Estimation and variable selection in high dimensional regression models with Information Matrix priors

Mayetri Gupta[1*] and Joseph G. Ibrahim[2]

[1] Department of Biostatistics, Boston University
[2] Department of Biostatistics, University of North Carolina at Chapel Hill

1 Introduction

The development of modern scientific techniques, including large scale genomic technologies, has led to the generation of enormous amounts of data often characterized by high dimensions and complex dependence structures. In many cases, the dimensionality of variables measured (d) exceeds the number of observations (n), leading to model non-identifiability and difficulties in parameter estimation. Variable selection procedures also fail due to the impossibility of enumerating and testing massive collections of models, as well as the inability to estimate the larger models by standard procedures. To overcome these situations, it has long been known that the specification of proper priors in a Bayesian framework alleviates such a nonidentifiability problem and leads to proper posterior distributions as long as one uses a valid probability density for the data, i.e., $\int f(y|\theta)\, dy < \infty$. For example, this is often done in factor analysis, where the number of parameters typically exceeds the number of observations, and suitable proper priors are elicited to obtain proper posterior distributions. Specification of proper priors in the $d > n$ context is not an easy problem since i) one is never guaranteed that the proper prior will lead to existence of prior or posterior moments, ii) theoretically checking the existence of prior or posterior moments is often not an easy task, iii) one desires to specify a proper prior that is relatively non-informative so that the data can essentially drive the inference, iv) it is desirable to specify a prior that is at least somewhat semi-automatic in nature requiring relatively little or minimal specification of hyper-parameters, and v) one desires priors that are easy to interpret and computationally feasible.

2 Methodology

In the $d > n$ paradigm, there has been very little work on the specification of such priors and in particular, priors that satisfy i) - v) above. When $d < n$, multivariate normal priors, such as $N_d(0, \gamma I)$, are often not desirable; since for small to moderate γ, they are typically too informative, and for large γ they often lead to computationally unstable posteriors as the model becomes weakly identified. Moreover, such priors do not capture the a priori correlation in the parameters, and eliciting a prior correlation or covariance matrix when $d > n$ is a monumental task.

We develop a framework for variable and model selection in regression-type models for high dimensional problems. First, we develop a process for the specification of a general class of prior distributions, called Information Matrix (IM) priors, focusing initially on linear, and generalized linear models (although these can be applied in any parametric regression context). In general, the functional form of the IM prior is immediately obtained once a parametric statistical model is specified for the data. The kernel of the IM prior is essentially specified through the Fisher Information matrix of the parameters.

To generalize the IM priors to a proper prior in the $d > n$ case, we introduce a scalar “ridge” parameter λ in the prior construction, leading to the Information Matrix Ridge (IMR) prior. The ridge parameter λ is motivated from the ideas of reducing effects of collinearity and introducing stability in regression models with high dimensional covariates (Hoerl and Kennard, 1970). The IM and IMR priors are based on a broad generalization of Zellner’s g-prior (Zellner, 1986) for Gaussian linear models.

3 Results

Theoretical and computational analyses of the use of this prior indicate a number of highly desirable properties, including existence of the prior and implied posterior moment generating functions for many popular generalized linear models under conditions that are simple to verify, and robustness arising from existence of heavier tails than Gaussian priors. In special limiting cases, these priors also reduce to many of the commonly used priors in the regression framework, such as the Gaussian, Jeffreys’ and Zellner’s g-priors.

Several simulation studies indicated many advantages of the IMR framework over Gaussian priors in high dimensional settings. In a high-dimensional logistic regression setting, use of the IMR prior and a Bayesian model averaging-based approach (Hoeting et al., 1999) led to virtually identical predictive performance, with the IMR approach being computationally less expensive than BMA. We also demonstrate the superior performance of the IMR prior in the context of the applications of (i) discovering gene regulatory networks from genomic sequence and gene expression microarray data in a yeast cell-cycle experiment and (ii) prediction of nucleosome positions using genomic sequence data in yeast.

References

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**:55–67.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statist. Sci.*, **14**(4):382–417.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti.* Goel, P.K. and Zellner, A. (Eds.), North-Holland, Amsterdam.