Haziq Jamil
Department of Statistics
London School of Economics and Political Science
http://haziqj.ml

Regression Modelling using Priors with Fisher Information Covariance Kernels (I-priors)

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Objectives

- Outline an efficient computational method for estimating the parameters of an I-prior model in the continuous responses case.
- Extend the I-prior methodology to categorical responses for classification and inference.
- Explore the usefulness of I-priors towards variable selection.

Introduction

Consider the following regression model for $i = 1, \dots, n$:

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

where $y_i \in \mathbb{R}$, $x \in \mathcal{X}$, and $f \in \mathcal{F}$. Let $\mathcal{F}$ be a reproducing kernel Hilbert space (RKHS) with kernel $h_\lambda : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The Fisher information for f is given by

$$\mathcal{I}\big(f(x), f(x')\big) = \sum_{k=1}^{n} \sum_{l=1}^{n} \Psi_{k,l} h_\lambda(x, x_k) h_\lambda(x', x_l). \tag{2}$$

The I-prior

The entropy maximising prior distribution for f, subject to constraints, is

$$\mathbf{f} = \big(f(x_1), \dots, f(x_n)\big)^\top \sim \mathrm{N}_n\left(\mathbf{f}_0, \mathcal{I}[f]\right).$$

Of interest are the

- Posterior distribution for the regression function
$$p(\mathbf{f}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f})}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{y}}; \text{ and}$$
- Posterior predictive distribution given new data
$$p(y_{\text{new}}|\mathbf{y}) = \int p(y_{\text{new}}|f_{\text{new}}, \mathbf{y}) p(f_{\text{new}}|\mathbf{y})\,\mathrm{d}\mathbf{y}.$$

Estimation

Model parameters (error precision Ψ, RKHS scale parameters λ, and any others) may be estimated via

- Maximum (marginal) likelihood, a.k.a. empirical-Bayes;
- Expectation-maximisation (EM) algorithm; or
- Markov chain Monte Carlo (MCMC) methods.

Under the normal model (1), the posterior for y, given some x and model parameters, is normal with mean

$$\hat{y}(x) = f_0(x) + \mathbf{h}_\lambda^\top(x)\Psi H_\lambda (H_\lambda \Psi H_\lambda + \Psi^{-1})^{-1}\big(y - f_0(x)\big)$$

and variance (3)

$$\hat{\sigma}^2(x) = \mathbf{h}_\lambda^\top(x)(H_\lambda \Psi H_\lambda + \Psi^{-1})^{-1}\mathbf{h}_\lambda(x) + \psi_x^{-1}.$$

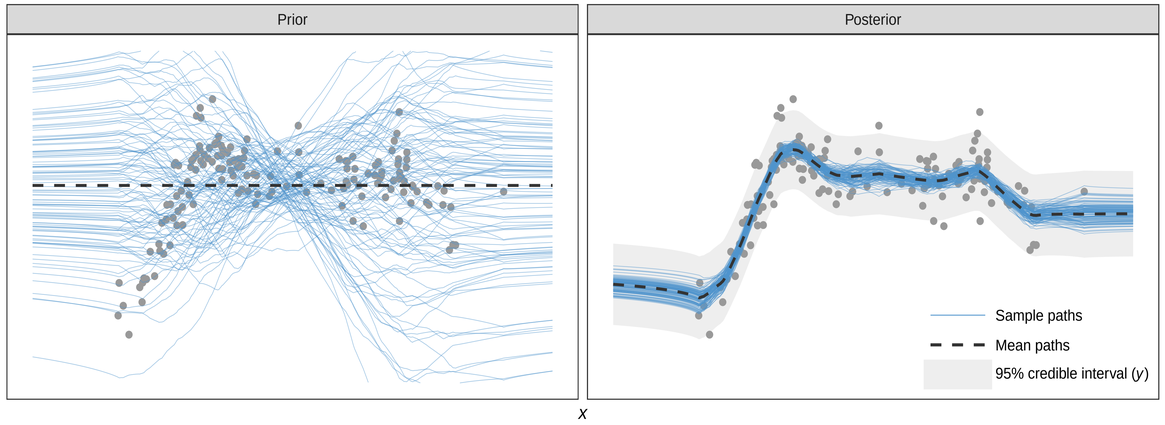

Figure 1: Sample paths from the fractional Brownian motion RKHS under an I-prior (left) and the posterior (right). There is somewhat controlled behaviour at the boundaries (compared to Gaussian process priors, say). Fewer information in this region pulls the function estimate towards the prior mean. The 95% credibility interval for posterior estimates of y are shaded grey.

Computational Hurdle

Computational complexity is dominated by the $n \times n$ matrix inversion in (3), which is $O(n^3)$. Suppose that $H_\lambda \Psi H_\lambda = QQ^\top$, where Q is a $n \times q$ matrix, is a valid low-rank decomposition. Then

$$(H_\lambda \Psi H_\lambda + \Psi^{-1})^{-1} = \Psi - \Psi Q (I_q + Q^\top \Psi Q)^{-1} Q^\top \Psi$$

is a much cheaper $O(nq^2)$ operation, especially if $q \ll n$. The Nyström method for low-rank matrix approximations is explored.

I-prior advantages

- Unifies methodology for various regressions models, including:
  - Multidimensional smoothing
  - Random effects/multilevel models
  - Longitudinal models
  - Functional linear/smooth regression
- Straightforward estimation and inference
- Often gives better prediction for new data

Categorical Responses

Suppose now that each $y_i \in \{1, \dots, m\}$ and that

$$y_i \sim \mathrm{Cat}(p_{i1}, \dots, p_{im})$$

with probability mass function

$$p(y_i) = \prod_{j=1}^{m} p_{ij}^{y_{ij}}, \qquad y_{ij} = [y_i = j],$$

satisfying $p_{ij} > 0$, $\sum_j p_{ij} = 1$ for $j \in \{1, \dots, m\}$. In the spirit of generalised linear models, take

$$\mathrm{E}[y_{ij}] = p_{ij} = g^{-1}\big(f_j(x_i)\big)$$

with some link function $g : \mathbb{R} \to [0, 1]$ and an I-prior on f_j.

Now, the marginal (on which the posterior depends),

$$p(\mathbf{y}) = \int \prod_{i=1}^{n} \prod_{j=1}^{m} \left[\left\{ g^{-1}\big(f_j(x_i)\big) \right\}^{[y_i = j]} \cdot \mathrm{N}_n(\mathbf{f}_{0j}, \mathcal{I}[f_j])\,\mathrm{d}\mathbf{f}_j \right],$$

cannot be found in closed form.

Variational Approximation

An approximation $q(\mathbf{f})$ to the true posterior density $p(\mathbf{f}|\mathbf{y})$ is considered, with q chosen to minimise the Kullback-Leibler divergence (under certain restrictions):

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

By working in a fully Bayes setting, we append model parameters to $\mathbf{f}$ and employ a variational approximation. The result is an iterative algorithm similar to the EM.

As this variational-EM works harmoniously with exponential family distributions, the ***probit*** link is preferred.

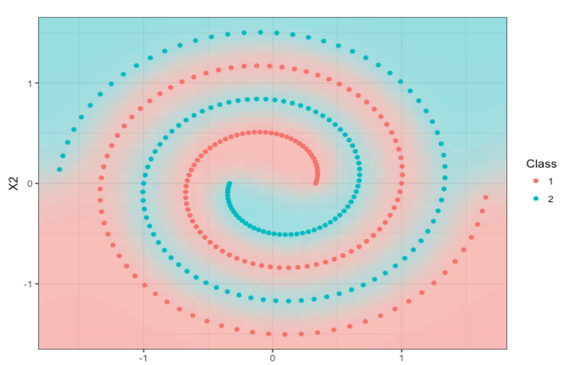

Figure 2: A toy example of binary classification using I-priors and the fBm-0.5 kernel over a two-dimensional predictor. Points indicate realisations, while background colours denote predicted classes.

Variable Selection for Linear Models

Model selection can easily be done by comparing likelihoods (empirical Bayes factors). However, with p variables to select, the 2^p comparisons will prove intractable.

For linear models of the form

$$(y_1, \dots, y_n)^\top \sim \mathrm{N}_n\left(\beta_0 \mathbf{1}_n + \sum_{j=1}^{p} \beta_j X_j, \Psi^{-1}\right),$$

the prior

$$(\beta_1, \dots, \beta_p)^\top \sim \mathrm{N}_p(0, \Lambda X^\top \Psi X \Lambda)$$

is an equivalent I-prior representation of (1) in the feature space of β under the linear kernel.

We employ a fully Bayesian treatment of the model in order to estimate *posterior model probabilities*

$$p(M|\mathbf{y}) \propto \int p(\mathbf{y}|M, \theta) p(\theta|M) p(M)\,\mathrm{d}\theta$$

where M is the model index and θ are model parameters.

	Signal-to-noise Ratio (SNR)				
False choices	**90%**	**75%**	**50%**	**25%**	**10%**
0-2	0.84	0.92	0.81	0.79	0.20
3-5	0.15	0.08	0.18	0.20	0.47
>5	0.01	0.00	0.01	0.01	0.33

Table 1: Simulation results (proportion of false choices) for experiments in selecting 100 pairwise-correlated variables using I-priors under differing SNR. Our method outperforms methods such as greedy selection, g-priors, and regularisation (ridge and Lasso).

Conclusions

- The dominating $O(n^3)$ step in (3) can be reduced to $O(nq^2)$, $q \ll n$, via low-rank matrix approximations.
- Advantages of I-priors in the continuous case extend well to binary and multinomial responses.
- Simulations and real-data examples indicate that I-priors work well for linear variable selection under multicollinearity.

References

[1] Wicher Bergsma.
Regression and classification with I-priors.
arXiv:1707.00274, July 2017.

[2] Christopher KI Williams and Matthias Seeger.
Using the Nyström method to speed up kernel machines.
In *Advances in Neural Information Processing Systems*, 2001.

[3] Mark Girolami and Simon Rogers.
Variational Bayesian multinomial probit regression with Gaussian process priors.
Neural Computation, 18(8), 2006.

[4] Lynn Kuo and Bani Mallick.
Variable selection for regression models.
Sankhyā: The Indian Journal of Statistics, Series B, 60(1), 1998.