Bull. Aust. Math. Soc. **93** (2016), 164–166
doi:10.1017/S0004972715001173

MODEL SELECTION AND ESTIMATING DEGREES OF FREEDOM IN BAYESIAN LINEAR AND LINEAR MIXED EFFECT MODELS

CHONG YOU

(Received 27 July 2015; first published online 13 November 2015)

2010 *Mathematics subject classification*: primary 62J05; secondary 62F15, 60J22, 62F07.

Keywords and phrases: consistency, Markov chain Monte Carlo, deviance, information criterion, resampling, bootstrap, mean field variational Bayes, spike and slab prior.

In practical applications of statistics, model selection is one of the most fundamental analysis tasks. The approaches used for selection in the context of the linear regression model have been extensively developed since the 1970s and are currently widely used. However, the ability of existing methods to produce efficient and effective analysis for high-dimensional data is very limited. Moreover, most of the current model selection techniques have been developed for the linear regression model only, and thus may not be appropriate in the context of other models, such as the linear mixed-effects model. This thesis explores the theory behind a fast, alternative method for model selection and investigates effective procedures for selecting linear mixed-effects models. The main emphasis of the paper is to construct desirable properties of variational Bayes estimates in Bayesian linear models and to develop a new generalised degrees-of-freedom measure for linear mixed-effects models.

This thesis comprises five chapters. Chapter 1 provides background information about the models and assumptions used, summarises some popular model selection techniques and, additionally, outlines a brief introduction to approximate Bayesian inference.

In Chapter 2, we introduce variational Bayes (VB) approximation, which is a fast alternative to Markov chain Monte Carlo methods for performing approximate Bayesian inference to analyse large datasets. VB is often criticised: this is typically based on empirical grounds, for being unable to produce valid statistical inferences [1, Section 1.6]. We contradict this criticism in the context of the Bayesian linear model for a particular choice of priors. In Chapter 2, we prove that under mild

Thesis submitted to the University of Sydney in August 2014; degree approved on 4 March 2015; supervisors Samuel Müller and John Ormerod.

regularity conditions, VB based estimators for the coefficients and the variance are consistent estimators of the true parameters. Furthermore, we propose two variational Bayes information criteria: the variational Akaike information criterion and the variational Bayesian information criterion. We show that the variational Akaike information criterion is asymptotically equivalent to the frequentist Akaike information criterion and that the variational Bayesian information criterion shares the same first-order asymptotic properties as the Bayesian information criterion in linear regression. Computationally, in the context of linear regression models, the variational Akaike information criterion and the variational Bayesian information criterion have no advantages over the Akaike information criterion and the Bayesian information criterion. However, they are naturally derived in Bayesian contexts with VB estimates and the asymptotic properties in linear regression models motivate the potential use of variational Bayes based information criteria for more complex models. In particular, the variational Akaike information criterion and the variational Bayesian information criterion can be used when the appropriate sample size and model size are not clear. In addition, we provide a new degrees-of-freedom measure for Bayesian linear models. These results are published in [2].

Encouraged by the consistency properties in Chapter 2, we extend the VB approach to more complicated priors in Chapter 3. Namely, we develop algorithms for model selection in the context of a Bayesian linear model with spike and slab priors and investigate the theoretical properties of the corresponding VB estimators. We show that, under mild regularity conditions, VB based estimators for the coefficients are consistent estimators of the true parameters. More importantly, we prove that the VB estimators of the model indicator variables shrink towards a probability of zero at an exponential rate if the corresponding true value of the coefficient is zero; if the true coefficient value is nonzero, then the convergence is to one at an exponential rate. This property allows us to use VB estimates to select high-dimensional models efficiently. We also investigate the selection of initial values to avoid local optimality problems. Simulation results support the hypothesis that our method is competitive in terms of efficiency and effectiveness in comparison with the various alternative model selection procedures considered.

In Chapter 4, we alter our focus from the linear model to the linear mixed-effects model, which is a commonly employed statistical model. It is highly flexible in dealing with a broad range of data but is much more complicated than the linear regression model. However, there is not yet consensus in the statistical community about which method of model selection to use. Chapter 4 considers the linear mixed-effects model selection problem by first establishing an appropriate model complexity measure. The number of unknown regression parameters, which are typically used as the degrees of freedom to measure complexity in linear regression models, may fail to work for linear mixed-effects models. Hence, we propose a new definition of generalised degrees of freedom. It is derived from a somewhat nonlikelihood point of view by using the sum of the sensitivities of the expected fitted values with respect to their underlying true means. We show that the proposed definition of generalised degrees of freedom satisfies some desirable properties: (1) it equals the number of unknown regression parameters in linear regression if the least squares estimates are used, and (2) empirically, it is nonnegative and monotone if models are nested. Furthermore, we explore and compare different solutions for approximating the generalised degrees of freedom. Two model selection procedures for linear mixed-effects models based on the generalised degrees of freedom are then proposed, which can select random effects and fixed effects simultaneously, rather than selecting fixed and random effects separately. The results of Chapter 4 are published in [3].

Chapter 5 provides a summary of the key contributions of this thesis as well as a discussion of some extensions.

References

[1] H. Rue, S. Martino and N. Chopin, 'Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations', *J. R. Stat. Soc.* B **71** (2009), 319–392.

[2] C. You, J. T. Ormerod and S. Müller, 'On variational Bayes estimation and variational information criteria for linear regression models', *Aust. N. Z. J. Stat.* **56** (2014), 73–87.

[3] C. You, S. Müller and J. T. Ormerod, 'On generalized degrees of freedom with application in linear mixed models selection', *Stat. Comput.* **25** (2015), 1–12.

CHONG YOU, School of Mathematics and Statistics,
University of Sydney, NSW 2006, Australia
e-mail: cyou6879@uni.sydney.edu.au

https://doi.org/10.1017/S0004972715001173 Published online by Cambridge University Press