Statistics and Probability Letters 79 (2009) 625–629

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A simplified adaptive fence procedure

Jiming Jiang*, Thuan Nguyen, J. Sunil Rao

University of California, Davis, United States
Oregon Health and Science University, United States
Case Western Reserve University, United States

ARTICLE INFO

Article history:
Received 27 April 2008
Received in revised form 8 August 2008
Accepted 14 October 2008
Available online 25 October 2008

ABSTRACT

In this short note, we propose a simplified adaptive fence procedure that reduces the computational burden of the adaptive fence procedure proposed by Jiang et al. [Jiang, J., Rao, J.S., Gu, Z., Nguyen, T., 2008. Fence methods for mixed model selection. Ann. Statist. 36, 1669–1692] for mixed model selection problems. The consistency property of the new procedure is established. Simulation results show that the new procedure performs very well in a small sample situation. The method is applied to a well-known data set in small area estimation.

1. Introduction

Mixed models are widely used in practice. While there is an extensive literature on inference about mixed models, including linear and generalized linear mixed models (e.g., Jiang (2007)), the literature on mixed model selection is rather sparse. Only recently have some useful results emerged. See Datta and Lahiri (2001), Jiang and Rao (2003), Fabrizi and Lahiri (2004), Meza and Lahiri (2005) and Vaida and Blanchard (2005), among others. As pointed out by Jiang et al. (2008), model selection in the context of mixed effects models is a nonconventional problem. The authors noted a number of limitations of the traditional model selection strategies when applied to mixed model situations. For example, the BIC procedure (Schwarz, 1978) relies on the effective sample size which is unclear in typical situations of mixed models. To overcome such difficulties, the authors developed a new strategy for model selection, called *fence methods*. The basic idea is to build a statistical fence to isolate a subgroup of what are called correct models. Once the fence is constructed, the optimal model is selected from those within the fence according to a criterion which can incorporate quantities of practical interest. Let $Q_M = Q_M(y, \theta_M)$ be a measure of lack-of-fit, where y represents the vector of observations, M indicates a candidate model, and θ_M denotes the vector of parameters under M. Here by lack-of-fit we mean that Q_M satisfies the basic requirement that $\mathrm{E}(Q_M)$ is minimized when M is a true model, and θ_M the true parameter vector under M. Then, a candidate model M is in the fence if

$$\hat{Q}_M \le \hat{Q}_{\tilde{M}} + c\hat{\sigma}_{M,\tilde{M}}, \tag{1}$$

where $\hat{Q}_M = \inf_{\theta_M \in \Theta_M} Q_M$, Θ_M being the parameter space under M, $\tilde{M}$ is a model that minimizes $\hat{Q}_M$ among $M \in \mathcal{M}$, the set of candidate models, and $\hat{\sigma}_{M,\tilde{M}}$ is an estimate of the standard deviation of $\hat{Q}_M - \hat{Q}_{\tilde{M}}$. The constant c on the right side of (1) can be chosen as a fixed number (e.g., $c = 1$) or adaptively.

The calculation of $\hat{Q}_M$ is usually straightforward. For example, in many cases Q_M can be chosen as the negative log-likelihood, or residual sum of squares. On the other hand, the computation of $\hat{\sigma}_{M,\tilde{M}}$ can be quite challenging. Sometimes, even if an expression can be obtained for $\hat{\sigma}_{M,\tilde{M}}$, its accuracy as an estimate of the standard deviation cannot be guaranteed

* Corresponding author at: University of California, Davis, United States. Tel.: +1 530 7548 589.
E-mail address: jiang@wald.ucdavis.edu (J. Jiang).

0167-7152/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2008.10.014

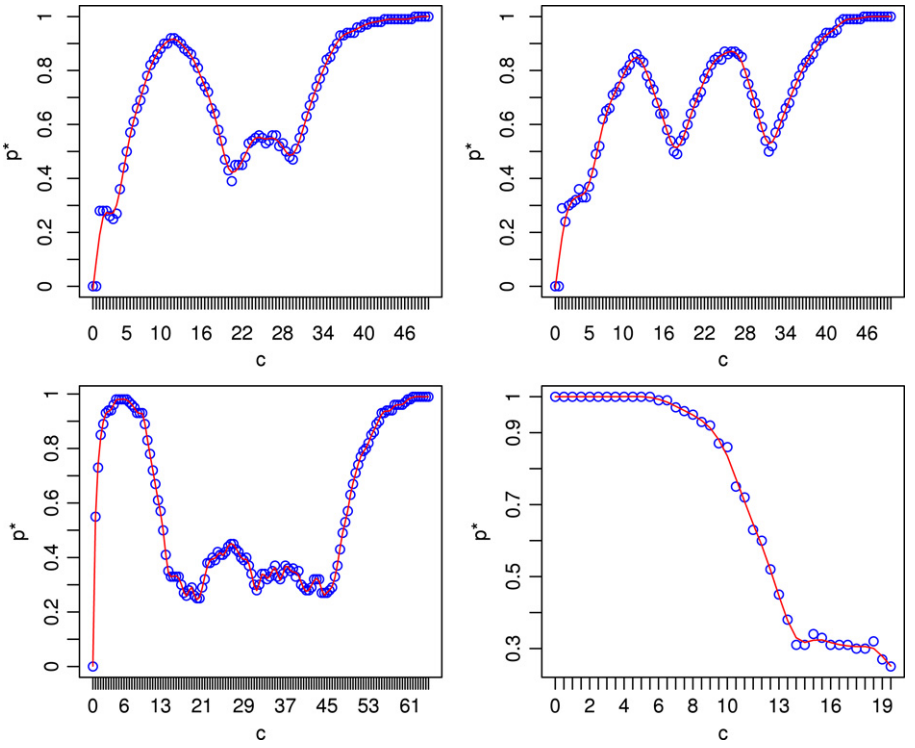

Fig. 1. Upper left: A plot of p^* based on the first simulated data set generated under Model I. Upper right: A plot of p^* based on the 35th simulated data set generated under Model I. Lower left: A plot of p^* based on the first simulated data set generated under Model II. Lower right: A plot of p^* based on the first simulated data set generated under Model II but without adjusting the baseline.

in a finite sample situation. For such a reason, this step of the fence method has complicated its applications. In this short note we propose a simplified procedure that avoids the calculation of $\hat{\sigma}_{M,\tilde{M}}$, and study the asymptotic and finite sample properties of this new procedure.

2. A simplified adaptive fence procedure

We assume that $\mathcal{M}$ contains a full model, $M_{\rm f}$, of which each candidate model is a submodel. Note that this is not a serious constraint because usually one can add a full model to $\mathcal{M}$, if it is not already included. For example, for selecting the fixed covariates one may include a model that contains all the candidate covariates, if such a model is not already under consideration. It follows that $\tilde{M} = M_{\rm f}$. To come up with the new procedure, we absorb the term $\hat{\sigma}_{M,\tilde{M}}$ on the right side of (1) into the constant c, which is to be determined adaptively. In other words, we let the adaptive constant take care the product $c\hat{\sigma}_{M,\tilde{M}}$ in the fence inequality (1). Under this simplified procedure, a model M is in the fence if

$$\hat{Q}_M - \hat{Q}_{M_{\rm f}} \le c, \tag{2}$$

where c is chosen adaptively as follows. For each $M \in \mathcal{M}$, let $p^*(M) = {\rm P}^*\{M_0(c) = M\}$ be the empirical probability of selection for M, where $M_0(c)$ denotes the model selected by the fence procedure based on (2) with the given c, and P* is obtained by bootstrapping under $M_{\rm f}$. For example, under a parametric model one can estimate the model parameters under $M_{\rm f}$ and then use a parametric bootstrap to draw samples under $M_{\rm f}$. Suppose that B samples are drawn; then $p^*(M)$ is simply the sample proportion (out of a total of B samples) for which M is selected by the fence procedure corresponding to (2) with the given c. Let $p^* = \max_{M\in\mathcal{M}} p^*(M)$. Note that p^* depends on c. Let c^* be the c that maximizes p^* and this is our choice.

Typically the optimal model is neither $M_{\rm f}$ nor M_*, the minimal model (dimensionwise; e.g., a model with only the intercept). However, these two extreme cases do need to be dealt with. Here we use the methods of baseline adjustment and threshold checking to deal with these two cases (see Jiang et al. (2008)). The baseline adjustment is done by generating an additional vector of covariates, say, $X_{\rm a}$, so that it is unrelated to the data. Then, define the model $M_{\rm f}^*$ as $M_{\rm f}$ plus $X_{\rm a}$, and replace $M_{\rm f}$ in (2) by $M_{\rm f}^*$, but let $\mathcal{M}$ remain unchanged. This way one knows for sure that the new full model, $M_{\rm f}^*$, is not an optimal model (because it is not a candidate model). The threshold checking inequality is given by $\hat{Q}_{M_*} - \hat{Q}_{M_{\rm f}^*} > d_*$, where d_* is the maximum of the left side of the threshold inequality computed under the bootstrap samples generated under M_*. If the threshold inequality holds, we ignore the right tail of the plot of p^* against c that eventually goes up and stays at 1 (see Fig. 1 for a demonstration).

Another adjustment is also considered. Notice that p^* is, in fact, a sample proportion (based on the bootstrap samples). Therefore, we construct a large sample 95% confidence lower bound,

$$p^* - 1.96\sqrt{p^*(1-p^*)/B} \tag{3}$$

where B is the bootstrap sample size. When selecting c that maximizes p^* we take (3) into account. More specifically, suppose that there are two peaks in the plot of p^* against c located at c_1 and c_2 such that $c_1 < c_2$. Let p_1^* and p_2^* be the heights of the peaks corresponding to c_1 and c_2. As long as p_1^* is greater than the confidence lower bound at p_2^*, that is, (3) with $p^* = p_2^*$, we choose c_1 over c_2. Clearly, the selection is in favor of smaller c in order to be more conservative. (In other words, we are more concerned with underfitting than overfitting.)

3. Consistency

In most practical problems there are a (large) number of candidate variables and only some of them are important. This means that the optimal model, M_{opt}, is neither the minimum model M_* (because some variables are important) nor the full model M_{f} (because not all variables are important). Therefore, without loss of generality we assume the following.

A1. There is a unique $M_{\text{opt}} \notin \{M_*, M_{\text{f}}\}$.

The next assumption states that there is a distributional separation between M_{opt} and the incorrect models that matters. Let $\mathcal{M}_-$ denote the subset of incorrect candidate models that has dimension $\le |M_{\text{opt}}|$ ($|M|$ represents the dimension of M). Write $d_M = \hat{Q}_M - \hat{Q}_{M_{\text{f}}}$, $M \in \mathcal{M}$, $d_{\text{opt}} = d_{M_{\text{opt}}}$, and $d_- = \min_{M\in\mathcal{M}_-} d_M$. Let F_{opt} and F_- be the cumulative distribution functions of d_{opt} and d_-, respectively. Let $M_0(c)$ denote the model selected by the fence method using (2) with the given c. Write $\text{P}(c) = \text{P}\{M_0(c) = M_{\text{opt}}\}$.

A2. For any $\epsilon > 0$, there are $0 < \delta < 0.1$, $c_1 < c_2 < c_3$, and $N \ge 1$ such that $\text{F}_{\text{opt}}(c_1) > 1-\epsilon$, $F_-(c_3) \le \epsilon$, $\text{P}(c_2) > 1-\delta$ and $1 - 4\delta < \text{P}(c_j) \le 1 - 3\delta$, $j = 1, 3$, if $n \ge N$.

Note that the c and c_j, $j = 1, 2, 3$, depend on n, and therefore should be denoted by c_n, $c_{n,1}$, etc., but for notational simplicity the subscript n is suppressed. The next assumption is about quality of the bootstrap approximation. Let $\text{P}^*(c)$ denotes the bootstrap version of $\text{P}(c)$.

A3. For any $\delta, \eta > 0$, there are N, N^* such that, when $n \ge N$ and $B \ge N^*$, we have

$$\text{P}\left\{\sup_{c>0} |\text{P}^*(c) - \text{P}(c)| < \delta\right\} > 1 - \eta.$$

The following theorem states the large sample behavior of c^*. The proof of the theorem is very similar to that of Theorem 3 of Jiang et al. (2008), and therefore omitted. Let M_0^* denote the model selected by the fence procedure using (2) with $c = c^*$. Note that c^* depends on the observed data, i.e., $c^* = c^*(y)$. Also let M_{opt} denote an optimal model defined as a true model with minimum dimension.

Theorem. *Under the regularity conditions* A1–A3 *there is c^* which is at least a local maximum and an approximate global maximum of p^*, such that the corresponding M_0^* is consistent in the sense that any $\delta, \eta > 0$, there are N, N^* such that* $\text{P}\{p^*(c^*) \ge 1-\delta\} \wedge \text{P}(M_0^* = M_{\text{opt}}) \ge 1 - \eta$, *if $n \ge N$ and $B \ge N^*$.*

4. A simulation study

We consider the following linear mixed model, also known as the nested error regression model:

$$y_{ij} = x_{ij}'\beta + v_i + e_{ij}, \quad i = 1, \ldots, m, j = 1, \ldots, n_i. \tag{4}$$

The number of clusters, m, is either 10 or 15. The n_i's are generated from a Poisson (3) distribution, and fixed throughout the simulations. The random effects, v_i, and errors, e_{ij}, are both generated independently from the $N(0, 1)$ distribution. The components of the covariates, x_{ij}, are to be selected from x_{ijk}, $k = 0, 1, \ldots, 5$, where $x_{ij0} = 1$; x_{ij1} and x_{ij2} are generated independently from $N(0, 1)$ and then fixed throughout; $x_{ij3} = x_{ij1}^2$, $x_{ij4} = x_{ij2}^2$, and $x_{ij5} = x_{ij1}x_{ij2}$.

The simulated data are generated under two models: I. the model that involves the linear terms only, i.e., x_{kij}, $k = 0, 1, 2$, with all the regression coefficients equal to 1; II. the model that involves both the linear and the quadratic terms, i.e., x_{kij}, $k = 0, \ldots, 5$, with all the regression coefficients equal to 1. We study the performance of the adaptive fence procedure introduced in Section 2 with or without using the confidence lower bound (3). Here Q_M is chosen as the negative log-likelihood function. The bootstrap sample size B is chosen as 100. The results based on 100 simulations are reported in Table 1 which presents the empirical probabilities of correct model selection. The results show that even in these cases of fairly small m, the performance of the adaptive fence is quite satisfactory. Note that for Model I, the method does not behave quite as well for the $m = 10$ case as for Model II, but that this problem quickly disappears by the time $m = 15$. The method that makes use of the confidence lower bound seems to perform better in the smaller m case but for the case of larger m the two methods are indistinguishable.

Fig. 1 displays some of the plots of p^* against c in various situations. The upper left plot is based on the first simulated data set generated under Model I. This plot is typical for most of the plots generated under Model I, where the highest peak in the middle corresponds to c^*. The upper right plot is based on the 35th simulated data set generated under Model II.

Table 1
Mixed model selection. Reported are empirical probabilities, in terms of percentage, based on 100 simulations for which the optimal model is selected.

Optimal model	# of clusters, m	Highest peak	Confidence lower bound
Model I	10	82	87
Model I	15	99	99
Model II	10	98	99
Model II	15	100	100

This data set is singled out because its plot is a little unusual compared to the typical situations. There are two peaks in the middle of almost the same height. In fact, the second peak (near $c = 27$) is slightly higher, although the difference is hardly distinguishable to the naked eye. However, if we choose the c^* corresponding to the second peak, we arrive at an incorrect model that has the intercept and one of the linear terms only. On the other hand, by using the confidence lower bound we will choose c^* corresponding to the first peak (near $c = 12$). This gives us Model I which is the optimal model. This plot is very helpful in illustrating the usefulness of confidence lower bound. The lower left plot is based on the first simulated data set generated under Model II, which is typical for plots generated under Model II. The lower right plot is based on the same data set but without adjusting the baseline. What happens is that, unlike for the lower left plot, there is no peak in the middle, which is typical for plots generated under Model II but without adjusting the baseline. This plot helps to explain the reason for the baseline adjustment. It should be pointed out that the threshold inequality holds in all these cases; therefore one should ignore the right tails of the plots.

5. Iowa crops data

One of the well-known problems in small area estimation was discussed in Battese et al. (1988), in which the authors presented data from 12 Iowa counties obtained from the 1978 June Enumerative Survey of the U.S. Department of Agriculture as well as data obtained from land observatory satellites on crop areas involving corn and soybeans. The objective was to predict mean hectares of corn and soybeans per segment for the 12 counties using the satellite information. In this paper, the authors introduced for the first time the nested error regression model that has since become popular in small area estimation (e.g., Rao (2003)). Their model can be expressed as (4) with $x_{ij}'\beta = \beta_0+\beta_1 x_{ij1}+\beta_2 x_{ij2}, i = 1, \ldots, 12, j = 1, \ldots, n_i$, where n_i ranges from 1 to 6. Here i represents county and j segment within the county; y_{ij} is the number of hectares of corn (or soybeans); x_{ij1} and x_{ij2} are the number of pixels classified as corn and soybeans, respectively, according to the satellite data. Furthermore, v_i is a small area specific random effect, and e_{ij} is the sampling error. It is assumed that the random effects are independent and distributed as $N(0, \sigma_v^2)$, the sampling errors are independent and distributed as $N(0, \sigma_e^2)$, and the random effects and sampling errors are uncorrelated. The authors discussed various model selection problems associated with the nested error regression, such as whether or not to include quadratic terms in the model. The model chosen by Battese et al. (1988), however, involved only linear terms.

We apply the simplified adaptive fence procedure of Section 2 to this data set. Here we consider the same group of variables as in our simulation study in Section 4, where $x_{ij0} = 1$; x_{ij1} and x_{ij2} are defined above; $x_{ij3} = x_{ij1}^2$, $x_{ij4} = x_{ij2}^2$ and $x_{ij5} = x_{ij1}x_{ij2}$. We use a predictive measure of lack-of-fit which is the squared Euclidean distance between the best linear predictor of the small area means and the empirical best predictor under the full model as Q_M. In addition, we consider models with or without the random effect v_i. The optimal models selected by the fence method are, for the corn data, $y_{ij} = \beta_0 + \beta_1 x_{ij1} + v_i + e_{ij}$; and, for the soybeans data, $y_{ij} = \beta_0 + \beta_2 x_{ij2} + v_i + e_{ij}$. (Note that the intercept may have different values under the two models even though the same notation is used.) In particular, both models have included the random effect v_i. This suggests extra variation in the data being present that is captured by the random effect. Also, both models have excluded the quadratic terms. These findings are in line with Battese et al. (1988). The main difference is that the models chosen by the fence method are simpler than those of Battese et al. (1988). Namely, the model for the corn data involves only the satellite information about the corn, while that for the soybeans data involves only the satellite information about the soybeans. Satellite data for both corn and soybeans were involved in both models of Battese et al. (1988). Fig. 2 shows the plot of p^* against c that led to the fence model selection. Note that the models selected by the fence method corresponds to the first significant peak, which is a more conservative choice in a small sample situation (Nguyen, 2008). Also note that, unlike Fig. 1, these plots are generated without the baseline adjustment proposed by Jiang et al. (2008), which was adopted in our simulation study in Section 4. As pointed out by Jiang et al. (2008, pp. 1679), in practice such an adjustment is usually unnecessary if the peak in the middle is obvious.

6. Concluding remarks

The simple modification of the adaptive fence method proposed in this short note has important practical implications. As mentioned, for complex problems the computation of $\hat{\sigma}_{M,\tilde{M}}$ is nontrivial. Even if a formula can be obtained for $\hat{\sigma}_{M,\tilde{M}}$, the computational burden significantly increases with the adaptive procedure due to the need for bootstrapping. Therefore, the proposed simplification is an important step towards making the fence method more suitable for a wide variety of

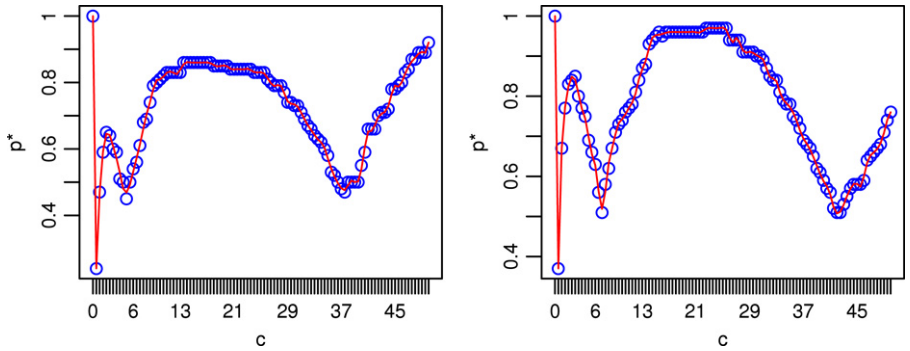

Fig. 2. Plots of p^* against c for the Iowa crops data. Left plot: Model selection for the corn data; Right plot: Model selection for the soybeans data. The first significant peak in each plot corresponds to the model selected by the fence.

problems. Furthermore, we show that the proposed simplification maintains consistency as well as the excellent finite sample performance of the original adaptive fence method (Jiang et al., 2008).

As pointed by Jiang et al. (2008), traditional methods such as the information criteria are not suitable for nonconventional problems such as mixed model selection. Still, these methods are being used in practice for selecting mixed effects models. A question of interest then is how does the fence method compare to the information criteria in mixed model selection, even though the latter methods may be considered *ad hoc* in such situations. A simulation study was recently carried out by Nguyen (2008), in which the author compared a version of adaptive fence method with different information criteria, including AIC, BIC, CAIC (consistent AIC, Bozdogan (1987)) and HQ (Hannan and Quinn, 1979), in selecting a linear mixed model for longitudinal data with many candidate covariates. (The intention was to develop a fence method for high-dimensional model selection problems.) The results showed that the fence method significantly outperformed all the traditional methods that it was being compared with in this case.

Acknowledgments

Jiming Jiang is partially supported by NSF grants DMS-0203676 and DMS-0402824. J. Sunil Rao is partially supported by NSF grants DMS-0203724, DMS-0405072 and NIH grant K25-CA89868.

References

Battese, G.E., Harter, R.M., Fuller, W.A., 1988. An error-components model for prediction of county crop areas using survey and satellite data. J. Amer. Statist. Assoc. 80, 28–36.
Bozdogan, H., 1987. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika 52, 345–370.
Datta, G.S., Lahiri, P., 2001. Discussions on a paper by Efron and Gous, in: Model Selection, IMS P. Lahiri (Ed.) in: Lecture Notes/Monograph 38.
Fabrizi, E., Lahiri, P., 2004. A new approximation to the Bayes information criterion in finite population sampling. Tech. Report. Dept. of Math., Univ. of Maryland.
Hannan, E.J., Quinn, B.G., 1979. The determination of the order of an autoregression. J. Roy. Statist. Soc. B 41, 190–195.
Jiang, J., 2007. Linear and Generalized Linear Mixed Models and Their Applications. Springer, New York.
Jiang, J., Rao, J.S., 2003. Consistent procedures for mixed linear model selection. Sankhya A 65, 23–42.
Jiang, J., Rao, J.S., Gu, Z., Nguyen, T., 2008. Fence methods for mixed model selection. Ann. Statist. 36, 1669–1692.
Meza, J., Lahiri, P., 2005. A note on the C_p statistic under the nested error regression model. Survey Methodology 31, 105–109.
Nguyen, T., 2008. New procedures of fence methods and their applications. Ph.D. Dissertation. Dept. of Statist., Univ. of Calif., Davis, CA.
Rao, J.N.K., 2003. Small Area Estimation. Wiley, New York.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.
Vaida, F., Blanchard, S., 2005. Conditional Akaike information for mixed effects models. Biometrika 92, 351–370.