Haziq Jamil
Department of Statistics
London School of Economics and Political Science
PhD thesis: 'Regression modelling using priors depending on Fisher information covariance kernels'
11 October 2018 (`v1.1@f378773`)

Chapter 7

Summary

The work done in this thesis explores the concept of regression modelling using priors with Fisher information covariance kernels (I-priors, Bergsma, 2018). It is best seen as a flexible regression technique which is able to fit both parametric and nonparametric models, and bears similarity to Gaussian process regression. For the regression model (1.1) subject to (1.2), stated again here for convenience,

$$\begin{aligned} y_i &= \alpha + f(x_i) + \epsilon_i && \text{(from 1.1)} \\ (\epsilon_1, \ldots, \epsilon_n) &\sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi}^{-1}) && \text{(from 1.2)} \\ i &= 1, \ldots, n, \end{aligned}$$

and it is assumed that the regression function f lies in some reproducing kernel Hilbert or Kreĭn space (RKHS/RKKS) $\mathcal{F}$ with kernel h_η defined over the set of covariates $\mathcal{X}$. In Chapter 2, we built a primer on basic functional analysis, and described various interesting RKHS/RKKS for regression modelling.

We then ascertained the form of the Fisher information for f, treated as a parameter of the model to be estimated, and from Corollary 3.3.1 (p. 93), it is

$$\begin{aligned} \mathcal{I}\big(f(x), f(x')\big) &= \sum_{i,j=1}^{n} \psi_{ij} h_\eta(x, x_i) h_\eta(x', x_j) \\ &= \mathbf{h}_\eta(x)^\top \boldsymbol{\Psi} \mathbf{h}_\eta(x'), \end{aligned}$$

for any two points x, x' in the domain of f, obtained using appropriate calculus for topological spaces detailed in Chapter 3. An I-prior for f is defined as Gaussian with mean function f_0 chosen a priori, and covariance function equal to the Fisher information.

The I-prior for f has the simple representation

$$f(x_i) = f_0(x_i) + \sum_{k=1}^{n} h_\eta(x_i, x_k) w_k$$
$$(w_1, \ldots, w_n)^\top \sim \mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi})$$
$$i = 1, \ldots, n,$$

and is written equivalently as the Gaussian process prior

$$\big(f(x_1), \ldots, f(x_n)\big)^\top \sim \mathrm{N}_n(\mathbf{f}_0, \mathbf{H}_\eta \boldsymbol{\Psi} \mathbf{H}_\eta),$$

where $\mathbf{H}_\eta = \big(h_\eta(x_i, x_j)\big)_{i,j=1}^n$.

In Chapter 4, we looked how the I-prior model has wide-ranging applications, from multilevel modelling, to longitudinal modelling, and modelling with functional covariates. Estimation was conducted mainly using a simple EM algorithm, although direct optimisation and Bayesian estimation using Markov chain Monte Carlo (MCMC) are also possible. In the case of polytomous responses, we used a latent variable framework in Chapter 5 to assign I-priors to latent propensities which drive the outcomes under a probit-transform scheme. An extension of the EM algorithm was considered, in which the E-step was replaced with variational inference, so as to overcome the intractability brought about by the conditional distributions. For both continuous and categorical response I-prior models, we find advantages of using I-priors, namely that model building and estimation is simple, inference straightforward, and predictions comparable, if not better, to similar state-of-the-art techniques.

Finally, in Chapter 6, we dealt with the problem of model selection, specifically for linear regression models. There, we used a fully Bayesian approach for estimating model probabilities in which regression coefficients are assigned an I-prior. We devised a model that requires minimal tuning on the part of the user, yet performs well in simulated and real-data examples, even if multicollinearity exists among the covariates.

7.1 Summary of contributions

We give a summary of the novel contributions of this thesis.

- **Fisher information for infinite-dimensional parameters**. When the RKHS/ RKKS $\mathcal{F}$ is infinite dimensional (e.g. covariates are themselves functions), then the Fisher information involves derivatives with respect to an infinite-dimensional vector. Finite-dimensional results using componentwise/partial derivatives may fail in infinite dimensions. The technology of Fréchet and Gâteaux differentials

accommodate for the fact that f may be infinite dimensional, which, at minimum, requires $\mathcal{F}$ to be a normed vector space. We foresee the work of Section 3.2 being applicable elsewhere, such as learning in (reproducing kernel) Banach spaces (H. Zhang et al., 2009; H. Zhang and J. Zhang, 2012), or in the theory of parameter estimation for general exponential family type distributions of the form

$$p(y|\theta) = B(y)\exp\big(\langle\theta, T(y)\rangle_{\mathcal{H}} - A(\theta)\big),$$

in which θ lies in some inner-product space $\mathcal{H}$ which might be infinite dimensional (Sriperumbudur et al., 2017).

- **Efficient estimation methods for normal I-prior models**. The preferred estimation method for normal I-prior models for stability is the EM algorithm. Implementing the EM algorithm can be computationally costly, due to the squaring and inversion of the kernel matrices in the Q function in (4.18) on page 113. Unfortunately, not much can be done about the inversion, but we explored systematic ways in which to perform the squaring. Combining a "front-loading method" of the kernel matrices (Section 4.3.2, p. 119) and an exponential family ECM (expectation conditional maximisation) algorithm (Meng and Rubin, 1993), the estimation procedure is streamlined. Our computational work culminated in the publicly available and well-documented R package **iprior** (Jamil, 2017) published on CRAN.

- **Methodological extension of I-priors to categorical responses**. An extension of the I-prior methodology to fit categorical responses was studied. We proposed a latent variable framework, in which there corresponds latent propensities for each category of the observations. Instead of modelling the responses directly, the latent propensities are modelled using an I-prior, and class probabilities obtained using a normal integral. We named this model the I-probit model. The challenge of estimation was overcoming said integral, and we used a variational EM algorithm in which the E-step uses a variational approximation to intractable conditional density. The variational EM algorithm was preferred over a fully Bayesian variational inference algorithm for two main reasons: 1) the work done in the normal I-prior EM algorithm applies directly; and 2) prior specification for hyperparameters can be dispensed with. Classification, meta-analysis and spatio-temporal modelling are specific examples of the applications of I-probit models.

- **Some distributional results for truncated normals**. In deriving the variational algorithm, some properties related to the conically truncated multivariate independent normal distribution (as defined in Appendix C.4, p. 281) were required. A small contribution of ours was to derive the closed-form expressions for

its first and second moments, and its entropy (Lemma C.5, p. 283). We have only seen closed-form expressions of the mean of such a distribution being used before (Girolami and Rogers, 2006) but not for the variance, nor an explicit derivation of these quantities.

- **Bayesian variable selection under collinearity**. Model comparison using likelihood ratio tests or Bayes factors is fine when the number of models under consideration is fairly small. Under a fully Bayesian scheme, we use MCMC to approximate posterior model probabilities of competing linear models. At the outset, we sought a model which required minimal intervention on the part of the user. The I-prior achieved this, with the added advantage of performing well under multicollinearity.

7.2 Open questions

In closing, we briefly discuss several questions which remain open during the course of completing this project.

- **Initialisation of EM or gradient-based methods**. Figure 4.1 (p. 112) indicates the impact that starting values can have on gradient-based optimisation. One can end up at a local optima on one of the two ridges. Usually, one of the ridges will have a higher maximum than the other, but it is not clear how to direct the algorithm in the direction of the "correct" ridge.

  Importantly, the interpretation of a flat ridge in the likelihood is that there is insufficient information coming from the data to inform parameter estimation. In the EM algorithm, estimation is usually characterised by a fast increase in likelihood in the first few steps (as it climbs up the ridge), and then later iterations only improve the likelihood ever so slightly (as it moves along the ridge in search of the maximum). In some real-data cases (e.g. Tecator data set), we noticed that the EM sequence veers to the boundary of the parameter space, where the likelihood is infinite (e.g. $L(\psi) \to \infty$ as $\psi \to 0, \infty$).

  Ill-posed problems similar to this are resolved by adding penalty terms to the log-likelihood. As to what penalty terms are appropriate remains an open question.

- **Standard errors for variational approximation**. Under a variational scheme, the log-likelihood function $L(\theta)$ is replaced with the evidence lower bound (ELBO) $\mathcal{L}_q(\theta)$ which serves as a conservative approximation to it. The question we have is whether the approximation degrades the asymptotic properties of the estimators obtained via variational inference? In particular, are the standard errors obtained

from the information matrix involving $\mathcal{L}_q(\theta)$ reliable? This question has also been posed by Bickel et al. (2013), Chen et al. (2018), and Hall et al. (2011).

Variational methods for maximum likelihood learning can be seen as a deliberate misspecification of the model to achieve tractibility. As such, the variational EM has been referred to as obtaining pseudo- or quasi-ML estimates. The quasi-likelihood literature has results relating to efficiency of parameter estimates (adjustments to the information matrix is needed), and we wonder if these are applicable for variational inference.

Also, obtaining standard errors directly from an EM algorithm is of interest, especially under a variational EM setting. Though this is described in McLachlan and Krishnan (2007, Ch. 4), we have not seen this implemented widely.

- **Comparison of logistic and probit links**. For general binary and multinomial models, the logistic link function sees more prevalent use than its probit counterpart. Of course, we chose the probit as it has distributional advantages which we can exploit for estimation using variational inference. However, is there a difference between the behaviour of the probit and logistic model? We know that there is a difference between the logistic and normal distribution, especially in scaling and behaviour in the tails, but do these affect the outcome of I-prior models?

- **Consistency of I-prior Bayesian variable selection**. We wondered about model selection consistency for I-priors in Bayesian variable selection. That is, assuming that model M_{true} is actually behind the true data generative process, do

$$\lim_{n\to\infty} \mathrm{P}(M_{\text{true}}|\mathbf{y}) = 1 \quad \text{and} \quad \lim_{n\to\infty} \mathrm{P}(M_k|\mathbf{y}) = 0, \forall M_k \neq M_{\text{true}}$$

  hold for the I-prior Bayesian variable selection methodology? In machine learning, this property is referred to as the *oracle property*. For the g-prior specifically, model consistency results were obtained by Fernández et al. (2001) and Liang et al. (2008). Casella et al. (2009) also looks at consistency of Bayesian procedures for a wide class of prior distributions, but we have yet to examine whether the I-prior falls under the remit of their work.

Bibliography

Bergsma, Wicher (2018). *Regression and classification with I-priors*. Manuscript in submission. arXiv: 1707.00274 [math.ST].

Bickel, Peter, David Choi, Xiangyu Chang, and Hai Zhang (2013). "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels". In: *The Annals of Statistics* 41.4, pp. 1922–1943. DOI: 10.1214/13-AOS1124.

Casella, George, F. Javier Girón, M. Lina Martínez, and Elías Moreno (2009). "Consistency of Bayesian procedures for variable selection". In: *The Annals of Statistics* 37.3, pp. 1207–1228. DOI: 10.1214/08-AOS606.

Chen, Yen-Chi, Y. Samuel Wang, and Elena A. Erosheva (2018). "On the use of bootstrap with variational inference: Theory, interpretation, and a two-sample test example". In: *Annals of Applied Statistics* to appear. arXiv: 1711.11057 [stat.ME].

Fernández, Carmen, Eduardo Ley, and Mark F. J. Steel (2001). "Benchmark priors for Bayesian model averaging". In: *Journal of Econometrics* 100.2, pp. 381–427. DOI: 10.1016/S0304-4076(00)00076-2.

Girolami, Mark and Simon Rogers (2006). "Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors". In: *Neural Computation* 18.8, pp. 1790–1817. DOI: 10.1162/neco.2006.18.8.1790.

Hall, Peter, Tung Pham, Matt P. Wand, and Shen S. J. Wang (2011). "Asymptotic normality and valid inference for Gaussian variational approximation". In: *The Annals of Statistics* 39.5, pp. 2502–2532. DOI: 10.1214/11-AOS908.

Jamil, Haziq (2017). ***iprior**: Regression Modelling using I-Priors*. R package version 0.7.1. URL: https://cran.r-project.org/web/packages/iprior.

Liang, Feng, Rui Paulo, German Molina, Merlise A. Clyde, and James O. Berger (2008). "Mixtures of g Priors for Bayesian Variable Selection". In: *Journal of the American Statistical Association* 103.481, pp. 410–423. DOI: 10.1198/016214507000001337.

McLachlan, Geoffrey and Thriyambakam Krishnan (2007). *The EM Algorithm and Extensions*. 2nd ed. John Wiley & Sons. ISBN: 978-0-471-20170-0. DOI: 10.1002/9780470191613.

Meng, Xiao-Li and Donald B. Rubin (1993). "Maximum likelihood estimation via the ECM algorithm: A general framework". In: *Biometrika* 80.2, pp. 267–278. DOI: 10.1093/biomet/80.2.267.

Sriperumbudur, Bharath, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar (2017). "Density Estimation in Infinite Dimensional Exponential Fami-

lies". In: *Journal of Machine Learning Research* 18.57, pp. 1–59. arXiv: 1312.3516 [math.ST].

Zhang, Haizhang, Yuesheng Xu, and Jun Zhang (2009). "Reproducing Kernel Banach Spaces for Machine Learning". In: *Journal of Machine Learning Research* 10, pp. 2741–2775.

Zhang, Haizhang and Jun Zhang (2012). "Regularized learning in Banach spaces as an optimization problem: representer theorems". In: *Journal of Global Optimization* 54.2, pp. 235–250. doi: 10.1007/s10898-010-9575-z.