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

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:

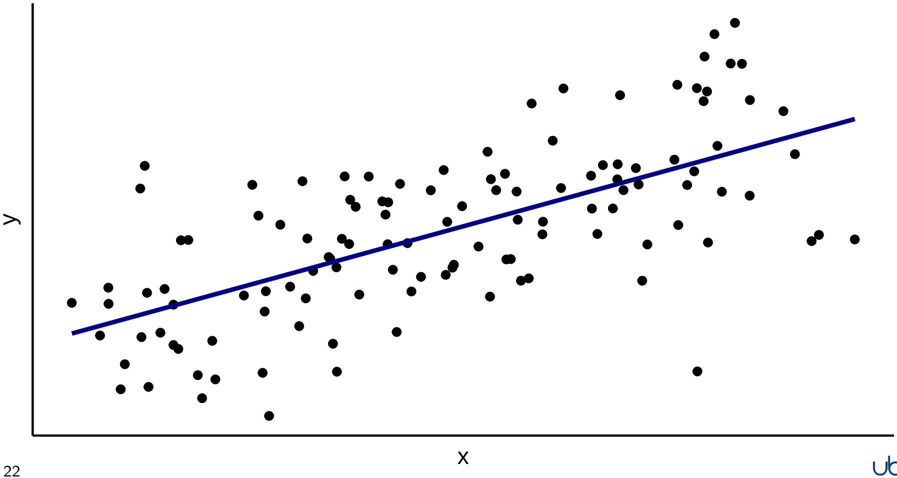
1. Ordinary linear regression when f is parameterised linearly.
2. Varying intercepts/slopes model when $\mathcal{X}$ is grouped.
3. Smoothing models when f is a smooth function.
4. Functional regression when $\mathcal{X}$ is functional.

Goal

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

Ordinary linear regression

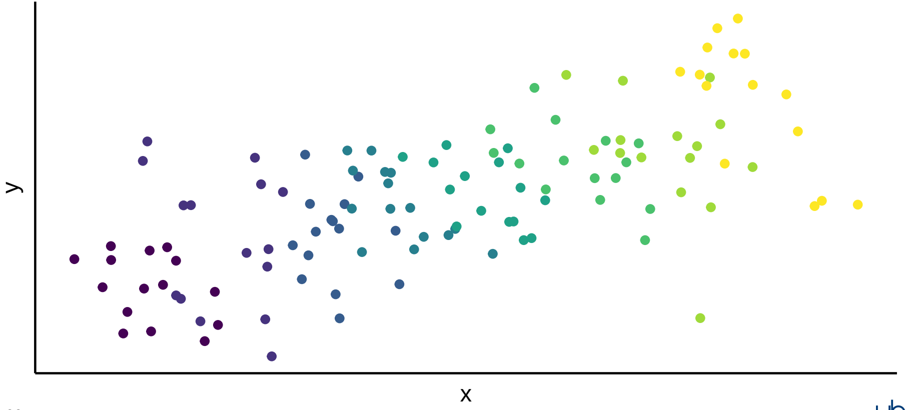
Suppose $f(x_i) = x_i^\top \beta$ for $i = 1, \dots, n$, where $x_i, \beta \in \mathbb{R}^p$.



Varying intercepts/slopes model

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

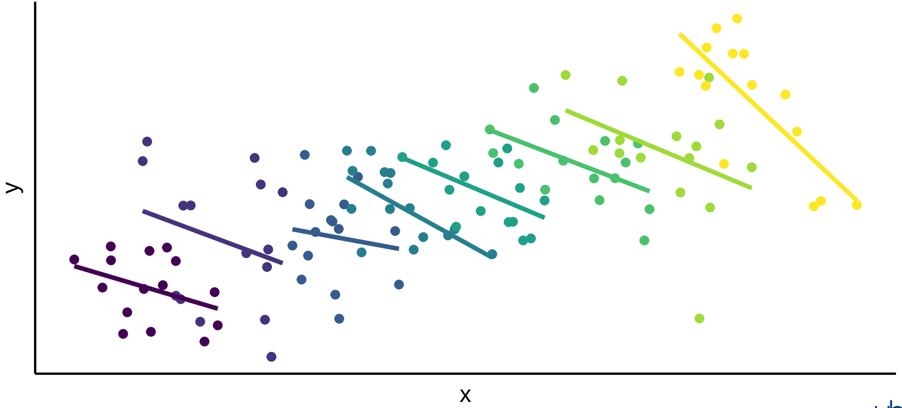
$$f(x_{kj}, j) = f_1(x_{kj}) + f_2(j) + f_{12}(x_{kj}, j).$$



Varying intercepts/slopes model

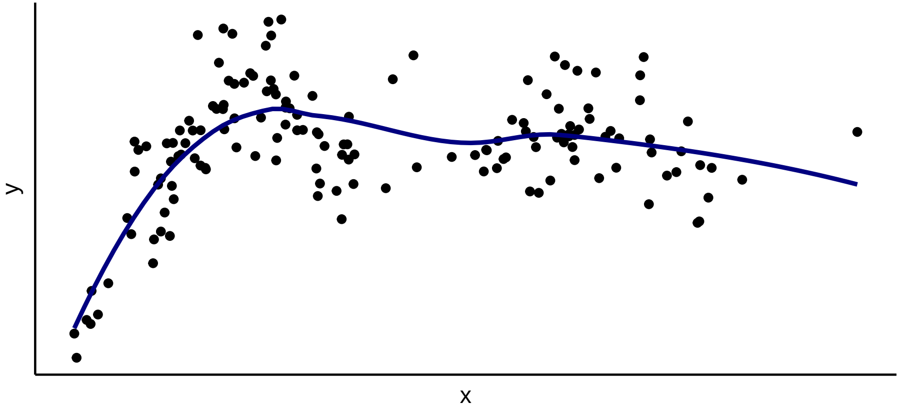
Suppose each unit $i = 1, \dots, n$ relates to the kth observation in group $j \in \{1, \dots, m\}$. Model the function f additively:

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



Smoothing models

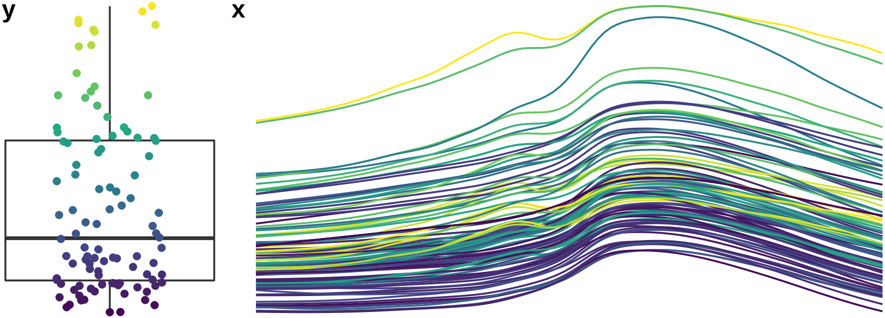
Suppose $f \in \mathcal{F}$ where $\mathcal{F}$ is a space of "smoothing functions" (models like LOESS, kernel regression, smoothing splines, etc.).



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

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

Therefore, the covariance kernel of $\mathbf{f} = \big(f(x_1), \dots, f(x_n)\big)^\top$ is determined by the function

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

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,

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

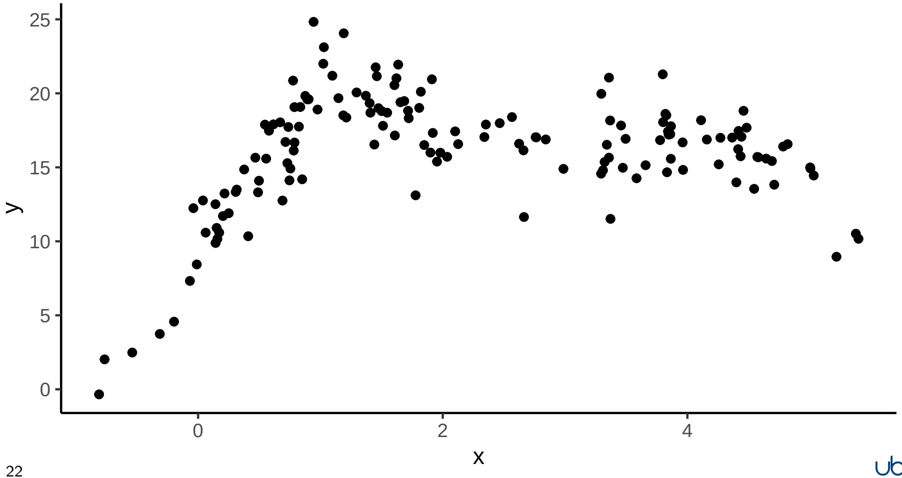
2. Posterior predictive distribution (given a new data point x_{new})

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

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

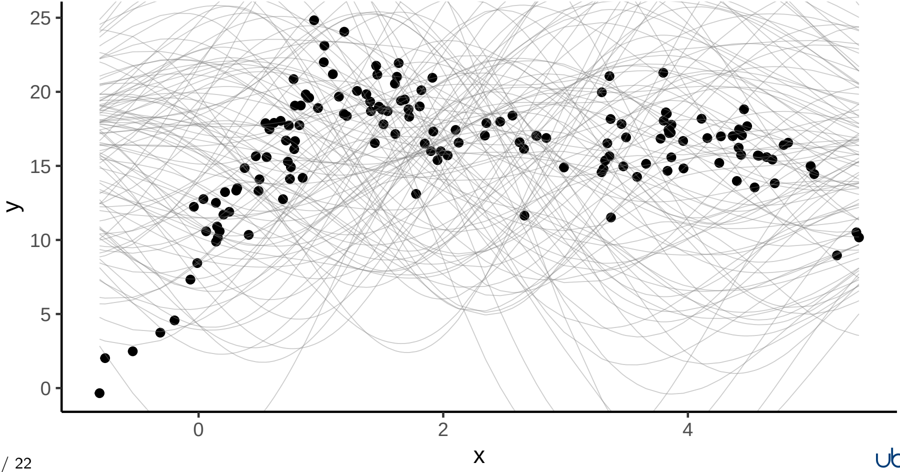
Introduction (cont.)

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



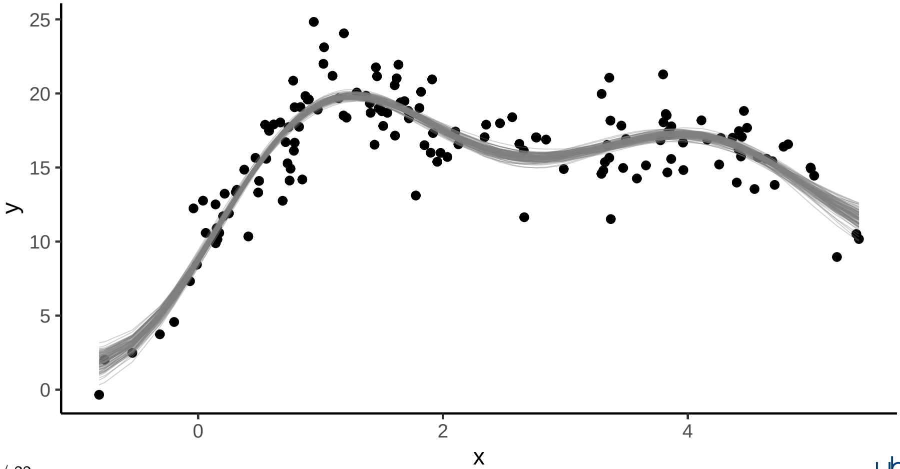
Introduction (cont.)

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



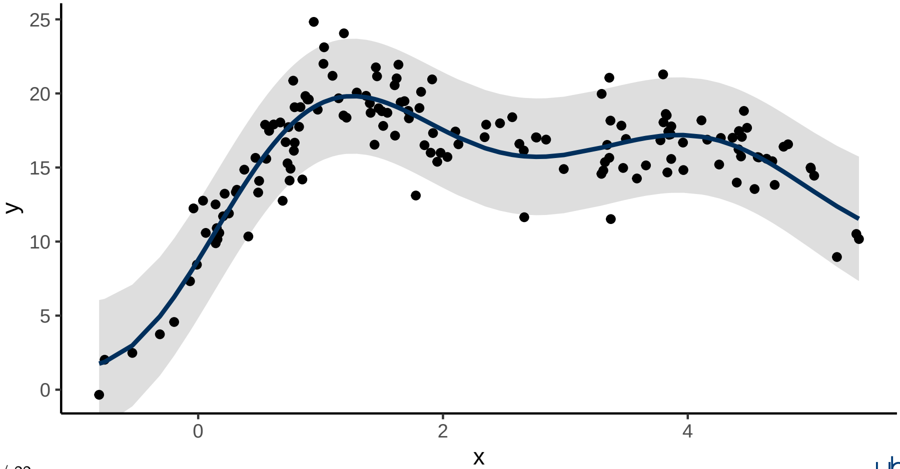
Introduction (cont.)

Sample paths from the posterior of f:



Introduction (cont.)

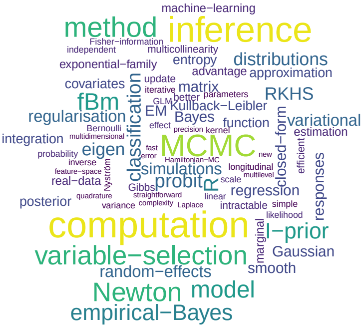
Posterior mean estimate for $y = f(x)$ and its 95% credibility interval.



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

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

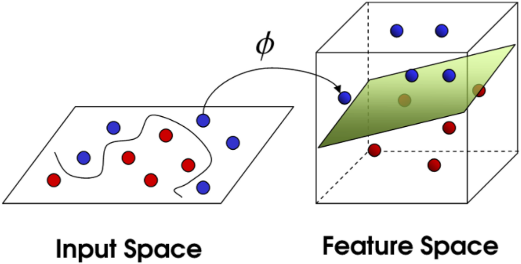
In particular, $\forall x, x' \in \mathcal{F}$, $h(x, x') = \langle h(\cdot, x), h(\cdot, x') \rangle_{\mathcal{F}}$.

Reproducing kernel Hilbert spaces (cont.)

- In ML literature, Mercer's Theorem states

$$h(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{V}} \quad \Leftrightarrow \quad h \text{ is semi p.d.}$$

where $\phi : \mathcal{X} \to \mathcal{V}$ is a mapping from $\mathcal{X}$ to the *feature space* $\mathcal{V}$.
- In many ML models, need not specify ϕ explicitly; computation is made simpler by the use of kernels.



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.

Corollary 5

Any $f \in \mathcal{F}$ can be approximated arbitrarily well by functions of the form

$$\tilde{f}(x) = \sum_{i=1}^{n} w_i h(x, x_i)$$

for some constants $w_1, \ldots, w_n \in \mathbb{R}$, because $\mathcal{F}$ is the completion of the vector space $\tilde{\mathcal{F}} = \operatorname{span}\{h(\cdot, x) \mid x \in \mathcal{X}\}$ equipped with the squared norm $\|\tilde{f}\|^2 = \sum_{i,j=1}^{n} w_i w_j h(x_i, x_j)$.

Examples of RKHSs

The Fisher information

Lemma 6 (Fisher information for regression function)

The Fisher information for f is given by

$$\mathcal{I}_f = \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} h(\cdot, x_i) \otimes h(\cdot, x_j)$$

where $\otimes$ is...

The Fisher information (cont.)

It's helpful to think of $\mathcal{I}_f$ as a bilinear form $\mathcal{I}_f : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ defined by

$$\mathcal{I}_f = -\operatorname{E} \nabla^2 L(f|y),$$

so the Fisher information between two point evaluation functionals of is

$$\mathcal{I}\big(f(x), f(x')\big) = \mathcal{I}_f\big(h(\cdot, x), h(\cdot, x')\big) = \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} h(x, x_i) h(x', x_j).$$

Example 7

Let

Suppose further that $f \in \mathcal{F}$ where $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) with reproducing kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then (1) can be expressed as

$$\begin{gathered} y_i = \big\langle f, h(\cdot, x_i) \big\rangle_{\mathcal{F}} + \epsilon_i \\ (\epsilon_1, \ldots, \epsilon_n)^\top \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi}^{-1}) \end{gathered} \tag{3}$$

The Fisher information for f is given by

$$\mathcal{I}_f = \sum_{i=1}^n \sum_{j=1}^n \psi_{ij} h(\cdot, x_i) \otimes h(\cdot, x_j)$$

where each $y_i \in \mathbb{R}$,, and $f \in \mathcal{F}$ a reproducing kernel Hilbert space (RKHS) with kernel $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The I-prior (Bergsma, 2019) for the regression function f is the random function defined

$$
\begin{aligned}
f(x_i) &= f_0(x_i) + \sum_{k=1}^{n} h(x_i, x_k) w_k \\
(w_1, \ldots, w_n)^\top &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi})
\end{aligned}
\tag{4}
$$

where f_0 is some prior mean for the regression function.

Further research

Hello