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

Introduction

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:

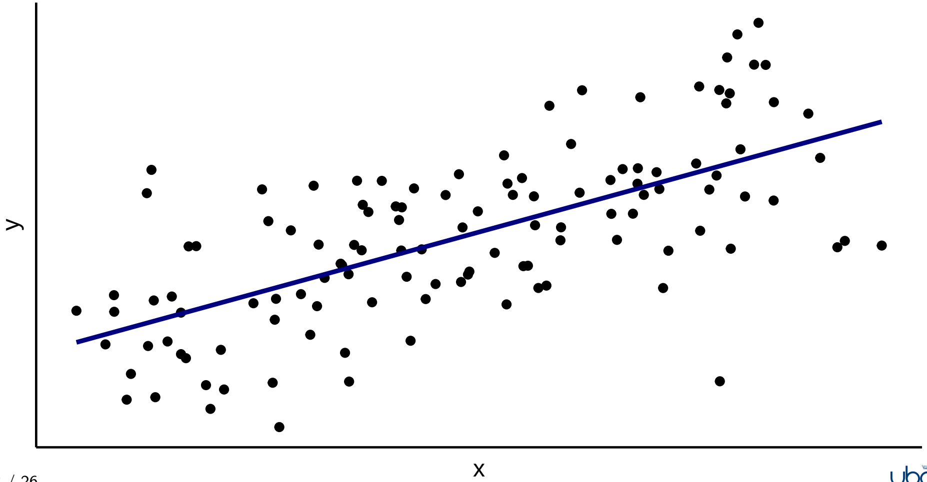
1. Ordinary linear regression when f is parameterised linearly.
2. Varying intercepts/slopes model when $\mathcal{X}$ is grouped.
3. Smoothing models when f is a smooth function.
4. Functional regression when $\mathcal{X}$ is functional.

Goal

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

Ordinary linear regression

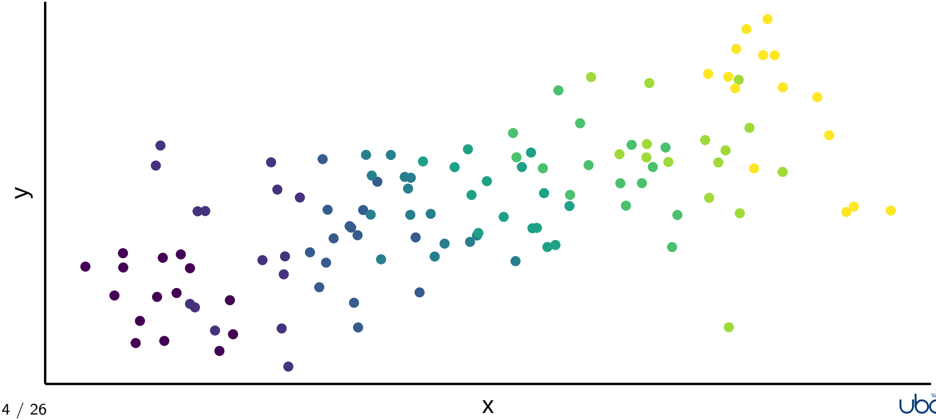
Suppose $f(x_i) = x_i^\top \beta$ for $i = 1, \dots, n$, where $x_i, \beta \in \mathbb{R}^p$.



Varying intercepts/slopes model

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

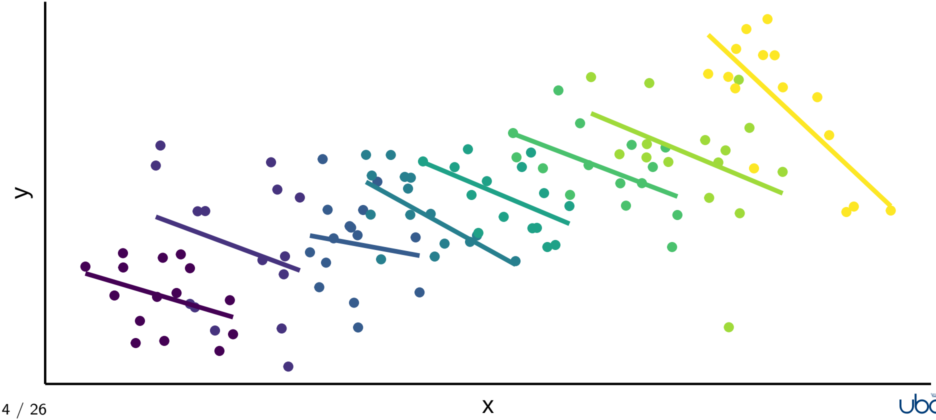
$$f(x_{kj}, j) = f_1(x_{kj}) + f_2(j) + f_{12}(x_{kj}, j).$$



Varying intercepts/slopes model

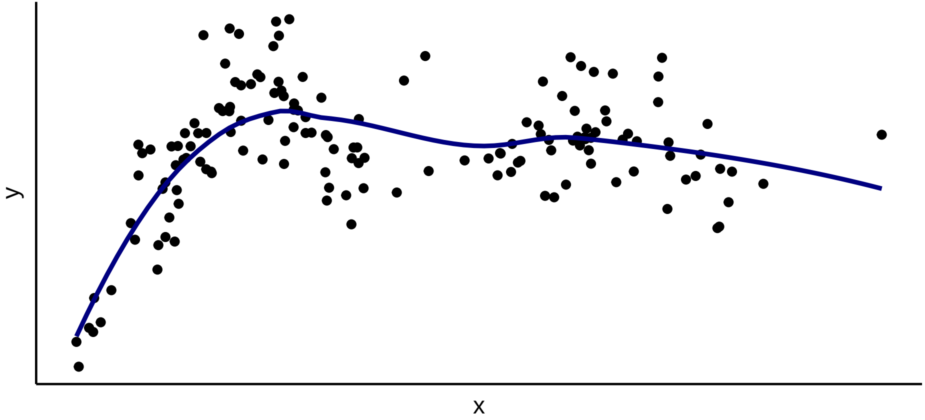
Suppose each unit $i = 1, \ldots, n$ relates to the kth observation in group $j \in \{1, \ldots, m\}$. Model the function f additively:

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



Smoothing models

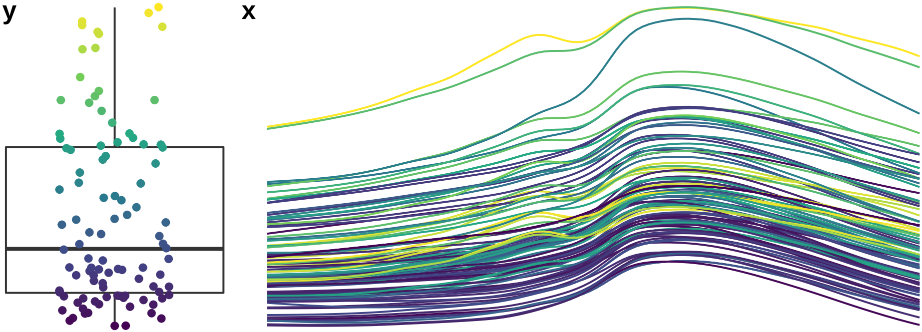
Suppose $f \in \mathcal{F}$ where $\mathcal{F}$ is a space of "smoothing functions" (models like LOESS, kernel regression, smoothing splines, etc.).



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 regression 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)

The entropy maximising prior distribution for f, subject to constraints, is

$$f(x) = \sum_{i=1}^{n} h(x, x_i) w_i$$
$$(w_1, \dots, w_n)^\top \sim \mathrm{N}_n(0, \Psi) \tag{2}$$

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

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

which happens to be **Fisher information** between two linear forms 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 (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 (and vice versa).

Of interest then are

1. Posterior distribution for the regression function,

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

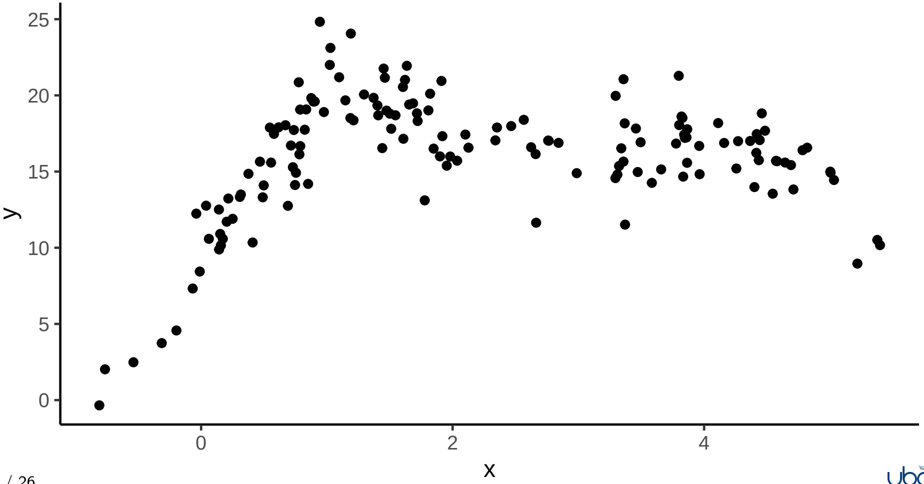
2. Posterior predictive distribution (given a new data point x_{new})

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

where $f_{new} = f(x_{new})$.

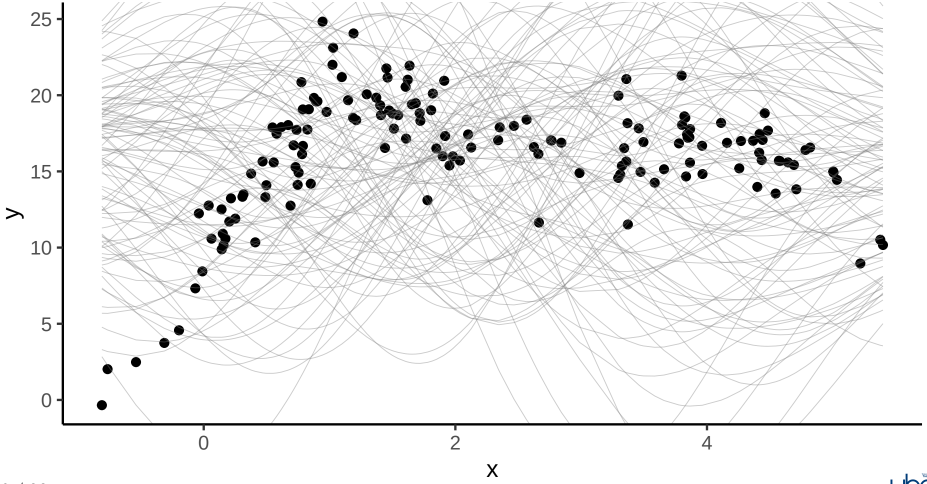
Introduction (cont.)

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



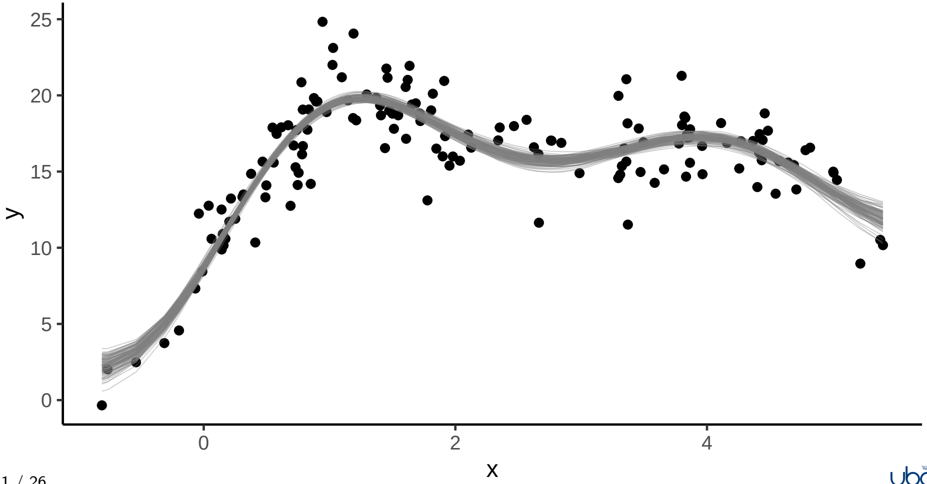
Introduction (cont.)

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



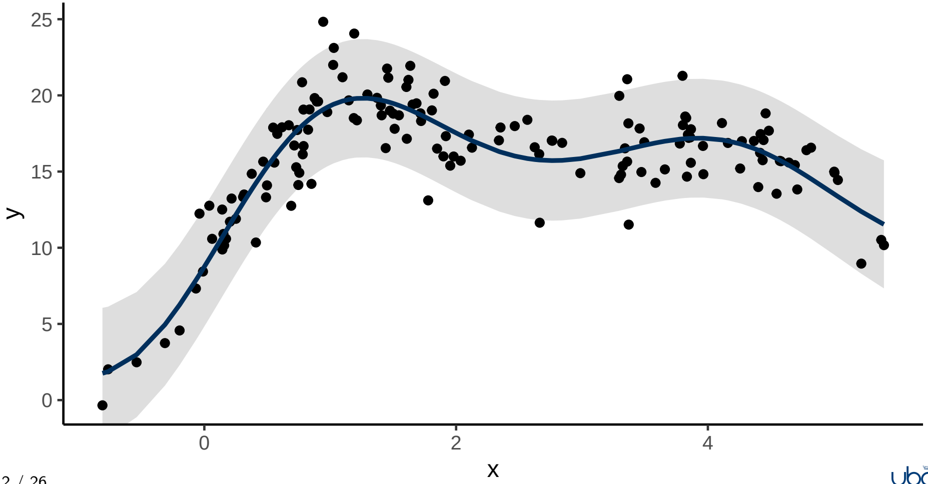
Introduction (cont.)

Sample paths from the posterior of f:



Introduction (cont.)

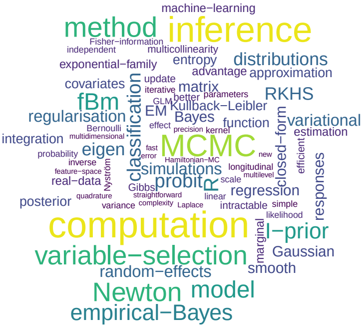
Posterior mean estimate for $y = f(x)$ and its 95% credibility interval.



Why I-priors?

Advantages

- Provides a unifying methodology for regression.
- Simple and parsimonious model specification and estimation.
- Often yield comparable (or better) predictions than competing ML algorithms.



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

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: Let $f \in \mathcal{F}$ be an RKHS with kernel h over a set $\mathcal{X}$.

Definition 2 (Hilbert spaces)

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

Definition 3 (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 $\forall x \in \mathcal{X}$,

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

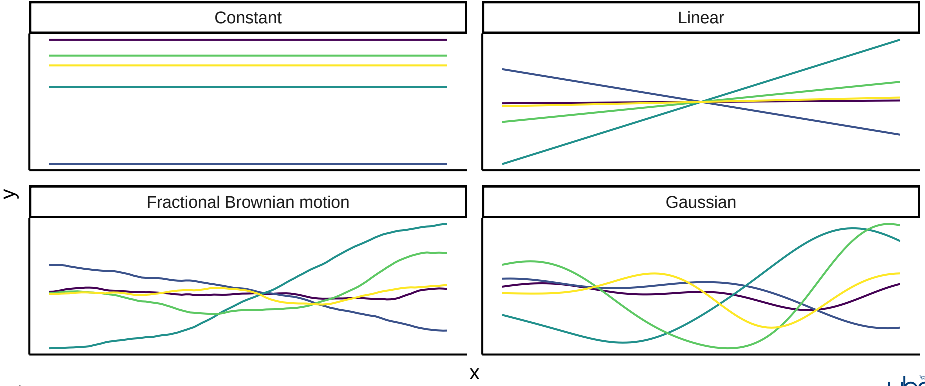
In particular, $\forall x, x' \in \mathcal{F}$, $h(x, x') = \langle h(\cdot, x), h(\cdot, x') \rangle_{\mathcal{F}}$.

Reproducing kernel Hilbert spaces (cont.)

Theorem 4

There is a bijection between

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



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 5 (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 regression model (1), the log-likelihood of f is given by

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

Lemma 6 (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 $\mathcal{I}_{f_g} = \langle \mathcal{I}_f, g \otimes g \rangle_{\mathcal{F}\otimes\mathcal{F}}$.

In particular, between two points $f_x := f(x)$ and $f_{x'} := f(x')$ [since $f_x = \langle f, h(\cdot, x) \rangle_\mathcal{F}$] we have:

$$\begin{aligned}
\mathcal{I}_f(x, x') &= \big\langle \mathcal{I}_f, h(\cdot, x) \otimes h(\cdot, x') \big\rangle_{\mathcal{F}\otimes\mathcal{F}} \\
&= \left\langle \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} h(\cdot, x_i) \otimes h(\cdot,_j) \ , \ h(\cdot, x) \otimes h(\cdot, x') \right\rangle_{\mathcal{F}\otimes\mathcal{F}} \\
&= \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} \langle h(\cdot, x), h(\cdot, x_i) \rangle_\mathcal{F} \big\langle h(\cdot, x'), h(\cdot, x_j) \big\rangle_\mathcal{F} \\
&= \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} h(x, x_i) h(x', x_j) =: k(x, x') \qquad (3)
\end{aligned}$$

The I-prior

Lemma 7

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 8 (I-prior)

Let ν be a volume measure induced by the norm above. The solution to

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

subject to the constraint

$$\mathrm{E}_{f \sim p} \|f\|^2_{\mathcal{F}_n} = \text{constant}$$

is the Gaussian distribution whose covariance function is $k(x, x')$.

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

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

Parameters of the model

$$\begin{aligned} y_i &= f_0(x_i) + \lambda \sum_{j=1}^n h(x_i, x_j) w_j + \epsilon_i \\ (\epsilon_1, \dots, \epsilon_n)^\top &\sim \mathrm{N}_n(0, \boldsymbol{\Psi}^{-1}) \\ (w_1, \dots, w_n)^\top &\sim \mathrm{N}_n(0, \boldsymbol{\Psi}) \end{aligned} \tag{4}$$

Further assumptions

1. The error variance Ψ is known up to a low-dimensional parameter, e.g. $\boldsymbol{\Psi} = \psi \mathbf{I}_n$, $\psi > 0$.
2. Each RKHS $\mathcal{F}$ of function is defined by the kernel $h_\lambda = \tilde{h}$, where $\lambda \in \mathbb{R}$ is a scale[1] parameter.
3. Certain kernels also require parameters themselves, e.g. the Hurst coefficient of the fBm or the lengthscale of the Gaussian kernel.
4. A prior mean function $f_0(x)$ may be set by the user.

[1] This necessitates the use of reproducing kernel Krein spaces.

Marginal likelihood

Denote by

- $\mathbf{y} = (y_1, \dots, y_n)^\top$
- $\mathbf{f} = \big(f(x_1), \dots, f(x_n)\big)^\top$
- $\mathbf{f}_0 = \big(f_0(x_1), \dots, f_0(x_n)\big)^\top$
- $\mathbf{w} = (w_1, \dots, w_n)^\top$
- $\mathbf{H}_\lambda = \big(h_\lambda(x_i, x_j)\big)_{i,j=1}^n \in \mathbb{R}^{n \times n}$

(1) + an I-prior on f implies

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

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

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

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

- Direct optimisation using e.g. conjugate gradients or Newton methods.
- Numerical stability issues (workaround: Cholesky or eigen decomp.).
- Prone to local optima.

EM algorithm

An alternative view of the model:

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

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

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

Choose starting values $\lambda^{(0)}$ and $\mathbf{\Psi}^{(0)}$. The E-step entails computing

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

EM algorithm (cont.)

The following quantities are needed and are easily obtained:

$$\tilde{\mathbf{w}} := \mathrm{E}(\mathbf{w} \mid \mathbf{y}, \lambda, \boldsymbol{\Psi}) \qquad \text{and} \qquad \tilde{\mathbf{W}} := \mathrm{E}(\mathbf{w}\mathbf{w}^\top \mid \mathbf{y}, \lambda, \boldsymbol{\Psi}) = \tilde{\mathbf{V}_w} + \tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top$$

Supposing $\boldsymbol{\Psi}$ but not $\mathbf{H}_\lambda$ depends on ψ; and $\mathbf{H}_\lambda$ depends on λ but not ψ, the M-step entails solving the following equations set to zero:

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

- 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 $\lambda^{(t)} \rightarrow \boldsymbol{\Psi}^{(t+1)} \rightarrow \lambda^{(t+1)} \rightarrow \cdots$ (expectation conditional maximisation, Meng and Rubin, 1993).

Introduction

Regression using I-priors

Estimation

Examples

Further research

Further research

Hello

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