Regression modelling using I-priors

Haziq Jamil
Supervisors: Dr. Wicher Bergsma & Prof. Irini Moustaki

Social Statistics (Year 1)
London School of Economics & Political Science

19 May 2015

PhD Presentation Event

Outline

Linear regression

- Consider a set of data points $\{(y_1, x_1), \ldots, (y_n, x_n)\}$.
- A model is linear if the relationship between y_i and the independent variables is linear.
  - $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ ✓
  - $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ ✓
  - $y_i = \beta_0 x_i^{\beta_1 + 2\beta_2} + \epsilon_i$ ✗
  - In other words, the equations must be linear in the parameters.

Linear regression

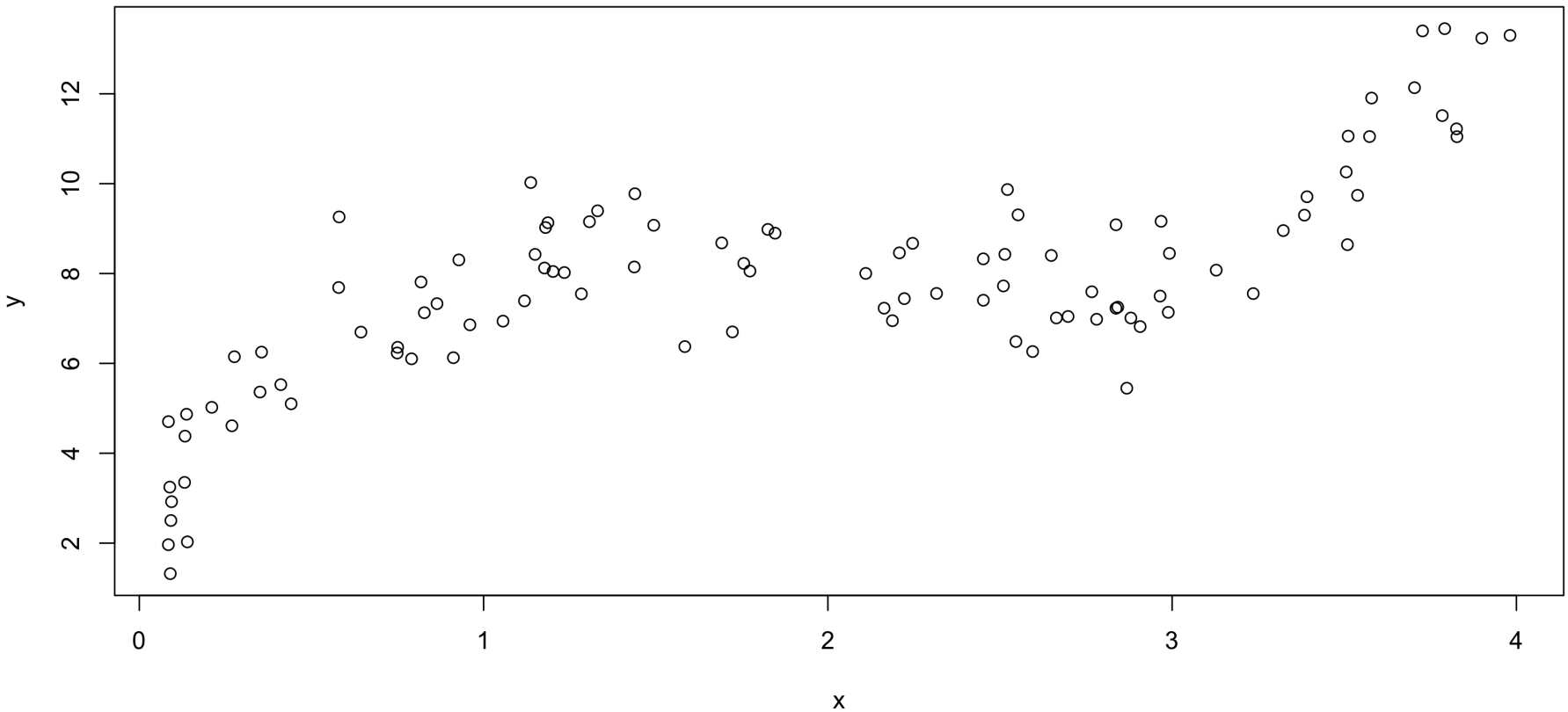
- Definition **(The linear regression model)**

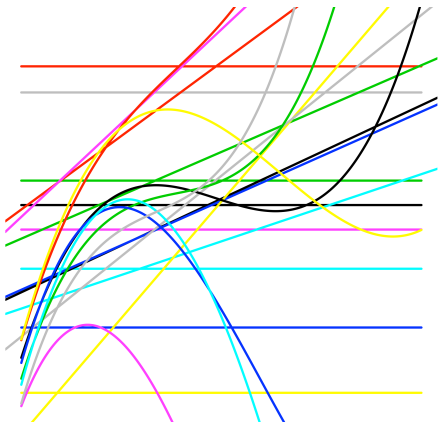
$$y_i = f(x_i) + \epsilon_i$$

$$\begin{aligned} y_i &\in \mathbb{R}, \text{ real-valued observations} \\ x_i &\in \mathcal{X}, \text{ a set of characteristics for unit } i \\ f &\in \mathcal{F}, \text{ a vector space of functions over the set } \mathcal{X} \\ (\epsilon_1, \ldots, \epsilon_n) &\sim N(\mathbf{0}, \boldsymbol{\Psi}^{-1}) \\ i &= 1, \ldots, n \end{aligned} \tag{1}$$

Note: For iid observations, $\boldsymbol{\Psi} = \psi \mathbf{I}_n$. In general, $\boldsymbol{\Psi} = (\psi_{ij})$.

Linear regression





Estimation methods

How to pick the best line from the bag of stuff?

- Many ways - Least squares, maximum likelihood, Bayesian...
- When dimensionality is large, may overfit. Solutions:
  - Dimension reduction
  - Random effects models
  - Regularization

  ...all require additional assumptions
- **I-priors**
  An I-prior on f is a distribution π on f such that its covariance matrix is the Fisher information of f. Also, assign a "best guess" on the prior mean, e.g. $f_0 = 0$.

Example: multiple regression

$$\mathbf{y} = \overbrace{\alpha + \mathbf{X}\beta}^{\mathbf{f}} + \epsilon$$
$$\epsilon \sim N(\mathbf{0}, \psi^{-1}\mathbf{I}_n)$$

We know from linear regression theory that $I[\beta] = \psi\mathbf{X}^T\mathbf{X}$. An I-prior on β is then

$$\beta \sim N(\beta_0, \lambda^2\psi\mathbf{X}^T\mathbf{X}).$$

Equivalently,

$$\beta = \beta_0 + \lambda\mathbf{X}^T\mathbf{w}$$
$$\mathbf{w} \sim N(\mathbf{0}, \psi\mathbf{I}_n).$$

Thus, an I-prior on $\mathbf{f}$ is

$$\mathbf{f} = \alpha + \mathbf{X}\beta_0 + \lambda\mathbf{X}\mathbf{X}^T\mathbf{w}$$
$$\mathbf{w} \sim N(\mathbf{0}, \psi\mathbf{I}_n).$$

Inner products

Functional vector spaces

Kernel methods

Reproducing kernels

Hilbert spaces

Gaussian random vectors

I-prior theory

Fisher Information

Krein spaces

Means of random functions

Feature maps

Variances of random functions

Moore-Aronszajn Theorem

Random functions

Definitions & theorem

- Definition **(Inner products)**
  Let $\mathcal{F}$ be a vector space $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ is said to be an inner product on $\mathcal{F}$ if all of the following are satisfied:
  - **Symmetry:** $\langle f, g \rangle_{\mathcal{F}} = \langle g, f \rangle_{\mathcal{F}}$
  - **Linearity:** $\langle af_1 + bf_2, g \rangle_{\mathcal{F}} = a\langle f_1, g \rangle_{\mathcal{F}} + b\langle f_2, g \rangle_{\mathcal{F}}$
  - **Non-degeneracy:** $\langle f, g \rangle_{\mathcal{F}} = 0 \Rightarrow f = 0$

  for all $f, f_1, f_2, g \in \mathcal{F}$ and $a, b \in \mathbb{R}$. Additionally, an inner product is positive definite (negative definite) if $\langle f, f \rangle_{\mathcal{F}} \geq 0$ (≤ 0). An inner product is indefinite if it is neither positive nor negative definite.

- Definition **(Hilbert space)**
  A positive definite inner product space which is complete, i.e. contains the limits of all Cauchy sequences.

- Definition **(Krein space)**
  An (indefinite) inner product space which generalizes Hilbert spaces by dropping the positive definite restriction.

Definitions & theorem

- Definition **(Kernels)**
  Let $\mathcal{X}$ be a non-empty set. A function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel if there exists a Hilbert space $\mathcal{F}$ and a map $\phi : \mathcal{X} \to \mathcal{F}$ such that $\forall x, x' \in \mathcal{X}$,
  $$h(x, x') = \langle \phi(x), \phi(x') \rangle.$$

- Definition **(Reproducing kernels)**
  Let $\mathcal{F}$ be a Hilbert/Krein space of functions over a non-empty set $\mathcal{X}$. A function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{F}$, and $\mathcal{F}$ a RKHS/RKKS, if h satisfies
  - $\forall x \in \mathcal{X}$, $h(\cdot, x) \in \mathcal{F}$
  - $\forall x \in \mathcal{X}$, $f \in \mathcal{F}$, $\langle f, h(\cdot, x) \rangle_{\mathcal{F}} = f(x)$.

- Kernel algorithms have many important uses in Machine Learning literature, such as pattern recognition, kernel PCA, finding distances of means in feature space, and many more.

Definitions & theorem

- Theorem **(Gaussian I-priors)** [Bergsma, 2014]
  For the linear regression model (1)*, let* $\mathcal{F}$ *be the RKKS with kernel* $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$*. Then, assuming it exists, the Fisher information for* f *is given by*

$$I[f](x_i, x_i') = \sum_{k=1}^{n}\sum_{l=1}^{n} \psi_{kl} h(x_i, x_k) h(x_i', x_l).$$

  Let π *be a Gaussian I-prior on* f *with prior mean* f_0 *and variance* $I[f]$*. Then* π *is called an I-prior for* f*, and a random vector* $f \sim \pi$ *has the random effect representation*

$$f(x_i) = f_0(x_i) + \sum_{k=1}^{n} h(x_i, x_k) w_k$$
$$(w_1, \ldots, w_n) \sim N(\mathbf{0}, \mathbf{\Psi}).$$

Back to the multiple regression example

- We saw the I-prior method applied to multiple regression:

$$f(\mathbf{x}_i) = \overbrace{\alpha + \mathbf{x}_i\boldsymbol{\beta}_0}^{f_0(\mathbf{x}_i)} + \overbrace{\lambda(\mathbf{X}\mathbf{X}^T)_i\mathbf{w}}^{\sum_{k=1}^n h(\mathbf{x}_i,\mathbf{x}_k)w_k}$$
$$\mathbf{w} := (w_1, \ldots, w_n) \sim N(\mathbf{0}, \psi\mathbf{I}_n).$$

Back to the multiple regression example

- We saw the I-prior method applied to multiple regression:

$$f(\mathbf{x}_i) = \overbrace{\alpha + \mathbf{x}_i\boldsymbol{\beta}_0}^{f_0(\mathbf{x}_i)} + \overbrace{\lambda(\mathbf{X}\mathbf{X}^T)_i\mathbf{w}}^{\sum_{k=1}^n h(\mathbf{x}_i,\mathbf{x}_k)w_k}$$
$$\mathbf{w} := (w_1, \ldots, w_n) \sim N(\mathbf{0}, \psi\mathbf{I}_n).$$

- Choose different RKHS/RKKS $\mathcal{F}$ and corresponding h to suit the type/characteristic of the $\mathbf{x}$s in order to do regression modelling.

Back to the multiple regression example

- We saw the I-prior method applied to multiple regression:

$$f(\mathbf{x}_i) = \overbrace{\alpha + \mathbf{x}_i\boldsymbol{\beta}_0}^{f_0(\mathbf{x}_i)} + \overbrace{\lambda(\mathbf{X}\mathbf{X}^T)_i\mathbf{w}}^{\sum_{k=1}^n h(\mathbf{x}_i,\mathbf{x}_k)w_k}$$
$$\mathbf{w} := (w_1, \ldots, w_n) \sim N(\mathbf{0}, \psi\mathbf{I}_n).$$

- Choose different RKHS/RKKS $\mathcal{F}$ and corresponding h to suit the type/characteristic of the $\mathbf{x}$s in order to do regression modelling.

Toolbox of RKHS/RKKS

$\mathcal{X} = \{x_i\}$	Characteristic/Uses	Vector space $\mathcal{F}$	Kernel $h(x_i, x_k)$
Nominal	1) Categorical covariates; 2) In a multilevel setting, x_i = group no. of unit i.	Pearson	$\frac{\mathbb{I}[x_i = x_k]}{p_i} - 1$ where $p_i = \mathbb{P}[X = x_i]$
Real	As in classical regression, x_i = real-valued covariate associated with unit i.	Canonical	$x_i x_k$
Real	As in (1-dim) smoothing, x_i = data point associated with observation y_i.	Fractional Brownian Motion (FBM)	$\lvert x_i\rvert^{2\gamma} + \lvert x_k\rvert^{2\gamma} - \lvert x_i - x_k\rvert^{2\gamma}$ with $\gamma \in (0, 1)$

- We can construct new RKHS/RKKS from existing ones.
  - Example **(ANOVA RKKS)** Set of $x_i = (x_{1i}, x_{2i})$ of Nominal + Real characteristics. Then

$$h(x_i, x_i') = h_1(x_{1i}, x_{1i}') + h_2(x_{2i}, x_{2i}') + h_1(x_{1i}, x_{1i}')h_2(x_{2i}, x_{2i}')$$

Parameters to be estimated

- Let's choose a prior mean of zero (or set an overall constant/intercept to be estimated).

- For the I-prior linear model

$$\begin{aligned} y_i &= \alpha + \sum_{k=1}^{n} h_\lambda(x_i, x_k) w_k + \epsilon_i \\ \epsilon_i &\sim N(0, \psi^{-1}) \\ w_i &\sim N(0, \psi) \\ i &= 1, \ldots, n, \end{aligned} \tag{2}$$

  the parameters to be estimated are $\boldsymbol{\theta} = (\alpha, \lambda, \psi)^T$.

- λ is introduced to resolve the arbitrary scale of an RKKS/RKHS $\mathcal{F}$ over a set $\mathcal{X}$. Number of λ parameters = number of kernels used, not interactions nor covariates.

EM algorithm

- For the I-prior model in (2), treat the w_is as missing.
- The distributions are easy enough to obtain:
  - $\mathbf{y} \sim N(\boldsymbol{\alpha}, \mathbf{V}_y)$, where $\mathbf{V}_y := \mathbf{H}_\lambda \boldsymbol{\Psi} \mathbf{H}_\lambda + \boldsymbol{\Psi}^{-1}$
  - $\mathbf{w} \sim N(\mathbf{0}, \boldsymbol{\Psi})$
  - $\begin{pmatrix} \mathbf{y} \\ \mathbf{w} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\alpha} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{V}_y & \mathbf{H}_\lambda \boldsymbol{\Psi} \\ \boldsymbol{\Psi} \mathbf{H}_\lambda & \boldsymbol{\Psi} \end{pmatrix} \right)$
  - $\mathbf{w}|\mathbf{y} \sim N\left(\boldsymbol{\Psi} \mathbf{H}_\lambda \mathbf{V}_y^{-1} (\mathbf{y} - \boldsymbol{\alpha}), \mathbf{V}_y^{-1} \right)$

  where $\mathbf{H}_\lambda(i,j) = h_\lambda(x_i, x_j)$ and $\boldsymbol{\Psi} = \psi \mathbf{I}_n$.
- E-step: Calculate $Q(\boldsymbol{\theta}) = \mathrm{E}_{\mathbf{w}} \left[\log f(\mathbf{y}, \mathbf{w}; \boldsymbol{\theta}) | \mathbf{y}; \boldsymbol{\theta}_t \right]$.
- M-step: $\boldsymbol{\theta}_{t+1} \leftarrow \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})$.

Generalised least square estimator for α

- Write Model (2) as $\mathbf{y} = \alpha\mathbf{1} + \mathbf{H}_\lambda\mathbf{w} + \boldsymbol{\epsilon}$, where $\mathbf{y} \sim N(\boldsymbol{\alpha}, \mathbf{V}_y)$.
- Assume values for λ and ψ are known, and thus too $\mathbf{V}_y(\lambda, \psi) = \mathbf{H}_\lambda\boldsymbol{\Psi}\mathbf{H}_\lambda + \boldsymbol{\Psi}^{-1}$.
- GLS estimator for α is

$$\hat{\alpha} = (\mathbf{1}^T\mathbf{V}_y^{-1}\mathbf{1})^{-1}(\mathbf{1}^T\mathbf{V}_y^{-1}\mathbf{y}).$$

- This turns out to be identical to the MLE.

Exponential family EM algorithm

- Consider a density function belonging to the exponential family with the (canonical) form $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\theta}) = \exp[\boldsymbol{\theta}\cdot\mathbf{T}(\mathbf{x}) - A(\boldsymbol{\theta})]h(\mathbf{x})$.
  - The MLE is found by solving the set of equations $\mathbf{T}(\mathbf{x}) = A'(\boldsymbol{\theta})$.
  - It is also know that $A'(\boldsymbol{\theta}) = \mathrm{E}[\mathbf{T}(\mathbf{x});\boldsymbol{\theta}]$.

- In the EM algorithm, the "full" data is $\mathbf{x} = (\mathbf{y}, \mathbf{w})$. The E-step involves calculating $Q(\boldsymbol{\theta})$, and for the exponential family, this turns out to be

$$Q(\boldsymbol{\theta}) = \mathrm{E}_{\mathbf{w}}\left[\boldsymbol{\theta}\cdot\mathbf{T}(\mathbf{y},\mathbf{w}) - A(\boldsymbol{\theta}) + \log h(\mathbf{y},\mathbf{w})|\mathbf{y};\boldsymbol{\theta}_t\right].$$

- Maximising this over $\boldsymbol{\theta}$, we arrive at the FOC

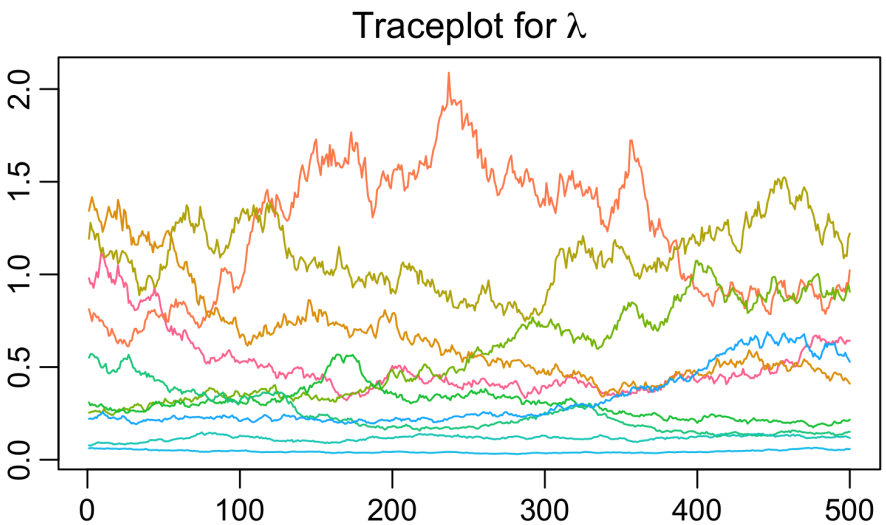
$$\begin{aligned} Q'(\boldsymbol{\theta}) = \mathrm{E}_{\mathbf{w}}\left[\boldsymbol{\theta}\cdot\mathbf{T}(\mathbf{y},\mathbf{w})|\mathbf{y};\boldsymbol{\theta}_t\right] - A'(\boldsymbol{\theta}) &= 0 \\ \Rightarrow \mathrm{E}_{\mathbf{w}}\left[\boldsymbol{\theta}\cdot\mathbf{T}(\mathbf{y},\mathbf{w})|\mathbf{y};\boldsymbol{\theta}_t\right] &= \mathrm{E}[\mathbf{T}(\mathbf{y},\mathbf{w});\boldsymbol{\theta}]. \end{aligned}$$

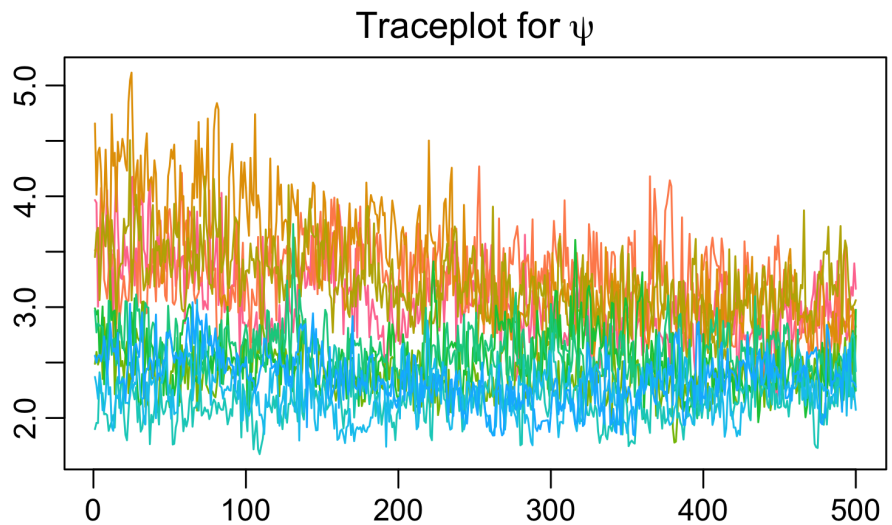
Full Bayesian approach

- Assign prior distributions to the parameters, for example
  - $\alpha \sim N(a, b^2)$
  - $\lambda \sim U(0, c)$
  - $\psi \sim \Gamma(d, e)$

- Draw from the posterior densities $f(\boldsymbol{\theta}|\mathbf{y})$ using Metropolis-Hastings algorithm. Estimates for the parameters are the posterior means.

- Easy to implement in `R` using JAGS (`rjags` or `R2jags`), but...





Example: Simple linear regression

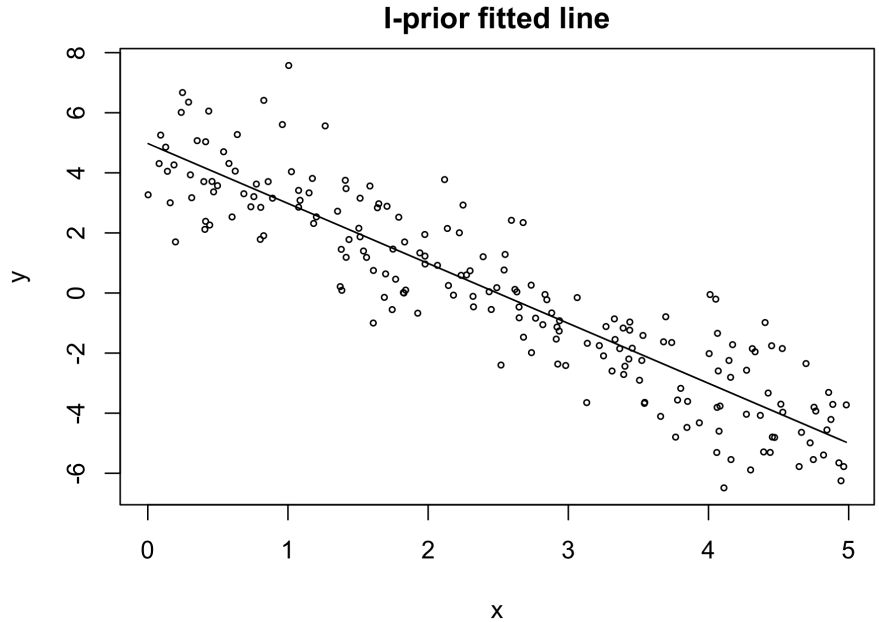
Classical model

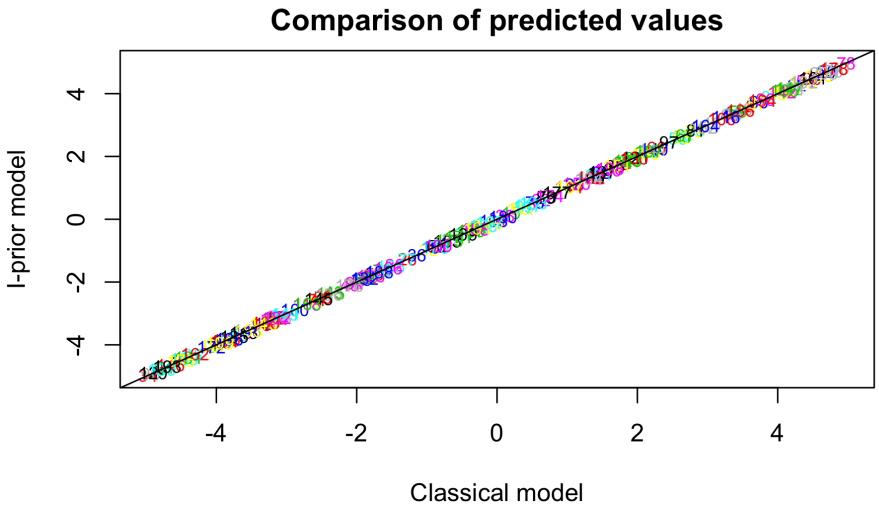
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma)$$

I-prior model

$$y_i = \alpha + \sum_{k=1}^{n} h_\lambda(x_i, x_k) w_k + \epsilon_i$$
$$\epsilon_i \sim N(0, \psi^{-1})$$
$$w_i \sim N(0, \psi)$$

h_λ is the Canonical kernel





MSE(classical) = 1.770 MSE(I-prior) = 1.770

Example: 1-dimensional smoothing

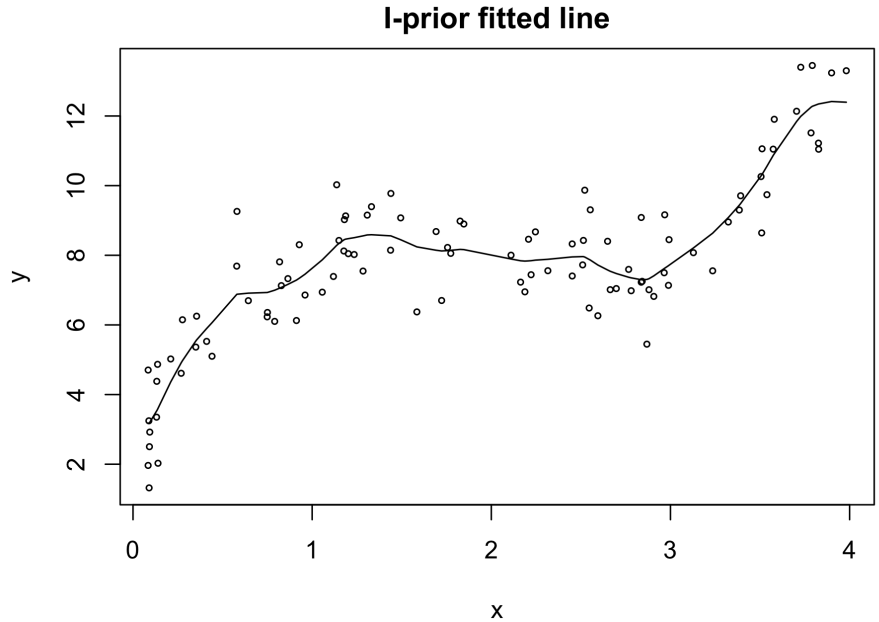
Classical model

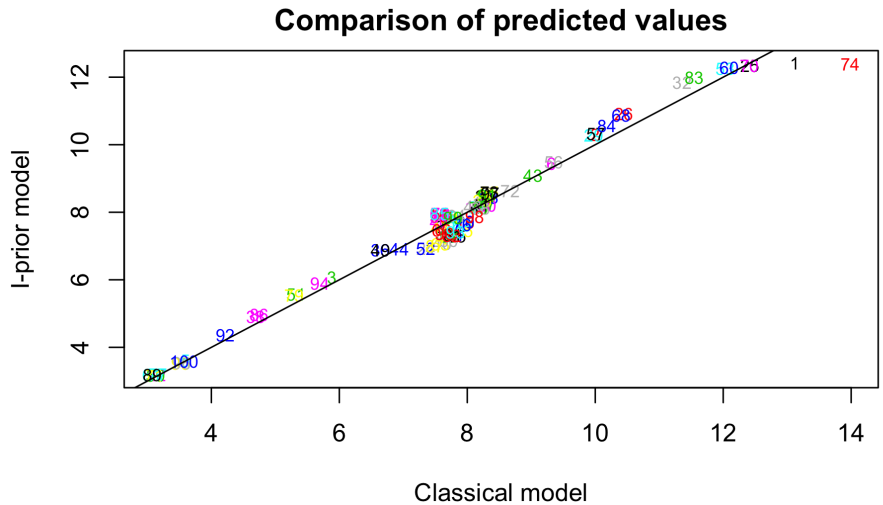
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$$
$$\epsilon_i \sim N(0, \sigma)$$

I-prior model

$$y_i = \alpha + \sum_{k=1}^{n} h_{\lambda,\gamma}(x_i, x_k) w_k + \epsilon_i$$
$$\epsilon_i \sim N(0, \psi^{-1})$$
$$w_i \sim N(0, \psi)$$

$h_{\lambda,\gamma}$ is the FBM kernel





MSE(classical) = 0.987 MSE(I-prior) = 0.836

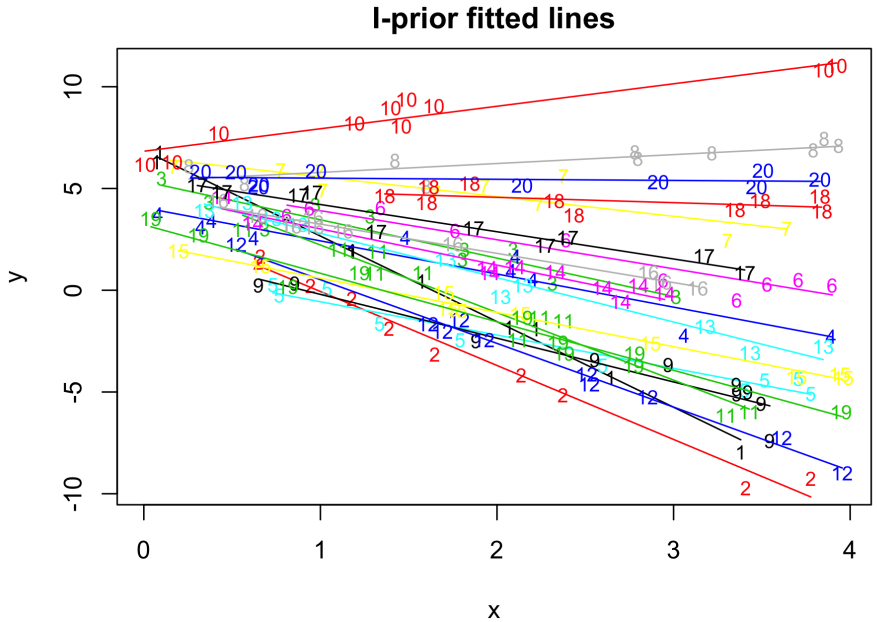
Example: Multilevel modelling

Classical model

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + \epsilon_{ij}$$

$$\begin{pmatrix}\beta_{0j}\\ \beta_{1j}\end{pmatrix} \sim N\left(\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \begin{pmatrix}\phi_0 & \phi_{01}\\ \phi_{01} & \phi_1\end{pmatrix}\right)$$

$$\epsilon_{ij} \sim N(0, \sigma)$$



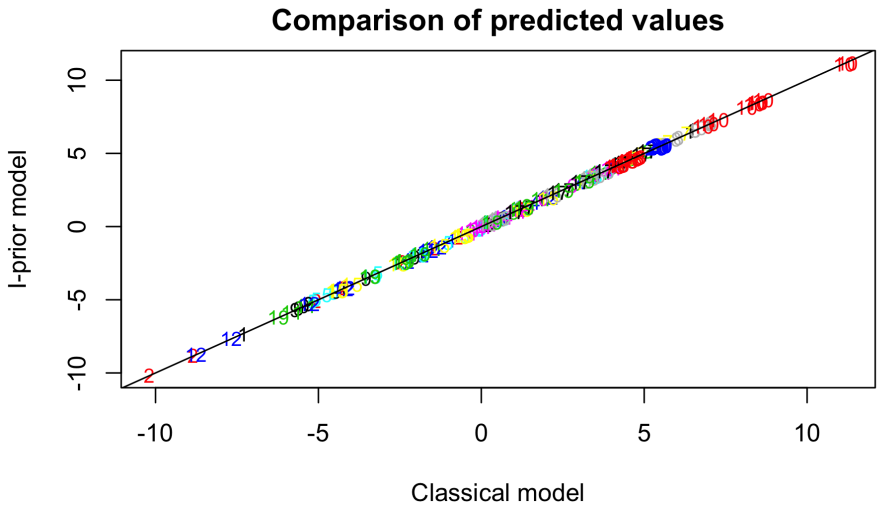
I-prior model

$$y_i = \alpha + \sum_{k=1}^{n} h_\lambda(x_i, x_k)w_k + \epsilon_i$$

$$\epsilon_i \sim N(0, \psi^{-1})$$

$$w_i \sim N(0, \psi)$$

h_λ is the ANOVA kernel



MSE(classical) = 0.227 MSE(I-prior) = 0.226

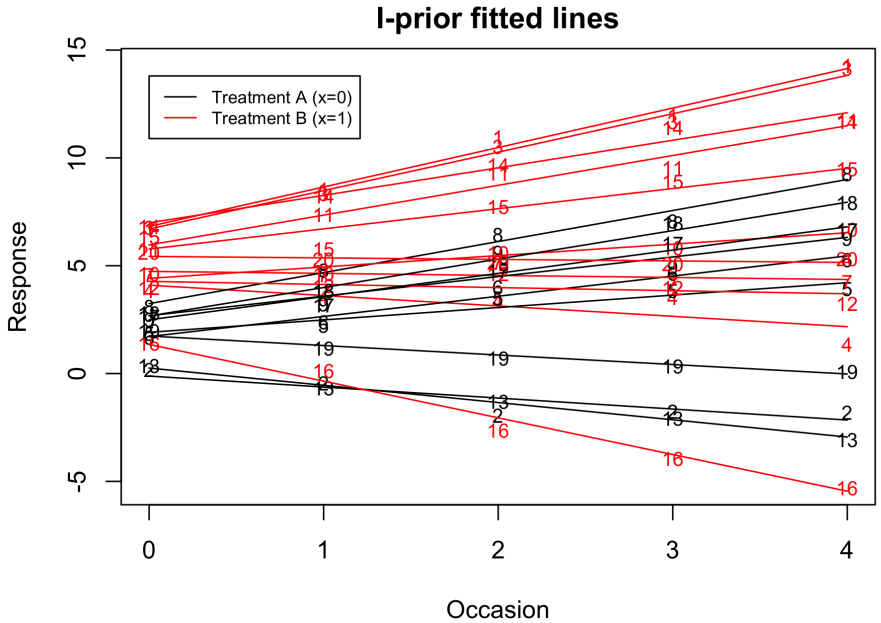
Example: Longitudinal modelling

Classical model

$$y_{ij} = \beta_{0j} + \beta_{1j} t_{ij} + \beta_3 x_{ij} + \epsilon_{ij}$$

$$\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} \phi_0 & \phi_{01} \\ \phi_{01} & \phi_1 \end{pmatrix} \right)$$

$$\epsilon_{ij} \sim N(0, \sigma)$$



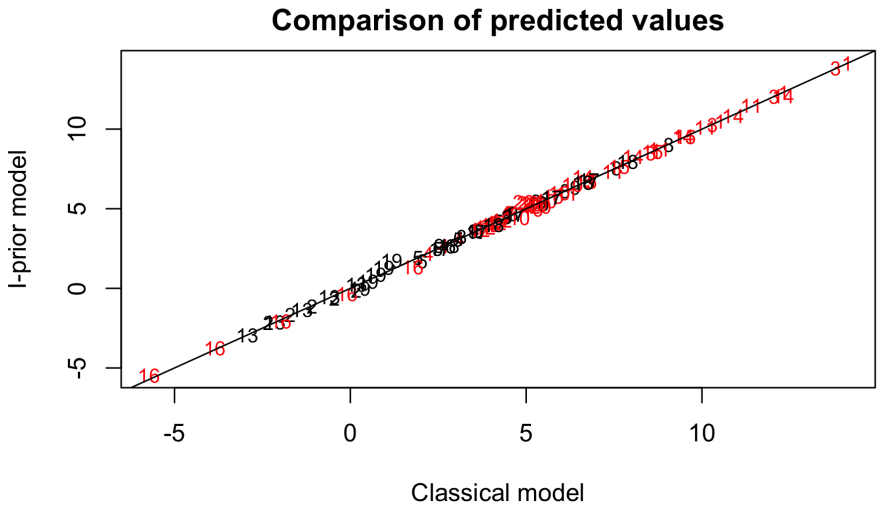
I-prior model

$$y_i = \alpha + \sum_{k=1}^{n} h_\lambda(x_i, x_k) w_k + \epsilon_i$$

$$\epsilon_i \sim N(0, \psi^{-1})$$

$$w_i \sim N(0, \psi)$$

h_λ is the ANOVA + Pearson kernel



MSE(classical) = 0.138 MSE(I-prior) = 0.114

Further work: Structural Equation Models

- The 1-factor model

$$x_{ij} = \mu_j + \lambda_j f_i + \delta_{ij}$$
$$f_i \sim N(0, 1)$$
$$\delta_{ij} \sim N(0, \theta_j)$$

- Relationship to longitudinal random intercept model:
  - Set $\mu_j = \mu$, $\forall j$.
  - Set $\lambda_j = 1$, , $\forall j$ and estimate variance of f_i instead.
  - Set $\theta_j = \theta$, $\forall j$ We already know how to estimate this model using I-prior.

- Further work:
  - Uses of this very restricted CFA model? Rasch model?
  - Post estimation work, e.g. obtaining factor scores.
  - Can we estimate both the λ_js and f_i simultaneously?

Further work: Structured error covariances

- Sometimes, the responses may be correlated in a way that the model specification can't account for completely. Extend model to allow for dependence between errors, such as autocorrelations.
- Example: AR(1) covariance matrix with equal gaps between observations:

$$\boldsymbol{\Psi} = \frac{\sigma^2}{1-\phi^2}\begin{pmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{n-1} \\ \phi & 1 & \phi & \cdots & \phi^{n-2} \\ \phi^2 & \phi & 1 & \cdots & \phi^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \phi^{n-3} & \cdots & 1 \end{pmatrix}$$

- Others: Heteroskedastic errors?

Further work: Logistic models

- Extending the I-prior methodology to GLMs, e.g. logit models:

$$y_i \sim \text{Bern}(\pi_i)$$

$$\text{logit } \pi_i = \alpha + \sum_{k=1}^{n} h_\lambda(x_i, x_k) w_k$$

$$w_i \sim N(0, \pi_i(1-\pi_i))$$

$$i = 1, \ldots, n$$

  i.e. putting an I-prior on the linear predictor, and setting the Fisher information as the variance.

- Difficulties faced
  - Unable to estimate this model using JAGS due to a circular dependence of the parameters.
  - Performing ML yields a high-dimensional intractable integral. Poor results from approximation methods like Laplace and Gauss-Hermite Quadrature.

Summary

- The I-prior methodology is a modelling technique that guards against overfitting linear models when dimensionality is large relative to sample size, with advantages such as
  - Model parsimony
  - Requires no additional assumptions
  - Simpler estimation
- Many models shown to work with using I-priors such as multiple regression, smoothing models, random effects models and growth curve models.
- Areas of research include
  - Extension to GLMs
  - Structural Equation Models
  - Models with structured error covariances

End

Thank you!