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:

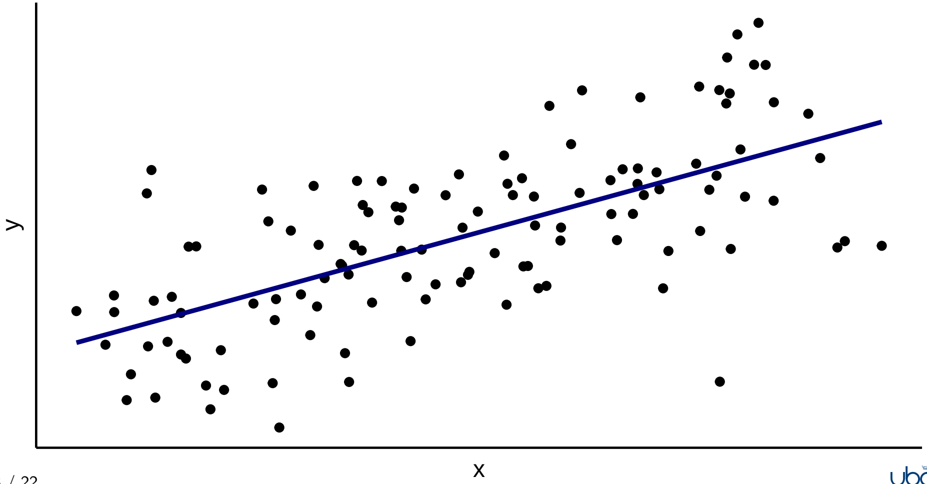
1. Ordinary linear regression when f is parameterised linearly.
2. Varying intercepts/slopes model when $\mathcal{X}$ is grouped.
3. Smoothing models when f is a smooth function.
4. Functional regression when $\mathcal{X}$ is functional.

Goal

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

Ordinary linear regression

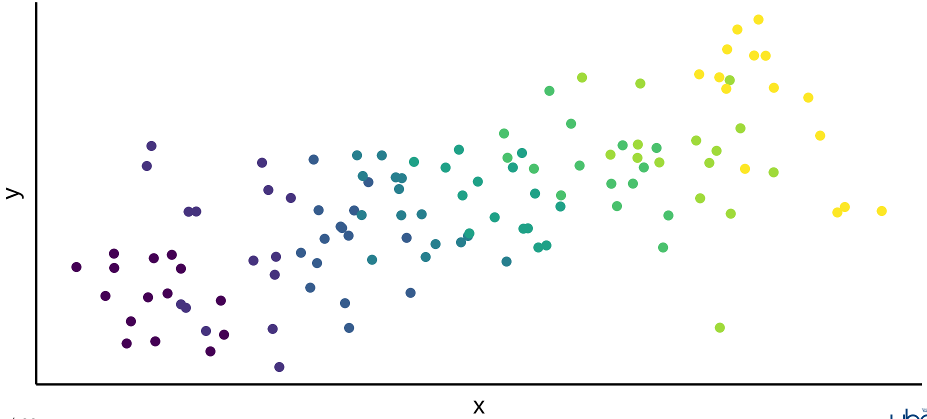
Suppose $f(x_i) = x_i^\top \beta$ for $i = 1, \dots, n$, where $x_i, \beta \in \mathbb{R}^p$.



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:

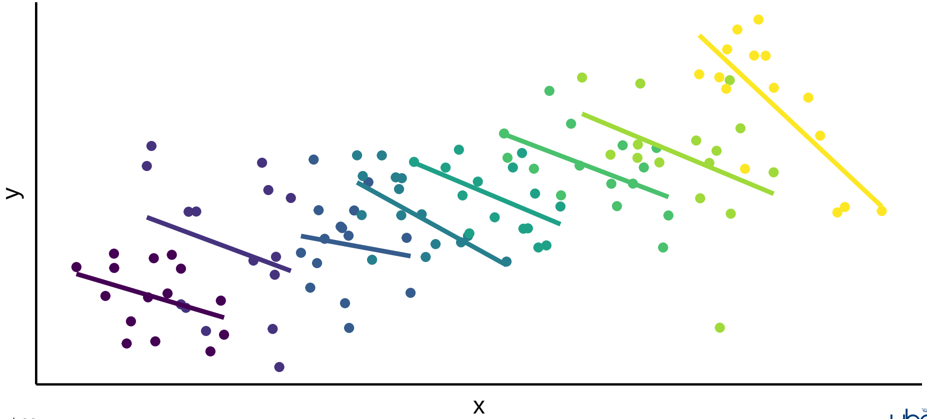
$$f(x_{kj}, j) = f_1(x_{kj}) + f_2(j) + f_{12}(x_{kj}, j).$$



Varying intercepts/slopes model

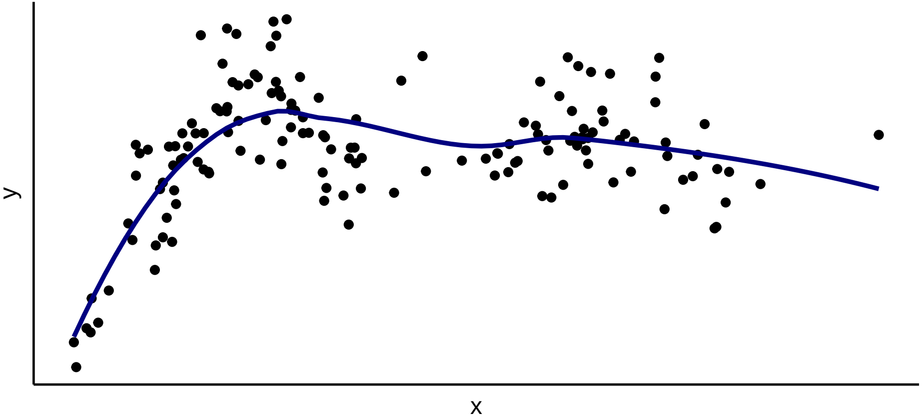
Suppose each unit $i = 1, \dots, n$ relates to the kth observation in group $j \in \{1, \dots, m\}$. Model the function f additively:

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



Smoothing models

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

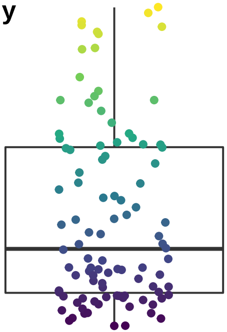


Functional regression

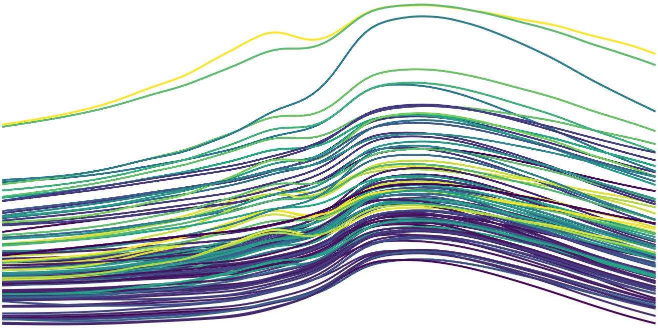
Suppose the input set $\mathcal{X}$ is functional. The (linear) regression aims to estimate a coefficient function $\beta : \mathcal{T} \to \mathbb{R}$

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

y



x



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

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

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

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,

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

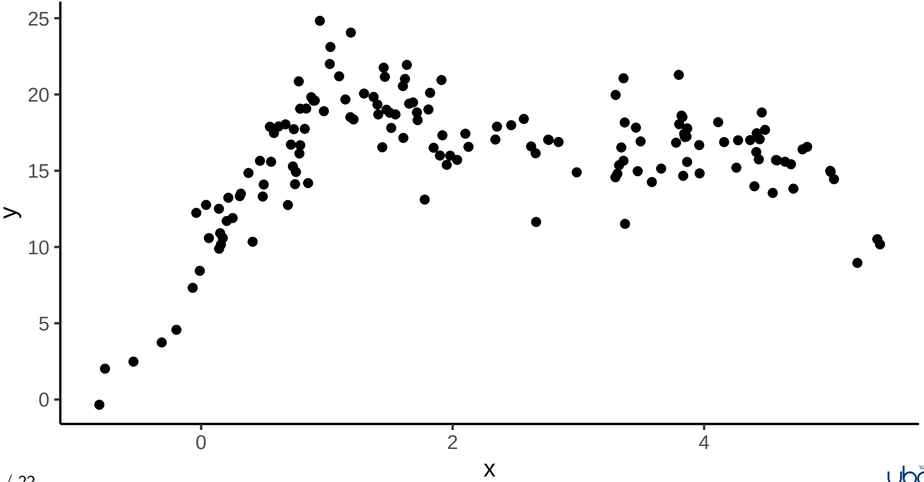
2. Posterior predictive distribution (given a new data point x_{new})

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

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

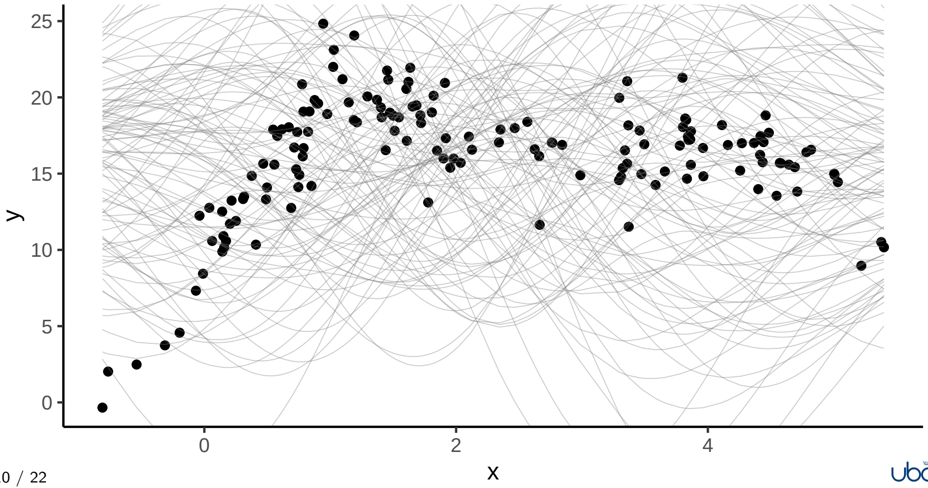
Introduction (cont.)

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



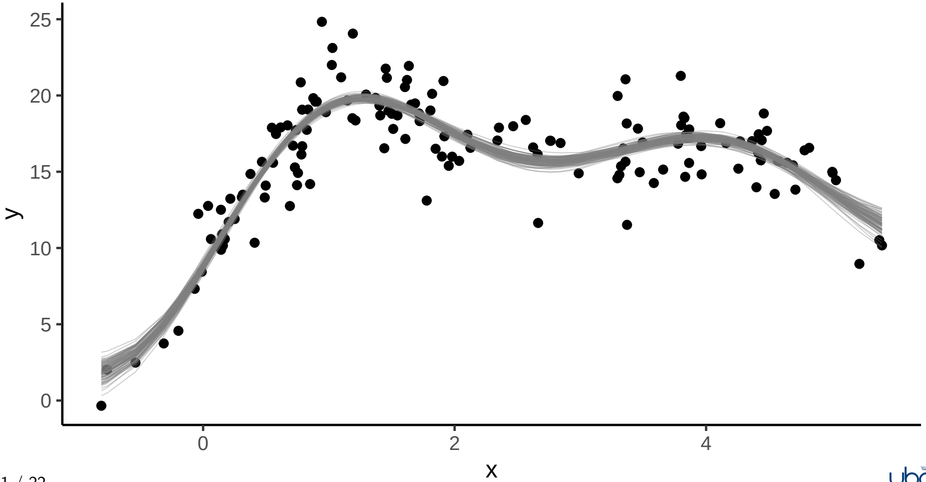
Introduction (cont.)

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



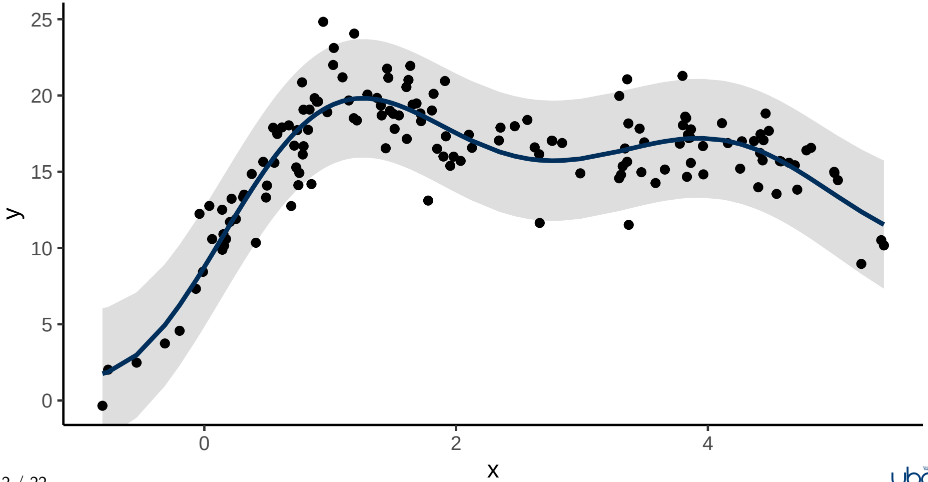
Introduction (cont.)

Sample paths from the posterior of f:



Introduction (cont.)

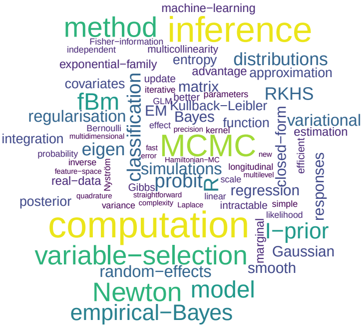
Posterior mean estimate for $y = f(x)$ and its 95% credibility interval.



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

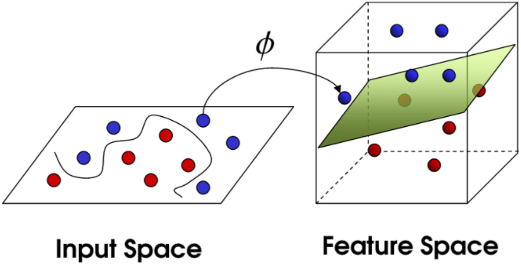
In particular, $\forall x, x' \in \mathcal{F}$, $h(x, x') = \langle h(\cdot, x), h(\cdot, x') \rangle_{\mathcal{F}}$.

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.

Examples of RKHSs

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 5 (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 6

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 7 (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

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

Further research

Hello