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

1. Introduction
2. I-prior theory
3. Estimation methods
4. Examples of I-prior modelling
   - Simple linear regression
   - 1-dimensional smoothing
   - Multilevel modelling
   - Longitudinal modelling
5. Further work
   - Structural Equation Models
   - Models with structured error covariances
   - Logistic models
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)*

\[ y_i = f(x_i) + \epsilon_i \]

\[ y_i \in \mathbb{R}, \text{ real-valued observations} \]

\[ x_i \in \mathcal{X}, \text{ a set of characteristics for unit } i \]

\[ f \in \mathcal{F}, \text{ a vector space of functions over the set } \mathcal{X} \]

\[ (\epsilon_1, \ldots, \epsilon_n) \sim \mathcal{N}(\mathbf{0}, \Psi^{-1}) \]

\[ i = 1, \ldots, n \]

Note: For iid observations, \( \Psi = \psi I_n \). In general, \( \Psi = (\psi_{ij}) \).
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 $\pi$ 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

\[ y = \alpha + X\beta + \epsilon \]
\[ \epsilon \sim N(0, \psi^{-1}I_n) \]

We know from linear regression theory that \( I[\beta] = \psi X^T X \). An I-prior on \( \beta \) is then

\[ \beta \sim N(\beta_0, \lambda^2 \psi X^T X) \]

Equivalently,

\[ \beta = \beta_0 + \lambda X^T w \]
\[ w \sim N(0, \psi I_n) \]

Thus, an I-prior on \( f \) is

\[ f = \alpha + X\beta_0 + \lambda XX^T w \]
\[ w \sim N(0, \psi I_n) \]
I-prior theory

Functional vector spaces
Reproducing kernels
Hilbert spaces
Krein spaces
Kernel methods
Fisher Information
Means of random functions
Feature maps
Variances of random functions
Random functions
Gaussian random vectors
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 $\langle f, f \rangle_{\mathcal{F}} \geq 0$ ($\leq 0$). 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 $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a kernel if there exists a Hilbert space $\mathcal{F}$ and a map $\phi : \mathcal{X} \rightarrow \mathcal{F}$ 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 $h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a reproducing kernel of $\mathcal{F}$, and $\mathcal{F}$ a RKHS/RKKS, if $h$ satisfies
  1. $\forall x \in \mathcal{X}, h(\cdot, x) \in \mathcal{F}$
  2. $\forall x \in \mathcal{X}, f \in \mathcal{F}, \langle f, h(\cdot, x) \rangle_{\mathcal{F}} = f(x)$.

• 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 \( h : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R} \). 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 \( \pi \) be a Gaussian I-prior on \( f \) with prior mean \( f_0 \) and variance \( I[f] \). Then \( \pi \) is called an I-prior for \( f \), and a random vector \( f \sim \pi \) has the random effect representation

  \[
  f(x_i) = f_0(x_i) + \sum_{k=1}^{n} h(x_i, x_k) w_k
  \]

  \((w_1, \ldots, w_n) \sim N(0, \Psi)\).
Back to the multiple regression example

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

\[
\begin{align*}
  f(x_i) &= f_0(x_i) + \alpha + x_i \beta_0 + \sum_{k=1}^n h(x_i,x_k) w_k \\
  w &:= (w_1, \ldots, w_n) \sim N(0, \psi I_n).
\end{align*}
\]
We saw the I-prior method applied to multiple regression:

\[
    f(x_i) = \underbrace{f_0(x_i)}_{\alpha + x_i\beta_0} + \underbrace{\lambda(XX^T)w}_{\sum_{k=1}^n h(x_i, x_k)w_k}
\]

\[
    w := (w_1, \ldots, w_n) \sim N(0, \psi I_n).
\]

Choose different RKHS/RKKS $\mathcal{F}$ and corresponding $h$ to suit the type/characteristic of the $x$s in order to do regression modelling.
We saw the I-prior method applied to multiple regression:

\[
f(x_i) = f_0(x_i) + \sum_{k=1}^{n} h(x_i, x_k)w_k + \lambda(XX^T)w
\]

where \( x := (w_1, \ldots, w_n) \sim N(0, \Psi I_n) \).

Choose different RKHS/RKKS \( \mathcal{F} \) and corresponding \( h \) to suit the type/characteristic of the \( x \)s in order to do regression modelling.
**Toolbox of RKHS/RKKS**

<table>
<thead>
<tr>
<th>$\mathcal{X} = {x_i}$</th>
<th>Characteristic/Uses</th>
<th>Vector space $\mathcal{F}$</th>
<th>Kernel $h(x_i, x_k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>1) Categorical covariates; 2) In a multilevel setting, $x_i =$ group no. of unit $i$.</td>
<td>Pearson</td>
<td>$\mathbb{1}_{[x_i=x_k]} - 1$ where $p_i = \mathbb{P}[X = x_i]$</td>
</tr>
<tr>
<td>Real</td>
<td>As in classical regression, $x_i =$ real-valued covariate associated with unit $i$.</td>
<td>Canonical</td>
<td>$x_ix_k$</td>
</tr>
<tr>
<td>Real</td>
<td>As in (1-dim) smoothing, $x_i =$ data point associated with observation $y_i$.</td>
<td>Fractional Brownian Motion (FBM)</td>
<td>$</td>
</tr>
</tbody>
</table>

- 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

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

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

\[ w_i \sim N(0, \psi) \]

\[ i = 1, \ldots, n, \]

(2)

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

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

- For the I-prior model in (2), treat the $w_i$s as missing.

- The distributions are easy enough to obtain:
  - $\mathbf{y} \sim N(\alpha, \mathbf{V}_y)$, where $\mathbf{V}_y := \mathbf{H}_\lambda \mathbf{\Psi} \mathbf{H}_\lambda + \mathbf{\Psi}^{-1}$
  - $\mathbf{w} \sim N(\mathbf{0}, \mathbf{\Psi})$
  - $\begin{pmatrix} \mathbf{y} \\ \mathbf{w} \end{pmatrix} \sim N \left( \begin{pmatrix} \alpha \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{V}_y & \mathbf{H}_\lambda \mathbf{\Psi} \\ \mathbf{\Psi} \mathbf{H}_\lambda & \mathbf{\Psi} \end{pmatrix} \right)$
  - $\mathbf{w} | \mathbf{y} \sim N \left( \mathbf{H}_\lambda \mathbf{V}_y^{-1} (\mathbf{y} - \alpha), \mathbf{V}_y^{-1} \right)$ where $\mathbf{H}_\lambda(i, j) = h_\lambda(x_i, x_j)$ and $\mathbf{\Psi} = \mathbf{\psi} \mathbf{I}_n$.

- E-step: Calculate $Q(\theta) = \mathbb{E}_{\mathbf{w}} [\log f(\mathbf{y}, \mathbf{w}; \theta) | \mathbf{y}; \theta_t]$.

- M-step: $\theta_{t+1} \leftarrow \arg \max_\theta Q(\theta)$. 
Generalised least square estimator for $\alpha$

- Write Model (2) as $\mathbf{y} = \alpha \mathbf{1} + \mathbf{H}_\lambda \mathbf{w} + \mathbf{\epsilon}$, where $\mathbf{y} \sim N(\alpha, \mathbf{V}_y)$.

- Assume values for $\lambda$ and $\psi$ are known, and thus too $\mathbf{V}_y(\lambda, \psi) = \mathbf{H}_\lambda \Psi \mathbf{H}_\lambda + \Psi^{-1}$.

- GLS estimator for $\alpha$ 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 $f_X(x; \theta) = \exp[\theta \cdot T(x) - A(\theta)]h(x)$.
  
  ▶ The MLE is found by solving the set of equations $T(x) = A'(\theta)$.
  
  ▶ It is also know that $A'(\theta) = E[T(x); \theta]$.

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

$$Q(\theta) = E_w[\theta \cdot T(y, w) - A(\theta) + \log h(y, w)|y; \theta_t].$$

• Maximising this over $\theta$, we arrive at the FOC

$$Q'(\theta) = E_w[\theta \cdot T(y, w)|y; \theta_t] - A'(\theta) = 0$$

$$\Rightarrow E_w[\theta \cdot T(y, w)|y; \theta_t] = E[T(y, w); \theta].$$
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 $f(\theta|y)$ 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**

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

\( h_\lambda \) is the Canonical kernel

\[ \text{MSE(classical)} = 1.770 \]
\[ \text{MSE(I-prior)} = 1.770 \]
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

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

\( h_\lambda \) 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**

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

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

\[
w_i \sim N(0, \psi)
\]

\(h_\lambda\) is the ANOVA + Pearson kernel

---

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

---

Haziq Jamil (LSE)  
Haziq Jamil (LSE)
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 \( \lambda_j \)s 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:

\[ \Psi = \frac{\sigma^2}{1 - \phi^2} \begin{pmatrix} 1 & \phi & \phi^2 & \ldots & \phi^{n-1} \\ \phi & 1 & \phi & \ldots & \phi^{n-2} \\ \phi^2 & \phi & 1 & \ldots & \phi^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \phi^{n-3} & \ldots & 1 \end{pmatrix} \]

- 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
Thank you!