يونيبرسيتي بروني دارالسلام
UNIVERSITI BRUNEI DARUSSALAM

Regression modelling using I-priors

NUS Department of Statistics & Data Science Seminar

Haziq Jamil
Mathematical Sciences, Faculty of Science, UBD
https://haziqj.ml

Wednesday, 16 November 2022

Overview

Regression analysis

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

$$\begin{gathered} y_i = f(x_i) + \epsilon_i \\ (\epsilon_1, \ldots, \epsilon_n)^\top \sim \mathrm{N}_n(0, \Psi^{-1}) \end{gathered} \tag{1}$$

where each $y_i \in \mathbb{R}$, $x_i \in \mathcal{X}$ (some set of covariates), and f is a regression function.

Regression analysis

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

$$\begin{gathered} y_i = f(x_i) + \epsilon_i \\ (\epsilon_1, \ldots, \epsilon_n)^\top \sim \mathrm{N}_n(0, \Psi^{-1}) \end{gathered} \tag{1}$$

where each $y_i \in \mathbb{R}$, $x_i \in \mathcal{X}$ (some set of covariates), and f is a regression function. This forms the basis for a multitude of statistical models:

1. Ordinary linear regression when f is parameterised linearly.

Regression analysis

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

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

where each $y_i \in \mathbb{R}$, $x_i \in \mathcal{X}$ (some set of covariates), and f is a regression function. This forms the basis for a multitude of statistical models:

1. Ordinary linear regression when f is parameterised linearly.
2. Varying intercepts/slopes model when $\mathcal{X}$ is grouped.

Regression analysis

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

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

where each $y_i \in \mathbb{R}$, $x_i \in \mathcal{X}$ (some set of covariates), and f is a regression function. This forms the basis for a multitude of statistical models:

1. Ordinary linear regression when f is parameterised linearly.
2. Varying intercepts/slopes model when $\mathcal{X}$ is grouped.
3. Smoothing models when f is a smooth function.

Regression analysis

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

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

where each $y_i \in \mathbb{R}$, $x_i \in \mathcal{X}$ (some set of covariates), and f is a regression function. This forms the basis for a multitude of statistical models:

1. Ordinary linear regression when f is parameterised linearly.
2. Varying intercepts/slopes model when $\mathcal{X}$ is grouped.
3. Smoothing models when f is a smooth function.
4. Functional regression when $\mathcal{X}$ is functional.

Regression analysis

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

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

where each $y_i \in \mathbb{R}$, $x_i \in \mathcal{X}$ (some set of covariates), and f is a regression function. This forms the basis for a multitude of statistical models:

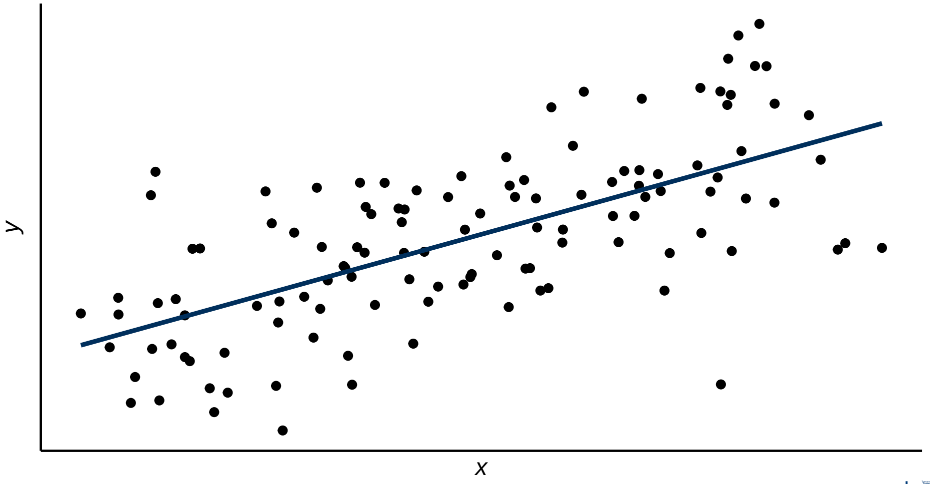
1. Ordinary linear regression when f is parameterised linearly.
2. Varying intercepts/slopes model when $\mathcal{X}$ is grouped.
3. Smoothing models when f is a smooth function.
4. Functional regression when $\mathcal{X}$ is functional.

Goal

To estimate the regression function f given the observations $\{(y_i, x_i)\}_{i=1}^n$.

1. Ordinary linear regression

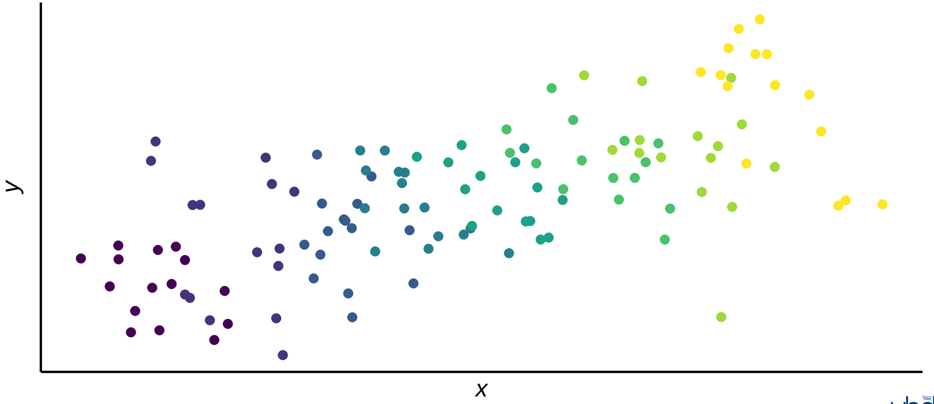
Suppose $f(x_i) = x_i^\top \beta$ for $i = 1, \dots, n$, where $x_i, \beta \in \mathbb{R}^p$.



2. Varying intercepts/slopes model

Suppose each unit $i = 1, \dots, n$ relates to the kth observation in group $j \in \{1, \dots, m\}$. Model the function f additively:

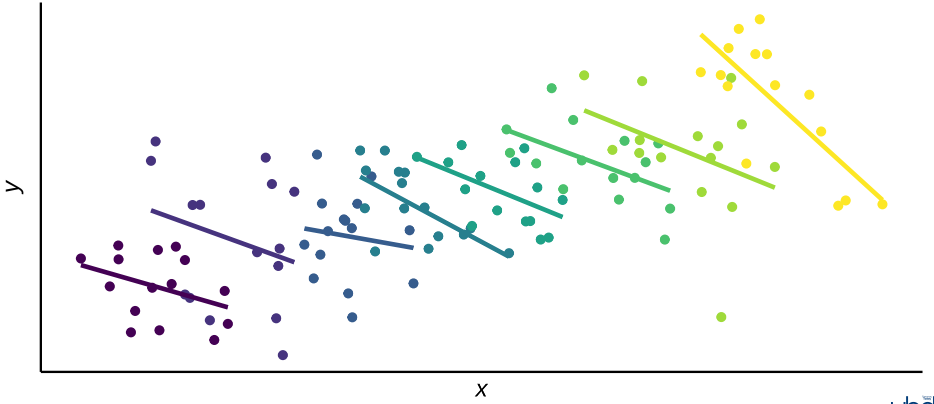
$$f(x_{kj}, j) = f_1(x_{kj}) + f_2(j) + f_{12}(x_{kj}, j).$$



2. Varying intercepts/slopes model

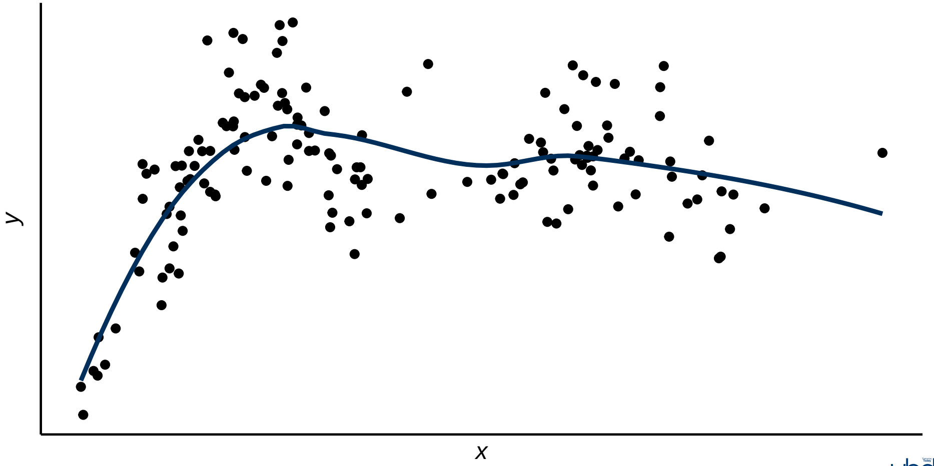
Suppose each unit $i = 1, \dots, n$ relates to the kth observation in group $j \in \{1, \dots, m\}$. Model the function f additively:

$$f(x_{kj}, j) = \underbrace{x_{kj}^\top \beta_1}_{f_1} + \underbrace{\beta_{0j}}_{f_2} + \underbrace{x_{kj}^\top \beta_{1j}}_{f_{12}}$$



3. Smoothing models

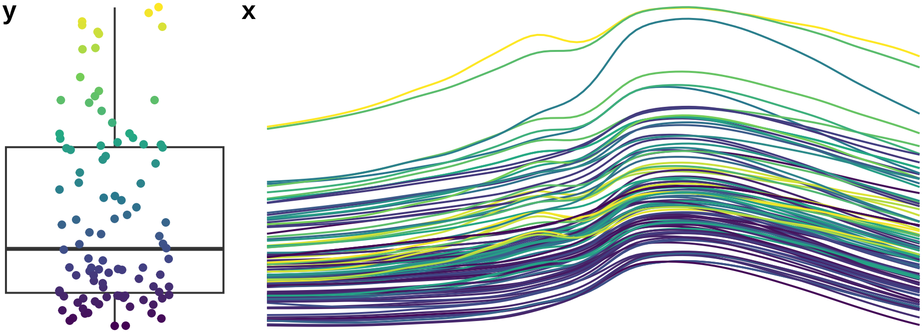
Suppose $f \in \mathcal{F}$ where $\mathcal{F}$ is a space of "smoothing functions" (models like LOESS, kernel regression, smoothing splines, etc.).



4. Functional regression

Suppose the input set $\mathcal{X}$ is functional. The (linear) regression aims to estimate a coefficient function $\beta : \mathcal{T} \to \mathbb{R}$

$$y_i = \underbrace{\int_{\mathcal{T}} x_i(t)\beta(t)\,\mathrm{d}t}_{f(x_i)} + \epsilon_i$$



The I-prior

For the normal model stated in (1), we assume that f lies in some RKHS of functions $\mathcal{F}$, with reproducing kernel h over $\mathcal{X}$.

The I-prior

For the normal model stated in (1), we assume that f lies in some RKHS of functions $\mathcal{F}$, with reproducing kernel h over $\mathcal{X}$.

Definition 1 (I-prior)

With $f_0 \in \mathcal{F}$ a prior guess, the entropy maximising prior distribution for f, subject to constraints, is

$$f(x) = f_0(x) + \sum_{i=1}^{n} h(x, x_i) w_i \qquad (2)$$

$$(w_1, \dots, w_n)^\top \sim \mathrm{N}_n(0, \Psi)$$

The I-prior

For the normal model stated in (1), we assume that f lies in some RKHS of functions $\mathcal{F}$, with reproducing kernel h over $\mathcal{X}$.

Definition 1 (I-prior)

With $f_0 \in \mathcal{F}$ a prior guess, the entropy maximising prior distribution for f, subject to constraints, is

$$f(x) = f_0(x) + \sum_{i=1}^n h(x, x_i) w_i \tag{2}$$

$$(w_1, \dots, w_n)^\top \sim \mathrm{N}_n(0, \Psi)$$

Therefore, the covariance kernel of $f(x)$ is determined by the function

$$k(x, x') = \sum_{i=1}^n \sum_{j=1}^n \Psi_{ij} h(x, x_i) h(x', x_j), \tag{3}$$

which happens to be the *Fisher information* between evaluations of f.

The I-prior (cont.)

Interpretation:

> The more information about f, the larger its prior variance, and hence the smaller the influence of the prior mean f_0 (and vice versa).

The I-prior (cont.)

Interpretation:

> The more information about f, the larger its prior variance, and hence the smaller the influence of the prior mean f_0 (and vice versa).

Of interest then are

1. Posterior distribution for the regression function,

$$p(f \mid y) = \frac{p(y \mid f)p(f)}{\int p(y \mid f)p(f)\,\mathrm{d}f}.$$

The I-prior (cont.)

Interpretation:

> The more information about f, the larger its prior variance, and hence the smaller the influence of the prior mean f_0 (and vice versa).

Of interest then are

1. Posterior distribution for the regression function,

$$p(f \mid y) = \frac{p(y \mid f)p(f)}{\int p(y \mid f)p(f)\,\mathrm{d}f}.$$

2. Posterior predictive distribution (given a new data point x_*)

$$p(y_* \mid y) = \int p(y_* \mid f_*)p(f_* \mid y)\,\mathrm{d}f_*,$$

where $f_* = f(x_*)$.

Posterior regression function

Denote by

- $\mathbf{y} = (y_1, \dots, y_n)^\top$
- $\mathbf{f} = \big(f(x_1), \dots, f(x_n)\big)^\top$
- $\mathbf{f}_0 = \big(f_0(x_1), \dots, f_0(x_n)\big)^\top$
- $\mathbf{H} = \big(h(x_i, x_j)\big)_{i,j=1}^n \in \mathbb{R}^{n \times n}$

(1) + an I-prior on f implies

$$\mathbf{y} \mid \mathbf{f} \sim \mathrm{N}_n(\mathbf{f}, \boldsymbol{\Psi}^{-1})$$
$$\mathbf{f} \sim \mathrm{N}_n(\mathbf{f}_0, \mathbf{H}\boldsymbol{\Psi}\mathbf{H})$$

Thus, $\mathbf{y} \sim \mathrm{N}_n(\mathbf{f}_0, \mathbf{V}_y := \mathbf{H}\boldsymbol{\Psi}\mathbf{H} + \boldsymbol{\Psi}^{-1})$.

Lemma 2

The posterior distribution for f is Gaussian with mean and covariance

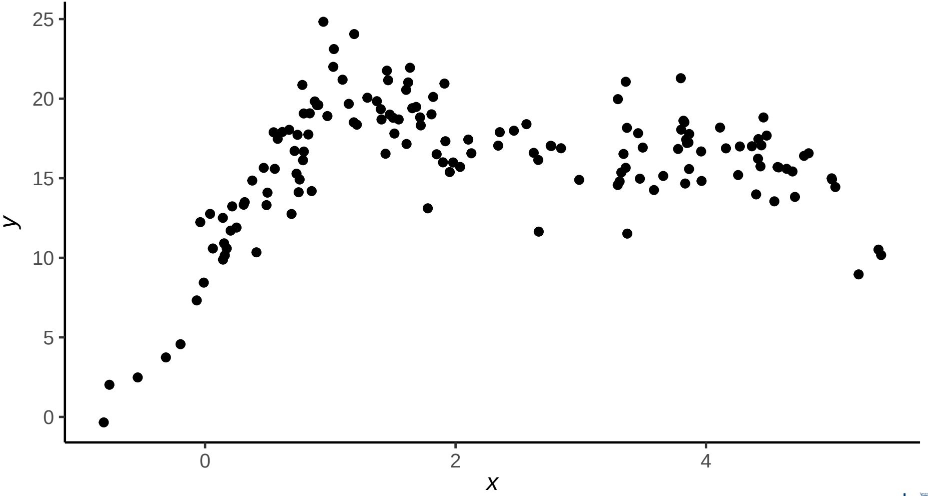
$$\mathrm{E}\big(f(x) \mid \mathbf{y}\big) = f_0(x) + \sum_{i=1}^n h(x, x_i)\hat{w}_i \tag{4}$$

$$\mathrm{Cov}\big(f(x), f(x') \mid \mathbf{y}\big) = \sum_{i=1}^n \sum_{j=1}^n \big(\mathbf{V}_y^{-1}\big)_{ij} h(x, x_i) h(x', x_j) \tag{5}$$

where $\hat{w}_1, \dots, \hat{w}_n$ are given by $\hat{\mathbf{w}} := \mathrm{E}\big(\mathbf{w} \mid \mathbf{y}\big) = \boldsymbol{\Psi}\mathbf{H}\mathbf{V}_y^{-1}(\mathbf{y} - \mathbf{f}_0)$.

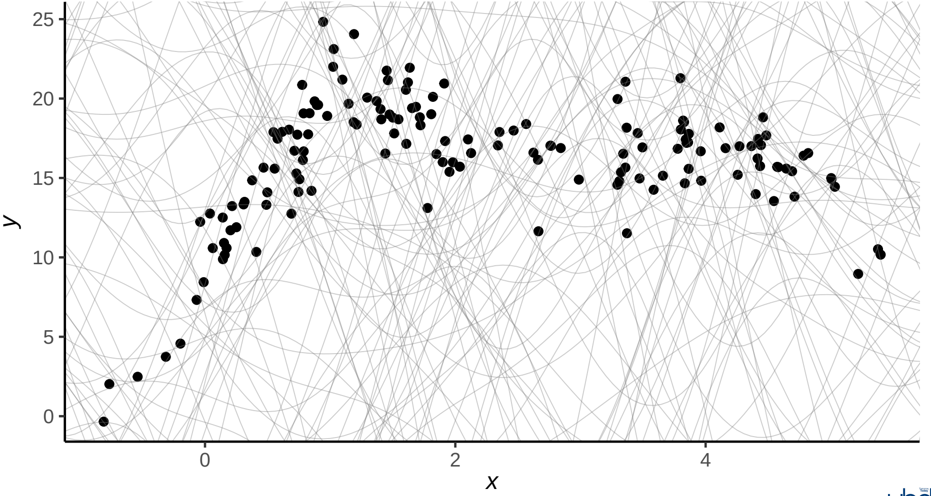
Illustration

Observations $\{(y_i, x_i) \mid y_i, x_i \in \mathbb{R} \; \forall i = 1, \ldots, n\}$.



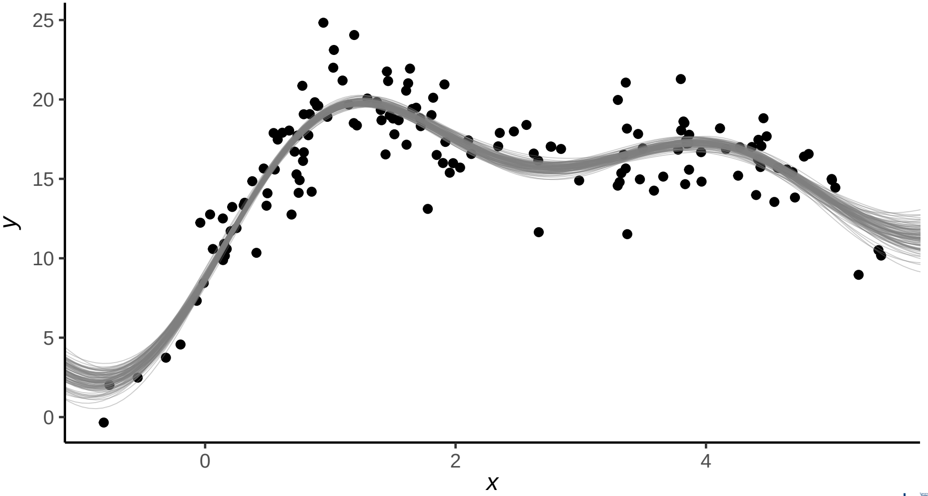
Illustration

Choose $h(x, x') = e^{-\frac{\|x-x'\|^2}{2}}$ (Gaussian kernel). Sample paths from I-prior:



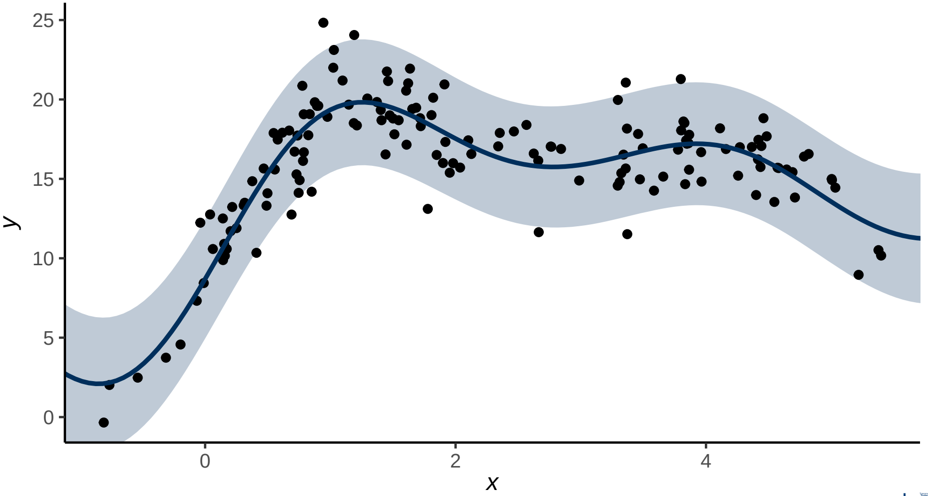
Illustration

Sample paths from the posterior of f:



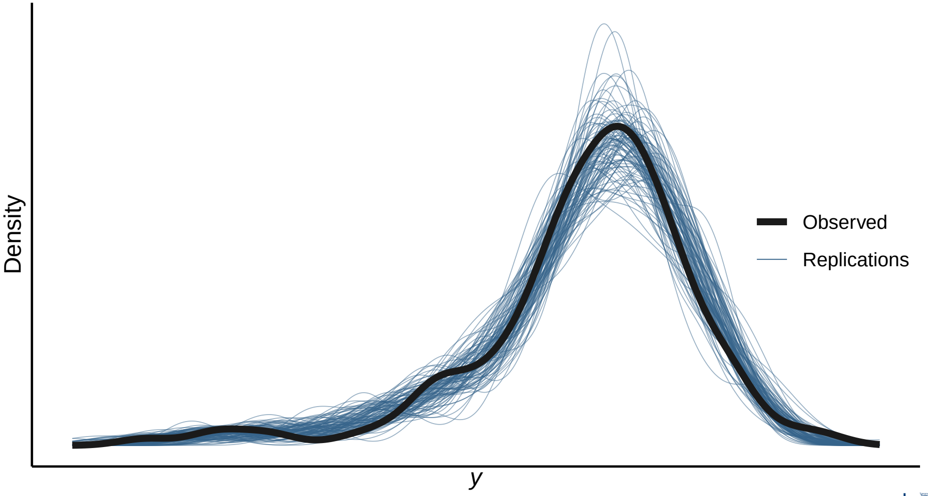
Illustration

Posterior mean estimate for $y = f(x)$ and its 95% credibility interval:



Illustration

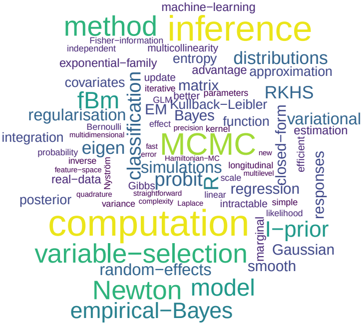
Other Bayesian stuff e.g. posterior predictive checks for $\{y_1, \ldots, y_n\}$:



Why I-priors?

Highlights

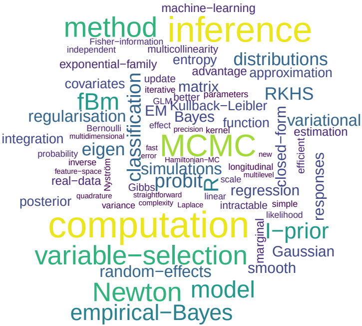
- An objective, data-driven prior. No user input required.
- The I-prior is proper; posterior estimates are thus *admissible*.
- Intuitive regression approach– model purpose is effected by kernel choices.



Why I-priors?

Highlights

- An objective, data-driven prior. No user input required.
- The I-prior is proper; posterior estimates are thus *admissible*.
- Intuitive regression approach–model purpose is effected by kernel choices.



Competitors:

- Tikhonov regulariser (e.g. cubic spline smoother)

$$\hat{f} = \arg\min_{f} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int f''(x)^2 \, \mathrm{d}x$$

- Gaussian process regression (Rasmussen & Williams, 2006)

State of the art

Professor Wicher Bergsma
London School of Economics and Political Science

1. Jamil, H. (2018). *Regression modelling using priors depending on Fisher information covariance kernels (I-priors)* [Doctoral dissertation, London School of Economics and Political Science].
2. Bergsma, W. (2019). Regression with I-priors. *Journal of Econometrics and Statistics.* https://doi.org/10.1016/j.ecosta.2019.10.002
3. Jamil, H., & Bergsma, W. (2019). `iprior`: An R Package for Regression Modelling using I-priors. *arXiv:1912.01376 [stat]*
4. Bergsma, W., & Jamil, H. (2020). Regression modelling with I-priors: With applications to functional, multilevel and longitudinal data. *arXiv:2007.15766 [math, stat]*
5. Jamil, H., & Bergsma, W. (2021). Bayesian Variable Selection for Linear Models Using I-priors. In S. A. Abdul Karim (Ed.), *Theoretical, modelling and numerical simulations toward industry 4.0* (pp. 107–132). Springer
6. Bergsma, W., & Jamil, H. (2022). Additive interaction modelling using I-priors. *Manuscript in prepration*

Reproducing kernel Hilbert spaces

Assumption: $f \in \mathcal{F}$ *where* $\mathcal{F}$ *is an RKHS with kernel* h *over* $\mathcal{X}$.

Definition 3 (Hilbert spaces)

A *Hilbert space* $\mathcal{F}$ is a vector space equipped with a positive definite inner product $\langle \cdot, \cdot \rangle_\mathcal{F} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$.

Definition 4 (Reproducing kernels)

A symmetric, bivariate function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *kernel*, and it is a *reproducing kernel* of $\mathcal{F}$ if h satisfies

i. $\forall x \in \mathcal{X}$, $h(\cdot, x) \in \mathcal{F}$;
ii. $\forall x \in \mathcal{X}$ and $\forall f \in \mathcal{F}$, $\langle f, h(\cdot, x) \rangle_\mathcal{F} = f(x)$.

In particular, $\forall x, x' \in \mathcal{F}$, $h(x, x') = \langle h(\cdot, x), h(\cdot, x') \rangle_\mathcal{F}$.

Reproducing kernel Hilbert spaces (cont.)

Theorem 5 (Moore-Aronszajn, etc.)

There is a bijection between

i. the set of positive definite functions; and
ii. the set of RKHSs.

Reproducing kernel Hilbert spaces (cont.)

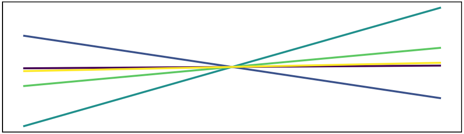
Theorem 5 (Moore-Aronszajn, etc.)

There is a bijection between

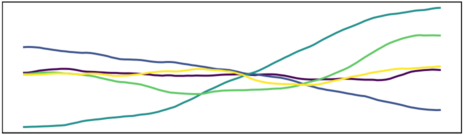
i. the set of positive definite functions; and
ii. the set of RKHSs.

$h(x, x') = 1$ (constant)

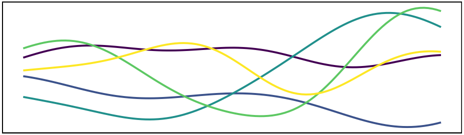
$h(x, x') = \langle x, x' \rangle_{\mathcal{X}}$ (linear)



$h(x, x') = -\frac{1}{2}\big(\|x - x'\|_{\mathcal{X}}^{2\gamma} - \|x\|_{\mathcal{X}}^{2\gamma} - \|x'\|_{\mathcal{X}}^{2\gamma}\big)$ (fBm)



$h(x, x') = \exp\left(-\frac{\|x - x'\|_{\mathcal{X}}^{2\gamma}}{2s^2}\right)$ (Gaussian)



Building more complex RKHSs

We can build complex RKHSs by adding and multiplying kernels:

- $\mathcal{F} = \mathcal{F}_1 \oplus \mathcal{F}_2$ is an RKHS defined by $h = h_1 + h_2$.
- $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2$ is an RKHS defined by $h = h_1 h_2$.

Building more complex RKHSs

We can build complex RKHSs by adding and multiplying kernels:

- $\mathcal{F} = \mathcal{F}_1 \oplus \mathcal{F}_2$ is an RKHS defined by $h = h_1 + h_2$.
- $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2$ is an RKHS defined by $h = h_1 h_2$.

Example 6 (ANOVA RKHS)

Consider RKHSs $\mathcal{F}_k$ with kernel h_k, $k = 1, \dots, p$. The ANOVA kernel over the set $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$ defining the ANOVA RKHS $\mathcal{F}$ is

$$h(x, x') = \prod_{k=1}^{p} \big(1 + h_k(x, x')\big).$$

For $p = 2$ let $\mathcal{F}_k$ be linear RKHS of functions over $\mathbb{R}$. Then $f \in \mathcal{F}$ where $\mathcal{F} = \mathcal{F}_\emptyset \oplus \mathcal{F}_1 \oplus \mathcal{F}_2 \oplus \mathcal{F}_1 \otimes \mathcal{F}_2$ are of the form

$$f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

The Fisher information

For the normal model (1), the log-likelihood of f is given by

$$\ell(f|y) = \text{const.} - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \psi_{ij}\big(y_i - \langle f, h(\cdot, x_i)\rangle_{\mathcal{F}}\big)\big(y_j - \langle f, h(\cdot, x_j)\rangle_{\mathcal{F}}\big)$$

The Fisher information

For the normal model (1), the log-likelihood of f is given by

$$\ell(f|y) = \text{const.} - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \psi_{ij}\big(y_i - \langle f, h(\cdot, x_i)\rangle_{\mathcal{F}}\big)\big(y_j - \langle f, h(\cdot, x_j)\rangle_{\mathcal{F}}\big)$$

Variational calculus leads us to the following result:

Lemma 7 (Fisher information for regression function)

The Fisher information for f is

$$\mathcal{I}_f = -\operatorname{E}\nabla^2\ell(f|y) = \sum_{i=1}^n\sum_{j=1}^n \psi_{ij} h(\cdot, x_i) \otimes h(\cdot, x_j)$$

where '$\otimes$' is the tensor product of two vectors in $\mathcal{F}$.

The Fisher information (cont.)

It's helpful to think of $\mathcal{I}_f$ as a bilinear form $\mathcal{I}_f : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$, making it possible to compute the Fisher information on linear functionals $f_g = \langle f, g \rangle_{\mathcal{F}}, \forall g \in \mathcal{F}$ as $\mathcal{I}_{f_g} = \langle \mathcal{I}_f, g \otimes g \rangle_{\mathcal{F} \otimes \mathcal{F}}$.

The Fisher information (cont.)

It's helpful to think of $\mathcal{I}_f$ as a bilinear form $\mathcal{I}_f : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$, making it possible to compute the Fisher information on linear functionals $f_g = \langle f, g \rangle_\mathcal{F}, \forall g \in \mathcal{F}$ as $\mathcal{I}_{f_g} = \langle \mathcal{I}_f, g \otimes g \rangle_{\mathcal{F} \otimes \mathcal{F}}$.

In particular, between two points $f_x := f(x)$ and $f_{x'} := f(x')$ we have:

$$
\begin{aligned}
\mathcal{I}_f(x, x') &= \big\langle \mathcal{I}_f, h(\cdot, x) \otimes h(\cdot, x') \big\rangle_{\mathcal{F} \otimes \mathcal{F}} \\
&= \left\langle \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} h(\cdot, x_i) \otimes h(\cdot,_j) \ , \ h(\cdot, x) \otimes h(\cdot, x') \right\rangle_{\mathcal{F} \otimes \mathcal{F}} \\
&= \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} \left\langle h(\cdot, x), h(\cdot, x_i) \right\rangle_\mathcal{F} \big\langle h(\cdot, x'), h(\cdot, x_j) \big\rangle_\mathcal{F}
\end{aligned}
$$

$$
\boxed{= \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} h(x, x_i) h(x', x_j) =: k(x, x')} \qquad \text{(from 3)}
$$

The I-prior

Lemma 8

The kernel (3) induces a finite-dimensional RKHS $\mathcal{F}_n < \mathcal{F}$, consisting of functions of the form $\tilde{f}(x) = \sum_{i=1}^n h(x, x_i) w_i$ (for some real-valued w_is) equipped with the squared norm

$$\|\tilde{f}\|^2_{\mathcal{F}_n} = \sum_{i,j=1}^n \psi^-_{ij} w_i w_j,$$

where ψ^-_{ij} is the (i,j)th entry of Ψ^{-1}.

- Let $\mathcal{R}$ be the orthogonal complement of $\mathcal{F}_n$ in $\mathcal{F}$. Then $\mathcal{F} = \mathcal{F}_n \oplus \mathcal{R}$, and any $f \in \mathcal{F}$ can be uniquely decomposed as $f = \tilde{f} + r$, with $\tilde{f} \in \mathcal{F}_n$ and $r \in \mathcal{R}$.
- The Fisher information for g is zero iff $g \in \mathcal{R}$. The data only allows us to estimate $f \in \mathcal{F}$ by considering functions in $\tilde{f} \in \mathcal{F}_n$.

The I-prior (cont.)

Theorem 9 (I-prior)

Let ν be a volume measure induced by the norm above, and let

$$\tilde{p} = \arg\max_p \left\{ - \int_{\mathcal{F}_n} p(f) \log p(f) \, \nu(\mathrm{d}f) \right\}$$

subject to the constraint

$$\mathrm{E}_{f \sim p} \| f - f_0 \|^2_{\mathcal{F}_n} = \text{constant}, \qquad f_0 \in \mathcal{F}.$$

Then $\tilde{p}$ is the Gaussian with mean f_0 and covariance function $k(x, x')$.

Equivalently, under the I-prior, f can be written in the form

$$f(x) = f_0(x) + \sum_{i=1}^n h(x, x_i) w_i, \qquad (w_1, \ldots, w_n)^\top \sim \mathrm{N}(0, \Psi)$$

Model hyperparameters

$$
\begin{gathered}
y_i = f_0(x_i) + \sum_{j=1}^n h_\lambda(x_i, x_j) w_j + \epsilon_i \\
(\epsilon_1, \dots, \epsilon_n)^\top \sim \mathrm{N}_n(0, \mathbf{\Psi}^{-1}) \\
(w_1, \dots, w_n)^\top \sim \mathrm{N}_n(0, \mathbf{\Psi})
\end{gathered}
\tag{6}
$$

A number of hyperparameters remain undetermined.

Model hyperparameters

$$
\begin{aligned}
y_i &= f_0(x_i) + \sum_{j=1}^{n} h_\lambda(x_i, x_j) w_j + \epsilon_i \\
(\epsilon_1, \dots, \epsilon_n)^\top &\sim \mathrm{N}_n(0, \mathbf{\Psi}^{-1}) \\
(w_1, \dots, w_n)^\top &\sim \mathrm{N}_n(0, \mathbf{\Psi})
\end{aligned}
\tag{6}
$$

A number of hyperparameters remain undetermined. Further assumptions:

1. The error variance $\mathbf{\Psi}$ is known up to a low-dimensional parameter, e.g. $\mathbf{\Psi} = \psi \mathbf{I}_n$, $\psi > 0$ (iid errors).

[1] This necessitates the use of reproducing kernel Krein spaces, as the kernels may no longer be positive definite.

Model hyperparameters

$$
\begin{aligned}
y_i &= f_0(x_i) + \sum_{j=1}^n h_\lambda(x_i, x_j) w_j + \epsilon_i \\
(\epsilon_1, \dots, \epsilon_n)^\top &\sim \mathrm{N}_n(0, \mathbf{\Psi}^{-1}) \\
(w_1, \dots, w_n)^\top &\sim \mathrm{N}_n(0, \mathbf{\Psi})
\end{aligned}
\tag{6}
$$

A number of hyperparameters remain undetermined. Further assumptions:

1. The error variance $\mathbf{\Psi}$ is known up to a low-dimensional parameter, e.g. $\mathbf{\Psi} = \psi \mathbf{I}_n$, $\psi > 0$ (iid errors).
2. Each RKHS $\mathcal{F}$ is defined by the kernel $h_\lambda = \lambda \tilde{h}$, where $\lambda \in \mathbb{R}$ is a scale[1] parameter.

[1] This necessitates the use of reproducing kernel Krein spaces, as the kernels may no longer be positive definite.

Model hyperparameters

$$
\begin{aligned}
y_i &= f_0(x_i) + \sum_{j=1}^n h_\lambda(x_i, x_j) w_j + \epsilon_i \\
(\epsilon_1, \dots, \epsilon_n)^\top &\sim \mathrm{N}_n(0, \mathbf{\Psi}^{-1}) \\
(w_1, \dots, w_n)^\top &\sim \mathrm{N}_n(0, \mathbf{\Psi})
\end{aligned} \tag{6}
$$

A number of hyperparameters remain undetermined. Further assumptions:

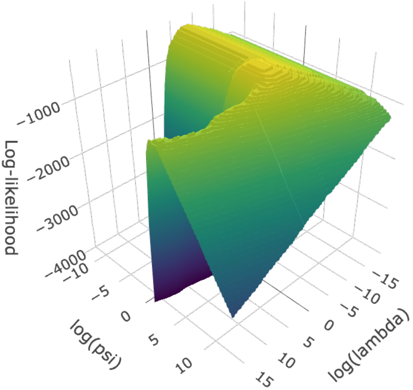
1. The error variance $\mathbf{\Psi}$ is known up to a low-dimensional parameter, e.g. $\mathbf{\Psi} = \psi \mathbf{I}_n$, $\psi > 0$ (iid errors).
2. Each RKHS $\mathcal{F}$ is defined by the kernel $h_\lambda = \lambda \tilde{h}$, where $\lambda \in \mathbb{R}$ is a scale[1] parameter.
3. Certain kernels also require tuning, e.g. the Hurst coefficient of the fBm or the lengthscale of the Gaussian. For now, assume fixed.

[1] This necessitates the use of reproducing kernel Krein spaces, as the kernels may no longer be positive definite.

Direct optimisation of (marginal) log-likelihood

The marginal log-likelihood of $(\lambda, \boldsymbol{\Psi})$ is

$$\ell(\lambda, \boldsymbol{\Psi} \mid \mathbf{y}) = \text{const.} - \frac{1}{2}\log|\mathbf{V}_y| - \frac{1}{2}(\mathbf{y} - \mathbf{f}_0)^\top \mathbf{V}_y^{-1}(\mathbf{y} - \mathbf{f}_0),$$



- Direct optimisation using e.g. conjugate gradients or Newton methods.
- Numerical stability issues–workaround: Cholesky or eigen decomposition.
- Prone to local optima.
- Possible to also optimise kernel hyperparameters.

EM algorithm

An alternative view of the model:

$$\begin{aligned} \mathbf{y} \mid \mathbf{w} &\sim \mathrm{N}_n(\mathbf{f}_0 + \mathbf{H}_\lambda \mathbf{w}, \boldsymbol{\Psi}^{-1}) \\ \mathbf{w} &\sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi}) \end{aligned}$$

in which the **w** are "missing".

EM algorithm

An alternative view of the model:

$$\begin{aligned} \mathbf{y} \mid \mathbf{w} &\sim \mathrm{N}_n(\mathbf{f}_0 + \mathbf{H}_\lambda \mathbf{w}, \boldsymbol{\Psi}^{-1}) \\ \mathbf{w} &\sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi}) \end{aligned}$$

in which the **w** are "missing". The full data log-likelihood is

$$\begin{aligned} L(\lambda, \boldsymbol{\Psi} \mid \mathbf{y}, \mathbf{w}) = \text{const.} - \frac{1}{2}(\mathbf{y} - \mathbf{f}_0)^\top \boldsymbol{\Psi}(\mathbf{y} - \mathbf{f}_0) - \frac{1}{2} \operatorname{tr}\left(\mathbf{V}_y \mathbf{w}\mathbf{w}^\top\right) \\ + (\mathbf{y} - \mathbf{f}_0)^\top \boldsymbol{\Psi} \mathbf{H}_\lambda \mathbf{w} \end{aligned}$$

EM algorithm

An alternative view of the model:

$$\begin{aligned}\mathbf{y} \mid \mathbf{w} &\sim \mathrm{N}_n(\mathbf{f}_0 + \mathbf{H}_\lambda \mathbf{w}, \boldsymbol{\Psi}^{-1}) \\ \mathbf{w} &\sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi})\end{aligned}$$

in which the $\mathbf{w}$ are "missing". The full data log-likelihood is

$$\begin{aligned}L(\lambda, \boldsymbol{\Psi} \mid \mathbf{y}, \mathbf{w}) = \text{const.} - \frac{1}{2}(\mathbf{y} - \mathbf{f}_0)^\top \boldsymbol{\Psi}(\mathbf{y} - \mathbf{f}_0) - \frac{1}{2}\operatorname{tr}\left(\mathbf{V}_y \mathbf{w}\mathbf{w}^\top\right) \\ + (\mathbf{y} - \mathbf{f}_0)^\top \boldsymbol{\Psi} \mathbf{H}_\lambda \mathbf{w}\end{aligned}$$

The E-step entails computing

$$Q_t(\lambda, \boldsymbol{\Psi}) = \mathrm{E}\left\{ L(\lambda, \boldsymbol{\Psi} \mid \mathbf{y}, \mathbf{w}) \,\middle|\, \mathbf{y}, \lambda^{(t)}, \boldsymbol{\Psi}^{(t)} \right\}$$

in which the following posterior quantities are needed

$$\hat{\mathbf{w}} := \mathrm{E}(\mathbf{w} \mid \mathbf{y}, \lambda, \boldsymbol{\Psi}) \quad \text{and} \quad \hat{\mathbf{W}} := \mathrm{E}(\mathbf{w}\mathbf{w}^\top \mid \mathbf{y}, \lambda, \boldsymbol{\Psi}) = \mathbf{V}_y^{-1} + \hat{\mathbf{w}}\hat{\mathbf{w}}^\top.$$

EM algorithm (cont.)

Let $\tilde{\mathbf{w}}^{(t)}$and $\tilde{\mathbf{W}}^{(t)}$ be versions of $\hat{\mathbf{w}}$ and $\hat{\mathbf{W}}$ computed using $\lambda^{(t)}$ and $\boldsymbol{\Psi}^{(t)}$. The M-step entails solving

$$\begin{aligned}
\frac{\partial Q_t}{\partial \lambda} &= -\frac{1}{2}\operatorname{tr}\left(\frac{\partial \mathbf{V}_y}{\partial \lambda}\tilde{\mathbf{W}}^{(t)}\right) + (\mathbf{y}-\mathbf{f}_0)^\top \boldsymbol{\Psi}\frac{\partial \mathbf{H}_\lambda}{\partial \lambda}\tilde{\mathbf{w}}^{(t)} &= 0 \\
\frac{\partial Q_t}{\partial \psi} &= -\frac{1}{2}\operatorname{tr}\left(\frac{\partial \mathbf{V}_y}{\partial \psi}\tilde{\mathbf{W}}^{(t)}\right) - \frac{1}{2}(\mathbf{y}-\mathbf{f}_0)^\top \left(\mathbf{y}-\mathbf{f}_0 - 2\mathbf{H}_\lambda\tilde{\mathbf{w}}^{(t)}\right) &= 0
\end{aligned}$$

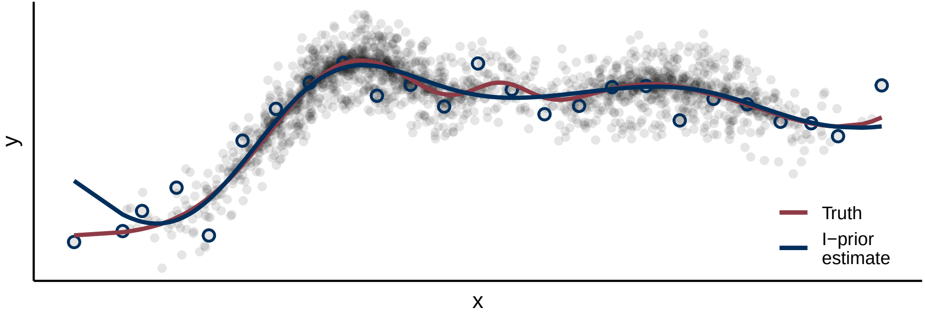
- This scheme admits a closed-form solution for ψ and (sometimes) for λ too (e.g. linear addition of kernels $h_\lambda = \lambda_1 h_1 + \cdots + \lambda_p h_p$).
- Sequential updating $\lambda^{(t)} \to \boldsymbol{\Psi}^{(t+1)} \to \lambda^{(t+1)} \to \cdots$ (expectation conditional maximisation, Meng and Rubin, 1993).
- Computationally unattractive for optimising kernel hyperparameters.

Computational bottleneck

In either estimation method, V_y^{-1} is computed and takes $O(n^3)$ time.

Computational bottleneck

In either estimation method, V_y^{-1} is computed and takes $O(n^3)$ time.



Trick: low-rank matrix approximations. Suppose $H \approx QQ^\top$, where $Q \in \mathbb{R}^{n \times m}$, $m \ll n$. Then, using the Woodbury matrix identity,

$$V_y^{-1} = (H\Psi H + \Psi^{-1})^{-1} \approx \Psi - \Psi Q\big((Q^\top \Psi Q)^{-1} + Q^\top \Psi Q\big)^{-1} Q^\top \Psi$$

is a much cheaper $O(nm^2)$ operation (Williams & Seeger, 2001).

Longitudinal analysis of cow growth data

Aim: Discern whether there is a difference between two treatments given to cows, and whether this effect varies among individual cows.

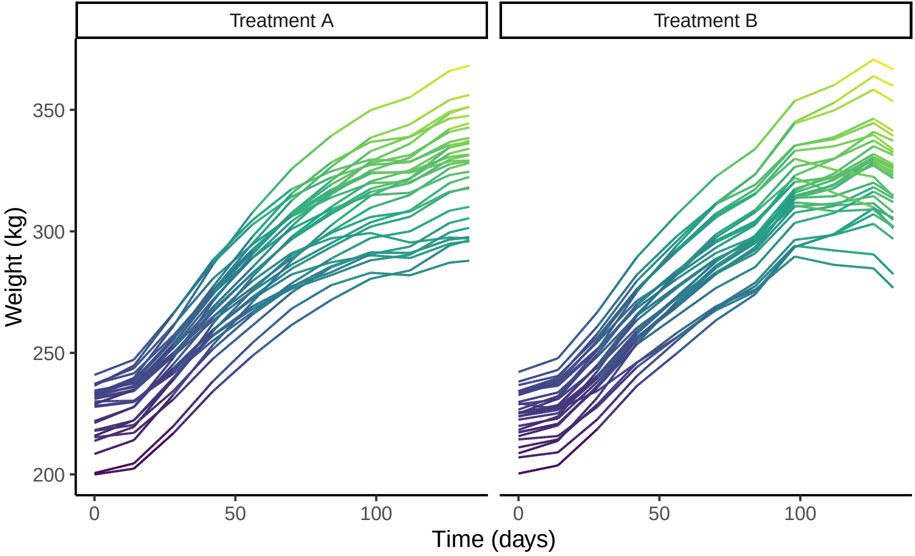
Data consists of a balanced longitudinal set of weights y_{it} for 60 cows. The herd were randomly split between two treatment groups (x_i). Model

$$y_{it} = f_{1t}(i) + f_{2t}(x_i) + f_{12t}(i, x_i) + \epsilon_{it}$$

assuming smooth effect of time, and nominal effect of cow index and treatment group.

	Explanation	Model	Log-lik.	No. of param.
1	Growth due to cows only	f_{1t}	−2792.2	3
2	Growth due to treatment only	f_{2t}	−2295.2	3
3	Growth due to both	$f_{1t} + f_{2t}$	−2270.9	4
4	Growth due to both with cow-treatment variation	$f_{1t} + f_{2t} + f_{12t}$	−2250.9	4

Growth curve



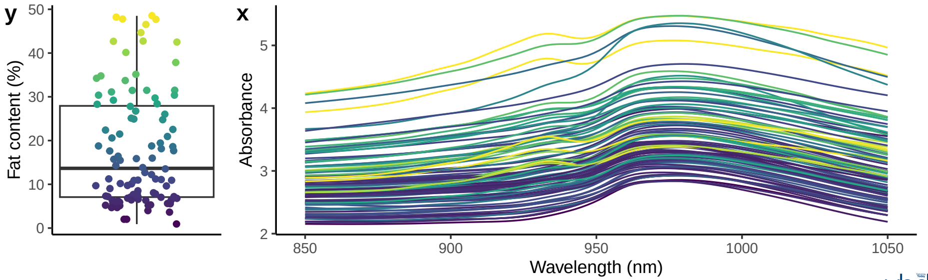
Predicting fat content in meat samples

Aim: Predict fat content of meat samples from its spectrometric curves (Tecator data set).

For each meat sample i, data consist of 100 channel spectrum of absorbances ($x_i(t)$) and its corresponding fat content (y_i). Train/test split is $160 + 55$. Model

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

where x_i is the ith spectral curve.



Results

Model	RMSE Train	RMSE Test
I-prior		
Linear	2.89	2.89
Quadratic	0.72	0.97
Smooth (fBm-0.70)	0.19	0.63
Others		
Linear functional regression		2.78
Quadratic functional regression		0.80
Gaussian process regression		2.93
Neural networks		0.36
Kernel smoothing		1.49
Multivariate adaptive regression splines (MARS)		0.88
Functional additive regression (CSEFAM)		0.85

Summary

A novel methodology for fitting a wide range of parameteric and nonparametric regression models.

- Parsimonious model specification and simple estimation.
- Inference is straightforward.
- Often yield comparable predictions to competing ML algorithms.

Further work

- Extension to non-Gaussian errors (e.g. classification or count data).
- $O(n^3)$ computational bottleneck.

End

Thank you!

References

Bergsma, W. (2019). Regression with I-priors. *Journal of Econometrics and Statistics*. https://doi.org/10.1016/j.ecosta.2019.10.002

Bergsma, W., & Jamil, H. (2020). Regression modelling with I-priors: With applications to functional, multilevel and longitudinal data. *arXiv:2007.15766 [math, stat]*.

Bergsma, W., & Jamil, H. (2022). Additive interaction modelling using I-priors. *Manuscript in prepration*.

Jamil, H. (2018). *Regression modelling using priors depending on Fisher information covariance kernels (I-priors)* [Doctoral dissertation, London School of Economics and Political Science].

Jamil, H., & Bergsma, W. (2019). `iprior`: An R Package for Regression Modelling using I-priors. *arXiv:1912.01376 [stat]*.

References

Jamil, H., & Bergsma, W. (2021). Bayesian Variable Selection for Linear Models Using I-priors. In S. A. Abdul Karim (Ed.), *Theoretical, modelling and numerical simulations toward industry 4.0* (pp. 107–132). Springer.

Meng, X.-L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, *80*(2), 267–278. https://doi.org/10.1093/biomet/80.2.267

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. The MIT Press. http://www.gaussianprocess.org/gpml/

Williams, C. K. I., & Seeger, M. (2001). Using the nyström method to speed up kernel machines. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems 13 (nips 2000)* (pp. 682–688).