يونيۏرسيتي 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

Regression analysis

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

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

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:

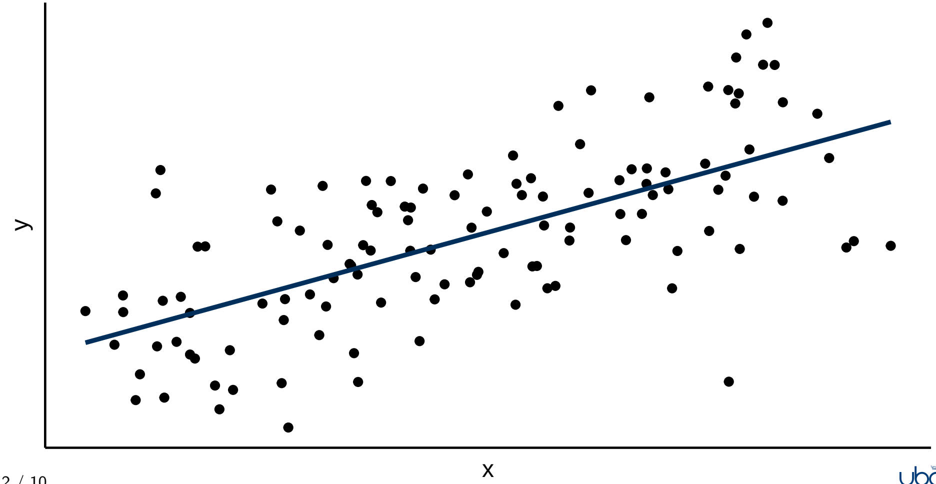
1. Ordinary linear regression when f is parameterised linearly.
2. Varying intercepts/slopes model when $\mathcal{X}$ is grouped.
3. Smoothing models when f is a smooth function.
4. Functional regression when $\mathcal{X}$ is functional.

Goal

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

Ordinary linear regression

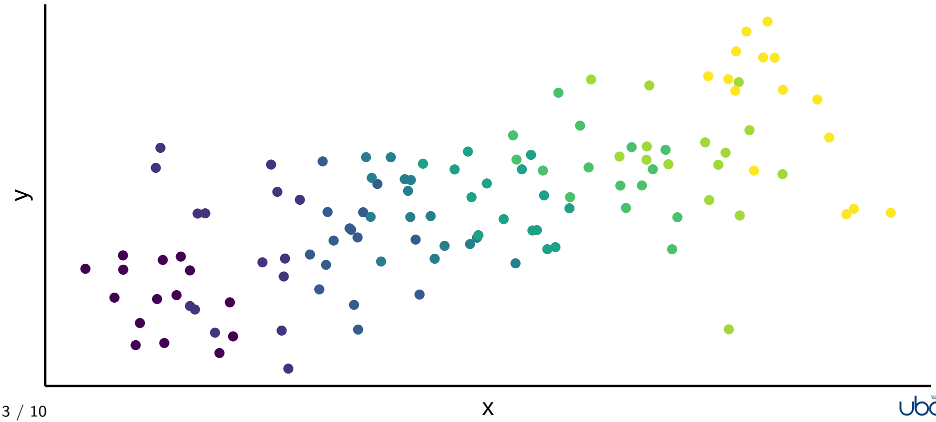
Suppose $f(x_i) = x_i^\top \beta$ for $i = 1, \ldots, n$, where $x_i, \beta \in \mathbb{R}^p$.



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:

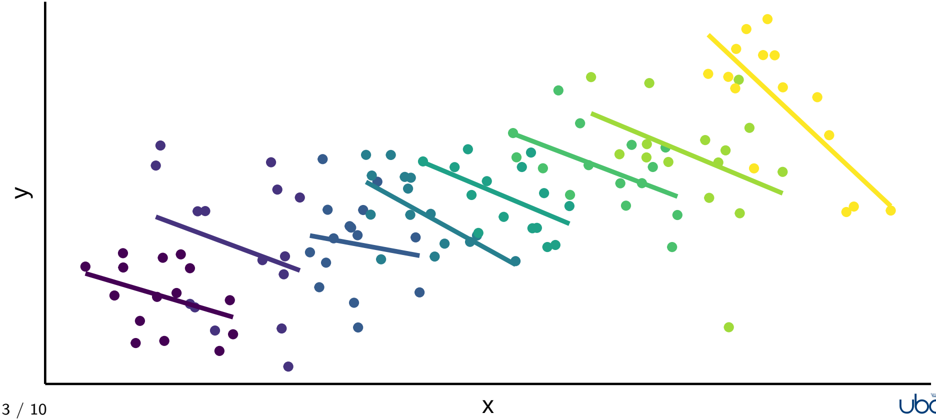
$$f(x_{kj}, j) = f_1(x_{kj}) + f_2(j) + f_{12}(x_{kj}, j).$$



Varying intercepts/slopes model

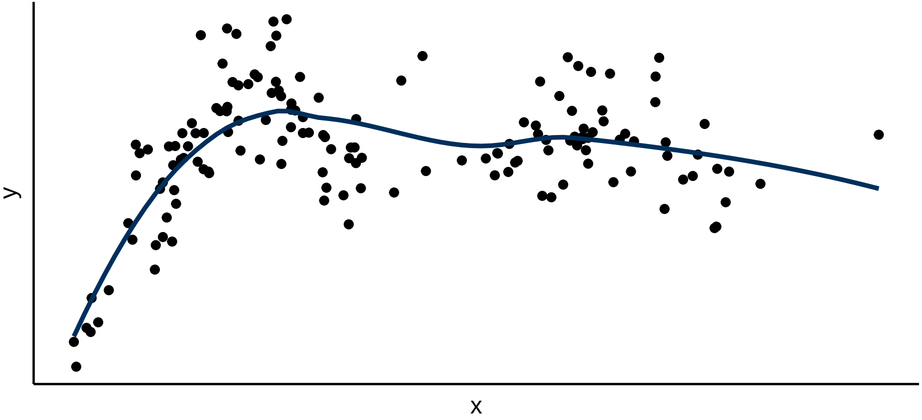
Suppose each unit $i = 1, \ldots, n$ relates to the kth observation in group $j \in \{1, \ldots, m\}$. Model the function f additively:

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



Smoothing models

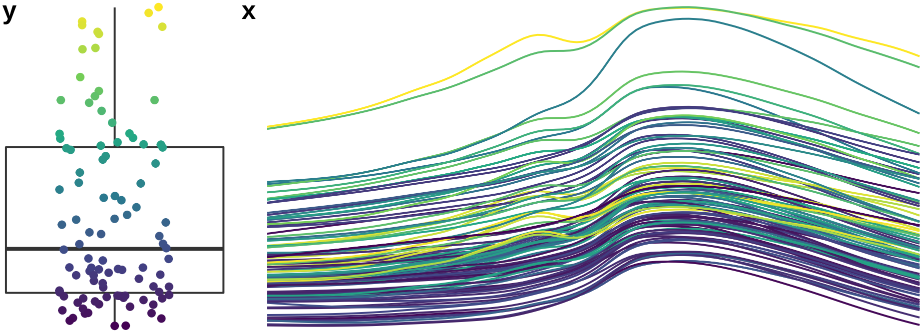
Suppose $f \in \mathcal{F}$ where $\mathcal{F}$ is a space of "smoothing functions" (models like LOESS, kernel regression, smoothing splines, etc.).



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

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

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:

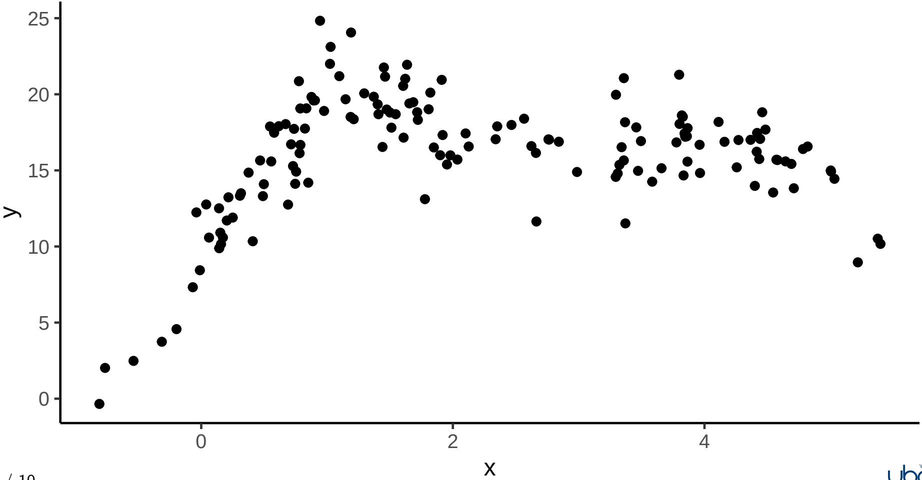
> The more information about f, the larger its prior variance, and hence the smaller the influence of the prior mean (and vice versa).

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

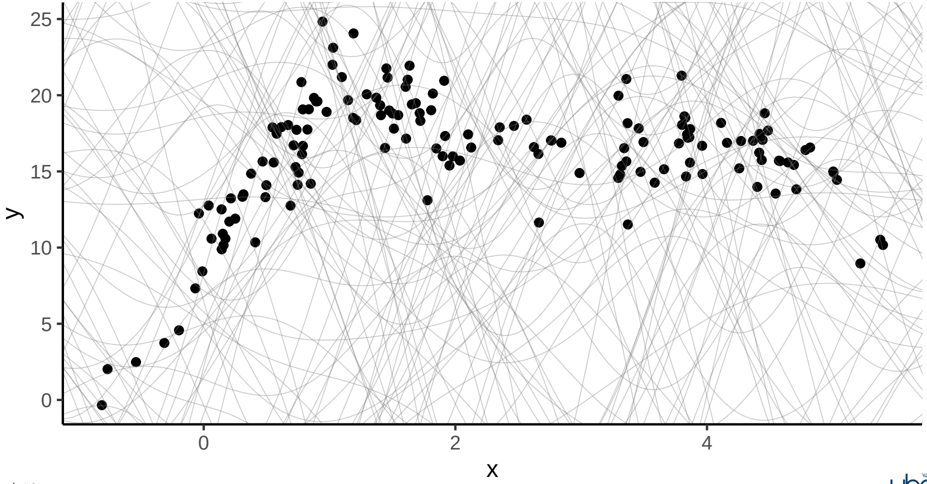
Illustration

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



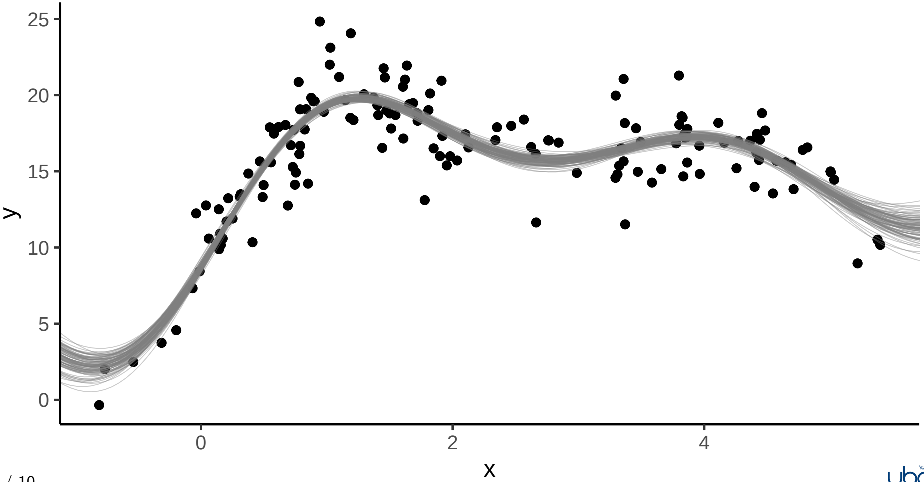
Illustration

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



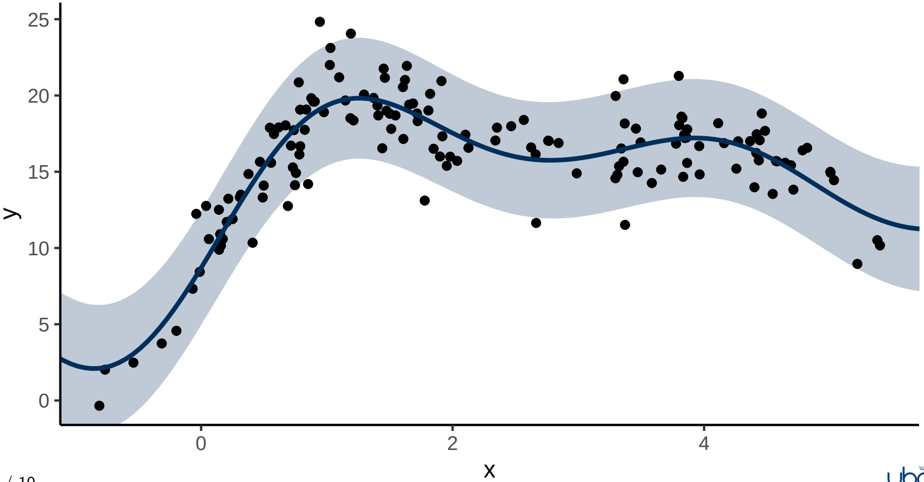
Illustration

Sample paths from the posterior of f:



Illustration

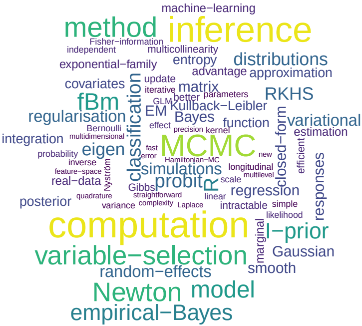
Posterior mean estimate for $y = f(x)$ and its 95% credibility interval.



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 (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*