A Diagnostic for Assessing the Influence of Cases on the Prediction of Random Effects in a Mixed Model

by

Joseph E. Cavanaugh [1]

e-mail: joe-cavanaugh@uiowa.edu

Department of Biostatistics C22-GH, College of Public Health, 200 Hawkins Drive,
The University of Iowa, Iowa City, IA 52242

and

Junfeng Shang

e-mail: js7a4@mizzou.edu

Department of Statistics, 146 Middlebush Hall,
University of Missouri, Columbia, MO 65211

Abstract: A diagnostic defined in terms of the Kullback-Leibler directed divergence is developed for identifying cases which impact the prediction of the random effects in a mixed model. The diagnostic compares two conditional densities governing the prediction of the random effects: one based on parameter estimates computed using the full data set, the other based on parameter estimates computed using a case-deleted data set. We present the definition of the diagnostic and derive a formula for its evaluation. Its performance is investigated in an application where exam scores are modeled using a mixed model containing a fixed exam effect and a random subject effect.

Key words: Case-deletion diagnostic; Kullback-Leibler divergence; influence diagnostic; predictive influence function.

Running title: Assessing Predictive Influence in a Mixed Model

[1] Corresponding author. Phone: (319) 384-5024. Fax: (319) 384-5018.

1. Introduction

Diagnostics for the detection of influential cases are widely used in statistical modeling. In linear regression, measures such as Cook's distance, DFBETAS, DFFITS, studentized residuals, and COVRATIO are routinely applied to flag cases which substantially affect the fitted model and associated results (see Belsley, Kuh, and Welsch, 1980; Cook and Weisberg, 1982). The purpose of such diagnostics is to gauge the impact of a case on a specific inferential objective.

In a Bayesian context, Johnson and Geisser (1982, 1983) proposed a measure for the detection of cases that influence the prediction of future values in linear modeling applications. Their measure is based on evaluating the disparity between the predictive densities formulated using the full data set and a case-deleted data set. The disparity between densities will often be reflected by differences in both location and shape, and can be characterized via the Kullback-Leibler (1951) directed divergence. Johnson and Geisser referred to their diagnostic as a *predictive influence function*. In a frequentist context, Cavanaugh and Johnson (1999) developed such a function for detecting cases that affect the prediction of the unobserved states in a state-space modeling framework. Extending this premise, Cavanaugh and Oleson (2001) proposed such a function for flagging cases that alter the recovery of missing values in missing data applications.

In this paper, our goal is to assess the impact of a specific case on the prediction of the random effects in a mixed model. Following Cavanaugh and Oleson (2001), we define the predictive influence function in this setting. We then derive a computational formula for its evaluation. This formula depends on parameter estimates based on both the full data set and a case-deleted data set. Such estimates may be conveniently obtained via the EM algorithm (Dempster, Laird, and Rubin, 1977) or some other fitting procedure for mixed models. One advantage of employing the EM algorithm is that the predicted values for the random effects are produced via the E-step.

Our paper is organized as follows. In Section 2, the diagnostic is introduced and a formula for its computational evaluation is presented. (The derivation of the formula is provided in the Appendix.) In Section 3, the performance of the diagnostic is investigated in an application where exam scores are modeled using a mixed model containing a fixed exam effect and a random subject effect. Section 4 concludes.

2. Derivation of the Predictive Influence Function in the Framework of Mixed Models

In this section, we introduce the diagnostic, present a computational formula for its case-by-case evaluation, and discuss its fundamental properties.

The general linear mixed model can be defined as

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i, \qquad i = 1, \ldots, m, \tag{2.1}$$

where y_i, β, b_i, and ε_i are all vectors, and X_i and Z_i are matrices. Specifically, y_i denotes an $n_i \times 1$ vector of n_i responses observed on the ith subject; X_i and Z_i are $n_i \times p$ and $n_i \times q$ design matrices, respectively; β is a $p \times 1$ fixed effects parameter vector; b_i is a $q \times 1$ random effects vector distributed as $N(0, G)$; and ε_i is an $n_i \times 1$ error vector distributed as $N(0, \sigma^2 R_i)$. G and R_i are $q \times q$ and $n_i \times n_i$ positive definite matrices, respectively, and σ^2 is a positive scalar. It is assumed that the vectors $b_1, \ldots, b_m, \varepsilon_1, \ldots, \varepsilon_m$ are distributed independently.

We regard case i as the response vector for the ith subject, y_i. Thus, the total number of cases is m. The total number of observations will be denoted by $N = \sum_{i=1}^{m} n_i$.

In the preceding model, the fixed effects parameters β need to be estimated, and the random effects b_i need to be predicted. Generally, G will consist of variance parameters which need to be estimated; R_i will be known for $i = 1, \ldots, m$, but σ^2 will need to be estimated.

A more succinct representation of model (2.1) can be obtained by combining all m subject-specific models into one overall model. This model will have the form

$$Y = X\beta + Zb + \varepsilon. \tag{2.2}$$

Here, Y denotes the $N \times 1$ response vector $({y_1}', \ldots, {y_m}')'$; X is an $N \times p$ design matrix defined as $X = [{X_1}' \cdots {X_m}']'$; Z is an $N \times mq$ block diagonal design matrix comprised of the m blocks $Z_1, \ldots, Z_m$; β is the $p \times 1$ fixed effects parameter vector; b is the $mq \times 1$ random effects vector $({b_1}', \ldots, {b_m}')'$; and ε is the $N \times 1$ error vector $({\varepsilon_1}', \ldots, {\varepsilon_m}')'$. We assume $b \sim N(0, D)$ and $\varepsilon \sim N(0, \sigma^2 R)$, with b and ε distributed independently. Here, R and D are positive definite block diagonal matrices: R is $N \times N$ and comprised of the m blocks $R_1, \ldots, R_m$, and D is $mq \times mq$ and comprised of m identical blocks, each of which is G.

The conditional distribution of Y given b and the marginal distribution of Y follow directly from model (2.2). We have

$$\begin{aligned} Y \,|\, b &\sim N(X\beta + Zb, \sigma^2 R), \qquad \text{and} \\ Y &\sim N(X\beta, ZDZ' + \sigma^2 R). \end{aligned}$$

Moreover, with model (2.2), the "posterior" distribution of b given Y can be derived by the use of Bayes' rule. With

$$V = ZDZ' + \sigma^2 R,$$

we have

$$b \,|\, Y \sim N(DZ'V^{-1}(Y - X\beta), (Z'(\sigma^2 R)^{-1} Z + D^{-1})^{-1}). \tag{2.3}$$

Let θ denote the unknown parameter vector, consisting of elements of the vector β and the matrix D along with the scalar σ^2. The evaluation of our diagnostic will require estimates of θ based on both the full data set and the data set with a specific case y_i deleted ($i = 1, \ldots, m$). The vector Y denotes the full data set; let the vector

Y^i denote Y with case y_i removed. Let $\hat{\theta}$ denote an estimate of θ based on Y, and let $\hat{\theta}^i$ denote an estimate of θ based on Y^i. The estimates $\hat{\theta}, \hat{\theta}^1, \ldots, \hat{\theta}^m$ could be obtained using maximum likelihood (e.g., via the EM algorithm) or some alternate fitting procedure for mixed models (e.g., restricted maximum likelihood).

The prediction of the vector of random effects b is governed by the conditional density $f_{\hat{\theta}}(b \mid Y)$. The influence of case y_i on the prediction of b might then be evaluated by gauging the disparity between $f_{\hat{\theta}}(b \mid Y)$ and $f_{\hat{\theta}^i}(b \mid Y)$. The Kullback-Leibler (1951) directed divergence assesses this disparity by reflecting the difference in expectation between $\log f_{\hat{\theta}}(b \mid Y)$ and $\log f_{\hat{\theta}^i}(b \mid Y)$, where the expectation is taken with respect to $f_{\hat{\theta}}(b \mid Y)$.

We define the *predictive influence function* (PIF) for measuring the influence of case y_i on the prediction of b as

$$\mathrm{PIF}(i) = \int \left\{ \log \left[\frac{f_{\hat{\theta}}(b \mid Y)}{f_{\hat{\theta}^i}(b \mid Y)} \right] \right\} f_{\hat{\theta}}(b \mid Y) db. \tag{2.4}$$

$\mathrm{PIF}(i)$ is always nonnegative (Kullback, 1968, pp. 14-15). Moreover, the magnitude of $\mathrm{PIF}(i)$ will reflect the degree to which the ith case is influential.

In order to apply the predictive influence function to identify unusual cases among the response vectors $y_1, \ldots, y_m$, the estimates $\hat{\theta}, \hat{\theta}^1, \ldots, \hat{\theta}^m$ must be obtained. The diagnostic (2.4) is then evaluated using a computational formula that depends upon Y, $\hat{\theta}$, and $\hat{\theta}^i$. This formula is presented and discussed in what follows. A plot of the $\mathrm{PIF}(i)$ values versus the case index i can then be constructed. Peaks in the plot correspond to cases which are potentially influential.

For the evaluation of $\mathrm{PIF}(i)$, the diagnostic can be partitioned into two components. Using $E_{\hat{\theta}}[\,\cdot \mid Y\,]$ to denote the expectation under $f_{\hat{\theta}}(b \mid Y)$, we can write

$$\begin{aligned} \mathrm{PIF}(i) &= \int [\log f_{\hat{\theta}}(b \mid Y)] f_{\hat{\theta}}(b \mid Y) db - \int [\log f_{\hat{\theta}^i}(b \mid Y)] f_{\hat{\theta}}(b \mid Y) db \\ &= E_{\hat{\theta}}[\log f_{\hat{\theta}}(b \mid Y) \mid Y] - E_{\hat{\theta}}[\log f_{\hat{\theta}^i}(b \mid Y) \mid Y]. \end{aligned} \tag{2.5}$$

Prior to the presentation of the computational formula for $\mathrm{PIF}(i)$, we comment on the form of (2.5). The density $f_\theta(\,b\,|\,Y\,)$ depends on the conditional mean vector $E_\theta[\,b\,|\,Y\,]$ and the conditional covariance matrix $\mathrm{Var}_\theta[\,b\,|\,Y\,]$. $E_\theta[\,b\,|\,Y\,]$ is the predictor of the random effect vector b; the elements of $\mathrm{Var}_\theta[\,b\,|\,Y\,]$ reflect the accuracy of the predictor. If case i influences the prediction of b (either through the predictor itself or through the predictor accuracy), the densities $f_{\hat\theta}(\,b\,|\,Y\,)$ and $f_{\hat\theta^i}(\,b\,|\,Y\,)$ will be discrepant. This discrepancy should be exhibited in the difference between the expectations $E_{\hat\theta}[\,\log f_{\hat\theta}(\,b\,|\,Y\,)\,|\,Y\,]$ and $E_{\hat\theta}[\,\log f_{\hat\theta^i}(\,b\,|\,Y\,)\,|\,Y\,]$, and therefore reflected in the value of $\mathrm{PIF}(i)$.

To present the computational formula for $\mathrm{PIF}(i)$, we require notation for parameter estimates based on both the full data set and a case-deleted data set. We will use $\hat\beta$, $\hat D$, and $\hat\sigma^2$ to denote the estimates of β, D, and σ^2 based on Y. Additionally, let

$$
\begin{aligned}
\hat V &= Z\hat D Z' + \hat\sigma^2 R, \\
B &= \hat D Z' \hat V^{-1}(Y - X\hat\beta), \qquad \text{and} \\
A &= Z'(\hat\sigma^2 R)^{-1} Z + \hat D^{-1}.
\end{aligned}
$$

Note that B represents $E_{\hat\theta}[\,b\,|\,Y\,]$, or equivalently, the predicted value of b under the model $f_{\hat\theta}(\,b\,|\,Y\,)$. A denotes the inverse of the covariance matrix $\mathrm{Var}_{\hat\theta}[\,b\,|\,Y\,]$.

Analogously, we will use $\hat\beta^i$, $\hat D^i$, and $(\hat\sigma^2)^i$ to denote the estimates of β, D, and σ^2 based on Y^i, and let

$$
\begin{aligned}
\hat V^i &= Z\hat D^i Z' + (\hat\sigma^2)^i R, \\
B^i &= \hat D^i Z' (\hat V^i)^{-1}(Y - X\hat\beta^i), \qquad \text{and} \\
A^i &= Z'((\hat\sigma^2)^i R)^{-1} Z + (\hat D^i)^{-1}.
\end{aligned}
$$

Note that B^i represents $E_{\hat\theta^i}[\,b\,|\,Y\,]$, or equivalently, the predicted value of b under the model $f_{\hat\theta^i}(\,b\,|\,Y\,)$. A^i denotes the inverse of the covariance matrix $\mathrm{Var}_{\hat\theta^i}[\,b\,|\,Y\,]$.

Our computational formula for $\mathrm{PIF}(i)$ results from simplifying each of the conditional expectations in (2.5) to expressions based on only B, A, B^i, and A^i. The derivation is presented in the Appendix. We obtain

$$\begin{aligned} \mathrm{PIF}(i) &= E_{\hat\theta}[\,\log f_{\hat\theta}(\,b\,|\,Y\,)\,|\,Y\,] - E_{\hat\theta}[\,\log f_{\hat\theta^i}(\,b\,|\,Y\,)\,|\,Y\,] \\ &= \frac{1}{2}(\log|A| - \log|A^i|) - \frac{1}{2}[mq - tr(A^i A^{-1})] \\ &\quad + \frac{1}{2}[(B - B^i)' A^i (B - B^i)]. \end{aligned} \tag{2.6}$$

Note that the density $f_{\hat\theta}(\,b\,|\,Y\,)$ is determined by B and A, whereas the density $f_{\hat\theta^i}(\,b\,|\,Y\,)$ is determined by B^i and A^i. Thus, PIF(i) assesses the discrepancy between $f_{\hat\theta}(\,b\,|\,Y\,)$ and $f_{\hat\theta^i}(\,b\,|\,Y\,)$ by providing a composite reflection of the differences between the mean vectors and covariance matrices that characterize these densities.

Note also that B, A, B^i, and A^i depend only on the observed data Y, the estimate based on the full data set $\hat\theta$, and the estimate based on a case-deleted data set $\hat\theta^i$. Thus, differences between B and B^i and between A and A^i originate from differences between $\hat\theta$ and $\hat\theta^i$. However, the goal of the diagnostic is not to directly assess estimative influence, yet rather to assess predictive influence. Obviously, the prediction of the random effects is governed by the estimates of the model parameters, yet the prediction may be more heavily influenced by certain perturbations in the parameter estimates than by others. The objective of PIF(i) is to determine whether a difference between $\hat\theta$ and $\hat\theta^i$ translates to a substantive impact on the prediction of b, either through the predictor itself or through the predictor accuracy.

Once the estimates $\hat\theta, \hat\theta^1, \ldots, \hat\theta^m$ are determined, the calculation of the PIF values PIF$(1), \ldots,$ PIF(m) can be accomplished.

3. Application

To illustrate the utility of the diagnostic in locating influential cases, we apply the predictive influence function in the analysis of a set of exam scores. The scores are modeled using a mixed model containing a fixed exam effect and a random subject effect. Specifically, the two-way balanced ANOVA mixed model is employed. We have

$$y_{ijk} = \mu_j + \tau_i + \varepsilon_{ijk}, \qquad i = 1, \ldots, m, \qquad j = 1, \ldots, J, \qquad k = 1, \ldots, K, \quad (3.1)$$

where the y_{ijk}, μ_j, τ_i, and ε_{ij} are all scalars. Here, y_{ijk} denotes the kth response observed on the ith subject under the jth treatment, μ_j is a fixed treatment effect, τ_i is a random subject effect, and ε_{ijk} is an error term. Note that the overall sample size is $m \times J \times K$. We assume that the τ_i are distributed as i.i.d. $N(0, \sigma_\tau^2)$, and that the ε_{ijk} are distributed as i.i.d. $N(0, \sigma^2)$. Further, we assume that τ_i and ε_{jkl} are independent for all i, j, k, l.

We can re-express model (3.1) in the format of model (2.1) by writing

$$y_i = X_i \mu + z_i \tau_i + \varepsilon_i, \qquad i = 1, \ldots, m, \quad (3.2)$$

where y_i, μ, z_i, and ε_i are all vectors, X_i is a matrix, and τ_i is a scalar. Specifically, y_i denotes a $JK \times 1$ vector of JK responses observed on the ith subject; X_i is a $JK \times J$ block diagonal matrix comprised of J identical blocks, each a $K \times 1$ vector consisting of all 1's; μ is the $J \times 1$ vector $(\mu_1, \ldots, \mu_J)'$; z_i is a $JK \times 1$ vector consisting of all 1's; τ_i is as defined previously; and ε_i is a $JK \times 1$ error vector distributed as $N(0, \sigma^2 I)$. We assume that $\tau_1, \ldots, \tau_m, \varepsilon_1, \ldots, \varepsilon_m$ are distributed independently.

Model (3.2) can be easily represented in the form of model (2.2). The PIF(i) values can then be obtained via the computational formula (2.6). In our implementation of (2.6), the EM algorithm is used to obtain parameter estimates.

We consider a data set consisting of midterm examination scores in a mathematical statistics course held in the fall of 1998 at the University of Missouri-Columbia. This data set consists of 3 exam scores for each of 72 students. For such a data set, it is reasonable to assume that the scores for a student are correlated, and yet sets of scores for different students are uncorrelated. The student effect can be regarded as random and the exam effect as fixed. Thus, the data is amenable to the mixed model (3.1). (Note that $m = 72, J = 3, K = 1$.) The purpose of our analysis will be to explore which cases have a substantial impact on the prediction of the random effects.

The PIF(i) values are evaluated using formula (2.6) for $i = 1, \ldots, 72$. In Figure 1, these values are plotted against the case index i. Based on the relative heights of the peaks in the plot, we designate cases 48, 55, and 71 as influential. The peaks corresponding to these cases are labelled accordingly.

An alternative approach to Figure 1 for graphically displaying the PIF results would be to sort the diagnostic values in descending order, and to plot the sorted values against an index corresponding to the magnitude. Such a plot appears in Figure 2. In format, this type of graph resembles a *scree plot*, which is often used in principal components analysis to choose an optimal number of components. Here, the prominent peaks appear on the left-hand side of the graph. The point of separation between these peaks and the more typical diagnostic values, often called the *elbow*, specifies the boundary for designating influential cases.

<INSERT FIGURE 1 AND FIGURE 2 NEAR HERE.>

At present, we have not developed a method for objectively determining a baseline to gauge unusual values of PIF(i). Thus, we subjectively flag influential cases by using the relative heights of the peaks in the plot of PIF(i) versus i (Figure 1), or by using the elbow of the sorted PIF plot (Figure 2). In order to permit our diagnostic to be applied in a more automatic fashion, we hope to develop a baseline by investigating the distribution of (2.4) under the assumption that no cases are anomalous. Such benchmarks are often proposed for case-deletion diagnostics to eliminate any ambiguity in determining which cases should be flagged.

Table 1 lists the estimates of μ_1, μ_2, μ_3, σ_τ^2, and σ^2 based on the full data set and the case-deleted data sets corresponding to cases 48, 55, 71. Since PIF(3) is quite small, for the purpose of comparison, the table also includes the parameter estimates based on the exclusion of case 3. Note that the removal of case 3 results in virtually no change in the estimates. In contrast, the exclusion of the influential cases results in substantial modifications. The estimate of μ_i $(i = 1, 2, 3)$ is marginally altered when one influential case is removed. However, the estimate of σ_τ^2 is markedly reduced when case 48 or 71 is deleted, and the estimate of σ^2 is significantly decreased with the exclusion of case 55.

<Insert Table 1 near here.>

Note that σ_τ^2 and σ^2 reflect different sources of variation: σ_τ^2 reflects between subject (case) variability, whereas σ^2 reflects within subject (case) variability. The exam scores for student 48 are 45, 50, 48; the scores for student 55 are 32, 86, 92; and the scores for student 71 are 20, 69, 37. Therefore, for case 48 or 71, the difference between the case mean ($\overline{y}_{i.}$) and the grand mean ($\overline{y}_{..}$) is more substantial than that for case 55. On the other hand, the differences among the observations comprising case 55 are more substantial than those among the observations comprising case 48 or 71. Thus, the omission of case 48 or case 71 has a more profound impact on the estimate of σ_τ^2 than the omission of case 55. The exclusion of case 55 conspicuously alters the estimate of σ^2, whereas the exclusion of case 48 or case 71 has only a marginal effect on this estimate.

In Figure 3 and 4, the case-deleted estimates $(\hat{\mu}_1)^i$, $(\hat{\mu}_2)^i$, $(\hat{\mu}_3)^i$, $(\hat{\sigma}_\tau^2)^i$ and $(\hat{\sigma}^2)^i$. are plotted against the case index i. The influential cases are less concealed in the variance plots than in the mean plots.

<Insert Figure 3 and Figure 4 near here.>

Overall estimative influence may be better reflected in scatterplots where one set of case-deleted estimates is plotted against another. Figure 5 features a 2×2 array of such graphs. The first three plots feature $(\hat{\mu}_k)^i$ versus $(\hat{\mu}_j)^i$ $(1 \leq j < k \leq 3)$; the last features $(\hat{\sigma}_\tau^2)^i$ versus $(\hat{\sigma}^2)^i$. In each graph, the influential cases are highlighted with solid dots; the remaining cases are designated with hollow dots. In the mean plots, note that the coordinates corresponding to influential cases often appear on the periphery. These coordinates are especially prominent in the variance plot: since each of the influential cases has a substantial impact on one of the two variances, each coordinate is outlying in either the horizontal or the vertical direction.

<Insert Figure 5 near here.>

The two-dimensional plots in Figure 5 may be more efficacious at flagging unusual cases than the one-dimensional plots in Figure 3 and 4. However, in either set of plots, it would be difficult to consolidate the information to identify cases that may indirectly impact the prediction of the random effects by influencing the parameter estimates. In such settings, the utility of the diagnostic becomes readily apparent.

In the present setting, we assume that the primary goal of our analysis is to predict the random effects $\tau_1, \ldots, \tau_{72}$. It is therefore of interest to investigate differences in the predicted values produced under the full data set Y and the case-deleted data sets Y^i. Table 2 features the predicted values for τ_1 to τ_5 based on both the full data set and the case-deleted data sets corresponding to cases 48, 55, and 71. Since PIF(3) is quite small, for the purpose of comparison, predicted values based on the omission of case 3 are also included.

As mentioned previously, the predictor of $\tau = (\tau_1, \ldots, \tau_{72})'$ based on the full data set is the expected value of τ with respect to $f_{\hat{\theta}}(\tau \,|\, Y)$; the predictor based on the case-deleted data set is the expected value of τ with respect to $f_{\hat{\theta}^i}(\tau \,|\, Y)$. Note that the removal of case 3 results in virtually no change in the recovery of the random effects, whereas the removal of any of the three influential cases markedly alters the magnitude of the predicted values. Further investigations indicate that no other case deletion affects the recovery of the τ_i as substantially as the case deletions considered here.

<Insert Table 2 near here.>

We emphasize that Table 2 reflects changes in only the predicted values, not in the variability estimates associated with the predicted values. As previously mentioned, overall predictive influence is characterized not only by changes in the predictors of the random effects, but also by changes in the accuracy of the predictors.

4. Conclusion and Further Directions

In statistical modeling, it is important to identify cases that have a substantial impact on key inferential results. Such cases may indicate recording errors or anomalies in the phenomenon that produced the data. Such cases may also serve as an indication that the underlying model is too simplistic; thus, the problems of model selection and influential case detection must be addressed jointly.

The application in Section 3 illustrates that the predictive influence function is effective in flagging cases that impact the prediction of random effects in a mixed model. Such cases are often not easily identified either by visually inspecting the data or by assessing estimative influence on parameters.

We note that the diagnostic could be used with either single or multiple case deletion. The latter approach could be beneficial in detecting possible masking effects (i.e., where the influence of one case is obscured by the presence of another case).

We also note that the diagnostic could be formulated for other modeling applications which involve unobservable quantities that are routinely predicted: e.g., censored survival times in survival analysis models, latent factors in dynamic factor analysis models, future values in time series models (see Cavanaugh and Oleson, 2001, Section 4). We hope to pursue some of these formulations in future work.

Acknowledgments

We wish to extend our appreciation to the referee for carefully reading the original version of our paper, and for providing insightful suggestions which served to improve the exposition and content. We also extend our thanks to the editor, Dr. Min-Te Chao.

Table 1: Parameter estimates for exam data.

Data set	μ_1	μ_2	μ_3	σ_τ^2	σ^2
Full data	69.25	84.58	73.61	125.22	120.71
Case 3 deleted	69.07	84.65	73.79	127.75	120.01
Case 48 deleted	69.59	85.07	73.97	115.81	121.96
Case 55 deleted	69.77	84.56	73.35	130.37	110.79
Case 71 deleted	69.94	84.80	74.13	112.03	118.28

Table 2: Predicted values of random effects for exam data.

τ_i	Full data	Case 3 deleted	Case 48 deleted	Case 55 deleted	Case 71 deleted
τ_1	8.21	8.24	7.74	8.39	7.67
τ_2	−13.23	−13.33	−13.20	−13.69	−13.28
τ_3	−1.12	−1.14	−1.38	−1.22	−1.45
τ_4	3.92	3.93	3.54	3.98	3.48
τ_5	11.49	11.54	10.95	11.77	10.88
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

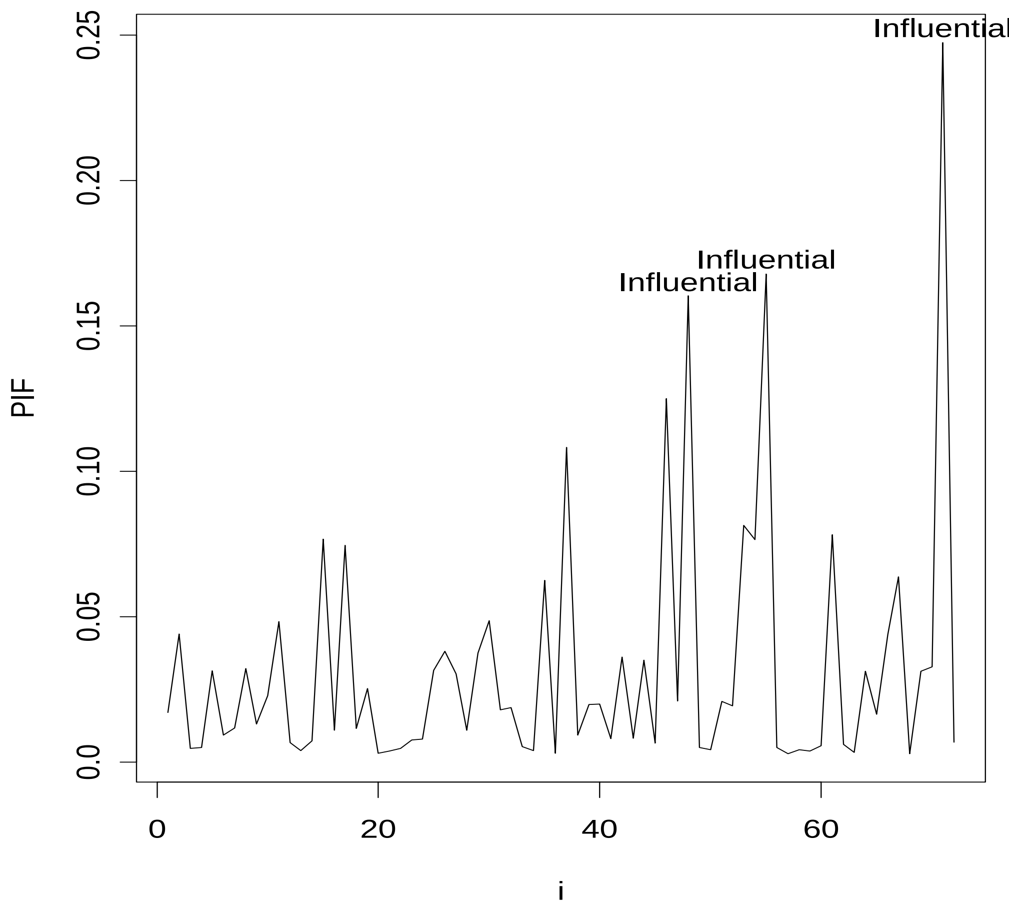

Figure 1: PIF(i) vs. case index i for exam data.

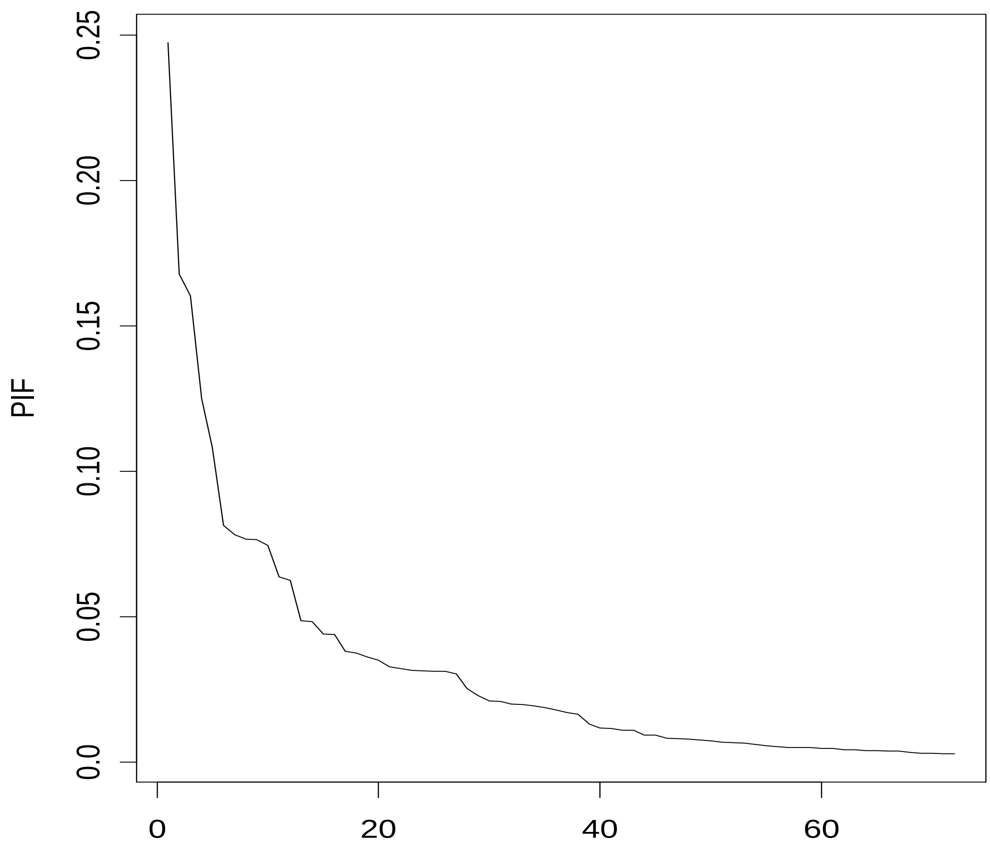

Figure 2: Sorted PIF(i) for exam data.

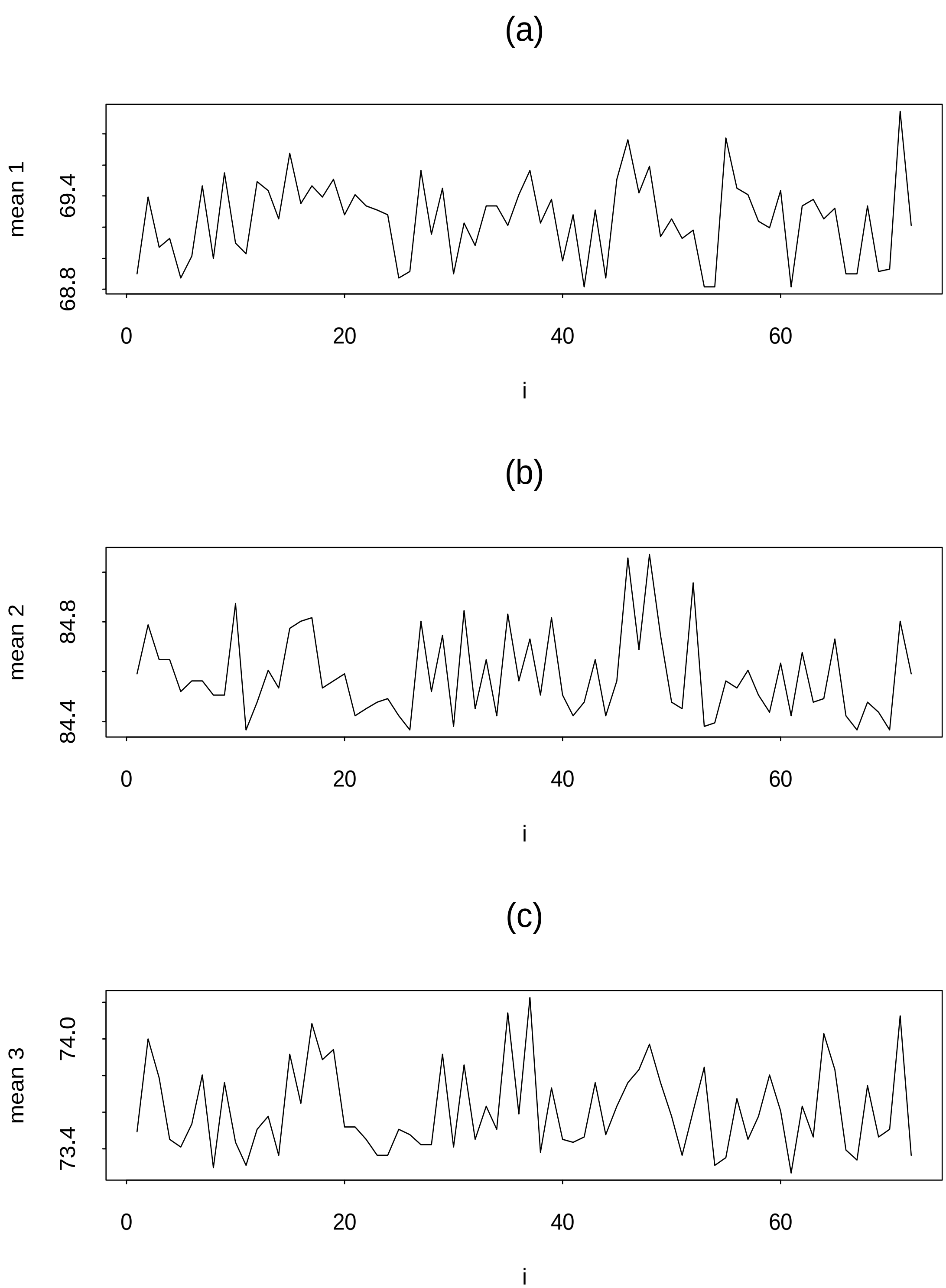

Figure 3: Case-deleted mean estimates for exam data. (a) $(\hat{\mu}_1)^i$ vs. case index i; (b) $(\hat{\mu}_2)^i$ vs. case index i; (c) $(\hat{\mu}_3)^i$ vs. case index i.

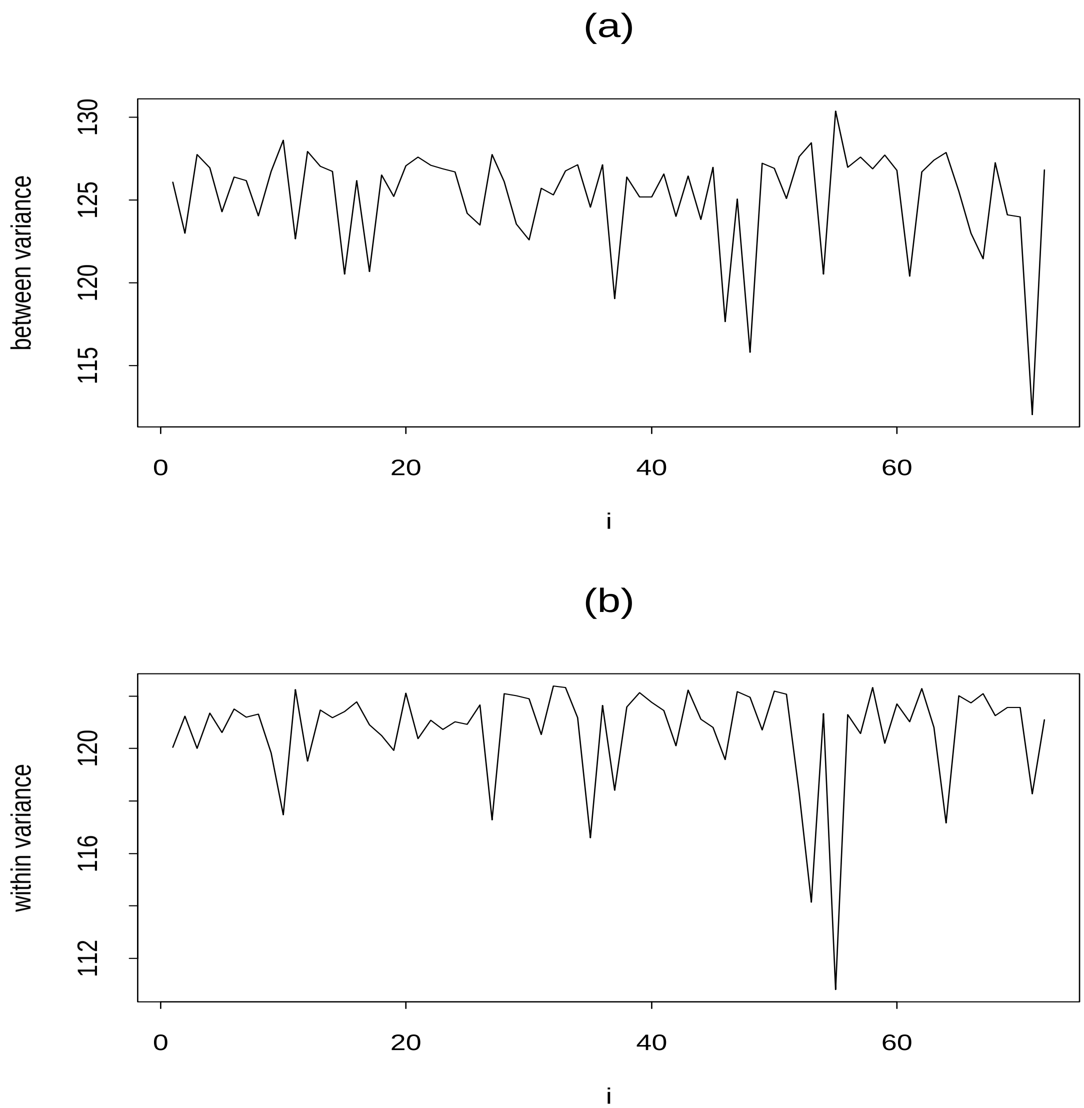

Figure 4: Case-deleted variance estimates for exam data. (a) $(\hat{\sigma}^2_\tau)^i$ vs. case index i; (b) $(\hat{\sigma}^2)^i$ vs. case index i.

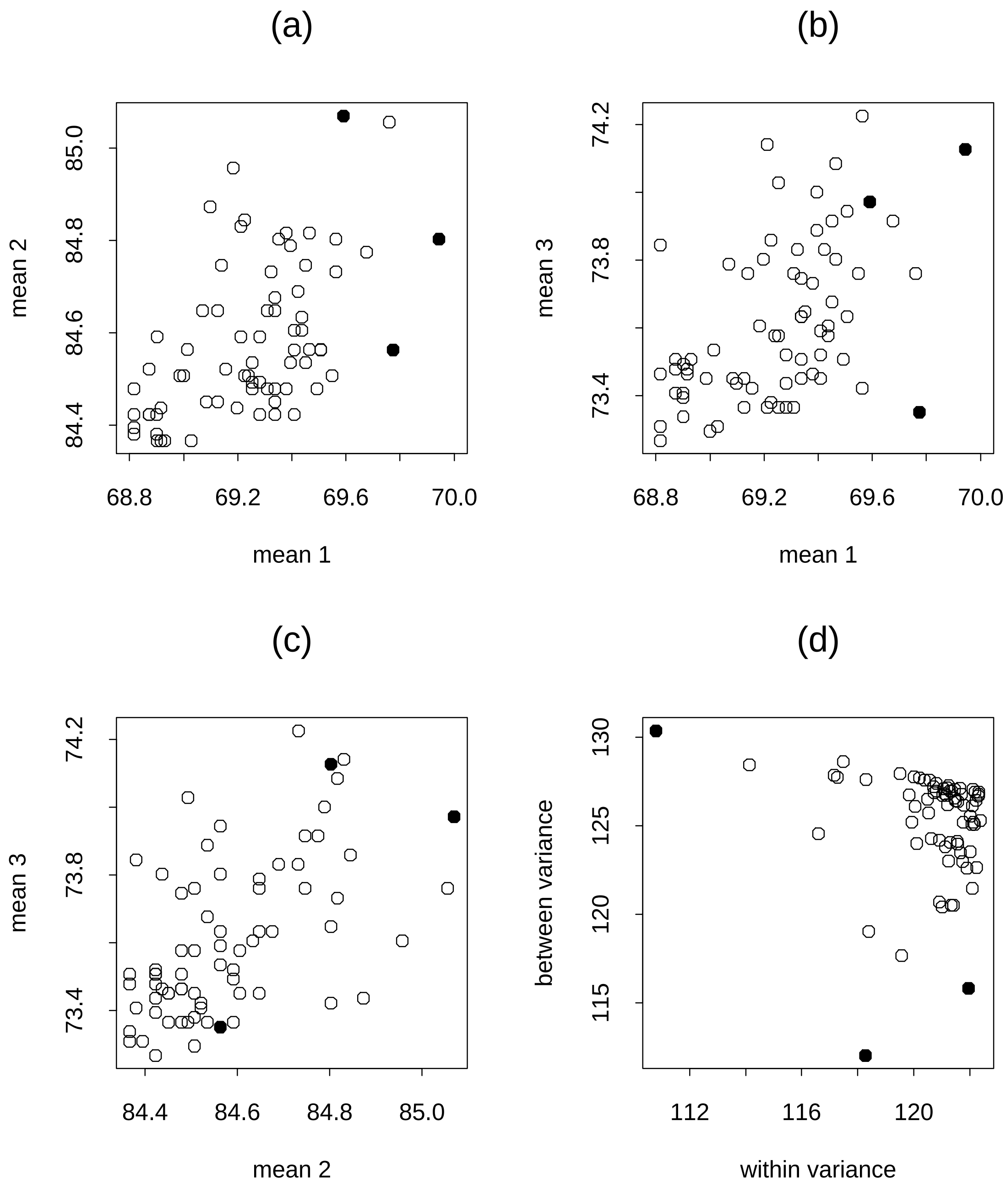

Figure 5: Scatterplots of case-deleted estimates for exam data. (a) $(\hat{\mu}_2)^i$ vs. $(\hat{\mu}_1)^i$; (b) $(\hat{\mu}_3)^i$ vs. $(\hat{\mu}_1)^i$; (c) $(\hat{\mu}_3)^i$ vs. $(\hat{\mu}_2)^i$; (d) $(\hat{\sigma}_\tau^2)^i$ vs. $(\hat{\sigma}^2)^i$.

Appendix: Derivation of Computational Formula (2.6)

By (2.3), the log of $f_{\hat\theta}(b\,|\,Y)$ and the log of $f_{\hat\theta^i}(b\,|\,Y)$ are given by

$$\begin{aligned}\log f_{\hat\theta}(b\,|\,Y) \;=\;& -\frac{1}{2}\log|(Z'(\hat\sigma^2R)^{-1}Z+\hat D^{-1})^{-1}| \\ & -\frac{1}{2}(b-\hat DZ'\hat V^{-1}(Y-X\hat\beta))'(Z'(\hat\sigma^2R)^{-1}Z+\hat D^{-1}) \\ & \quad (b-\hat DZ'\hat V^{-1}(Y-X\hat\beta)),\end{aligned} \tag{A.1}$$

and

$$\begin{aligned}\log f_{\hat\theta^i}(b\,|\,Y) \;=\;& -\frac{1}{2}\log|(Z'((\hat\sigma^2)^iR)^{-1}Z+(\hat D^i)^{-1})^{-1}| \\ & -\frac{1}{2}(b-\hat D^iZ'(\hat V^i)^{-1}(Y-X\hat\beta^i))'(Z'((\hat\sigma^2)^iR)^{-1}Z+(\hat D^i)^{-1}) \\ & \quad (b-\hat D^iZ'(\hat V^i)^{-1}(Y-X\hat\beta^i)).\end{aligned} \tag{A.2}$$

First, we simplify $E_{\hat\theta}[\log f_{\hat\theta^i}(b\,|\,Y)\,|\,Y]$.

To obtain the expectation of (A.2) with respect to $f_{\hat\theta}(b\,|\,Y)$, we derive the expectation of the term in (A.2) which involves the quadratic form in b. (Conditional on Y, b is the only random quantity in (A.2).) We have

$$\begin{aligned}& E_{\hat\theta}[\,(b-\hat D^iZ'(\hat V^i)^{-1}(Y-X\hat\beta^i))'(Z'((\hat\sigma^2)^iR)^{-1}Z+(\hat D^i)^{-1}) \\ & \quad (b-\hat D^iZ'(\hat V^i)^{-1}(Y-X\hat\beta^i))\,|\,Y\,] \\ =\;& E_{\hat\theta}[(b'-(B^i)')A^i(b-B^i)\,|\,Y\,] \\ =\;& tr(A^iE_{\hat\theta}[b\,b'\,|\,Y\,])-E_{\hat\theta}[\,b'\,|\,Y\,]A^iB^i-(B^i)'A^iE_{\hat\theta}[\,b\,|\,Y\,]+(B^i)'A^iB^i \\ =\;& tr(A^i(\mathrm{Var}_{\hat\theta}[\,b\,|\,Y\,]+E_{\hat\theta}[\,b\,|\,Y\,]E_{\hat\theta}[\,b\,|\,Y\,]')) \\ & \quad -B'A^iB^i-(B^i)'A^iB+(B^i)'A^iB^i \\ =\;& tr(A^iA^{-1}+A^iBB')-B'A^iB^i-(B^i)'A^iB+(B^i)'A^iB^i \\ =\;& tr(A^iA^{-1})+B'A^iB-B'A^iB^i-(B^i)'A^iB+(B^i)'A^iB^i \\ =\;& tr(A^iA^{-1})+(B-B^i)'A^i(B-B^i).\end{aligned} \tag{A.3}$$

Thus, for the expectation of (A.2) with respect to $f_{\hat\theta}(b\,|\,Y)$, we have

$$\begin{aligned}& E_{\hat\theta}[\,\log f_{\hat\theta^i}(b\,|\,Y)\,|\,Y\,] \\ =\;& \frac{1}{2}\log|A^i|-\frac{1}{2}[tr(A^iA^{-1})+(B-B^i)'A^i(B-B^i)].\end{aligned} \tag{A.4}$$

Next, we simplify $E_{\hat\theta}[\log f_{\hat\theta}(b\,|\,Y)\,|\,Y]$.

To obtain the expectation of (A.1) with respect to $f_{\hat\theta}(b\,|\,Y)$, we require the expectation of the term in (A.1) which involves the quadratic form in b. This expectation follows from (A.3) with $\hat\beta^i = \hat\beta$, $\hat D^i = \hat D$, $(\hat\sigma^2)^i = \hat\sigma^2$, $\hat V^i = \hat V$, $A^i = A$, and $B^i = B$. We have

$$\begin{aligned} & E_{\hat\theta}[(b - \hat D Z'\hat V^{-1}(Y - X\hat\beta))'(Z'(\hat\sigma^2 R)^{-1}Z + \hat D^{-1})(b - \hat D Z'\hat V^{-1}(Y - X\hat\beta))\,|\,Y] \\ = \; & tr(AA^{-1}) \\ = \; & tr(I) \\ = \; & mq. \end{aligned}$$

Thus, for the expectation of (A.1) with respect to $f_{\hat\theta}(b\,|\,Y)$, we have

$$E_{\hat\theta}[\log f_{\hat\theta}(b\,|\,Y)\,|\,Y] = \frac{1}{2}\log|A| - \frac{1}{2}mq. \tag{A.5}$$

Computational formula (2.6) follows from (A.4) and (A.5).

References

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley and Sons, New York.

Cavanaugh, J.E. and Johnson, W.O. (1999). Assessing the predictive influence of cases in a state-space process. *Biometrika* **86**, 183–190.

Cavanaugh, J.E. and Oleson, J.J. (2001). A diagnostic for assessing the influence of cases on the prediction of missing data. *Journal of the Royal Statistical Society, Series D* **50**, 427–440.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman and Hall, London.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Johnson, W. and Geisser, S. (1982). Assessing the predictive influence of observations. In *Statistics and Probability: Essays in Honor of C.R. Rao*, edited by G. Kallianpour, et al., 343–358. North Holland, Amsterdam.

Johnson, W. and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *Journal of the American Statistical Association* **78**, 137–144.

Kullback, S. (1968). *Information Theory and Statistics.* Dover, New York.

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79–86.