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

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

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

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

Placeholder

Image

Figure 1: Figure caption

Computational Hurdles

For models with many scale parameters,

- Newton methods are problematic due to the presence of multiple local optima; while
- Gibbs sampling suffer from severe autocorrelation in the posterior samples.

The EM provides a reliable and straightforward framework for estimating the parameters of I-prior models.

Regardless, computational complexity is dominated by the $n \times n$ matrix inversion in (X), 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 $q \ll n$.

We explore the Nyström method for low-rank matrix approximations, which we find works well for the fractional Brownian motion RKHS.

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 the parameters to $\mathbf{f}$ and employ the 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.

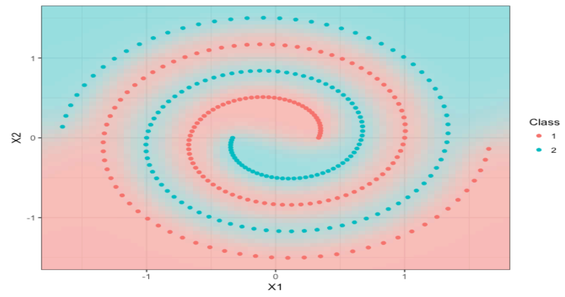

Figure 2: A toy example of binary classification using I-priors and the fBm-0.5 kernel over a two-dimensional predictor. Points indicate realisations, while background colours denote predicted classes.

Variable Selection for Linear Models

In the original I-prior model, model selection can 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.

Simulation studies together with real-data applications show promising results, outperforming methods such as greedy selection, g-priors, and regularisation (ridge and Lasso) under multicollinearity.

	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 of choosing 100 pairwise correlated variables using I-priors. Under differing SNR, proportions of false choices were recorded.

Conclusions

- The EM algorithm provides a straightforward and efficient method of estimating I-prior models.
- The dominating $O(n^3)$ step in the algorithm 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.
- I-priors work well for linear variable selection under multicollinearity.

References