Journal of Data Science **9**(2011), 15-21

Asymptotic Equivalence between Cross-Validations and Akaike Information Criteria in Mixed-Effects Models

Yixin Fang
Georgia State University

Abstract: For model selection in mixed effects models, Vaida and Blanchard (2005) demonstrated that the marginal Akaike information criterion is appropriate as to the questions regarding the population and the conditional Akaike information criterion is appropriate as to the questions regarding the particular clusters in the data. This article shows that the marginal Akaike information criterion is asymptotically equivalent to the leave-one-cluster-out cross-validation and the conditional Akaike information criterion is asymptotically equivalent to the leave-one-observation-out cross-validation.

Key words: AIC, degrees of freedom, functional data, model selection.

1. Introduction

Linear mixed effects models (Laird and Ware, 1982) are powerful for the analysis of clustered data, longitudinal data, meta data analysis, and recently for the functional data analysis (e.g., Brumback and Rice, 1998; Rice and Wu, 2001). Since the beginning, many model selection procedures have been proposed for the linear mixed effects models. Among them, Akaike information criteria (AIC; Akaike, 1973) are most popular, and they are of a similar formula, AIC = -2log likelihood + $2K$, where K is the number of effective degrees of freedom measuring the model complexity.

In the traditional AIC criteria, K simply counts the number of fixed parameters; see for example, Pinheiro and Bates (2000) and Ngo and Brand (2002). Vaida and Blanchard (2005) demonstrated that the marginal Akaike information criterion (mAIC) is appropriate as to the questions regarding the population and the conditional Akaike information criterion (cAIC) is appropriate as to the questions regarding the particular clusters in the data, where in the mAIC the effective degrees of freedom are the number of fixed parameters and in the cAIC the effective degrees of freedom are the ρ proposed by Hodges and Sargent (2001). Without assuming that the scaled variance-covariance matrix of random effects

is known, Liang *et al.* (2008) developed a general conditional AIC. Actually, the general cAIC developed by Liang *et al.* (2008) coincides with the concept of generalized degrees of freedom developed by Ye (1998). This finding clearly classifies the AIC criteria for mixed effects models existing in the literature into two main streams, the mAIC and the cAIC.

To further investigate the mAIC and the cAIC, I attempt to find connections between the cross-validation procedures and the AIC criteria for linear mixed effects models. It is well known that for ordinary linear regression models, the leaving-one-out cross-validation is asymptotically equivalent to the AIC criterion (Stone, 1977). In this article I show that the leave-one-cluster-out cross-validation (CLCV) is asymptotically equivalent to the mAIC and the leave-one-observation-out cross-validation (OBCV) is asymptotically equivalent to the cAIC. The CLCV and the OBCV were applied by Wu and Zhang (2002) for bandwidth selection in the local polynomial mixed effects models for longitudinal data (there they are named respectively as SJCV and PTCV), but the intricate difference between the CLCV and the OBCV was not discussed. After establishing the asymptotic equivalences between the CLCV and the mAIC and between the OBCV and the cAIC, I can conclude that the CLCV is appropriate as to the questions regarding the population and the OBCV is appropriate as to the questions regarding the particular clusters in the data. This conclusion applies to other model selection problems, such as bases selection in functional data analysis, and bandwidth selection in the local polynomial mixed effects models.

2. Main Results

As in Laird and Ware (1982) and Vaida and Blanchard (2005), assume the observation y_{ij} of subject j in cluster i can be modeled by

$$y_{ij} = x_{ij}^T\beta + z_{ij}^T b_i + \epsilon_{ij}, \tag{2.1}$$

where $i = 1, \cdots, m$, $j = 1, \cdots n_i$, β is the p-vector of fixed effects, b_i is the q-vector of random effects for cluster i following $N(0, \sigma^2 G)$ independently, ϵ_{ij} is following $N(0, \sigma^2)$ independently of one another and b_i, and G is $q \times q$ positive definite. At the cluster level, the model is

$$y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{2.2}$$

where $X_i = (x_{i1}, \cdots, x_{in_i})^T$ is $n_i \times p$ matrix of rank of p, $Z_i = (z_{i1}, \cdots, z_{in_i})^T$ is $n_i \times q$ matrix of rank q, and $\epsilon_i = (\epsilon_{i1}, \cdots, \epsilon_{in_i})^T$ is following $N(0, \sigma^2 I_{n_i})$. Furthermore, letting N be the total number of observations, the model can be written as

$$y = X\beta + Zb + \epsilon, \tag{2.3}$$

where $X = (X_1^T, \cdots, X_m^T)^T$ is $N \times p$ matrix of rank p, $Z = \text{diag}(Z_1, \cdots, Z_m)$ is $N \times r$ block-diagonal of rank $r = mq$, $b = (b_1^T, \cdots, b_m^T)^T$ is r-vector following $N(0, \sigma^2 G_0)$ with the block-diagonal matrix $G_0 = \text{diag}(G, \cdots, G)$, and $\epsilon = (\epsilon_1^T, \cdots, \epsilon_m^T)^T$ is following $N(0, \sigma^2 I_N)$.

For the population focus, the interest is in the fixed effects (i.e. β) and the model can be assessed by the prediction error evaluated at a new cluster,

$$\text{Err}_{\text{CL}} = E(y_{m+1} - X_{m+1}\hat{\beta})^T (I_{n_{m+1}} + Z_{m+1} G Z_{m+1}^T)^{-1} (y_{m+1} - X_{m+1}\hat{\beta})/n_{m+1}, \tag{2.4}$$

where X_{m+1} and Z_{m+1} are the predictors of the new cluster, y_{m+1} is the outcome of the new cluster, and $\hat{\beta}$ is the estimate of β by the EM algorithm in Laird and Ware (1982) based on all the training data. This can be estimated by the leave-one-cluster-out cross-validation,

$$\text{CLCV} = \sum_{i=1}^{m} (y_i - X_i \hat{\beta}^{[i]})^T (I_{n_i} + Z_i \hat{G}^{[i]} Z_i^T)^{-1} (y_i - X_i \hat{\beta}^{[i]})/(mn_i), \tag{2.5}$$

where $\hat{\beta}^{[i]}$ and $\hat{G}^{[i]}$ are respectively the estimates of β and G by the same method based on the training data without cluster i.

For the cluster focus, the interest is in the cluster effects (i.e. b_i) and the model can be assessed by the prediction error evaluated at a new observation in each cluster,

$$\text{Err}_{\text{OB}} = \sum_{i=1}^{m} E(y_{i(n_i+1)} - x_{i(n_i+1)}^T \hat{\beta} - z_{i(n_i+1)}^T \hat{b}_i)^2/m, \tag{2.6}$$

where $y_{i(n_i+1)}$, $x_{i(n_i+1)}$ and $z_{i(n_i+1)}$ are respectively the outcome and predictors of the new observation in cluster i, and $\hat{\beta}$ and $\hat{b}_i$ are respectively the estimates of β and b_i based on all the training data. This can be estimated by the leave-one-observation-out cross-validation,

$$\text{OBCV} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - x_{ij}^T \hat{\beta}^{[i,j]} - z_{ij}^T \hat{b}_i^{[i,j]})^2/N, \tag{2.7}$$

where $\hat{\beta}^{[i,j]}$ and $\hat{b}_i^{[i,j]}$ are respectively the estimates of β and b_i based on the training data without subject j in cluster i.

Following the arguments in Stone (1977), I obtain Theorem 1, which implies the asymptotical equivalence between the CLCV and the mAIC as model selection procedures.

Theorem 1. Assume that G is known, $n_i \equiv n$, and (y_i, X_i, Z_i) are i.i.d. in a common distribution. Let $W_i = I_{n_i} + Z_i G Z_i^T$. As m goes to infinity,

$$\begin{aligned} & \sum_{i=1}^{m} (y_i - X_i \hat{\beta}^{[i]})^T W_i^{-1} (y_i - X_i \hat{\beta}^{[i]}) \\ = & \sum_{i=1}^{m} (y_i - X_i \hat{\beta})^T W_i^{-1} (y_i - X_i \hat{\beta}) + 2p\sigma^2 + o_p(1). \end{aligned} \tag{2.8}$$

If G is unknown and W_i is estimated by $I_{n_i} + Z_i \hat{G}^{[i]} Z_i^T$, the penalty term $2p\sigma^2$ in the RHS of (2.8) becomes $2(p+d)\hat{\sigma}^2$, where d is number of parameters in G. Still, as model selection procedures, the CLCV and the mAIC are asymptotically equivalent. The assumption $n_i \equiv n$ can be replaced by $\sum_{i=1}^{m} n_i/m \to n_0$ as $m \to \infty$, because under this new assumption the difference between (2.5) and (2.8) is $o_p(1)$.

Motivated by the proof of Lemma 1 in Wang *et al.* (2000), I obtain Theorem 2, which implies the asymptotical equivalence between the OBCV and the cAIC. The proof is in Appendix. To state Theorem 2, let H_1 be the hat matrix mapping the observed data vector y in model (2.3) into the fitted vector $\hat{y} = X\hat{\beta} + Z\hat{b}$, that is $\hat{y} = H_1 y$. This matrix was first proposed by Hodges and Sargent (2001) and also used in Vaida and Blanchard (2005) and Liang *et al.* (2008); see Appendix for details.

Theorem 2. Assume that G is known. Letting $k = k(i,j) = \sum_{v=1}^{i-1} n_v + j$, we have

$$\text{OBCV} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\frac{y_{ij} - x_{ij}^T \hat{\beta} - z_{ij}^T \hat{b}_i}{1 - h_{kk}} \right)^2, \tag{2.9}$$

where h_{kk} is the $(k(i,j), k(i,j))$ component of H_1.

As Golub *et al.* (1979), replacing h_{kk} in (2.9) by $tr(H_1)/N$ leads to

$$\text{GCV} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left[\frac{y_{ij} - x_{ij}^T \hat{\beta} - z_{ij}^T \hat{b}_i}{1 - tr(H_1)/N} \right]^2. \tag{2.10}$$

Noting that $\sum_{k=1}^{N} h_{kk} = tr(H_1)$, we see that the difference between the GCV and the OBCV is $o_p(1/N)$, under the assumption that $\sum_{i=1}^{m} n_i/m \to n_0$ as $m \to \infty$.

Furthermore, noting that $1/(1-x)^2 \approx 1 + 2x$, we see that the difference between the GCV in (2.10) and the cAIC in Vaida and Blanchard (2005) is $o_p(1/N)$. Therefore, this implies the asymptotical equivalence between the OBCV and the mAIC as model selection procedures, where K is equal to $tr(H_1)$, denoted as ρ in Hodges and Sargent (2001).

3. Practical Issues

In Theorems 1 and 2, I assume that G is known temporally. This assumption is not needed in constructions of the CLCV and the OBCV. Therefore, the OBCV procedure is more general than the cAIC (Vaida and Blanchard 2005), where G is assumed to be known, and it is more straightforward than the general cAIC (Liang *et al.* 2008), which requires the Monte Carlo method to compute the penalty term.

For the purpose of model selection, the CLCV is asymptotically equivalent to the mAIC and the OBCV is asymptotically equivalent to the cAIC. Therefore, the CLCV is appropriate as to the questions regarding the population and the OBCV is appropriate as to the questions regarding the particular clusters in the data.

Here the asymptotical equivalence has two meanings. Let me take the CLCV and the mAIC for example. For estimating the prediction error $\mathrm{Err}_{\mathrm{CL}}$, both the CLCV and the mAIC, as estimates, are asymptotically unbiased. For selecting one appropriate, parsimonious model including only a subset of predictors, both the CLCV and the mAIC, as criteria, are asymptotically the same.

Moreover, the idea of choosing between the mAIC and the cAIC (or equivalently between the CLCV and the OBCV) is an example of the Focused Information Criterion (FIC) proposed by Claeskens and Hjort (2003). As pointed out by Claeskens and Hjort (2003), a model selector should focus on the parameters singled out for interest, instead of on selecting a single model with good overall properties.

4. Discussion

This article shows connections, respectively, between the leave-one-cluster-out cross-validation and the marginal AIC, and between the leave-one-observation-out cross-validation and conditional the AIC. The results can be extended to functional data analysis, where one curve is one cluster. In functional data analysis, when predicting a new curve is of interest, leave-one-curve-out cross-validation is appropriate, and when predicting future observations in particular curves is of interest, leave-one-point-out cross-validation is appropriate.

Appendix

Lemma 1. Assume that G is known, and for some $r \times r$ matrix Δ we have $G_0 = (\Delta^T\Delta)^{-1}$. Let $M = \begin{pmatrix} X & Z \\ \mathbf{0} & -\Delta \end{pmatrix}$ and $H_1 = (X : Z)(M^T : M)^{-1}(X : Z)^T$.

Then we have

$$(\hat{\beta}^T : \hat{b}^T)^T = (M^T : M)^{-1}(X : Z)^T y, \text{ and } \hat{y} = X\hat{\beta} + Z\hat{b} = H_1 y.$$

Lemma 1 was proved in Hodges and Sargent (2001) and was used in Vaida and Blanchard (2005) and Liang *et al.* (2008). Here $\hat{\beta}$ coincides with the estimator of Harville (1977) and $\hat{b}$ with the empirical Bayes estimator $\hat{b} = E(b|y, \hat{\beta})$.□

Let $y^{[i,j]}$, $X^{[i,j]}$ and $Z^{[i,j]}$ be, respectively, the matrices y, X and Z without subject j in cluster i, let $\hat{\beta}^{[i,j]}$ and $\hat{b}^{[i,j]}$ be, respectively, the estimates of β and b based on the training data without y_{ij}. Now for a given subject (i, j), define an N-vector y^* such that $y^*_{i'j'} = y_{i'j'}$ for any $(i', j') \neq (i, j)$ and $y^*_{ij} = x_{ij}\hat{\beta}^{[i,j]} + z_{ij}\hat{b}^{[i,j]}$. Let $\hat{\beta}^*$ and $\hat{b}^*$ be respectively the estimates of β and b based on data (y^*, X, Z). Since

$$\begin{aligned} & \| y^* - X\beta - Zb \|^2 \geq \| y^{[i,j]} - X^{[i,j]}\beta - Z^{[i,j]}b \|^2 \\ \geq & \| y^{[i,j]} - X\hat{\beta}^{[i,j]} - Z\hat{b}^{[i,j]} \|^2 = \| y^* - X\hat{\beta}^{[i,j]} - Z\hat{b}^{[i,j]} \|^2, \end{aligned}$$

together with Lemma 1, we have the following leave-one-out lemma.

Lemma 2. (Leave-one-out Lemma) Following the above notation, we have

$$x_{ij}^T\hat{\beta}^* + z_{ij}^T\hat{b}^* = x_{ij}^T\hat{\beta}^{[i,j]} + z_{ij}^T\hat{b}^{[i,j]}.$$

Actually, Lemma 2 holds in many other settings; see for example, Wahba (1990), Hastie and Tibshirani (1990), and Wang *et al.* (2000).□

Proof of Theorem 2. Let $k = k(i, j) = \sum_{v=1}^{i-1} n_v + j$. By Lemma 1, we have $\hat{y} = H_1 y$, in particular, $\hat{y}_{ij} = \sum_{l=1}^{N} h_{kl} y_l$, where h_{kl} is the (k, l) component of H_1. By Lemma 2, we have

$$\hat{y}_{ij}^{[i,j]} = x_{ij}^T\hat{\beta}^{[i,j]} + z_{ij}^T\hat{b}^{[i,j]} = x_{ij}^T\hat{\beta}^* + z_{ij}^T\hat{b}^* = \sum_{l=1}^{N} h_{kl} y_l^* = \sum_{l \neq k} h_{kl} y_l + h_{kk} y_{ij}^{[i,j]}.$$

Combining the above two formulas, we have $\hat{y}_{ij} - \hat{y}^{[i,j]} = h_{kk}(y_{ij} - \hat{y}^{[i,j]})$. Then by some algebra, we have $y_{ij} - \hat{y}_{ij}^{[i,j]} = (y_{ij} - \hat{y}_{ij})/(1 - h_{kk})$. □

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, 267-281.

Brumback, B. A. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* **93**, 961-976.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association* **98**, 900-916.

Golub, G., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215-224.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall.

Hodges, J. S. and Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly parameterized models. *Biometrika* **88**, 367-379.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.

Liang, H., Wu, H., and Zou, G. (2008). A note on conditional AIC for linear mixed effects models. *Biometrika* **95**, 773-778.

Ngo, L. and Brand, R. (2002). *Model selection in linear mixed effects models using SAS Proc Mixed.* SUGI 2002; 22.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS.* Springer.

Rice, J. A. and Wu, C.O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253-259.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B* **39**, 44-47.

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM.

Wang, Y., Guo, W., and Brown, M. B. (2000). Spline smoothing for bivariate data with applications to association between hormones. *Statistica Sinica* **10**, 377-397

Wu, H. and Zhang, J. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of American Statistist Association* **97**, 883-897.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika* **92**, 351-370.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of American Statistist Association* **93**, 120-131.

Received April 29, 2009; accepted August 25, 2009.

Yixin Fang
750 COE, 7th floor, 30 Pryor Street
Atlanta, GA 30303, USA
matyxf@langate.gsu.edu