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

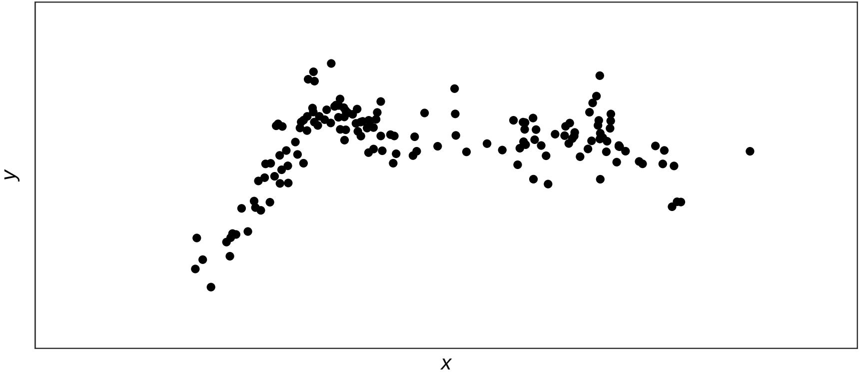
Outline

The regression model

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

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

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



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

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

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

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

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

I-priors (cont.)

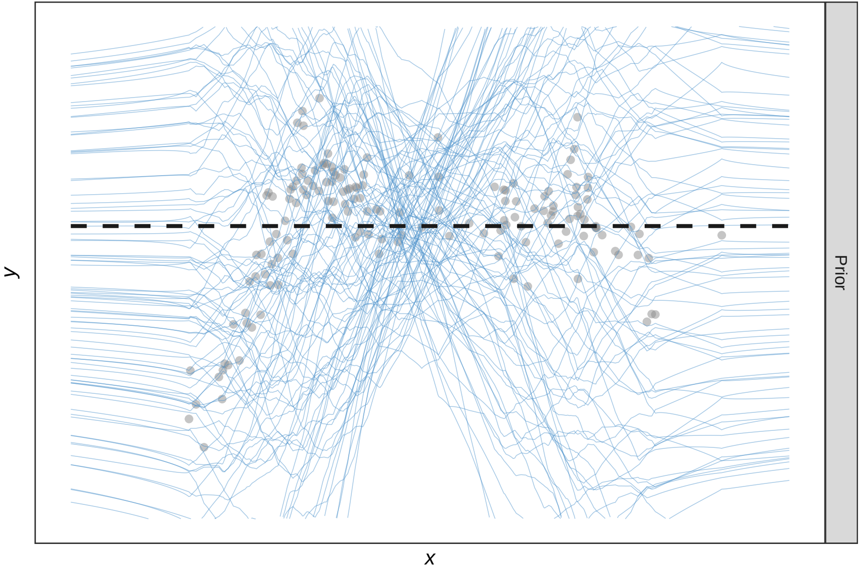
- Of interest is the posterior regression function characterised by the distribution
$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}},$$
and also the posterior predictive distribution for new data points x_{new}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

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

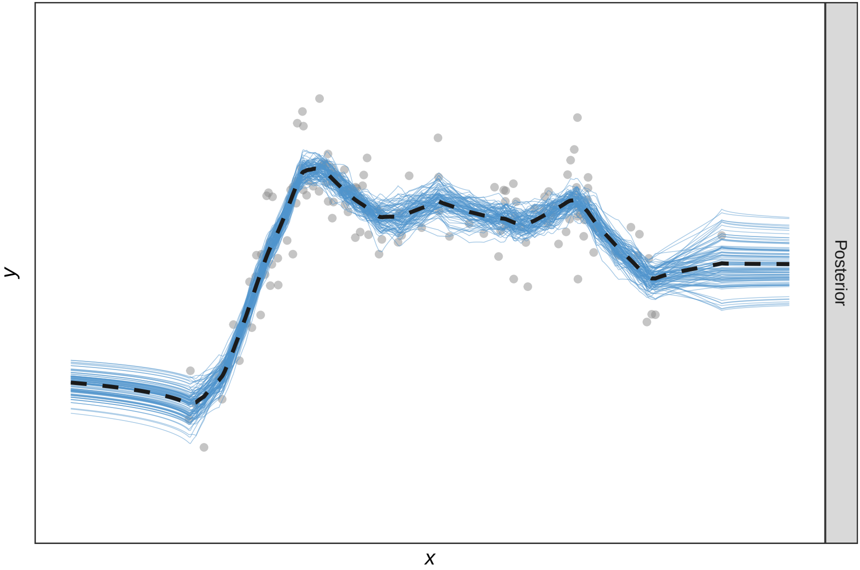
- Complete Bayesian estimation also possible.

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

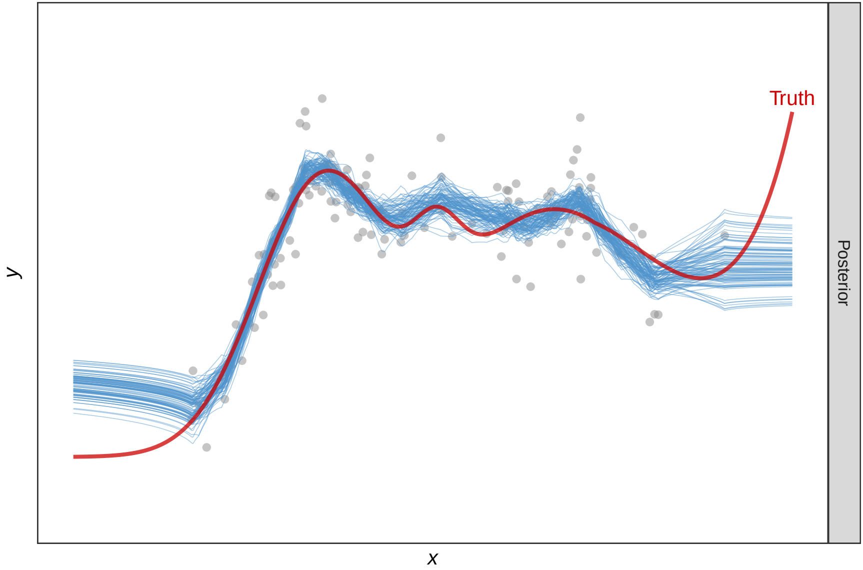
Fractional Brownian motion (FBM) RKHS



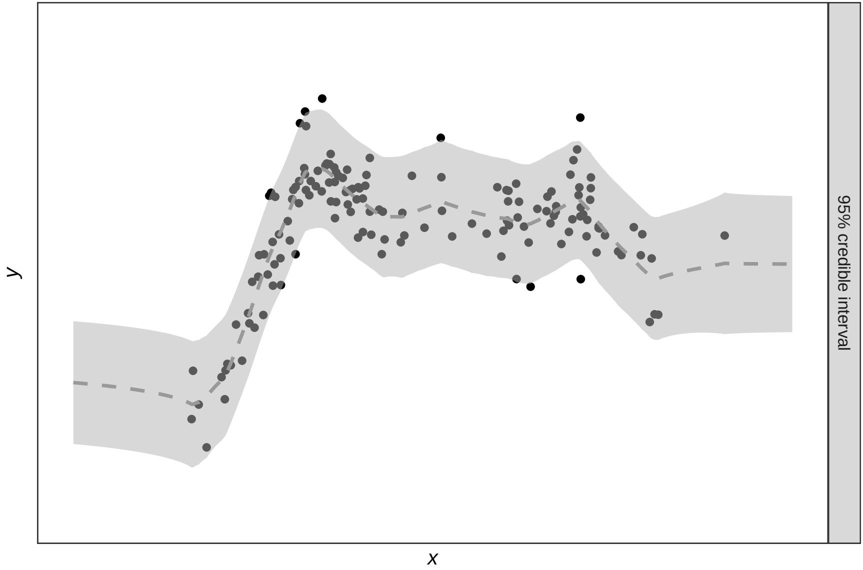
Fractional Brownian motion (FBM) RKHS



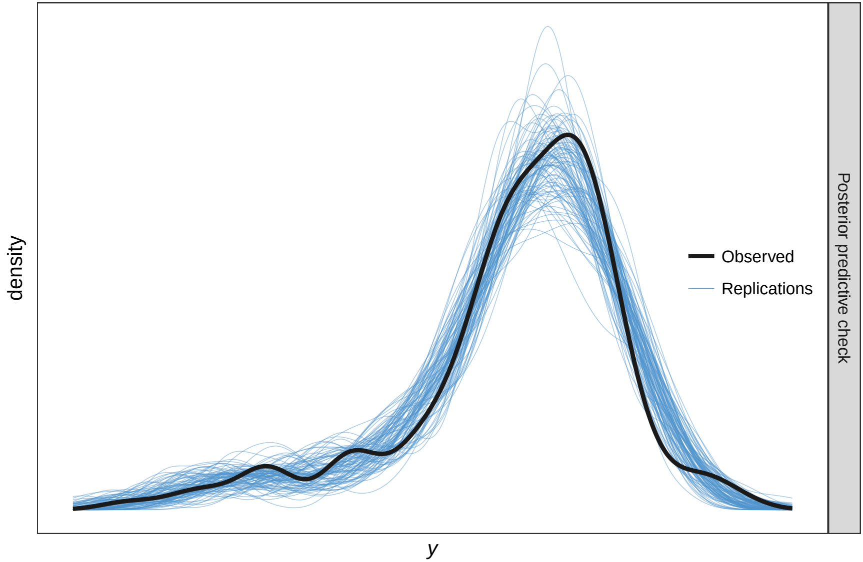
Fractional Brownian motion (FBM) RKHS



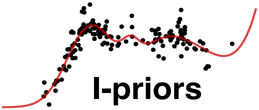
Posterior predictive distribution

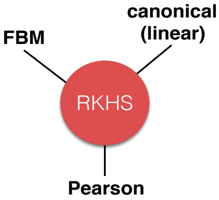


Posterior predictive distribution



PhD Roadmap





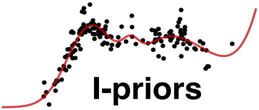
Unified methodology for

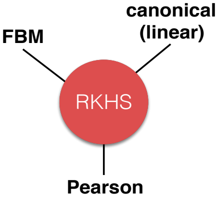
- additive models
- multilevel models
- models with functional covariates

Advantages

- Minimal assumptions
- Straightforward inference
- Performance competetive

PhD Roadmap





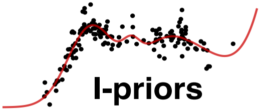
Unified methodology for
- additive models
- multilevel models
- models with functional covariates

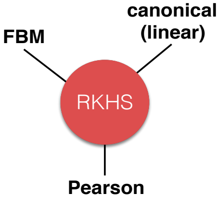
<u>Advantages</u>
- Minimal assumptions
- Straightforward inference
- Performance competetive

R/iprior

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

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)

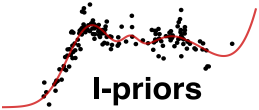
Bayesian Variable Selection

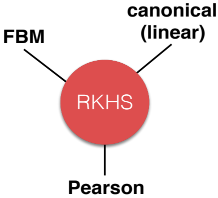
(using I-priors in the canonical RKHS)

✓	✓	✗	✗	✓
X_1	X_2	X_3	X_4	X_5

Good performance in cases with multicollinearity

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

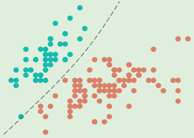
Good performance in cases with multicollinearity

R/iprior

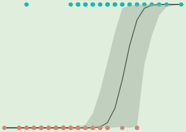
Estimation:
- Direct maximisation
- **<u>EM algorithm</u>**
- MCMC (Gibbs/HMC)

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,

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

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

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

Using I-priors

- Assume an I-prior on f. Then,

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

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

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

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

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

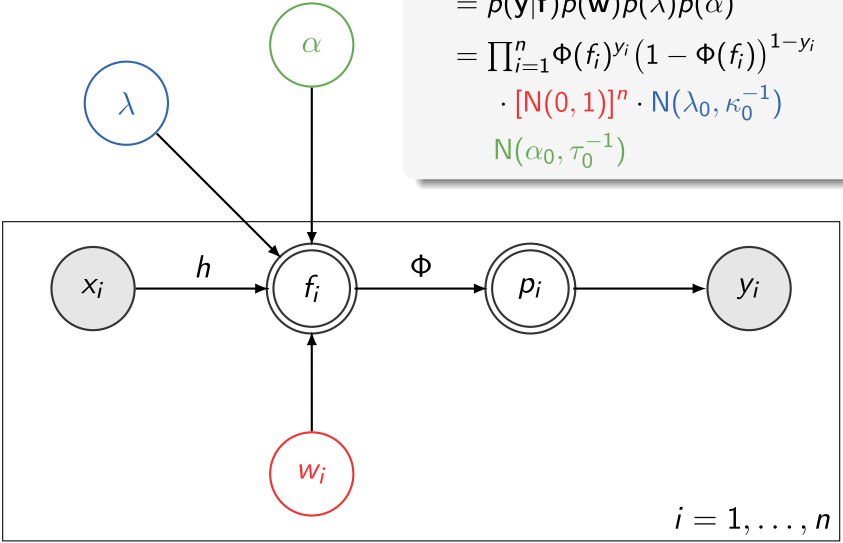
$$\begin{aligned} p_i = \mathrm{P}[y_i = 1] &= \mathrm{P}[y_i^* \geq 0] \\ &= \mathrm{P}[\epsilon_i \leq f(x_i)] \\ &= \Phi\Big(\psi^{1/2}(\alpha + \textstyle\sum_{k=1}^{n} h_\lambda(x_i, x_k) w_k)\Big) \end{aligned}$$

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 I-prior probit model

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



Estimation

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

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

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

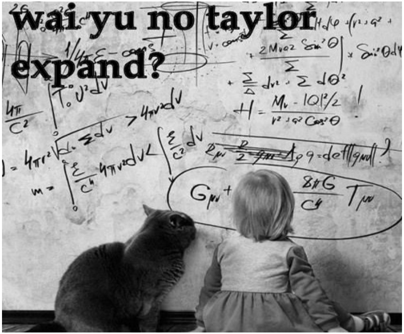
Estimation

- 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



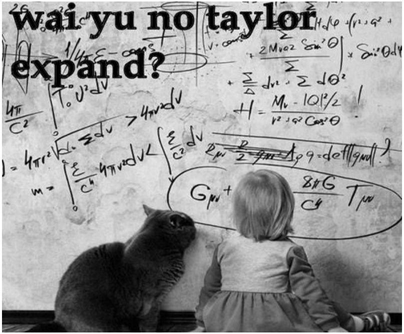
- Denote $f_i = f(x_i)$ for short.
- The marginal density

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

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
  - ✓ 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 $\tilde{\mathbf{f}}$.

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

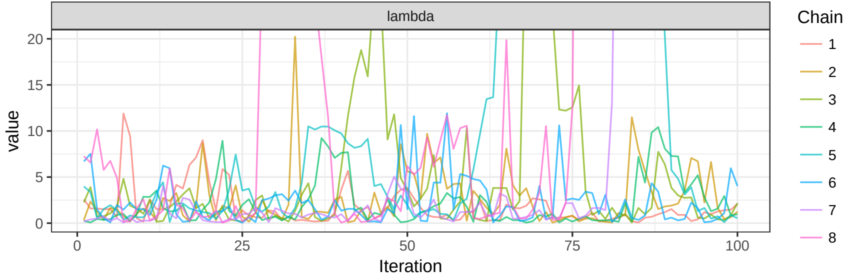
Full Bayesian analysis using MCMC

- 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.



End

Thank you!

References I

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