Errata for PhD Thesis:

Regression modelling using priors depending on Fisher information covariance kernels

Md. Haziq Md. Jamil

10 October 2018

These are the responses to the comments and questions raised by the examiners relating to the thesis submitted for examination (git version `master@54604a8`). All references and page numbers in this document pertain to the amended thesis (git version `master@XXXXX`).

1. **Additional details regarding maximum entropy priors.**

   The central tenet of the thesis is estimating regression functions using a novel methodology which improves upon Tikhonov regularisation using an objective, data-dependent prior for the regression function.

   For the regression problem (1.1), one could choose to assign a subjective prior on the regression function. For instance, using Gaussian priors, this is Gaussian process regression (Rasmussen and Williams, 2006).

   Instead, we focus on objective priors, specifically a prior on the regression function based on the principle of maximum entropy. Jaynes (1957a, 1957b, 2003) argues that in the absence of any prior knowledge, a probability distribution that maximises information entropy (as per Definition 3.5) should be advocated. Entropy-maximising priors is "uninformative" in the sense that it minimises the amount of prior information encoded into prior distributions.

   In Chapter 3, we show that the entropy maximising prior distribution for the regression function is Gaussian with covariance kernel proportional to its Fisher information (hence the name I-prior). This has the intuitive property that the more Fisher information available regarding the regression function, the larger its prior variance, and hence lesser influence of the prior mean and more of the data, and vice versa.

   Amendments to thesis: Reworded introductory chapter to better motivate I-priors (p. 33).

2. **How do I-priors benefit from the principle of maximum entropy? Compare to other priors.**

   Besides being based on the principle of maximum entropy, we have not discovered a definitive optimality criterion for which the I-prior satisfies. However, our small and

large sample simulations and real-data examples have been promising, especially in terms of predictive abilities (as shown by the examples in Section 4.5 (p. 127) and Section 5.7 (p. 178)).

The prior for the regression function can certainly be chosen based on other objective principles, and we discuss these briefly. An expanded discussion of these priors is given in Bergsma (2018).

- The g-prior (Zellner, 1986) for regression coefficients has covariance proportional to the *inverse* of its Fisher information. This is entirely different from I-priors, and a comparison is discussed in Chapter 6 and Appendix E.
- Jeffreys prior (Jeffreys, 1946) is proportional to the square root of the determinant of the Fisher information. For large or even potentially infinite-dimensional regression functions this poses a problem, so practically speaking, it is only suitable for low-dimensional problems.
- A particular kernel called the Fisher kernel has been used in kernel machines. Like the I-prior, it also uses Fisher information, but the similarities end there.

3. **On the choice of I-priors leading to a finite-dimensional estimator of the regression function.**

   In the conclusion section of Chapter 3 (p. 100), I wrote

   > The dimension of the function space $\mathcal{F}$ could be huge, infinite dimensional even, while the task of estimating $f \in \mathcal{F}$ only relies on a finite amount of data point. However, we are certain that the Fisher information for f exists only for the finite subspace $\mathcal{F}_n$ as defined in (3.4), and it is zero everywhere else. This suggests that the data only allows us to provide an estimation to the function $f \in \mathcal{F}$ by considering functions in an (at most) n-dimensional subspace instead.

   In the above, I have alluded to the fact that one need only consider functions in $\mathcal{F}_n$, i.e. functions of the form

   $$f_n(x) = \sum_{i=1}^{n} h(x, x_i) w_i, \tag{1}$$

   to estimate the regression function, thus providing an element of dimension reduction especially when $\dim(\mathcal{F}) \gg n$. The argument for this is as follows (adapter from Bergsma, 2018). By the orthogonal decomposition theorem, any $f \in \mathcal{F}$ may be decomposed into $f = f_n + r$, where $f_n \in \mathcal{F}_n \subset \mathcal{F}$, and r in its orthogonal complement $\mathcal{F}_n^{\perp}$. Since $r \in \mathcal{F}_n^{\perp}$ is orthogonal to each of the $h(\cdot, x_i) \in \mathcal{F}$, we have that by the reproducing property of h in $\mathcal{F}$, $r(x_i) = \langle r, h(\cdot, x_i) \rangle_{\mathcal{F}} = 0$.

The likelihood for f therefore does not depend on r, and since f_n is orthogonal to r, the data do not contain Fisher information regarding r. Thusly, it is not possible to perform inference on r using the data at hand, and one can only do statistical inference on f_n.

The main details of this argument has been provided in Section 3.4 (p. 93).

4. **How to motivate the choice of kernels?**

   Any kernel, and thus RKHS, can be used with the I-prior methodology. In this thesis, we chose to study the linear, fBm and Pearson RKHS as the building blocks, and using these RKHSs we build more complex RKHSs/RKKSs using the polynomial or ANOVA construction. A short discussion for motivating the choice of each of these kernels is given below:

   - The canonical (linear) kernel produces "straight-line" regression functions, similar to that of linear regression. This choice of kernel is obviously for linear problems.
   - The Pearson kernel takes discrete covariates as input. This is used for regressing independent variables against categorical covariates, or used in an ANOVA-type construction to deal with "group" information (e.g. in multilevel modelling). Other discrete kernels might be used (e.g. identity kernel or string kernel). The Pearson kernel uses inverse probability weights as a measure of similarity between two discrete inputs.
   - The fBm kernel is used for smoothing problems. Certainly other smooth kernels can be used, such as squared-exponential kernel, but the fBm kernel has several desirable properties
     (a) Requires only one smoothing parameter (the Hurst coefficient), which can usually be left at the default of 1/2;
     (b) Automatic boundary correction in the sense described in Section 4.1.5 (p. 107); and
     (c) Prior and posterior sample paths under an I-prior are suitably smooth and not as rough as fBm paths themselves or even functions in an fBm RKHS.

5. **On the connection with Generalised Additive Models (GAMs).**

   Apart from the additive nature of (4.2) (p. 103), the I-prior methodology is completely different from additive models or GAMs.

   In the I-prior methodology, the principle of decomposing the regression function into additive parts is the ANOVA functional decomposition. The advantage of this

is that we are able to also describe two-, three- or higher order interactions, and this is useful for describing multilevel or longitudinal models.

The focus of additive models or GAMs is to model the relationship of the independent variable y and the covariates x through a series one-dimensional smoothers nonparametericallly. Estimation is performed usually using a backfitting algorithm. There are no assumptions on the function space to which the regression function belongs.

A brief comparison with GAMs has already been provided in Section 5.8 (p. 195).

6. **Priors for RKHS scale parameters**

   On page 116, the stated priors for λ (and ψ) are only applicable in full Bayesian estimation of I-prior models. These priors are a suggestion to reflect the ignorance surrounding the true value of these parameters.

   Amendments to thesis: Stronger emphasis that these priors are only applicable to a full Bayesian analysis, and not the EM algorithm or direct optimisation method.

7. **Error in Table 4.1, p. 116**

   The 'Predictive RMSE' is not predictive. This has been changed to simply 'RMSE', with the formula now given, which is as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

   where $\hat{y}_i$ is the estimate of the i'th observation, as described in Section 4.4.

8. **Computational cost of GPR vs I-priors**

   A comment received from the examiners is as follows: For some univariate GPR models, computational cost is reduced to $O(n)$ or $O(n \log n)$ if the kernel matrix is of a specific structure such as tridiagonal or Toeplitz. Inversions of such kernel matrices would be less than $O(n^3)$.

   In our experience, we have not come across a kernel matrix with such structures before, and it is also unlikely to be the case. A Toeplitz-structured kernel matrix would suggest that the data points are somewhat equally spaced apart, so perhaps one might encounter such matrices when dealing with time series data.

9. **On Figure 4.6, p. 128**

   The posterior predictive density check allows us to compare the distribution of the observed data with the distribution that is predicted by the model. This is not for

a single y, but rather each line represents the distribution of all the data points (observed or replications).

Amendments to thesis: Added a sentence in the caption to clear the confusion.

10. **On recovering regression coefficients in a linear RKHS**

    I-prior modelling indeed allows us to perform model in a nonparametric manner. Choosing the RKHS implicitly sets the type of functions being used for regression modelling.

    Specifically for functions in the linear RKHS, one might be interested in obtaining the slope and intercept of the estimated regression function. One possible reason for this is to make comparisons to other types of models which uses the slope and intercept parameterisation explicitly (just like in this example). Ordinarily, one need not recover the slopes and intercepts in I-prior modelling, but this small example just indicates that one *may* do so if they wished.

11. **On the '99-dimensional' covariate, p. 138**

    The example in Section 4.5.3 (p. 137) concerns functional covariates, that is, each observed x_i is assumed to belong to a function space $\mathcal{X}$. As absorption data had not been measured continuously, 100 equally-spaced discretised points were measured and this makes up each data point x_i.

    As per Section 4.1.6 (p. 109), we make an implicit assumption that $\mathcal{X}$ is a Hilbert-Sobolev space with inner product given in that section. In order to apply the linear, fBm or any other kernels which make use of inner products, one should make the approximation as given by (4.11). It involves taking first differences of the 100-dimensional covariate, and this reduces to 99 dimensions.

    The above has been explained in the second paragraph on page 138.

12. **Would smoother GPR yield better results in the Tecator example?**

    As we understand it, squared exponential kernels are the de-facto kernels for Gaussian process regression. As seen from the results in Table 4.5 (p. 141), GPR does not perform well compared to I-priors or any other method for that matter.

    Perhaps performance can be improved by using 'smoother' kernels. For instance, the **kernlab** package in R provides options for the hyperbolic tangent kernel, Laplacian kernel, Bessel kernel, ANOVA Gaussian RBF kernel, and the spline kernel. It's not clear which one is best until all of them are tried on the data. Further, each of these kernels would have additional parameters which need to be tuned as well.

This highlights the advantage of I-prior regression using the fBm RKHS for smoothing: good performance with the 'defacto' fBm kernel which does not necessarily require optimising the Hurst coefficient, which simplifies estimation.

13. **Limiting form of I-priors in relation to integrated Brownian motion and cubic splines**

    The link between I-priors and integrated Brownian bridge, which has a close relationship with cubic spline smoothers, has been brought up in in Remark 4.3 (p. 109).

14. **The logit link with I-priors**

    In Chapter 5, the I-prior methodology is extended by way of a 'probit-like' link function on the regression functions. The probit link was chosen as this is compatible with a variational method of estimation—the normality of the distributions implied by the link function facilitates variational inference.

    The logit link is not compatible with the I-prior methodology. As per the latent variable motivation in Section 5.1, the I-prior is assigned to the latent regression problem. This latent regression problem is assumed to have normally distributed errors, which is one of the crucial assumptions of I-prior modelling. This, in turn, yields the probit link.

    A logit link would mean that the errors follow a logistic distribution. This, in theory, cannot motivate placing an I-prior on the regression function.

15. **On the Laplace approximation for I-probit models**

    The modes of the Laplace approximation was obtained using a quasi-Newton algorithm. No EM algorithm was used. A suggestion of using integrated nested Laplace approximations (INLA) was received.

    Amendments to thesis: The suggestion of INLA has been noted (p. 157).

16. **Variational inference for logit link**

    The logit link is not compatible with I-priors (see point 13). Even if it were used, a different form of variational inference would be required other than the mean-field variational approximation, e.g. a local variational bound (Bishop, 2006).

17. **Comparing Gaussian process priors**

    Gaussian process priors for categorical data is often used in a classification setting, because often, prediction is key. To this end, we did look at binary classification of cardiac arrhythmia using GPC and a squared exponential kernel. Additionally,

we will also run the GPC on the vowel recognition data set and include the results in Table 5.6 (p. 188).

Amendments to thesis: Added GPC to results to Table 5.6.

18. **Additional criticisms to Bayesian variable selection**

Amendments to thesis: On page 201, the following was added:

> The choice of priors for model parameters affects consistency of Bayesian model selection procedures. Specifically, improper priors cannot be used to calculate posterior model probabilities (Casella et al., 2009) — otherwise, one risks running into Lindley's paradox (Lindley, 1957).

In the footnotes, a short explanation of the Lindley paradox was included, which reads

> Briefly, in testing a point null hypothesis of the mean of a normally distributed parameter, the null hypothesis is increasingly accepted as the prior variance of the parameter approaches infinity, regardless of evidence for or against the null. The paradox is also termed Jeffreys-Lindley paradox (Robert, 2014).

19. **Clarification on model-averaged version of β**

On p. 207, a reference was made to "model-averaged" coefficients. In general, any quantity of interest Δ can be model-averaged (Madigan and Raftery, 1994), for which its posterior distribution is given by

$$p(\Delta|\mathbf{y}) = \sum_{M_k \in \mathcal{M}} p(\Delta|\mathbf{y}, M_k) p(M_k|\mathbf{y}). \tag{3}$$

The quantity Δ may be "effect size, a future observable, or the utility of a course of action" (Hoeting et al., 1999; Madigan and Raftery, 1994). Averaging in such a manner incorporates model uncertainty into analyses, which avoids overconfidence of inference surrounding Δ.

The original motivation for model averaging was to improve upon predictive accuracy. Predictions obtained in such a manner is understood to be the average prediction over a subset/all models as weighted by posterior model probabilities.

Explanatory inference using model-averaged coefficients on the other hand, is not practically meaningful. Banner and Higgs (2017) writes that "regression coefficients... may not hold equivalent interpretations across all of the models in which they appear", and one reason for this might be "interpretation of partial regression coefficients can depend on other variables that have been included in the model".

The use of model-averaged effect sizes may result in misleading inferences (Cade, 2015).

The aims of Bayesian variable selection using I-priors is two-fold: 1) to ascertain variable importance; and 2) prediction using model-averaged regression coefficients. To this end, we advocate the view of Hoeting et al. (1999) and Madigan and Raftery (1994) that quantities of interest should be model-averaged to avoid overconfidence, *and* we also agree with the assessment of Banner and Higgs (2017) and Cade (2015) that using model-averaged regression coefficients is meaningless.

To clarify, in our data applications of Sections 6.7.1 and 6.7.2 report the model-averaged regression coefficients, but no attempt of interpreting these coefficients were made. The data example in Section 6.7.3 highlights predictive accuracy of the choice of I-priors in Bayesian variable selection, in which predictive RMSE was made using model-averaged regression coefficients. In the simulation study, the aim was to assess the performance of the I-prior by looking at the discrepancy between what the BVS model thinks is the highest probability model, and what the actual model behind the data generating mechanism was.

Amendments to thesis: The following remark was added on p. 207:

> The intention of computing model-averaged regression coefficients $\boldsymbol{\theta}$ is solely for the inclusion of model uncertainty. There is a strong agreement in the Bayesian variable selection literature that that such coefficients are practically meaningless when it comes to explanatory inferences. Banner and Higgs (2017) writes that "regression coefficients... may not hold equivalent interpretations across all of the models in which they appear", and one reason for this might be "interpretation of partial regression coefficients can depend on other variables that have been included in the model". The use of model-averaged effect sizes may result in misleading inferences (Cade, 2015).