To-do list

Contents

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
March 12, 2018

Chapter 4

Regression modelling using I-priors

In the previous chapter, we defined an I-prior for the normal regression model (1.1) subject to (1.2) and f belonging to a reproducing kernel Hilbert or Krein space of functions. We also saw how new function spaces can be constructed via the polynomial and ANOVA RKKS. In this chapter, we shall describe various regression models, and connect them to an appropriate RKKS, so that an I-prior may be defined on it. Methods for estimating I-prior models will also be described. Finally, several examples of I-prior modelling are presented.

4.1 Various regression models

Without loss of generality, we assume a prior mean of zero for the I-prior distribution.

4.1.1 Multiple linear regression

Let $\mathcal{X} \equiv \mathbb{R}^p$ equipped with the regular dot product, and $\mathcal{F}_\lambda$ the scaled canonical RKHS of functions over $\mathcal{X}$ with kernel $h_\lambda(\mathbf{x}, \mathbf{x}') = \lambda \mathbf{x}^\top \mathbf{x}'$. Then, an I-prior on f implies that

$$\begin{aligned} f(\mathbf{x}_i) &= \sum_{j=1}^{n} \lambda \mathbf{x}_i^\top \mathbf{x}_j w_j \\ &= \sum_{j=1}^{n} \lambda \sum_{k=1}^{p} x_{ik} x_{jk} w_j \\ &= \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \end{aligned}$$

where each $\beta_k = \lambda \sum_{j=1}^{n} x_{jk} w_j$. This implies a multivariate normal prior distribution for the regression coefficients

$$\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p) \sim \mathrm{N}_p(\mathbf{0}, \lambda^2 \mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X}),$$

where $\mathbf{X}$ is the $n \times p$ design matrix for the covariates, excluding the column of ones at the beginning typically reserved for the intercept. The covariance matrix for $\boldsymbol{\beta}$ is recognised as the scaled Fisher information matrix for the regression coefficients.

The I-prior for $\boldsymbol{\beta}$ resembles the objective g-prior (Zellner, 1986) for regression coefficients,

$$\boldsymbol{\beta} \sim \mathrm{N}_p\left(\mathbf{0}, g(\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1}\right),$$

although they are quite different objects. The g-prior for $\boldsymbol{\beta}$ has the *inverse* (scaled) Fisher information matrix as its covariance matrix. This, in itself, has a much different and arguably counterintuitive meaning: large amounts of Fisher information about $\boldsymbol{\beta}$ corresponds to a small prior variance, and hence less deviation away from the prior mean of zero in estimating $\boldsymbol{\beta}$. The choice of the hyperparameter g has been the subject of much debate, with choices ranging from fixing $g = n$ (corresponding to the concept of *unit Fisher information*), to fully Bayesian and empirical Bayesian methods of estimating g from the data.

On the other hand, we note that the g-prior has an I-prior interpretation when argues as follows. Assume that the regression function f lies in the continual dual space of $\mathbb{R}^p$ equipped with the inner product $\langle \mathbf{x}, \mathbf{x}' \rangle_{\mathcal{X}} = \mathbf{x}^\top (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1} \mathbf{x}$. With this inner product and from (3.3), the Fisher information for $\boldsymbol{\beta}$ is

$$\begin{aligned}
\mathcal{I}_g(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1} \mathbf{x}_i \otimes (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1} \mathbf{x}_j \\
&= (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1} \cancel{(\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})} \cancel{(\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1}} \\
&= (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1},
\end{aligned}$$

and this rather than the usual $\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X}$ as the prior covariance matrix for $\boldsymbol{\beta}$ means that the I-prior is in fact the standard g-prior.

The metric induced by the inner product is actually the *Mahalanobis distance*, a scale-invariant natural distance if the covariates are measured on different scales. To expand on this idea, circle back to the regression function and write it as $f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta} \rangle_{\mathcal{X}}$. In

usual least squares regression, the choice of inner product is irrelevant, so the usual dot product is commonly used (however, as we have seen above, the choice of inner product determines the form of the Fisher information for $\boldsymbol{\beta}$). In particular, suppose that all the x_{ik}'s, $k = 1, \ldots, p$ for each unit $i = 1, \ldots, n$ are measured on the same scale; for instance, these could be measurements in centimetres. In this case, the dot product is reasonable, because $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \sum_{k=1}^{p} x_{ik} x_{jk}$ and the inner product has a coherent unit, namely the squared unit of the x_{ik}'s. However, if they were a mix of various scaled measurements, then obviously the inner product's unit is incoherent—one would be resorted to adding measurements in different units, for example, cm^2 and kg^2 and so on. In such a case, a unitless inner product is appropriate, like the Mahalonobis inner product, which technically rescales the x_{ik}'s to unity. In summary, if the covariates are all measured on the same scale, then the I-prior is appropriate, and if not, the g-prior is appropriate.

2. Can I just standardise x?

A different approach for covariate measurements of differing scales, without resorting to g-priors, is the ANOVA approach. By considering only the main effects, one decomposes the regression function into

$$f(\mathbf{x}_i) = \beta_0 + f_1(x_{i1}) + \cdots + f_p(x_{ip})$$

for which $(f - \beta_0) \in \mathcal{F}_\lambda \equiv \mathcal{F}_{\lambda_1} \oplus \cdots \oplus \mathcal{F}_{\lambda_p}$, and $\mathcal{F}_{\lambda_k}, k = 1, \ldots, p$ are unidimensional centred canonical RKKSs with kernels $h_{\lambda_k}(x_{ik}, x_{jk}) = \lambda_k (x_{ik} - \bar{x}_k)(x_{jk} - \bar{x}_k)$, where $\bar{x}_k = \sum_{i=1}^{n} x_{ik}/n$. In effect, we now have p scale parameters, one for each of the RKKSs associated with the p covariates. Denote $\tilde{x}_{ik} = x_{ik} - \bar{x}_k$, the centred covariates. The RKKS $\mathcal{F}_\lambda$ therefore has the kernel

$$h(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} \lambda_k \tilde{x}_{ik} \tilde{x}_{jk},$$

and hence each regression coefficient can now be written as $\beta_k = \sum_{j=1}^{n} \lambda_k \tilde{x}_{jk} w_j$. Thus, the corresponding I-prior for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \sim \mathrm{N}_p(\mathbf{0}, \lambda^2 \tilde{\mathbf{X}}^\top \boldsymbol{\Lambda} \boldsymbol{\Psi} \boldsymbol{\Lambda} \tilde{\mathbf{X}}),$$

with $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$. Note that the overall effect β_0 can be treated in a number of ways, the simplest of which is as an intercept to be estimated. This approach is disadvantageous when p is large, in which case there would be numerous scale parameters to estimate.

4.1.2 Multilevel linear modelling

Let $\mathcal{X} \equiv \mathbb{R}^p$, and suppose that alongside the covariates, there is information on group levels $\mathcal{M} = \{1, \dots, m\}$ for each unit i. That is, every observation for unit i is known to belong to a specific group j, and we write $\mathbf{x}_i^{(j)}$ to indicate this. Let n_j denote the sample size for cluster j, and the overall sample size be $n = \sum_{j=1}^m n_j$. When modelled linearly with the responses $y_i^{(j)}$, the model is known as a multilevel (linear) model, although it is known by many other names: random-effects models, random coefficient models, hierarchical models, and so on. As this model is seen as an extension of linear models, applications are plenty, especially in research designs for which the data varies at more than one level.

Consider a functional ANOVA decomposition of the regression function as follows:

$$f(\mathbf{x}_i^{(j)}, j) = f_0 + f_1(\mathbf{x}_i^{(j)}) + f_2(j) + f_{12}(\mathbf{x}_i^{(j)}, j). \tag{4.1}$$

To mimic the multilevel model, assume $f_1 \in \mathcal{F}_1$ the Pearson RKHS, $f_2 \in \mathcal{F}_2$ the centred canonical RKHS, and $f_{12} \in \mathcal{F}_{12} = \mathcal{F}_1 \otimes \mathcal{F}_2$, the tensor product space of $\mathcal{F}_1$ and $\mathcal{F}_2$. As we know, f_0 is the overall intercept, and the varying intercepts are given by the function f_2. f_1 is the (main) linear effect of the covariates, while f_{12} provides the varying linear effect per group of the covariates. The I-prior for $f - f_0$ is assumed to lie in the function space $\mathcal{F} - f_0$, which is an ANOVA RKKS with kernel

$$h_\lambda\big((\mathbf{x}_i^{(j)}, j), (\mathbf{x}_i^{(j')}, j')\big) = \lambda_1 h_1(\mathbf{x}_i^{(j)}, \mathbf{x}_{i'}^{(j')}) + \lambda_2 h_2(j, j') + \lambda_1\lambda_2 h_1(\mathbf{x}_i^{(j)}, \mathbf{x}_{i'}^{(j')}) h_2(j, j'),$$

with h_1 the centred canonical kernel and h_2 the Pearson kernel.

We can show that the regression function (4.1) corresponds to the standard way of writing the multilevel model,

$$f(\mathbf{x}_i^{(j)}, j) = \beta_0 + \mathbf{x}_i^{(j)\top}\boldsymbol{\beta}_1 + \beta_{0j} + \mathbf{x}_i^{(j)\top}\boldsymbol{\beta}_{1j}.$$

and determine the prior distributions on $(\beta_{0j}, \boldsymbol{\beta}_{1j})^\top$. Write $f_0 = \beta_0$, and for simplicity, assume iid errors, i.e., $\boldsymbol{\Psi} = \psi\mathbf{I}_n$. The form of $f \in \mathcal{F}$ is now $f(\mathbf{x}_i^{(j)}, j) = \sum_{i'=1}^{n_{j'}} \sum_{j'=1}^m h_\lambda\big((\mathbf{x}_i^{(j)}, j), (\mathbf{x}_{i'}^{(j')}, j')\big) w_{i'j'}$, where each $w_{i'j'} \sim \mathrm{N}(0, \psi^{-1})$. We have seen from the previous section that $f_1(\mathbf{x}_i^{(j)}) = \tilde{\mathbf{x}}_i^{(j)\top}\boldsymbol{\beta}$, with $\boldsymbol{\beta} = \lambda_1\tilde{\mathbf{X}}^\top\mathbf{w} \sim \mathrm{N}_p(\mathbf{0}, \lambda_1^2\psi\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})$. Here, $\tilde{\mathbf{X}}$ is the $(n_1 + \cdots + n_m) \times p$ matrix containing centred entries of $\mathbf{x}_i^{(j)}$. Now, functions

in the scaled RKHS $\mathcal{F}_2$ have the form

$$\begin{aligned} f_2(j) &= \sum_{i=1}^{n_{j'}} \sum_{j'=1}^{m} \lambda_2 \left(\frac{\delta_{jj'}}{p_j} - 1 \right) w_{ij'} \\ &= \lambda_2 \left(\frac{w_{+j}}{p_j} - w_{++} \right), \end{aligned}$$

where a '+' in the index of w_{ik} indicates a summation over that index, and p_j is the empirical distribution over $\mathcal{M}$, i.e. $p_j = n_j/n$. Clearly $f_2(j)$ is a variable depending on j, so write $f_2(j) = \beta_{0j}$. The distribution of β_{0j} is normal with zero mean and variance

$$\begin{aligned} \operatorname{Var} \beta_{0j} &= \lambda_2^2 \left(\frac{\cancel{n_j}\psi}{\cancel{n_j^2}/n^2} + n\psi \right) \\ &= n\psi\lambda_2^2 \left(\frac{1}{p_j} + 1 \right). \end{aligned}$$

The covariance between any two random intercepts β_{0j} and $\beta_{0j'}$ is

$$\begin{aligned} \operatorname{Cov}(\beta_{0j}, \beta_{0j'}) &= \operatorname{Cov}\left(\lambda_2 \left(\frac{w_{+j}}{p_j} - w_{++} \right), \lambda_2 \left(\frac{w_{+j'}}{p_{j'}} - w_{++} \right) \right) \\ &= \frac{\lambda_2^2}{p_j p_{j'}} \cancel{\operatorname{Cov}(w_{+j}, w_{+j'})}^{\,0} - \frac{\lambda_2^2}{p_j} \operatorname{Cov}(w_{+j}, w_{++}) - \frac{\lambda_2^2}{p_{j'}} \operatorname{Cov}(w_{++}, w_{+j'}) \\ &\quad + \lambda_2^2 \operatorname{Cov}(w_{++}, w_{++}) \\ &= -\frac{\lambda_2^2}{\cancel{n_j}/n} \cancel{n_j}\psi - \frac{\lambda_2^2}{\cancel{n_{j'}}/n} \cancel{n_{j'}}\psi + \lambda_2^2 n\psi \\ &= -n\psi\lambda_2^2. \end{aligned}$$

Functions in $\mathcal{F}_{12}$, on the other hand, have the form

$$\begin{aligned} f_{12}(\mathbf{x}_i, j) &= \sum_{i'=1}^{n_{j'}} \sum_{j'=1}^{m} \lambda_1\lambda_2 \cdot \tilde{\mathbf{x}}_i^{(j)\top} \tilde{\mathbf{x}}_{i'}^{(j')} \cdot \left(\frac{\delta_{jj'}}{p_j} - 1 \right) w_{i'j'} \\ &= \tilde{\mathbf{x}}_i^{(j)\top} \underbrace{\left(\frac{\lambda_1\lambda_2}{p_j} \sum_{i'=1}^{n_j} \tilde{\mathbf{x}}_{i'}^{(j)} w_{i'j} - \lambda_1\lambda_2 \sum_{i'=1}^{n_{j'}} \sum_{j'=1}^{m} \tilde{\mathbf{x}}_{i'}^{(j')} w_{i'j'} \right)}_{\boldsymbol{\beta}_{1j}}, \end{aligned}$$

and this is, as expected, a linear form dependent on cluster j. We can calculate the

variance for β_{1j} to be

$$\begin{aligned}
\operatorname{Var} \boldsymbol{\beta}_{1j} &= \lambda_1^2 \lambda_2^2 \operatorname{Var} \left(\frac{1}{p_j} \tilde{\mathbf{X}}_j^\top \mathbf{w}_j - \tilde{\mathbf{X}}^\top \mathbf{w} \right) \\
&= \lambda_1^2 \lambda_2^2 \left(\frac{\psi}{n_j^2/n^2} \tilde{\mathbf{X}}_j^\top \tilde{\mathbf{X}}_j + \psi \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} - \frac{1}{p_j} \tilde{\mathbf{X}}_j^\top \operatorname{Cov}(\mathbf{w}_j, \mathbf{w}) \tilde{\mathbf{X}}^\top \right) \\
&= n\psi \lambda_1^2 \lambda_2^2 \left(\frac{1}{p_j} \mathbf{S}_j + \mathbf{S} - \mathbf{S}_j \right) \\
&= n\psi \lambda_1^2 \lambda_2^2 \left(\left(\frac{1}{p_j} - 1 \right) \mathbf{S}_j + \mathbf{S} \right)
\end{aligned}$$

where $\mathbf{S}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} (\mathbf{x}_i^{(j)} - \bar{\mathbf{x}})^\top (\mathbf{x}_i^{(j)} - \bar{\mathbf{x}})$, $\mathbf{S} = \frac{1}{n} \sum_{i=1}^{n_j} \sum_{j=1}^{m} (\mathbf{x}_i^{(j)} - \bar{\mathbf{x}})^\top (\mathbf{x}_i^{(j)} - \bar{\mathbf{x}})$, and $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n_j} \sum_{j=1}^{m} \mathbf{x}_i^{(j)}$. The covariance between two vectors of the random slopes is

$$\begin{aligned}
\operatorname{Cov}(\boldsymbol{\beta}_{1j}, \boldsymbol{\beta}_{1j'}) &= \lambda_1^2 \lambda_2^2 \operatorname{Cov} \left(\frac{1}{p_j} \tilde{\mathbf{X}}_j^\top \mathbf{w}_j - \tilde{\mathbf{X}}^\top \mathbf{w}, \frac{1}{p_{j'}} \tilde{\mathbf{X}}_{j'}^\top \mathbf{w}_{j'} - \tilde{\mathbf{X}}^\top \mathbf{w} \right) \\
&= \psi \lambda_1^2 \lambda_2^2 \left(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} - \frac{1}{p_j} \tilde{\mathbf{X}}_j^\top \tilde{\mathbf{X}}_j - \frac{1}{p_{j'}} \tilde{\mathbf{X}}_{j'}^\top \tilde{\mathbf{X}}_{j'} \right) \\
&= n\psi \lambda_1^2 \lambda_2^2 \left(\mathbf{S} - \mathbf{S}_j - \mathbf{S}_{j'} \right).
\end{aligned}$$

Another quantity of interest is the covariance between the random intercepts and random slopes:

$$\begin{aligned}
\operatorname{Cov}(\beta_{0j}, \boldsymbol{\beta}_{1j}) &= \lambda_1 \lambda_2^2 \operatorname{Cov} \left(\frac{1}{p_j} \mathbf{1}_{n_j}^\top \mathbf{w}_j - \mathbf{1}_n^\top \mathbf{w}, \frac{1}{p_j} \tilde{\mathbf{X}}_j^\top \mathbf{w}_j - \tilde{\mathbf{X}}^\top \mathbf{w} \right) \\
&= \psi \lambda_1 \lambda_2^2 \left(\cancel{\mathbf{1}_n^\top \tilde{\mathbf{X}}}^{\,0} + \frac{1}{p_j^2} \mathbf{1}_{n_j}^\top \tilde{\mathbf{X}}_j - \frac{2}{p_j} \mathbf{1}_{n_j}^\top \tilde{\mathbf{X}}_j \right) \\
&= n\psi \lambda_1 \lambda_2^2 \left(\left(\frac{1}{p_j} - 2 \right) \frac{1}{n_j} \sum_{i=1}^{n_j} (\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}) \right) \\
&= n\psi \lambda_1 \lambda_2^2 \left(\frac{1}{p_j} - 2 \right) (\bar{\mathbf{x}}^{(j)} - \bar{\mathbf{x}})
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{Cov}(\beta_{0j}, \boldsymbol{\beta}_{1j'}) &= \lambda_1\lambda_2^2 \operatorname{Cov}\left(\frac{1}{p_j}\mathbf{1}_{n_j}^\top \mathbf{w}_j - \mathbf{1}_n^\top \mathbf{w}, \frac{1}{p_{j'}}\tilde{\mathbf{X}}_{j'}^\top \mathbf{w}_{j'} - \tilde{\mathbf{X}}^\top \mathbf{w}\right) \\
&= \psi\lambda_1\lambda_2^2 \left(\cancelto{0}{\mathbf{1}_n^\top \tilde{\mathbf{X}}} + \frac{1}{p_j p_{j'}}\mathbf{1}_{n_j}^\top \cancelto{0}{\operatorname{Cov}(\mathbf{w}_j, \mathbf{w}_{j'})} \tilde{\mathbf{X}}_{j'} - \frac{1}{p_j}\mathbf{1}_{n_j}^\top \tilde{\mathbf{X}}_j - \frac{1}{p_{j'}}\mathbf{1}_{n_{j'}}^\top \tilde{\mathbf{X}}_{j'}\right) \\
&= n\psi\lambda_1\lambda_2^2 \left(-\frac{1}{n_j}\sum_{i=1}^{n_j}(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}) - \frac{1}{n_{j'}}\sum_{i=1}^{n_{j'}}(\mathbf{x}_i^{(j')} - \bar{\mathbf{x}})\right) \\
&= n\psi\lambda_1\lambda_2^2 \left(2\bar{\mathbf{x}} - \bar{\mathbf{x}}^{(j)} - \bar{\mathbf{x}}^{(j')}\right).
\end{aligned}$$

The standard multilevel random effects assumption is that $(\beta_{0j}, \boldsymbol{\beta}_{1j})^\top$ is normally distributed with mean zero and covariance matrix $\boldsymbol{\Phi}$ of dimensions $(p+1)\times(p+1)$. In total, there are $p+1$ regression coefficients and $(p+1)(p+2)/2$ covariance parameters in Φ to be estimated. In contrast, the I-prior model is parameterised by only two RKKS scale parameters and the error precision ψ. While the estimation procedure for $\boldsymbol{\Phi}$ in the standard multilevel model can result in non-positive covariance matrices, the I-prior model has the advantage that positive definiteness is taken care of automatically. This is seen from the calculations for $\operatorname{Var}\beta_{0j}$, $\operatorname{Var}\boldsymbol{\beta}_{1j}$ and the respective covariances. An example of multilevel modelling using I-priors is given in Section 4.3.1.

As a remark, the following regression functions are nested

- $f(\mathbf{x}_i^{(j)}, j) = f_0 + f_1(\mathbf{x}_i^{(j)}) + f_2(j)$ (random intercept model);
- $f(\mathbf{x}_i^{(j)}, j) = f_0 + f_1(\mathbf{x}_i^{(j)})$ (linear regression model);
- $f(\mathbf{x}_i^{(j)}, j) = f_0 + f_2(j)$ (ANOVA model);
- $f(\mathbf{x}_i^{(j)}, j) = f_0$ (intercept only model),

and thus one may compare likelihoods to ascertain the best fitting model. In addition, one may add flexibility to the model in two possible ways:

1. **More than two levels**. The model can be easily adjusted to reflect the fact that that the data is structured in a hierarchy containing three or more levels. For the three level case, let the indices $j \in \{1, \ldots, m_1\}$ and $k \in \{1, \ldots, m_2\}$ denote the two

levels, and simply decompose the regression function accordingly.

$$f(\mathbf{x}_i^{(j,k)}, j, k) = f_0 + f_1(\mathbf{x}_i^{(j,k)}) + f_2(j) + f_3(k) + f_{12}(\mathbf{x}_i^{(j,k)}, j) + f_{13}(\mathbf{x}_i^{(j,k)}, k) + f_{23}(j, k) + f_{123}(\mathbf{x}_i^{(j,k)}, j, k).$$

2. **Group-level covariates**. Suppose now we would like to add group-level covariates to the model, i.e., covariates $\mathbf{z}_j$ that only vary across groups. The regression function would then be

$$f(\mathbf{x}_i^{(j)}, j, \mathbf{z}_j) = f_0 + f_1(\mathbf{x}_i^{(j)}) + f_2(j) + f_3(\mathbf{z}_j) + f_{12}(\mathbf{x}_i^{(j)}, j) + f_{13}(\mathbf{x}_i^{(j)}, \mathbf{z}_j) + f_{23}(j, \mathbf{z}_j) + f_{123}(\mathbf{x}_i^{(j)}, j, \mathbf{z}_j).$$

   Not all of these terms need to be included. For example, excluding f_{23} would imply that the regression coefficient for $\mathbf{z}_j$ does not vary across groups.

4.1.3 Longitudinal modelling

Longitudinal or panel data observes covariate measurements $x_i \in \mathcal{X}$ and responses $y_i(t) \in \mathbb{R}$ for individuals $i = 1, \ldots, n$ across a time period $t \in \{1, \ldots, T\} =: \mathcal{T}$. Often, the time indexing set $\mathcal{T}$ may be unique to each individual i, so measurements for unit i happens across a time period $\{t_{i1}, \ldots, t_{iT_i}\} =: \mathcal{T}_i$—this is known as an unbalanced panel. It is also possible that covariate measurements vary across time too, so appropriately they are denoted $x_i(t)$. For example, $x_i(t)$ could be repeated measurements of the variable x_i at time point $t \in \mathcal{T}_i$. The relationship between the response variables $y_i(t)$ at time $t \in \mathcal{T}_i$ is captured through the equation

$$y_i(t) = f\big(x_i, t\big) + \epsilon_i(t)$$

where the distribution of $\boldsymbol{\epsilon}_i = \big(\epsilon_i(t_{i1}), \ldots, \epsilon_i(t_{iT_i})\big)^\top$ is Gaussian with mean zero and covariance matrix $\boldsymbol{\Psi}_i$. Assuming $\boldsymbol{\Psi}_i = \psi_i \mathbf{I}_{T_i}$ or even $\boldsymbol{\Psi}_i = \psi \mathbf{I}_{T_i}$ are perfectly valid choices, even though this seemingly ignores any time dependence between the observations. In reality, the I-prior induces time dependence of the observations via the kernels in the prior covariance matrix for f. Additionally, the random vectors $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\epsilon}_{i'}$ are assumed to be independent for any two distinct $i, i' \in \{1, \ldots, n\}$.

Using the functional ANOVA decomposition on the regression function, we obtain

$$f(x_i, t) = f_0 + f_1(x_i) + f_2(t) + f_{12}(x_i, t), \tag{4.2}$$

where f_0 is an overall constant, $f_1 \in \mathcal{F}_1$, $f_2 \in \mathcal{F}_2$, and $f_{12} \in \mathcal{F}_1 \otimes \mathcal{F}_2$. Choices for $\mathcal{F}_1$ and $\mathcal{F}_2$ are plentiful. In fact, any of the RKHS/RKKS described in Chapter 3 can be used to either model a linear dependence (canonical RKHS), nominal dependence (Pearson RKHS), polynomial dependence (polynomial RKKS) or smooth dependence (fBm or SE RKHS) on the x_i's and t's on f.

Remark 4.1. Although (4.2) is a special case of the multilevel model decomposition (4.1) for which $x_i = x_i(t)$ (time-varying covariates), it is different to how longitudinal models are normally treated in using random coefficients. As a multilevel model, longitudinal models treat the individuals as the groups or clusters (level two), and the time points as the various measurements within the clusters (level one).

4.1.4 Smoothing models

Assume that, up to a constant, the regression function lies in the scaled, centred fBm RKHS $\mathcal{F}$ of functions over $\mathcal{X} \equiv \mathbb{R}$ with Hurst index $1/2$. Thus, with a centring with respect to the empirical distribution of $\{x_1, \ldots, x_n\}$ and using the absolute norm on $\mathbb{R}$, $\mathcal{F}$ has kernel

$$h_\lambda(x, x') = \frac{\lambda}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \big(|x - x_i| + |x' - x_j| - |x - x'| - |x_i - x_j| \big).$$

According to van der Vaart and van Zanten (2008, Section 10), $\mathcal{F}$ contains functions possessing a square integrable weak derivative. The posterior mean of f based on an I-prior is then a (one-dimensional) smoother for the data.

With $w_k \overset{\text{iid}}{\sim} \mathrm{N}(0, \psi)$, functions in $\mathcal{F}$ are of the form

$$\begin{aligned} f(x) &= \sum_{k=1}^{n} h(x, x_k) w_k \\ &= \sum_{k=1}^{n} \left(\frac{\lambda}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \big(|x - x_i| + |x_k - x_j| - |x - x_k| - |x_i - x_j| \big) \right) w_k. \end{aligned}$$

The derivative of $f(x)$ with respect to x is

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}f(x) &= \frac{\lambda}{2n^2}\sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\mathrm{d}}{\mathrm{d}x}|x-x_i| - \frac{\mathrm{d}}{\mathrm{d}x}|x-x_k|\right)w_k \\ &= \frac{\lambda}{2n^2}\sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\operatorname{sign}(x-x_i) - \operatorname{sign}(x-x_k)\right)w_k \\ &= \frac{\lambda}{2n}\sum_{k=1}^{n}\sum_{i=1}^{n}\left(\operatorname{sign}(x-x_i) - \operatorname{sign}(x-x_k)\right)w_k\end{aligned}$$

4.2 Estimation

4.3 Examples

4.4 Conclusion

The steps for I-prior modelling are basically three-fold:

1. Select an appropriate function space; equivalently, the kernels for which a specific effect is desired on the covariates. Several modelling examples are described in Section 4.1.
2. Estimate the hyperparameters (these included the RKHS scale parameter(s), error precision, and any other kernel parameters such as the Hurst index of fBm) of the I-prior model and obtain the posterior regression function.
3. Post-estimation procedures include
   - Posterior predictive checks;
   - Model comparison via log-likelihood ratio tests/empirical Bayes factors; and
   - Prediction of new data point.

The main sticking point with the estimation procedure is the involvement of the $n \times n$ kernel matrix, for which its inverse is needed. This requires $O(n^2)$ storage and $O(n^3)$ computational time. The Nyström method of approximating the kernel matrix reduces complexity to $O(nm)$ storage and approximately $O(nm^2)$, and is highly advantageous

if $m \ll n$. The computational issue faced by I-priors are mirrored in Gaussian process regression, so the methods to overcome these computational challenges in GPR can be explored further. However, most efficient computational solutions exploit the nature of the SE kernel structure, which is the most common kernel used in GPR.

One promising avenue to achieve efficient computation for I-prior models is by using variational methods. A sparse variational approximation (typically by using inducing points) or stochastic variational inference can greatly reduce computational storage and speed requirements. A recent paper by Cheng and Boots (2017) suggested a variational algorithm with linear complexity for GPR-type models.

Omitted

Bibliography

Cheng, C.-A. and B. Boots (2017). "Variational Inference for Gaussian Process Models with Linear Complexity". In: *Advances in Neural Information Processing Systems*, pp. 5190–5200.

van der Vaart, A. W. and van Zanten (2008). "Reproducing kernel Hilbert spaces of Gaussian priors". In: *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*. Institute of Mathematical Statistics, pp. 200–222.

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions". In: *Bayesian inference and decision techniques*.

List of Figures

List of Tables

List of Theorems

List of Definitions

List of Symbols

$\mathrm{N}_p(\mu, \Sigma)$	p-dimensional multivariate normal distribution with mean vector μ and covariance Σ.
$\sim$	Is distributed as.
$\delta_{xx'}$	The Kronecker delta: $\delta_{xx'} = 1$ if $x = x'$, and 0 otherwise.
$\otimes$	The tensor product.

Index