To-do list

Contents

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
PhD thesis: ‘Regression modelling using Fisher information covariance kernels (I-priors)’

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, 2017). It is best categorised as a form non-parametric regression, and bears similarity to Gaussian process regression. For the regression model (1.1) subject to (1.2), stated again here for convenience,

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

it is assumed that the regression function f lies in some reproducing kernel Hilbert or Krein space $\mathcal{F}$ 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 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, it is

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

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

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

and written equivalently as a Gaussian process prior

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

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 fully Bayesian estimation using MCMC is 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, especially 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 component-wise/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, Xu, 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

$$p(X|\theta) = B(X) \exp \big(\langle \theta, T(X) \rangle - A(\theta) \big),$$

for which θ lies in some inner-product space Θ which might be infinite-dimensional (Sriperumbudur et al., 2013).

- **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 of the kernel matrices in the Q function in (4.15). Combining a 'front-loading method' of the kernel matrices (Section 4.3.2) and an exponential family ECM (expectation conditional maximisation) algorithm (Meng and Rubin, 1993), the estimation procedure is streamlined. Our computational work is encapsulated in the publicly available and well-documented R package **iprior** published on CRAN.

- **Methodological extension of I-priors to categorical responses**. Extension of the I-prior methodology to fit categorical responses is of great interest. We proposed a latent variable framework, for which there corresponds latent propensities corresponding to each category of the observed response variable. 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 is 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 reasons: 1) the work done in the continuous case EM algorithm applies directly; and 2) prior specification for hyperparameter can be dispensed with. Classification, meta-analysis and spatio-temporal modelling are specific examples of the applications of the 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) 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). 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 completing this project, several questions remain open, and we discuss these briefly below.

- **Initialisation of EM or gradient-based methods**. Figure 4.1 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 $\psi \to 0, \infty$).

1. Rethink this.

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 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 Hall et al. (2011), Bickel et al. (2013), and Chen et al. (2017).

  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.

- **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 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 Fernandez 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.

7.3 Next steps

As far as next steps go, we identify the following to be concentrated on for immediate future work.

- **Estimation of $\boldsymbol{\Psi}$ in I-probit models**. As discussed in conclusion section of Chapter 5, estimation of $\boldsymbol{\Psi}$ would certainly add flexibility and is especially of interest for choice models. We discussed that estimation of $\boldsymbol{\Psi}$ is done by its inclusion as a free parameter in the M-step of the variational EM algorithm, subject to suitable constraints. If this is successful, then this would be a key feature to be added on to the ongoing development of the **iprobit** package in R. As an aside, the package also contains features related to the truncated multivariate normal, which, if exported, could prove to be useful for other statistical models such as tobit models, constrained linear regression, and others.

- **Extension of I-prior methodology to other model classes**. An immediate extension to I-probit models is for modelling ordinal responses. The underlying latent variable model in (5.1) is changed to

$$y_i = \begin{cases} 1 & \text{if } y_i^* \le \tau_1 \\ 2 & \text{if } \tau_1 < y_i^* \le \tau_2 \\ \vdots & \\ m & \text{if } y_i^* > \tau_{m-1}, \end{cases} \tag{7.1}$$

{eq:latentmodel2}

  where instead of m latent propensities, there is only one, but $m-1$ thresholds $\tau_1, \dots, \tau_{m-1}$ need to be estimated.

  Another extension of the I-prior methodology is to fit Poisson count data. A suggestion would be to model $y_i \sim \text{Pois}(\mu_i)$, where $\mu_i = g^{-1}\big(\alpha + f(x_i) + \epsilon_i\big)$ is modelled using the regression model of (1.1) subject to (1.2) and an I-prior. Since the mean of the Poisson is positive, the function g^{-1} should map real values to the positive reals. Examples include $g^{-1}(x) = e^x$ (which is the canonical link function in GLM theory), or simply $g^{-1}(x) = x^2$ (as per Lloyd et al., 2015).

  Both the ordinal probit and Poisson model should be very interesting to look at especially from an estimation perspective.

- **Variational inference for I-prior Bayesian variable selection**. The work of Ormerod et al. (2017) is encouraging in that the stochastic nature of BVS models can be replaced with a deterministic variational algorithm. Especially since Gibbs conditional densities are somewhat related to mean-field variational densities, there could be minimal effort required in switching between them. The benefits of course would come in terms of speed of estimation of the posterior model probabilities.

  Hyperparameters (κ, σ^2) in the I-prior BVS model can also be replaced with their posterior mode estimate, so as to impart an empirical Bayesian flavour to BVS, so perhaps a variational EM algorithm can be explored.

  We think it would also be interesting to extend the BVS model to linear categorical response models (I-probit with canonical RKHS).

Bibliography

bergsma2017 Bergsma, Wicher (2017). “Regression with I-priors”. In: *Unpublished manuscript.*

bickel2013asymptotic 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*, pp. 1922–1943.

casella2009consistency Casella, George, F Javier Girón, M Lina Martínez, and Elias Moreno (2009). “Consistency of Bayesian procedures for variable selection”. In: *The Annals of Statistics*, pp. 1207–1228.

chen2017use Chen, Yen-Chi, Y Samuel Wang, and Elena A Erosheva (2017). “On the use of bootstrap with variational inference: Theory, interpretation, and a two-sample test example”. In: *arXiv preprint arXiv:1711.11057.*

fernandez2001benchmark Fernandez, Carmen, Eduardo Ley, and Mark FJ Steel (2001). “Benchmark priors for Bayesian model averaging”. In: *Journal of Econometrics* 100.2, pp. 381–427.

girolami2006variational Girolami, Mark and Simon Rogers (2006). “Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors”. In: *Neural Computation* 18.8, pp. 1790–1817.

hall2011asymptotic Hall, Peter, Tung Pham, Matt P Wand, Shen SJ Wang, et al. (2011). “Asymptotic normality and valid inference for Gaussian variational approximation”. In: *The Annals of Statistics* 39.5, pp. 2502–2532.

liang2008mixtures Liang, Feng, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger (2008). “Mixtures of g priors for Bayesian variable selection”. In: *Journal of the American Statistical Association* 103.481, pp. 410–423.

lloyd2015variational

Lloyd, Chris, Tom Gunter, Michael Osborne, and Stephen Roberts (2015). “Variational inference for Gaussian process modulated Poisson processes”. In: *International Conference on Machine Learning*, pp. 1814–1822.

mclachlan2007algorithm

McLachlan, Geoffrey and Thriyambakam Krishnan (2007). *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons.

meng1993maximum

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.

ormerod2017variational

Ormerod, John T, Chong You, Samuel Müller, et al. (2017). “A variational Bayes approach to variable selection”. In: *Electronic Journal of Statistics* 11.2, pp. 3549–3594.

sriperumbudur2013density

Sriperumbudur, Bharath, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar (2013). “Density estimation in infinite dimensional exponential families”. In: *arXiv preprint arXiv:1312.3516*.

zhang2009reproducing

Zhang, Haizhang, Yuesheng Xu, and Jun Zhang (2009). “Reproducing kernel Banach spaces for machine learning”. In: *Journal of Machine Learning Research* 10.Dec, pp. 2741–2775.

zhang2012regularized

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.