>**EM algorithm**</u>
- MCMC (Gibbs/HMC)

Binary probit models with I-priors

Extension to binary responses
Estimation using variational inference



classification



inference / fitted probabilities

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

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

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

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

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,

$$f(x_i) = f_0(x_i) + \sum_{k=1}^{n} h_\lambda(x_i, x_k) w_k$$
$$(w_1, \ldots, w_n) \sim \mathrm{N}(\mathbf{0}, \mathbf{\Psi})$$

Using I-priors

- Assume an I-prior on f. Then,

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

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

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

Using I-priors

- Assume an I-prior on f. Then,

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

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

- 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.

Using I-priors

- Assume an I-prior on f. Then,

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

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

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

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

  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 $\psi = 1$.

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

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

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

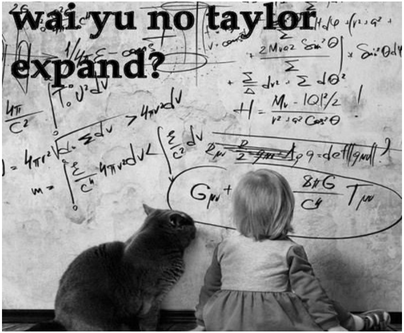
Estimation

- 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



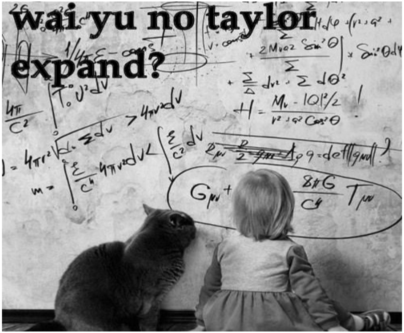
- Denote $f_i = f(x_i)$ for short.
- The marginal density

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

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

Variational inference

- Consider a statistical model where we have observations $(y_1, \dots, y_n)$ and also some latent variables $(z_1, \dots, z_n)$.

C. M. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer, Ch. 10
K. P. Murphy (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press, Ch. 21

Variational inference

- 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.

C. M. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer, Ch. 10
K. P. Murphy (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press, Ch. 21

Variational inference

- 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 (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press, Ch. 21

Decomposition of the log marginal

- Let $q(\mathbf{z})$ be some density function to approximate $p(\mathbf{z}|\mathbf{y})$.

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

$$\log p(\mathbf{y}) = \log p(\mathbf{y}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{y})$$

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

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

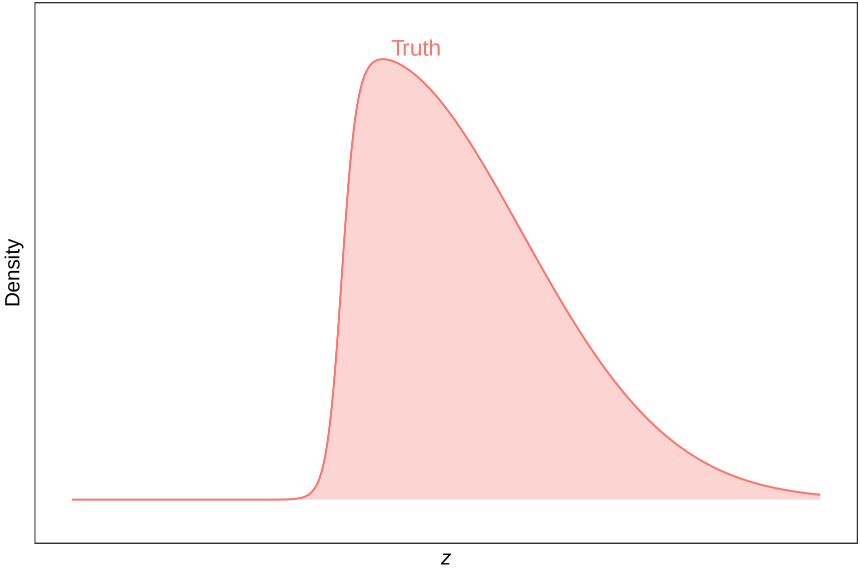
Decomposition of the log marginal

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

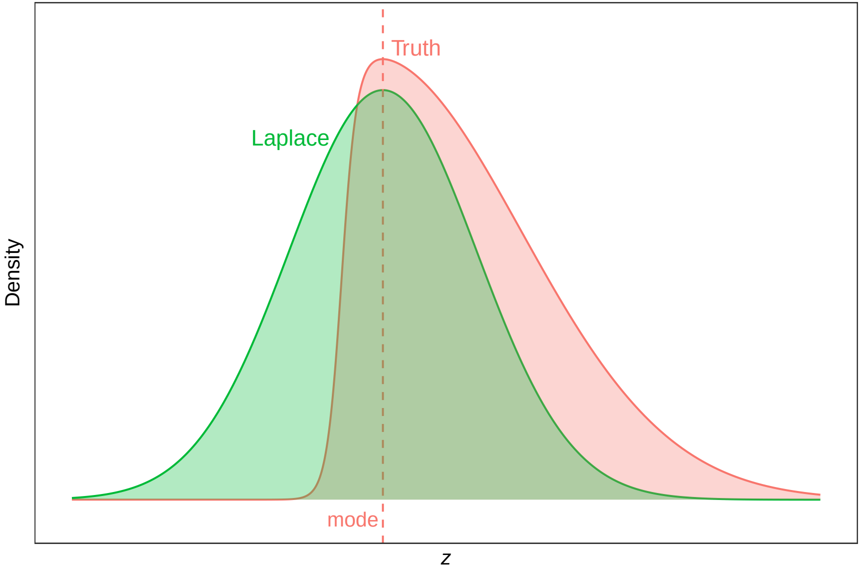
$$\begin{aligned}\log p(\mathbf{y}) &= \log p(\mathbf{y},\mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \\ &= \int \left\{ \log \frac{p(\mathbf{y},\mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z}) \, d\mathbf{z} \\ &= \mathcal{L}(q) + \mathrm{KL}(q\|p)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

- $\mathcal{L}$ is referred to as the "lower-bound", and it serves as a surrogate function to the marginal.
- Maximising $\mathcal{L}(q)$ 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 $p(\mathbf{z}|\mathbf{y})$.

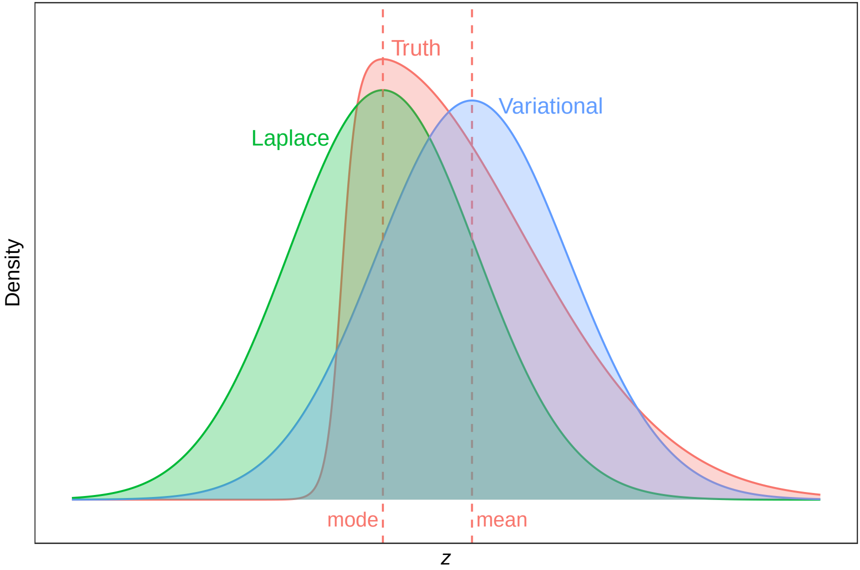
Comparison of approximations (density)



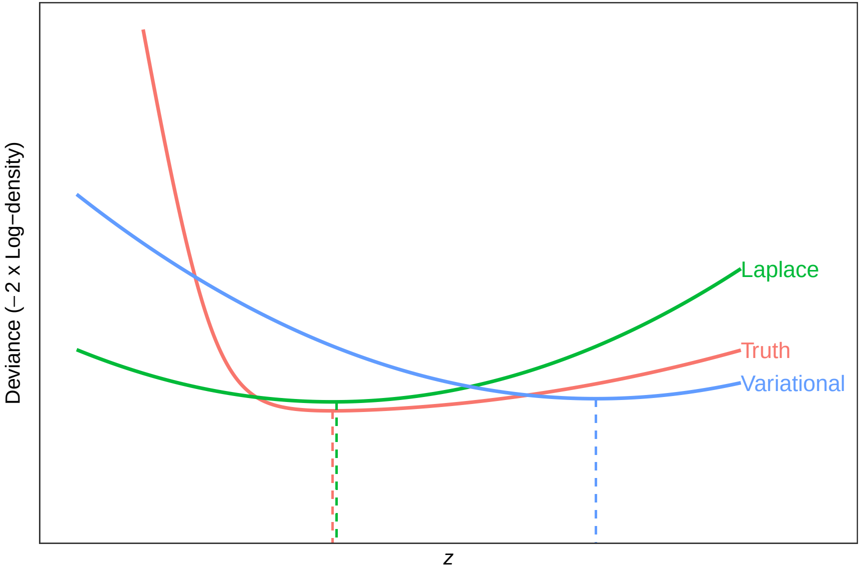
Comparison of approximations (density)



Comparison of approximations (density)



Comparison of approximations (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

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

D. M. Blei et al. (2016). "Variational Inference: A Review for Statisticians". arXiv: 1601.00670

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

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp\big(\operatorname{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\big) \tag{2}$$

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

D. M. Blei et al. (2016). "Variational Inference: A Review for Statisticians". arXiv: 1601.00670

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

$$\tilde{q}_j(\mathbf{z}^{(j)}) \propto \exp\big(\mathrm{E}_{-j}[\log p(\mathbf{y}, \mathbf{z})]\big) \tag{2}$$

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

D. M. Blei et al. (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

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

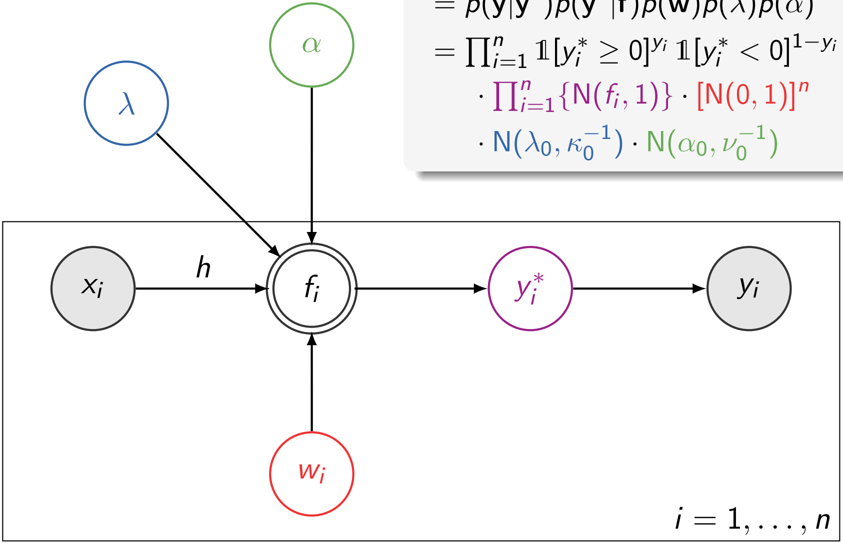
Algorithm 1 CAVI

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

example

Variational I-prior probit

$$
\begin{aligned}
& p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \alpha, \lambda) \\
& = p(\mathbf{y}|\mathbf{y}^*)p(\mathbf{y}^*|\mathbf{f})p(\mathbf{w})p(\lambda)p(\alpha) \\
& = \prod_{i=1}^n \mathbb{1}[y_i^* \geq 0]^{y_i}\, \mathbb{1}[y_i^* < 0]^{1-y_i} \\
& \quad \cdot \prod_{i=1}^n \{\mathrm{N}(f_i, 1)\} \cdot [\mathrm{N}(0,1)]^n \\
& \quad \cdot \mathrm{N}(\lambda_0, \kappa_0^{-1}) \cdot \mathrm{N}(\alpha_0, \nu_0^{-1})
\end{aligned}
$$



Posterior distribution

- Approximate the posterior by a mean-field variational density

$$p(\mathbf{y}^*, \mathbf{w}, \alpha, \lambda | \mathbf{y}) \approx \prod_{i=1}^{n} q(y_i^*) q(\mathbf{w}) q(\alpha) q(\lambda)$$

Posterior distribution

- Approximate the posterior by a mean-field variational density

$$p(\mathbf{y}^*, \mathbf{w}, \alpha, \lambda | \mathbf{y}) \approx \prod_{i=1}^{n} q(y_i^*) q(\mathbf{w}) q(\alpha) q(\lambda)$$

where

$$q(y_i^*) \equiv \begin{cases} \mathbb{1}[y_i^* \geq 0] \, \mathrm{N}(\tilde{f}_i, 1) & \text{if } y_i = 1 \\ \mathbb{1}[y_i^* < 0] \, \mathrm{N}(\tilde{f}_i, 1) & \text{if } y_i = 0 \end{cases} \qquad q(\mathbf{w}) \equiv \mathrm{N}(\tilde{\mathbf{w}}, \tilde{\mathbf{V}}_w)$$

$$q(\lambda) \equiv \mathrm{N}(\tilde{\lambda}, \tilde{v}_w) \qquad q(\alpha) \equiv \mathrm{N}(\tilde{\alpha}, 1/n)$$

Posterior distribution

- Approximate the posterior by a mean-field variational density

$$p(\mathbf{y}^*, \mathbf{w}, \alpha, \lambda | \mathbf{y}) \approx \prod_{i=1}^n q(y_i^*) q(\mathbf{w}) q(\alpha) q(\lambda)$$

where

$$q(y_i^*) \equiv \begin{cases} \mathbb{1}[y_i^* \geq 0] \, \mathrm{N}(\tilde{f}_i, 1) & \text{if } y_i = 1 \\ \mathbb{1}[y_i^* < 0] \, \mathrm{N}(\tilde{f}_i, 1) & \text{if } y_i = 0 \end{cases} \qquad q(\mathbf{w}) \equiv \mathrm{N}(\tilde{\mathbf{w}}, \tilde{\mathbf{V}}_w)$$

$$q(\lambda) \equiv \mathrm{N}(\tilde{\lambda}, \tilde{v}_w) \qquad q(\alpha) \equiv \mathrm{N}(\tilde{\alpha}, 1/n)$$

$$\tilde{f}_i = \tilde{\alpha} + \textstyle\sum_{k=1}^n h_{\tilde{\lambda}}(x_i, x_k) \tilde{w}_k \qquad \tilde{\alpha} = \frac{1}{n} \textstyle\sum_{k=1}^n \left(\mathrm{E}[y_i^*] - h_{\tilde{\lambda}}(x_i, x_k) \tilde{w}_k \right)$$

$$\tilde{\mathbf{w}} = \tilde{\mathbf{V}}_w \mathbf{H}_{\tilde{\lambda}} (\mathrm{E}[\mathbf{y}^*] - \tilde{\alpha} \mathbf{1}_n) \qquad \tilde{\mathbf{V}}_w^{-1} = \mathbf{H}_{\tilde{\lambda}}^2 + \mathbf{I}_n$$

$$\tilde{\lambda} = (\mathrm{E}[\mathbf{y}^*] - \tilde{\alpha} \mathbf{1}_n) \mathbf{H} \tilde{\mathbf{w}} / \tilde{v}_\lambda \qquad \tilde{v}_\lambda = \mathrm{tr}(\mathbf{H}^2 (\tilde{\mathbf{V}}_w + \tilde{\mathbf{w}} \tilde{\mathbf{w}}^\top))$$

Posterior predictive distribution

- Given new data points x_{new}, interested in

$$
\begin{aligned}
p(y_{\text{new}}|\mathbf{y}) &= \int p(y_{\text{new}}|y^*_{\text{new}}, \mathbf{y}) p(y^*_{\text{new}}|\mathbf{y}) \, \mathrm{d}y^*_{\text{new}} \\
&\approx \int p(y_{\text{new}}|y^*_{\text{new}}) q(y^*_{\text{new}}) \, \mathrm{d}y^*_{\text{new}} \\
&= \begin{cases} \Phi(\tilde{f}_{\text{new}}) & \text{if } y_{\text{new}} = 1 \\ 1 - \Phi(\tilde{f}_{\text{new}}) & \text{if } y_{\text{new}} = 0 \end{cases}
\end{aligned}
$$

  where $\tilde{f}_{\text{new}} = \tilde{\alpha} + \sum_{k=1}^n h_{\tilde{\lambda}}(x_{\text{new}}, x_k)\tilde{w}_k$.

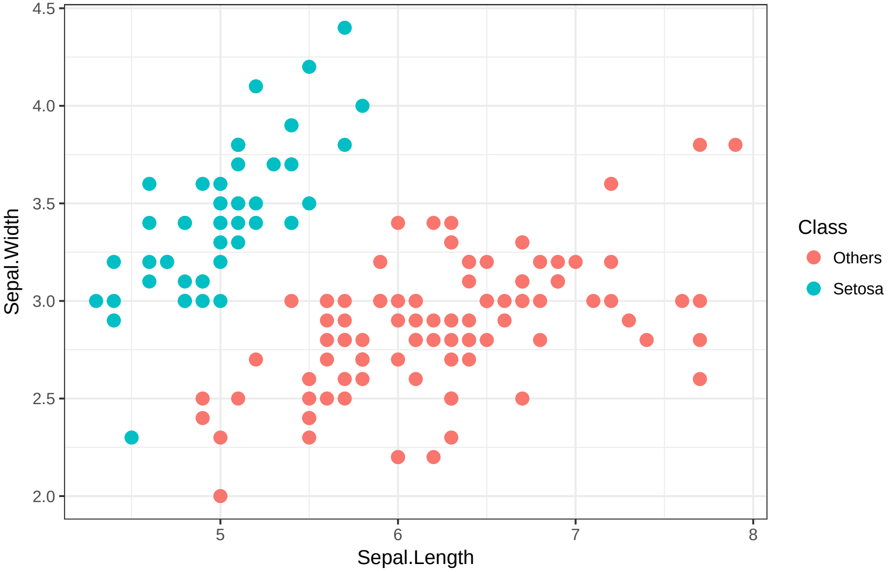
- f_{new} represents the estimate of the latent propensity for y_{new}, and its uncertainty is described by $q(y^*_{\text{new}})$.

Variational lower bound

- Since the solutions are coupled, we implement an iterative scheme (as per Algorithm 1)
- Assess convergence by monitoring the lower bound

$$
\begin{aligned}
\mathcal{L} &= \mathrm{E}_q[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \alpha, \lambda)] - \mathrm{E}_q[\log q(\mathbf{y}^*, \mathbf{w}, \alpha, \lambda)] \\
&= \text{const.} + \sum_{i=1}^{n} \Big(y_i \log \Phi(\tilde{f}_i) + (1 - y_i) \log \big(1 - \Phi(\tilde{f}_i)\big) \Big) \\
&\quad - \frac{1}{2} \Big(\operatorname{tr} \tilde{\mathbf{V}}_w + \operatorname{tr}(\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top) - \log |\tilde{\mathbf{V}}_w| + \log \tilde{v}_\lambda \Big)
\end{aligned}
$$

Fisher's Iris data set



Fisher's Iris data set - Model fitting

- Varitional inference for I-prior probit models implemented in R package iprobit (still lots of work to do!).

```
R> system.time(
+   (mod <- iprobit(y, X))
+ )


##
## |=================================                      |  61%
##  Converged after 6141 iterations.
##  Training error rate: 0 %
##     user   system elapsed
##   67.857   6.396  74.277
```

HJ (2017b). *iprobit: Binary Probit Regression with I-priors*. R Package version 0.1.0: GitHub

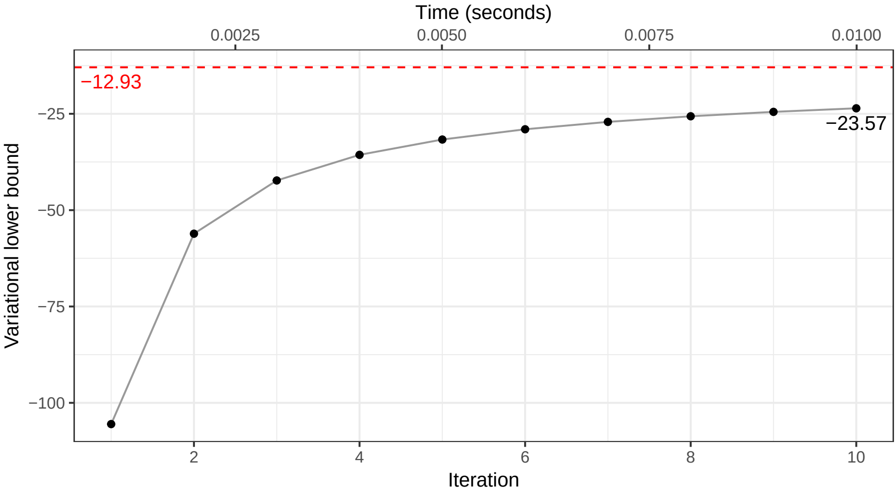
Fisher's Iris data set - Model summary

```
R> summary(mod)

##
## Call:
## iprobit(y = y, X, maxit = 10000)
##
## RKHS used: Canonical
##
##            Mean   S.E.    2.5%   97.5%
## alpha  -4.1730 0.0816 -4.3330 -4.0129
## lambda  1.2896 0.0142  1.2618  1.3175
##
## Converged to within 1e-05 tolerance. No. of iterations: 6141
## Model classification error rate (%): 0
## Variational lower bound: -12.93486
```

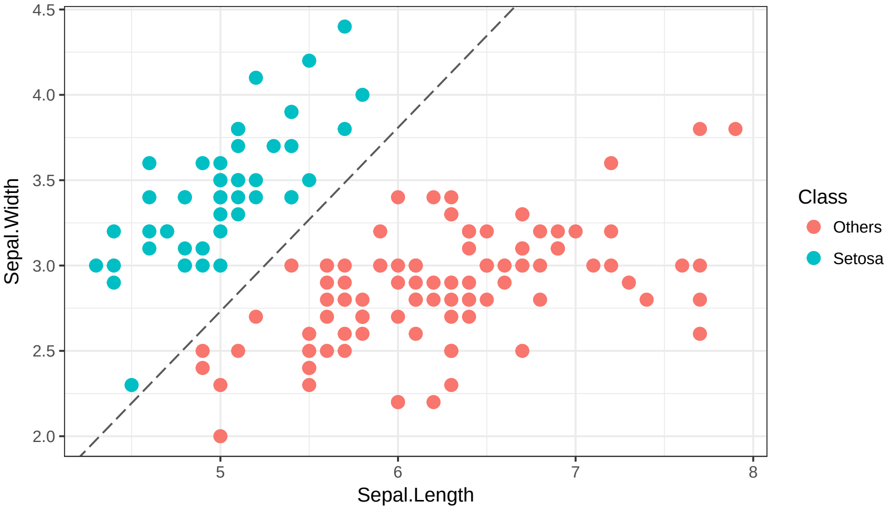
Fisher's Iris data set - Lower bound

```
R> iplot_lb(mod, niter.plot = 10)
```



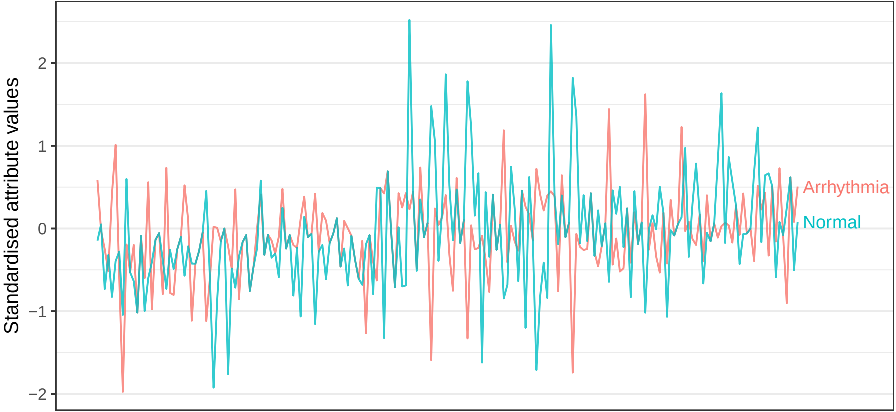
Fisher's Iris data set - Decision boundary

```
R> iplot_decbound(mod)
```



Cardiac arrhythmia data set

- Detect the presence of cardiac arrhythmia based on various ECG data and other attributes such as age and weight ($n = 451, p = 194$).



H. A. Guvenir et al. (1998). *UCI Machine Learning Repository: Arrhythmia Data Set*. URL: https://archive.ics.uci.edu/ml/datasets/Arrhythmia

Cardiac arrhythmia data set - Model fit

- Fit an I-prior probit model using Canonical and FBM kernels. The full data set takes about 35 seconds.

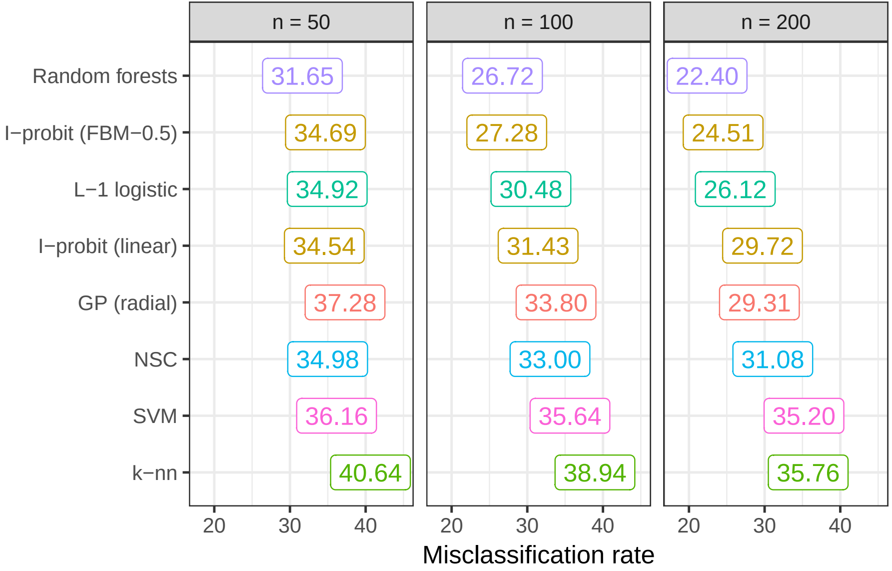
```
R> mod <- iprior(y, X, kernel = "FBM")
```

- Compare against popular classifiers: 1) *k*-nearest neighbours; 2) support vector machine; 3) Gaussian process classification; 4) random forests; 5) nearest shrunken centroids (Tibshirani et al. 2003); and 6) L-1 penalised logistic regression.

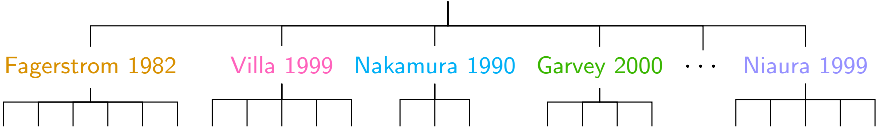
- Experiment set-up:
  - Form training set by sub-sampling $n_{\text{sub}} \in \{50, 100, 200\}$ data points.
  - Use remaining data as test set.
  - Fit model on training set and obtain test error rates.
  - Repeat 100 times.

T. I. Cannings and R. J. Samworth (2017). "Random-projection ensemble classification". *J. R. Stat. Soc. Ser. B: Stat. Methodol (w. discussion)*, to appear

Cardiac arrhythmia data set - Results



Meta-analysis of smoking cessation



- Data from 27 separate smoking cessation studies, where participants subjected to nicotine gum treatment or placed in control group.
- Some summary statistics:

	Min.	Avg.	Max.	Prop. quit	Odds quit
Control	20	101	617	0.207	0.261
Treated	21	117	600	0.320	0.470

- Raw odds ratio: 1.801.
- Random-effects analysis using a multilevel logistic model estimates this odds ratio as 1.768.

A. Skrondal and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC, §9.5

Meta-analysis of smoking cessation - model

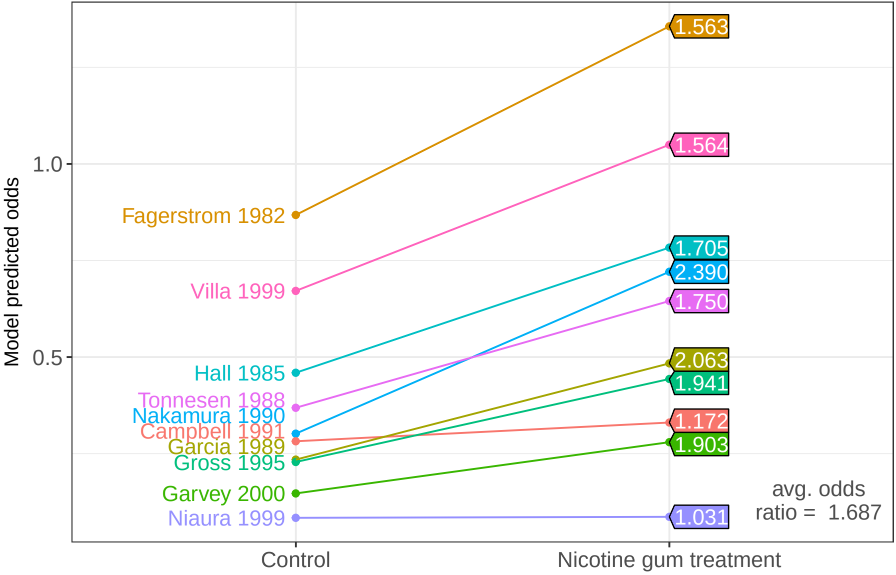
- Let $i = 1, \dots, n_j$ index the patients in study group $j \in 1, \dots, 27$.
- Denote y_{ij} as the binary response variable indicating `Quit` (1) or `Remain` (0), and x_{ij} as patient$_{ij}$'s treatment group indicator.
- Model binary data using I-probit model

$$\begin{aligned}\Phi^{-1}(p_{ij}) &= f(x_{ij}, j) \\ &= f_1(x_{ij}) + f_2(j) + f_{12}(x_{ij}, j)\end{aligned}$$

with $f_1, f_2 \in$ Pearson RKHS, and $f_{12} \in$ ANOVA RKHS.

	Model	Lower bound	Brier score	No. of RKHS param.
1	f_1	-3210.79	0.0311	1
2	$f_1 + f_2$	-3097.24	0.0294	2
3	$f_1 + f_2 + f_{12}$	-3091.21	0.0294	2

Meta-analysis of smoking cessation - results



Summary

- An extension of the I-prior methodology to binary responses.
- Variational inference used to approximate the intractable likelihood.
  - A deterministic approximation of the posterior density by a "close" (in the KL divergence sense), tractable density.
  - It's somewhere between Laplace's method and MCMC sampling.
- Several real-world examples demonstrated the use of I-probit models for classification and inference.
- Further work:
  - R package `iprobit`.
  - Extend to non-iid errors case.
  - Extend to multinomial probit models.
  - Other algorithms (e.g. expectation propagation).

Slides, source code and results are made available at: http://phd3.haziqj.ml

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.

Cannings, T. I. and R. J. Samworth (2017). "Random-projection ensemble classification". *Journal of the Royal Statistical Society. Series B: Statistical Methodology (with discussion)*, to appear.

Guvenir, H. A., M. Burak Acar, and H. Muderrisoglu (1998). *UCI Machine Learning Repository: Arrhythmia Data Set*. URL: https://archive.ics.uci.edu/ml/datasets/Arrhythmia.

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

References II

Jamil, H. (2017b). *iprobit: Binary Probit Regression with I-priors*. R Package version 0.1.0: GitHub.

Kass, R. and A. Raftery (1995). "Bayes Factors". *Journal of the American Statistical Association* 90.430, pp. 773–795.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

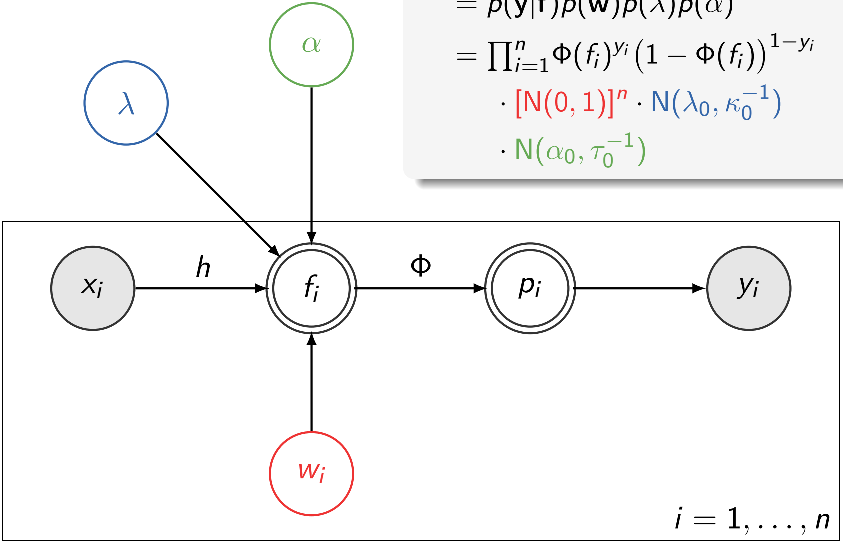
Skrondal, A. and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2003). "Class prediction by nearest shrunken centroids, with applications to DNA microarrays". *Statistical Science* 18.1, pp. 104–117.

6 Additional material

The I-prior probit model

$$
\begin{aligned}
&p(\mathbf{y}, \mathbf{w}, \alpha, \lambda) \\
&= p(\mathbf{y}|\mathbf{f})p(\mathbf{w})p(\lambda)p(\alpha) \\
&= \textstyle\prod_{i=1}^n \Phi(f_i)^{y_i}\big(1 - \Phi(f_i)\big)^{1-y_i} \\
&\qquad \cdot [\mathrm{N}(0,1)]^n \cdot \mathrm{N}(\lambda_0, \kappa_0^{-1}) \\
&\qquad \cdot \mathrm{N}(\alpha_0, \tau_0^{-1})
\end{aligned}
$$



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 $\tilde{\mathbf{f}}$.

R. Kass and A. Raftery (1995). "Bayes Factors". *Journal of the American Statistical Association* 90.430, pp. 773–795, §4.1, pp. 777-778.

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

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

R. Kass and A. Raftery (1995). "Bayes Factors". *Journal of the American Statistical Association* 90.430, pp. 773–795, §4.1, pp. 777-778.

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

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

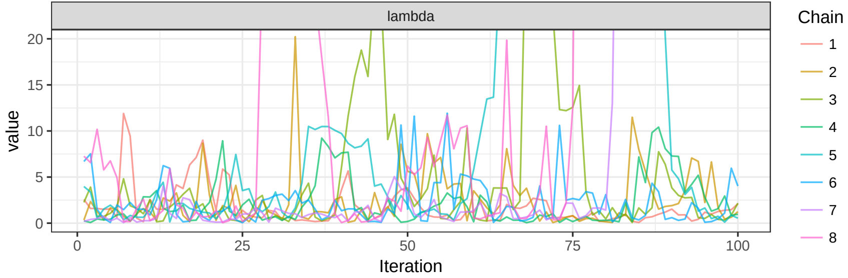
- Won't scale with large n; difficult to find modes in high dimensions.

R. Kass and A. Raftery (1995). "Bayes Factors". *Journal of the American Statistical Association* 90.430, pp. 773–795, §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.



Variational inference

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

Functions	$p : \theta \mapsto \mathbb{R}$	(standard calculus)
Functionals	$\mathcal{H} : p \mapsto \mathbb{R}$	(variational calculus)

Variational inference

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

Functions	$p : \theta \mapsto \mathbb{R}$	(standard calculus)
Functionals	$\mathcal{H} : p \mapsto \mathbb{R}$	(variational calculus)

- 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.

Variational inference

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

Functions	$p : \theta \mapsto \mathbb{R}$	(standard calculus)
Functionals	$\mathcal{H} : p \mapsto \mathbb{R}$	(variational calculus)

- 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.

Estimation of a 1-dim Gaussian mean and variance

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

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

back

Estimation of a 1-dim Gaussian mean and variance

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

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

$$\left.\begin{aligned} \mu|\psi &\sim \mathrm{N}\left(\mu_0, (\kappa_0\psi)^{-1}\right) \\ \psi &\sim \Gamma(a_0, b_0) \end{aligned}\right. \qquad \text{Priors}$$

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

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

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

back

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 (2), we can work out the solutions

back

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

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

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

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

- From (2), we can work out the solutions

back

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

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

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

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

- From (2), we can work out the solutions

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

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

back

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 $j \in \{1, \ldots, m\}$.

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

- From (2), we can work out the solutions

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

back

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

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

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

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

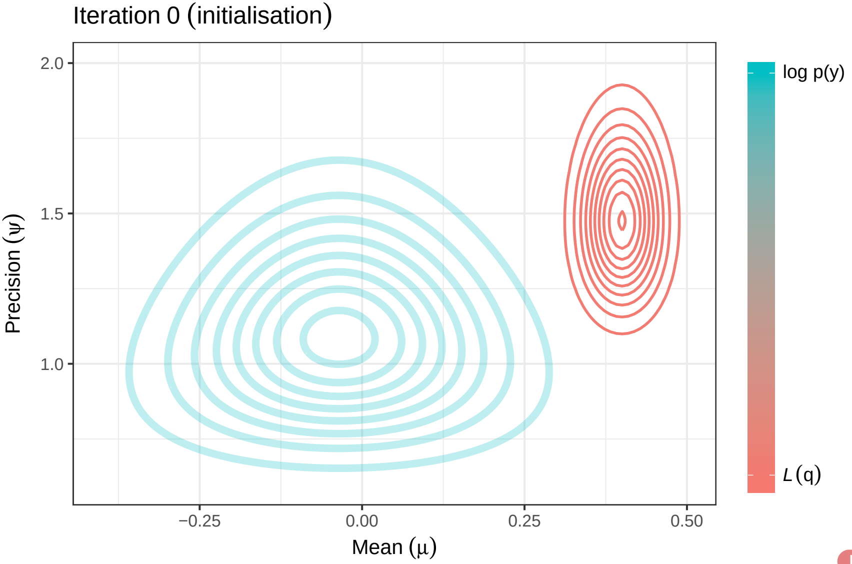
- From (2), we can work out the solutions

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

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

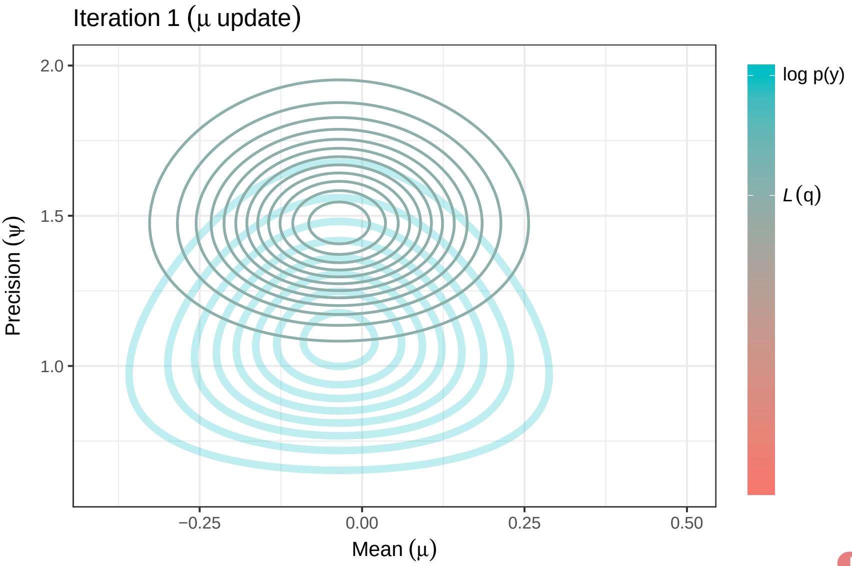
back

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



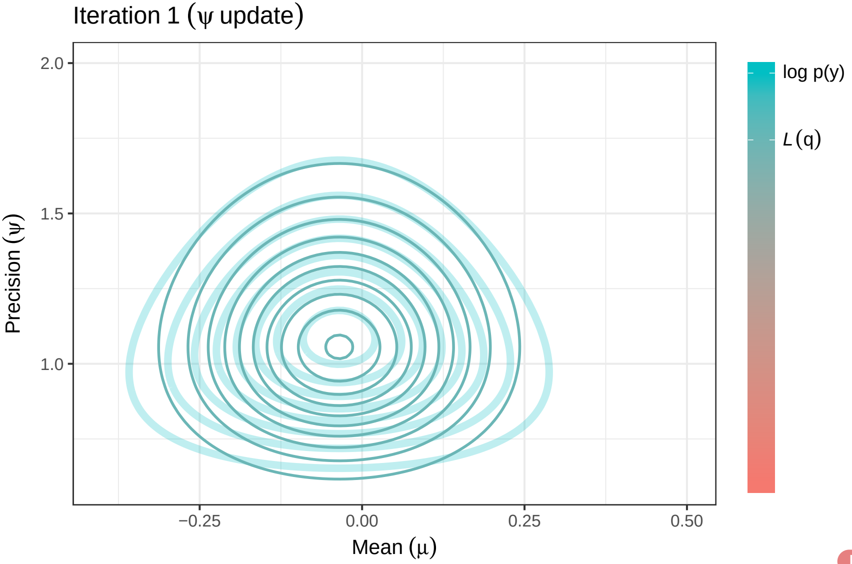
back

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



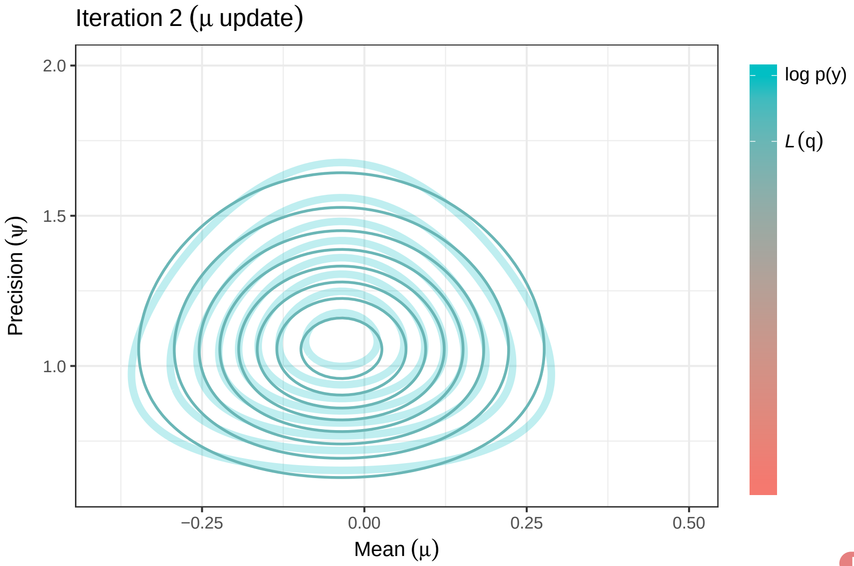
back

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



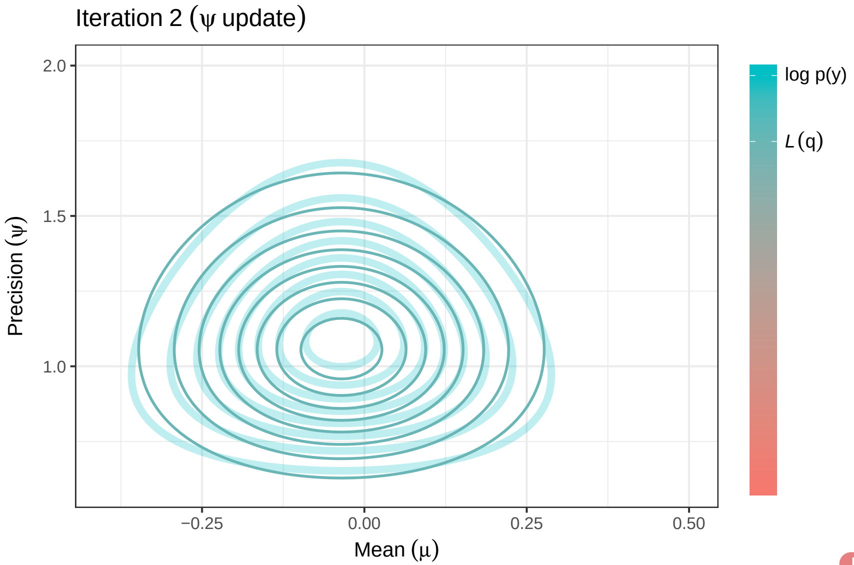
back

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



back

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



back