Errata for PhD Thesis:

Regression modelling using priors depending on Fisher information covariance kernels

Md. Haziq Md. Jamil
2 October 2018

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

   Why maximum entropy priors? What is entropy? Not the original motivation, but it was a nice consequence that was noticed. Motivate the principle of maximum entropy.

   - For the regression problem, one could choose to assign subjective prior on the regression function. Using Gaussian priors, this is Gaussian process regression. Gaussian and Levy process priors are described in Pillai, Wu, Liang, Mukerjee and Wolpert (2007).
   - Instead we focus on objective priors. E.g. include maximum-entropy priors, Jeffreys' priors, reference priors, g-priors, etc.

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

   A comparison is given in Bergsma (2018). Here, we make brief statements regarding advantage of maximum entropy prior over other objective priors.

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

   In the conclusion section of Chapter 3, 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.

Amendments to thesis: A linguistic argument is added at the end of Chapter 3 (page?). Most of the details have already been provided in various parts of the thesis (where?), but this additional paragraph should help make this clearer.

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

   Arbitrary kernels can be used within the I-prior methodology. But we choose specifically the linear, fBm, and Pearson RKHSs as the building blocks, and using these RKHSs we build more complex RKHSs/RKKSs using the polynomial or ANOVA construction.

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

   Apart from the additive nature of the equation in reference (Eq. 4.2, p. 103), the I-prior methodology is completely different from 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 GAM is to model the relationship of the independent variable y and the covariates x through a series one-dimensional smoothers nonparameterically. Estimation is performed usually using a backfitting algorithm. There are no assumptions on the function space to which the regression function belongs.

   Amendments to thesis: Added a short paragraph at the conclusion to Chapter 4.

6. **Priors for RKHS scale parameters**

   On p. 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 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, 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 the second equation on page 109. need to number this. It involves taking first differences of the 100-dimensional covariate, and this reduces to 99 dimensions.

    Amendments to thesis: Reference to appropriate section for clarity.

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 on 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**

    Also compare GPR with I-priors in terms of smoothness.

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 (Brümmer 2014). A suggestion of using INLA is received.

    Amendments to thesis: The suggestion of INLA will be noted in where.

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.

    Amendments to thesis: Add GPC to results in Table 5.6.

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

    impact of prior, Lindley paradox

19. **Clarification on model-averaged version of $\boldsymbol{\beta}$**

    Page 207. Yes, on their own it would probably not make sense... these betas are instead used for prediction etc. Need to read MAdigan and Raftery 1994. It is possible to do this, and may even be informative. High spike at zero tells us that this coefficient is almost zero all the time. Large confidence band means that it wasn't included in the model that often. However, don't bother interpreting these regression coefficients.

    Model averaged coefficients includes model uncertainty, and gives an indication of how important a variable is. However, interpretation not straight forward.