Ann. Inst. Statist. Math.
Vol. 56, No. 3, 415–433 (2004)

DETECTING STAGE-WISE OUTLIERS IN HIERARCHICAL BAYESIAN LINEAR MODELS OF REPEATED MEASURES DATA

MARIO PERUGGIA[1], THOMAS J. SANTNER[1] AND YU-YUN HO[2]

[1] *Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210-1247, U.S.A.*
[2] *Biostatistics & Statistical Reporting, Novartis Pharmaceuticals Corporation, One Health Plaza, East Hanover, NJ 07936-1080, U.S.A.*

(Received September 2, 2002; revised August 7, 2003)

Abstract. We propose numerical and graphical methods for outlier detection in hierarchical Bayes modeling and analyses of repeated measures regression data from multiple subjects; data from a single subject are generically called a "curve." The first-stage of our model has curve-specific regression coefficients with possibly autoregressive errors of a prespecified order. The first-stage regression vectors for different curves are linked in a second-stage modeling step, possibly involving additional regression variables. Detection of the *stage* at which the curve appears to be an outlier and the *magnitude and specific component* of the violation at that stage is accomplished by embedding the null model into a larger parametric model that can accommodate such unusual observations. We give two examples to illustrate the diagnostics, develop a BUGS program to compute them using MCMC techniques, and examine the sensitivity of the conclusions to the prior modeling assumptions.

Key words and phrases: Autoregressive errors, BUGS, graphical diagnostics, model-based diagnostics, outlier accommodation models, diagnostics for multi-stage models.

1. Introduction

This paper proposes outlier detection methods for a class of hierarchical Bayesian linear models that are widely used to analyze data consisting of repeated measurements on each of a set of subjects. In data from designed experiments, the measurements are usually taken at ordered time points or locations although in observational studies this need not be the case. In any event, each subject's data is referred to as a *curve*. The goal of our diagnostics is to determine, for each curve, that *either* there is no evidence of model violations at any of the hierarchical stages *or* to identify the stage(s) where model assumptions are violated and specific details of the model violation(s).

The analysis of repeated measures data occurs frequently in medicine, epidemiology, psychology, and many other disciplines. One popular frequentist method of analysis of such data is based on the specification of a two-stage random effects model (Laird and Ware (1982)). Several variations on this basic theme are presented in Crowder and Hand (1990), Lindstrom and Bates (1990), and Lindsey (1993); a tutorial review is given in Cnaan *et al.* (1997). The Bayesian analysis of repeated measures data is typically based on hierarchical generalizations of mixed effects models. Early references are Berger and Hui (1983) and Wakefield *et al.* (1994). In recent years such models have found increasingly fruitful application in both medicine and epidemiology (see for

example Palmer and Müller (1998); Joseph *et al.* (1999); Tan *et al.* (1999); Pauler and Laird (2000, 2002); Lambert *et al.* (2001); Berlin *et al.* (2002)).

There are several Bayesian approaches to outlier detection with respect to a given model. One approach identifies as outliers those observations whose realized errors with respect to that model have high posterior probabilities of being "large." The fundamental idea is contained in Zellner (1975) and is applied to the linear model and more general hierarchical models in Chaloner and Brant (1988) and Chaloner (1994), respectively. Weiss (1995) extends this approach to repeated measures data. Hodges (1998) exploits the geometric properties of a linear model representation of hierarchical Bayes models to derive analogues of classical diagnostics (see also Langford and Lewis (1998) for an exposition of frequentist ideas regarding outlier analysis in multilevel data).

Another approach is to embed the null model into a larger parametric model that can accommodate unusual observations. Outlier detection then consists of parametric inference based on the extended model. For example, using variance inflation and/or location-shift extensions of normal linear models, this approach is applied in Pettit and Smith (1985), Sharples (1990), and Verdinelli and Wasserman (1991). Carota *et al.* (1996) provide a comprehensive account of related model elaboration methodology.

The model extension techniques of the previous paragraph follow Principle *D2* of Weisberg (1983) by turning a problem of null model criticism into one of parametric inference. However, because they use more complex models, Bayesian outlier diagnostics are, in general, more computationally intensive than their frequentist counterparts. Thus Bayesian diagnostics need not display the computational simplicity of Weisberg's Principle *D3*.

This paper uses the approach of extending a null model to analyze a widely-used, three-stage hierarchical Bayesian linear model for repeated measures data. Stage I of the model (the likelihood) specifies a regression with curve-specific coefficients and allows autoregressive measurement errors of a known order to account for dependencies among the residuals on the same curve. Stage II of the model describes the variability in the (random) curve-specific regression coefficients using a second set of (Stage II) regressors and unexplained random variation. Lastly, Stage III of the model specifies priors for the hyperparameters in Stage II.

For each curve, our goal is to determine whether or not that curve shows evidence of a measurement error outlier (a Stage I violation) and whether or not the regression coefficient vector for that curve is, after possibly accounting for covariates, inconsistent with the regression vectors of the remaining curves (a Stage II violation). We provide numerical and graphical tools to identify specific sources of Stage I and II violations. This is accomplished by introducing two location-shift outlier indicators for each curve, one at each stage. The Stage I outlier indicator equals unity when a measurement error is present while the Stage II outlier indicator equals unity when a regression coefficient error is present; both indicators are equal to zero when the null model is adequate. The extended model also includes curve-specific vectors to measure the magnitude of model violations at each stage. Then, as an example, posterior values of these quantities, calculated for each curve, can be used as null model diagnostics. An initial version of this approach, based on a much simpler model, is given in Ho *et al.* (1995) where it is applied to an artificial example similar to the example of Section 4. The more complicated model of this paper is required to analyze the data presented in the substantive example of Section 5.

Our proposed diagnostics follow Principle *D4* of Weisberg (1983), emphasizing the

use of graphical summaries. These summaries are a contribution to the difficult problem of determining and describing what does or does not constitute a representative curve from a set of longitudinal data (Jones and Rice (1992); Segal (1994)).

The remainder of this paper is organized as follows. Section 2 states the hierarchical null model while Section 3 introduces the extended model and uses it to define diagnostics. Section 4 provides a simple example illustrating the diagnostic in an idealized set-up. Section 5 applies the technique to a set of bone data that was analyzed in Peruggia *et al.* (1994). Some extensions are discussed in Section 6. The proposed diagnostics can be evaluated using the output from a BUGS program (Spiegelhalter *et al.* (1996)) which is available from the first author.

2. The hierarchical Bayes null model

We use a hierarchical regression model with random coefficients to describe data in which each of a collection of subjects contributes a curve, typically in time or space. We allow for the random coefficients of the different curves to be related by their regression on curve-specific and coefficient-specific covariates.

To describe the model formally, let I denote the number of subjects, $\mathbf{0}_n$ the vector of zeroes of length n, $\mathbf{1}_n$ the vector of ones of length n, $\boldsymbol{I}_n$ the identity matrix of order n, and $\mathrm{diag}(\boldsymbol{v})$ the $m \times m$ diagonal matrix with elements on the main diagonal given by the vector $\boldsymbol{v} = (v_1, \ldots, v_m)$. Also, let $N_n(\cdot, \cdot)$ denote the multivariate normal distribution of dimension n, and $IG(a, b)$ denote the inverse gamma distribution with shape parameter a and scale parameter b whose density is given by $p(w \mid a, b) = (\Gamma(a) w^{a+1})^{-1} b^a \exp(-b/w)$, for $w > 0$. In the model specified below we assume conditional independence unless otherwise stated, i.e., the random variables at any level of the model are independent, given the hyperparameters at the next level.

Stage I. For $i = 1, \ldots, I$,

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_i + \boldsymbol{\eta}_i, \tag{2.1}$$

where $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,J_i})^\top$ is the vector of measurements for the i-th curve, $\boldsymbol{X}_i$ is a $J_i \times L$ design matrix, $\boldsymbol{\beta}_i$ is an $L \times 1$ vector of parameters, and $\boldsymbol{\eta}_i$ is a $J_i \times 1$ vector of measurement errors having the autoregressive structure

$$\eta_{i,j} = \phi_1 \eta_{i,j-1} + \cdots + \phi_{\min(j,P)} \eta_{i,j-\min(j,P)} + \varepsilon_{i,j}, \qquad \text{for} \quad j = 1, \ldots, J_i,$$

where the $\varepsilon_{i,j}$ are $N(0, \sigma_\varepsilon^2)$ innovations for $i = 1, \ldots, I$ and $j = P+1, \ldots, J_i$, while $\varepsilon_{i,1}, \ldots, \varepsilon_{i,P}$ are zero-mean normal innovations with arbitrary variances $\sigma^2_{\varepsilon_1}, \ldots, \sigma^2_{\varepsilon_P}$ for $i = 1, \ldots, I$ (this is done because we do not require conditional stationarity of the $\eta_{i,j}$ given the autoregressive parameters).

Stage II. Let $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_L^\top)^\top$, $\boldsymbol{\sigma}^2_\beta = (\sigma^2_\beta[1], \ldots, \sigma^2_\beta[L])^\top$, and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_P)^\top$. Then

$$\beta_{il} \mid \boldsymbol{\gamma}, \boldsymbol{\sigma}^2_\beta \sim N(\boldsymbol{z}_{il}^\top \boldsymbol{\gamma}_l, \sigma^2_\beta[l]), \qquad \text{for} \quad i = 1, \ldots, I \text{ and } l = 1, \ldots, L, \tag{2.2}$$

where $\boldsymbol{z}_{il}$ is a $K_l \times 1$ vector of covariates specific to the l-th component of the Stage I regression vector for curve i, $\boldsymbol{\gamma}_l$ is the corresponding $K_l \times 1$ hierarchical regression coefficient vector,

$$\boldsymbol{\phi} \sim N_P(\mathbf{0}_P, \boldsymbol{\Sigma}_\phi = \mathrm{diag}(\sigma^2_\phi[1], \ldots, \sigma^2_\phi[P])),$$

$$\begin{aligned} &\sigma_\varepsilon^2 \sim IG(a_\varepsilon, b_\varepsilon), \\ &\sigma_{\varepsilon_j}^2 \sim IG(a_{\varepsilon_j}, b_{\varepsilon_j}), \qquad \text{for} \quad j = 1, \ldots, P. \end{aligned} \tag{2.3}$$

The hyperparameters $\boldsymbol{\Sigma}_\phi$, a_ε, b_ε, a_{ε_j}, and b_{ε_j}, $j = 1, \ldots, P$, are assumed to be known.

Stage III.

$$\begin{aligned} &\boldsymbol{\gamma}_l \sim N_{K_l}(\boldsymbol{\mu}_\gamma[l], \operatorname{diag}(\sigma_\gamma^2[l,1], \ldots, \sigma_\gamma^2[l,K_l])), \qquad \text{for} \quad l = 1, \ldots, L, \\ &\sigma_\beta^2[l] \sim IG(a_l, b_l), \qquad \text{for} \quad l = 1, \ldots, L. \end{aligned} \tag{2.4}$$

The hyperparameters $\boldsymbol{\mu}_\gamma[l]$ (a $K_l \times 1$ vector), $\sigma_\gamma^2[l,k]$, a_l, and b_l, for $l = 1, \ldots, L$ and $k = 1, \ldots, K_l$, are assumed to be known.

This model assumes that the regression coefficients β_{il}, while curve-specific, are nevertheless related to one another. This is reflected in two aspects of the model. First, the means of the normal prior distributions for the β_{il} depend on curve-specific covariates, $\boldsymbol{z}_{il}$, through a Stage II regression structure; the vector of regression parameters used to model $(\beta_{1l}, \ldots, \beta_{Il})$ is $\boldsymbol{\gamma}_l$. Second, for a fixed regressor l, additional dependence is introduced among the parameters $\beta_{1l}, \ldots, \beta_{Il}$ (across curves) by assuming that they are conditionally independent given a common variance, $\sigma_\beta^2[l]$, drawn from an inverse gamma distribution.

An important special case of model (2.1)–(2.4) occurs when there is no other curve-specific covariate information. This case corresponds to setting $K_l = 1$ and $\boldsymbol{z}_{il}^\top \boldsymbol{\gamma}_l = \gamma_l$ in Stage II, as we do in the example of Section 4. In addition, setting the order of the autoregressive filter, P, to *zero* in Stage I yields a model with independent measurement errors.

3. Location-shift outlier detection

Our proposed diagnostics for detecting location-shift outliers are based on an extended version of model (2.1)–(2.4) that contains two outlier indicators for each curve, one indicator at each of Stages I and II, and two corresponding location shifts. The diagnostics use the posterior distributions of the outlier indicators and the location shifts. After presenting the extended model, we describe its features. As in the case of the null model, we always assume conditional independence.

Stage I. For $i = 1, \ldots, I$,

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_i + \delta_i^y \boldsymbol{c}_i + \boldsymbol{\eta}_i, \tag{3.1}$$

where δ_i^y takes value 0 or 1, $\boldsymbol{c}_i$ is a $J_i \times 1$ measurement-shift vector, and the distribution of the $\boldsymbol{\eta}_i$ is the same as in the null model (see equation (2.1)).

Stage II. Let $\boldsymbol{d} = (d_{11}, \ldots, d_{1L}, \ldots, d_{I1}, \ldots, d_{IL})^\top$, where d_{il} is a shift in the Stage I regression coefficient β_{il}, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_L^\top)^\top$, and $\boldsymbol{\delta}^\beta = (\delta_1^\beta, \ldots, \delta_I^\beta)$, where δ_i^β takes value 0 or 1. Then

$$\begin{aligned} \beta_{il} \mid \boldsymbol{\gamma}, \boldsymbol{d}, \boldsymbol{\delta}^\beta, \boldsymbol{\sigma}_\beta^2 \sim N(\boldsymbol{z}_{il}^\top \boldsymbol{\gamma}_l + \delta_i^\beta d_{il}, \sigma_\beta^2[l]), \\ \text{for} \quad i = 1, \ldots, I \text{ and } l = 1, \ldots, L, \end{aligned} \tag{3.2}$$

$$
\begin{aligned}
&\boldsymbol{c}_i \sim N_{J_i}(\mu_c \mathbf{1}_{J_i}, \sigma_c^2 \boldsymbol{I}_{J_i}), \quad \text{for} \quad i = 1, \ldots, I,\\
&\delta_i^y \mid p_y \sim Ber(p_y) \quad \text{for} \quad i = 1, \ldots, I,\\
&\boldsymbol{\phi} \sim N_P(\mathbf{0}_P, \boldsymbol{\Sigma}_\phi = \operatorname{diag}(\sigma_\phi^2[1], \ldots, \sigma_\phi^2[P])),\\
&\sigma_\varepsilon^2 \sim IG(a_\varepsilon, b_\varepsilon),\\
&\sigma_{\varepsilon_j}^2 \sim IG(a_{\varepsilon_j}, b_{\varepsilon_j}), \quad \text{for} \quad j = 1, \ldots, P.
\end{aligned} \tag{3.3}
$$

The hyperparameters μ_c, σ_c^2, $\boldsymbol{\Sigma}_\phi$, a_ε, b_ε, a_{ε_j}, and b_{ε_j}, $j = 1, \ldots, P$, are assumed to be known.

Stage III.

$$
\begin{aligned}
&\boldsymbol{\gamma}_l \sim N_{K_l}(\boldsymbol{\mu}_{\gamma[l]}, \operatorname{diag}(\sigma_\gamma^2[l,1], \ldots, \sigma_\gamma^2[l,K_l])), \quad \text{for} \quad l = 1, \ldots, L,\\
&d_{il} \sim N(\mu_d[l], \sigma_d^2[l]) \quad \text{for} \quad i = 1, \ldots, I \text{ and } l = 1, \ldots, L,\\
&\delta_i^\beta \mid p_\beta \sim Ber(p_\beta) \quad \text{for} \quad i = 1, \ldots, I,\\
&\sigma_\beta^2[l] \sim IG(a_l, b_l), \quad \text{for} \quad l = 1, \ldots, L,
\end{aligned}
$$

where the hyperparameters $\boldsymbol{\mu}_\gamma[l]$ and $\mu_d[l]$, for $l = 1, \ldots, L$, and $\sigma_\gamma^2[l,k]$, a_l, b_l, and $\sigma_d^2[l]$, for $l = 1, \ldots, L$ and $k = 1, \ldots, K_l$, are assumed to be known.

Stage IV.

$$
p_y \sim Beta(u_y, v_y), \quad p_\beta \sim Beta(u_\beta, v_\beta) \tag{3.4}
$$

where the hyperparameters u_y, v_y, u_β, v_β are assumed to be known.

The interpretation of model (3.1)–(3.4) is the following.

a. Each curve is allowed to have or not have a Stage I deviation (for $\delta_i^y = 1$ or 0, respectively) and have or not have a Stage II deviation (for $\delta_i^\beta = 1$ or 0, respectively).

b. A Stage I mean shift occurs for each curve independently with an unknown probability p_y; $\delta_i^y = 1$ indicates that *one or more* components of the i-th curve exhibit an anomaly. The magnitude of the shift, $\boldsymbol{c}_i$, can vary from curve to curve and its components identify the anomalies within each curve.

c. In Stage II, a shift occurs independently in the mean of each Stage I regression coefficient with an unknown probability p_β; $\delta_i^\beta = 1$ indicates that *one or more* of the Stage I regression coefficients of the i-th curve exhibit an anomaly after accounting for their Stage II regression structure. The magnitude of the shift, d_{il}, is specific to both the curve and regression component.

d. The prior means for the proportions of Stage I and Stage II outliers are $u_y/(u_y + v_y)$ and $u_\beta/(u_\beta + v_\beta)$, respectively, and the prior precision of these means increases in $u_y + v_y$ and $u_\beta + v_\beta$, respectively.

The prior means of the two shift vectors, defined by μ_c and $\mu_d[l]$, would ordinarily be taken to be zero although our BUGS program allows for nonzero choices should specific subject matter considerations suggest this be the case.

As a first diagnostic, we propose the examination of the posterior probabilities

$$
p_i^y = P\{\delta_i^y = 1 \mid \boldsymbol{y}\} \quad \text{and} \quad p_i^\beta = P\{\delta_i^\beta = 1 \mid \boldsymbol{y}\} \tag{3.5}
$$

for each curve i. These probabilities can be conveniently displayed in $[0,1]^2$ or simply tabulated. They convey a *global* assessment of the presence of outliers relative to model (2.1)–(2.4) in the light of the data. Curve i is judged to contain one or more Stage I (measurement) outlying components when p_i^y is high; curve i is judged to contain one or more Stage II (regression) outlying coefficients when p_i^β is high.

When there is evidence that the i-th curve is a Stage I outlier, our second diagnostic seeks to determine the component(s) of $\boldsymbol{Y}_i$ where the model violation(s) occur. We do this by examining the posterior distributions of the shift vector

$$\delta_i^y \times \boldsymbol{c}_i. \tag{3.6}$$

Similarly, when there is evidence that the i-th curve is a Stage II outlier, the posterior distribution of the shift vector

$$\delta_i^\beta \times \boldsymbol{d}_i, \tag{3.7}$$

where $\boldsymbol{d}_i = (d_{i1}, \ldots, d_{iL})^\top$, can be examined to gain insight about the specific regressors exhibiting deviations. Examples will be given in Sections 4 and 5.

Given M independent draws from the posterior distributions of the parameters in model (3.1)–(3.4), i.e., $\{\{\boldsymbol{\beta}_i^{(m)}, \boldsymbol{c}_i^{(m)}, \boldsymbol{d}_i^{(m)}, \delta_i^{y(m)}, \delta_i^{y(m)}\}_{i=1}^I, \boldsymbol{\phi}^{(m)}, \sigma_\varepsilon^{2(m)}, \boldsymbol{\gamma}^{(m)}, \sigma_\beta^{2(m)}\}_{m=1}^M$, estimators of the posterior probabilities (3.5) and of the means of the shift vectors (3.6) and (3.7) can be derived either by averaging the appropriate elements of the draws or by the Rao-Blackwellized method suggested by Gelfand and Smith (1990).

The next two sections illustrate the use of the proposed diagnostics in two examples. The first example is constructed to evaluate the performance of the diagnostics in a case that can be easily understood visually. We show how the sensitivity of the diagnostics to the prior can be assessed. The second example considers a set of thicknesses of cortical bone that was previously analyzed in Peruggia *et al.* (1994) and illustrates the use of the diagnostics in a situation in which it is not obvious which observations are outliers. Several static graphical methods are introduced to help interpret the estimated posterior quantities.

4. Worked example—Simple linear regression with repeated measures

This first example is constructed to illustrate the performance of the diagnostics in a case that can be explained graphically. The data set consists of 20 curves, each having 10 measurements. For $j = 1, \ldots, 10$, the i-th curve, $\boldsymbol{Y}_i$, was generated to have j-th component

$$Y_{i,j} = b_{i1} + b_{i2} \times j + \xi_{ij}, \tag{4.1}$$

where, *within* a curve, the ξ_{ij} follow a first-order autoregressive model with parameter 0.5 and have unit-variance innovations while the ξ_{ij} from *different* curves are mutually independent. The intercepts, $\{b_{i1}\}$, were generated independently as $N(0, 0.04)$ and the slopes, $\{b_{i2}\}$, were generated independently as $N(0.4, 0.01)$. We introduced Stage I errors in $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_{20}$ by adding 3.0 to the 3rd component of $\boldsymbol{Y}_1$ and the same amount to the 4th component of $\boldsymbol{Y}_{20}$. In addition, two Stage II errors were created; the intercept for curve $\boldsymbol{Y}_{19}$, $b_{19,1}$, was modified by adding 3.0 to its original value and the slope of curve $\boldsymbol{Y}_{20}$, $b_{20,2}$, was increased by 1.0. These modifications are clearly visible in Fig. 1 which shows a time series plot of the data.

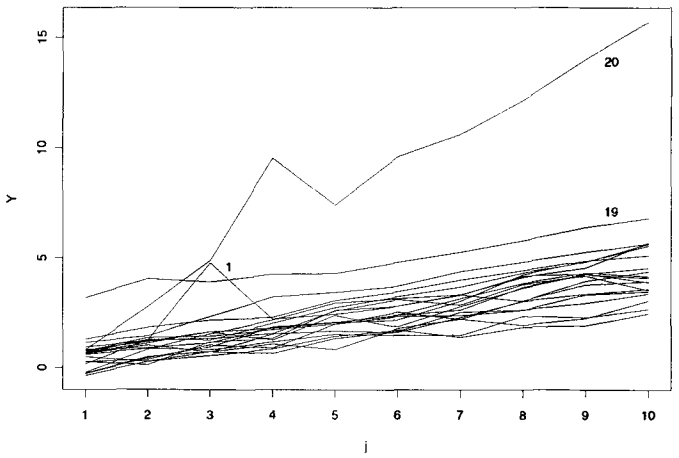

Fig. 1. Plot of $\{(j, Y_{i,j}) : i = 1, \ldots, 20; j = 1, \ldots, 10\}$. The third component of $\boldsymbol{Y}_1$ is unusually large relative to the majority of the curves; $\boldsymbol{Y}_{20}$ has a large slope *and* one component (the 4th) is large relative to the others; $\boldsymbol{Y}_{19}$ has a large intercept but no components have measurement outliers.

We based our analysis on model (2.1)–(2.4), with equation (2.1) specified as

$$Y_{i,j} = \beta_{i1} + \beta_{i2} \times j + \eta_{ij}.$$

The $\{\beta_{i1}\}$ were independent and identically normally distributed as were the $\{\beta_{i2}\}$. The $\{\eta_{ij}\}_j$ followed a first-order AR model. The shape and scale hyperparameters of the inverse gamma distributions for the variances of the intercepts β_{i1} and slopes β_{i2}, call them $\sigma^2_{\beta_1}$ and $\sigma^2_{\beta_2}$ in this example, were taken to be the informative choices $a_{\beta_1} = 3.0 = a_{\beta_2}$ and $b_{\beta_1} = 1.0 = b_{\beta_2}$ which have mean 0.5 and variance 0.25.

The prior variance for the autoregressive coefficient ϕ was set equal to 4.0 which allows both stationary and non-stationary models. The shape and scale hyperparameters of the inverse gamma distributions for the innovations of the η_i were set equal to $a_\varepsilon = 3.0$ and $b_\varepsilon = 1.0$, except for the first innovation that was assumed to follow an $IG(1.0, 1.0)$ distribution. The components of the measurement error shifts $\boldsymbol{c}_i$ were taken to have the common mean value zero and variance 4.0.

The prior means of the intercepts β_{i1} and slopes β_{i2}, γ_1 and γ_2, were set equal to zero and 0.5, respectively. Their variances were both set equal to 4.0. The vector $\boldsymbol{d}_i = (d_{i1}, d_{i2})^\top$ corresponds to the 2×1 shift for the intercept and slope of curve i. Each component of $\boldsymbol{d}_i$ was given prior mean zero, a null value representing no shift. The prior variances of both shift components were set equal to 4.0. These variances are eight times larger than the means of the prior variances of β_{i1} and β_{i2} and allow for substantial deviations should the data suggest that they are needed.

The values of the hyperparameters in the expanded version of the model were set as follows. In Stage IV, the prior mean probabilities of a Stage I error, p_y, and a Stage II error, p_β, were both set equal to 0.15 and the prior probability that either of these values exceeds 0.30 was set equal to 0.10; this yields $u_y = 1.5 = u_\beta$ and $v_y = 8.5 = v_\beta$.

Using our BUGS program, we obtained draws from the posterior distribution of the model parameters. We discarded the first 2,000 warm-up cycles and computed diagnostics based on the next 5,000 iterations. Table 1 lists the Bayes estimates of the

Table 1. Estimates of p_i^y and p_i^β for the example of Section 4. Overlapping batch mean estimates of standard error are shown in parentheses.

i	1	2–18		19	20
p_i^y	**1.00** (0.00)	min = 0.00	max = 0.00	**0.00** (0.00)	**1.00** (0.00)
p_i^β	**0.01** (0.00)	min = 0.01	max = 0.02	**0.80** (0.03)	**0.54** (0.04)

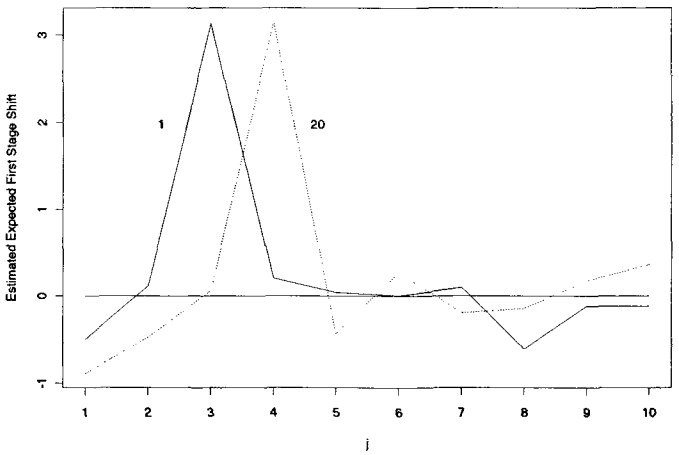

Fig. 2. Index plot of estimated posterior expected Stage I shifts, $\mathrm{E}\{\delta_i^y \times \boldsymbol{c}_i | \boldsymbol{y}\}$, for $i = 1, \ldots, 20$. The lines connecting the 10 expected shifts for $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_{20}$ are labeled "1" and "20," respectively. The lines connecting the expected shifts for curves $\boldsymbol{Y}_2$–$\boldsymbol{Y}_{19}$ are indistinguishable and essentially lie on the horizontal line through zero.

posterior probabilities p_i^y and p_i^β for each curve, along with overlapping batch mean estimates of standard error for curves $\boldsymbol{Y}_1$, $\boldsymbol{Y}_{19}$, and $\boldsymbol{Y}_{20}$ (Chen and Schmeiser (1993)). Curves $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_{20}$ are strongly indicated to contain one or more components with measurement errors. Curve $\boldsymbol{Y}_{19}$ is singled out as a regression coefficient outlier with curve $\boldsymbol{Y}_{20}$ having the second largest estimated p_i^β which is 0.54. No other curve shows evidence of either type of model inadequacy.

To provide more detailed information about the nature of the violations identified in Table 1, we estimated the means of the posterior distributions of the Stage I and Stage II shifts, $\mathrm{E}\{\delta_i^y \times \boldsymbol{c}_i \mid \boldsymbol{y}\}$ and $\mathrm{E}\{\delta_i^\beta \times \boldsymbol{d}_i \mid \boldsymbol{y}\}$. Figure 2 is an index plot in which the ten estimated components of $\mathrm{E}\{\delta_i^y \times \boldsymbol{c}_i \mid \boldsymbol{y}\}$ are connected. Clearly, $\boldsymbol{Y}_1$ is designated as a Stage I outlier because of its *third* component, $\boldsymbol{Y}_{20}$ is designated as a Stage I outlier because of its *fourth* component, while $\boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_{19}$ show no evidence of Stage I outlyingness.

Similarly, we estimated the means of the posterior distributions of the Stage II shifts for the intercept, β_{i1}, and the slope, β_{i2}, which are $\mathrm{E}\{\delta_i^\beta \times d_{i1} \mid \boldsymbol{y}\}$ and $\mathrm{E}\{\delta_i^\beta \times d_{i2} \mid \boldsymbol{y}\}$, respectively. The expected (intercept, slope) shifts are: $(0.01, -0.002)$ for $\boldsymbol{Y}_1$, $(1.95, -0.02)$ for $\boldsymbol{Y}_{19}$ and $(0.02, 0.60)$ for $\boldsymbol{Y}_{20}$. For $\boldsymbol{Y}_2$–$\boldsymbol{Y}_{18}$ the intercept estimates range between -0.01 and 0.01 and the slope estimates range between -0.02 and 0.02. These values show clear evidence that $\boldsymbol{Y}_{19}$ has a non-conforming *intercept* and moderate evidence that $\boldsymbol{Y}_{20}$ has a non-conforming *slope*.

The extent to which the model allows the curves and the individual measurements to vary has a significant impact in determining what the model regards as an outlier. The impact for our Bayesian model can be quantified by conducting a *sensitivity analysis* of the model conclusions to the prior assumptions by studying how the variances σ_ε^2, $\sigma_{\beta_1}^2$, and $\sigma_{\beta_2}^2$ affect the posterior probabilities of declaring individual curves to be outliers at each stage (p_i^y and p_i^β) and the posterior mean shifts at each stage ($\mathrm{E}\{\delta_i^y \times \boldsymbol{c}_i \mid \boldsymbol{y}\}$ and $\mathrm{E}\{\delta_i^\beta \times \boldsymbol{d}_i \mid \boldsymbol{y}\}$).

For this purpose, we decided to vary the prior parameters b_ε, b_{β_1}, and b_{β_2} which control the scale of the prior distributions for σ_ε^2, $\sigma_{\beta_1}^2$, and $\sigma_{\beta_2}^2$, respectively, while holding all other parameters fixed. The rationale for this choice begins by recalling that our model took $\sigma_\varepsilon^2 \sim IG(a_\varepsilon, b_\varepsilon)$, $\sigma_{\beta_1}^2 \sim IG(a_{\beta_1}, b_{\beta_1})$, and $\sigma_{\beta_2}^2 \sim IG(a_{\beta_2}, b_{\beta_2})$ where $(a_\varepsilon, b_\varepsilon) = (a_{\beta_1}, b_{\beta_1}) = (a_{\beta_2}, b_{\beta_2}) = (3.0, 1.0)$. This inverse gamma distribution has mean 0.5 and variance 0.25, so that its coefficient of variation (CV) is equal to 1.0. If an alternative $IG(a_\varepsilon, b_\varepsilon)$ prior for σ_ε^2 leads to a model favoring *larger* measurement errors than the $IG(3.0, 1.0)$ prior, then it should be *more* difficult to declare individual components of $\boldsymbol{Y}_i$ to be Stage I outliers (and vice versa).

We decided to investigate alternative inverse gamma prior distributions all having CV equal to 1.0, with means ranging from a minimum of 0.25 to a maximum of 1.0. An easy calculation shows that this is equivalent to keeping $a_\varepsilon = 3.0$ and studying $b_\varepsilon \in [0.5, 2.0]$ (with the parameterization of the inverse gamma distribution that we use, the mean increases with b_ε). In a similar way, if an alternative $IG(a_{\beta_1}, b_{\beta_1})$ prior for $\sigma_{\beta_1}^2$ leads to a model favoring *larger* departures of β_1 from its prior mean than the $IG(3.0, 1.0)$ prior, then it should be *more* difficult to declare individual intercepts to be outliers (and vice versa). An analogous statement holds for the slopes. For the same reason as described above we keep $a_{\beta_1} = a_{\beta_2} = 3.0$ and vary $b_{\beta_1} \in [0.5, 2.0]$ and $b_{\beta_2} \in [0.5, 2.0]$.

Having decided which prior parameters to vary and the ranges over which each is to be studied, one must determine the four diagnostics identified above as functions of the prior parameters. When only the single MCMC run at the original prior parameters is available, the most frequently used method for assessing the impact of alternative values of these parameters on the inferential conclusions is Importance Sampling (Robert and

Table 2. Bayes estimates of the posterior probabilities p_{19}^β for the sensitivity study of Section 4.

b_ε	b_{β_1}	b_{β_2}	p_{19}^β	s.e.
0.575	0.875	1.175	0.76	0.04
1.025	0.575	1.475	0.81	0.06
0.725	1.775	1.775	0.51	0.03
1.325	1.925	1.025	0.33	0.03
1.925	1.325	1.625	0.39	0.04
1.775	0.725	0.725	0.65	0.07
1.475	1.025	1.925	0.53	0.06
0.875	1.175	0.875	0.69	0.05
1.175	1.475	0.575	0.41	0.04
1.625	1.625	1.325	0.38	0.03
1.0	1.0	1.0	0.80	0.03

Casella (1999)); we use Importance Sampling for this purpose in the bone strength analysis of Section 5.

In the present application, as well as others of moderate computational expense, it is feasible to make a few additional MCMC runs using alternative b_ε, b_{β_1}, and b_{β_2} values and to interpolate the output to obtain estimates of the posterior quantities of interest over a dense grid of $(b_\varepsilon, b_{\beta_1}, b_{\beta_2})$-values in $[0.5, 2.0]^3$. The remainder of this section shows how we used this method to assess the sensitivity of p_{19}^{β} to $(b_\varepsilon, b_{\beta_1}, b_{\beta_2})$.

We selected ten vectors $(b_\epsilon, b_{\beta_1}, b_{\beta_2})$ in $[0.5, 2.0]^3$ (in addition to the original values). We then used our BUGS program to estimate the corresponding p_{19}^{β}. This provided $n = 11$ points to be used for interpolation which are listed in Table 2. We chose the ten additional $(b_\varepsilon, b_{\beta_1}, b_{\beta_2})$ input sites to be "space filling" over $[0.5, 2.0]^3$, in the sense that these 10 points maximize the minimum Euclidean interpoint distance among all possible ten-point Latin hypercube designs on $[0.5, 2.0]^3$ (Johnson *et al.* (1990) and McKay *et al.* (1979)). Latin hypercube designs allocate the $(b_\varepsilon, b_{\beta_1}, b_{\beta_2})$ points in such a way that the projection of the 10 design points onto any of the b_ε, b_{β_1} and b_{β_2} axes gives a uniformly distributed point spread over $[0.5, 2.0]$. We used the ACED software of Welch (1985) to compute the required input design.

Our p_{19}^{β} interpolator was an empirical kriging predictor based on a stationary Gaussian stochastic process with product power exponential correlation (Sacks *et al.*

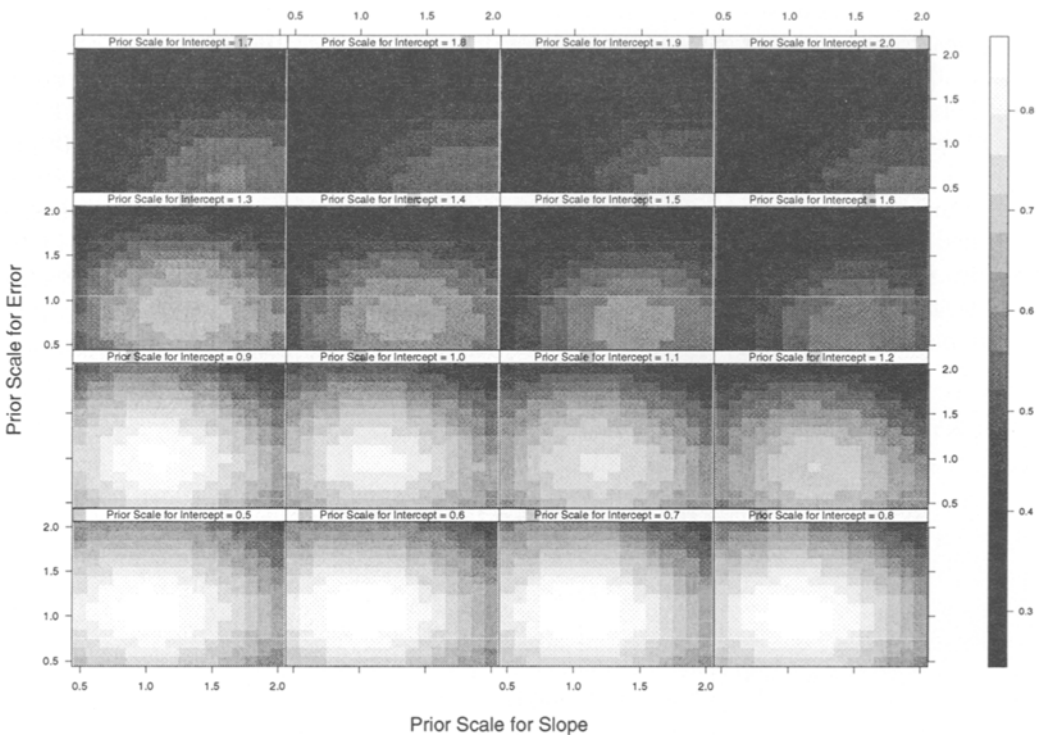

Fig. 3. Sensitivity analysis for the simple linear regression example. Estimated posterior probabilities p_{19}^{β} as a function of the scale parameters of the prior distributions for the variances of the innovation error (b_ϵ) and of the regression coefficients (b_{β_1} and b_{β_2}).

(1989); Koehler and Owen (1996)). We interpolated p_{19}^{β} at an equispaced grid of 4096 $(= 16^3)$ $(b_\varepsilon, b_{\beta_1}, b_{\beta_2})$ values; several tools can be used to study the results. Because there are only three parameters that are varied in this example, we used trellis plots (Becker *et al.* (1996)) to provide detailed information about the effects of each parameter on the diagnostic. Figure 3 shows the interpolated p_{19}^{β} as a function of $(b_\varepsilon, b_{\beta_1}, b_{\beta_2})$.

The primary conclusion is that p_{19}^{β} is more sensitive to the choice of b_{β_1} than to the choice of either b_{β_2} or b_ε. Consistent with our intuition, larger $\sigma_{\beta_1}^2$ values make deviations in the individual intercepts appear less like outliers, thus making p_{19}^{β} decrease. A secondary conclusion is that, for fixed b_{β_1}, the posterior probability that $\boldsymbol{Y}_{19}$ contains a Stage II outlier is relatively constant in b_{β_2} and b_ε, although large values of both b_{β_2} and b_ε generally yield lower posterior values for p_{19}^{β} than other choices. This is also intuitive because, if larger measurement errors and larger variation in slopes are possible under the model, then an aberration in the curve's intercept may simply be due to a "wild" slope that causes the curve to intersect the vertical axis at a point distant from the intercepts of the remaining curves making the intercept look less like an outlier.

We conclude by noting that the values plotted in Fig. 3 are affected by two sources of uncertainty. The first is uncertainty in the 11 estimates listed in Table 2 as quantified by the corresponding standard errors. The second source of uncertainty is introduced by the stochastic process prior that is at the heart of the kriging interpolator (see Sacks *et al.* (1989)); in our case the interquartile range of the standard errors due to interpolator uncertainty is $[0.04, 0.07]$ over the 4,096 grid values used to construct Fig. 3.

5. Worked example—Measuring bone strength

The data for this example come from an observational study of patients admitted for evaluation and possible hip replacement surgery at the Hospital for Special Surgery in New York City. As part of the evaluation process, a series of CAT scans were made of the cross section of each patient's femur in the area near the lesser trochanter (a section of the femur close to the hip joint). Figure 4 displays a typical bone cross-section with two clearly delineated regions: a central area of honeycomb-like trabecular bone and a surrounding area of denser cortical bone.

These data consist of 41 scans, which are a subset of those analyzed by Peruggia *et al.* (1994) to identify the factors associated with the distribution of bone strength. We illustrate the identification of outliers for one of the models of cortical bone thickness discussed by Peruggia *et al.* (1994).

The thickness was measured counterclockwise starting from an anatomical landmark along a series of 72 equally spaced rays emanating from the centroid of each section. Figure 4 shows, for a typical section, the landmark (denoted by a small circle), the centroid, and the rays. A set of risk factors was available to explain differences in the subject's cross-sectional thicknesses. The risk factors were the subject's age (AGE), gender (G), and diagnosis (DX). The variable AGE was the subject's age centered so that AGE $= 0$ corresponded to a 45 year-old subject and was scaled to range over $(-1, 1)$; the latter facilitated the prior modeling of the AGE-related regression coefficients. There were three DX categories: osteoarthritis (OA), juvenile rheumatoid arthritis (JRA), and other (OTH). OA is the most common cause of hip replacement surgery in older patients; its etiology is the breakdown of cartilage in the joint, typically due to injury or wear. JRA is a congenital disease which manifests itself early in life and affects the growth of

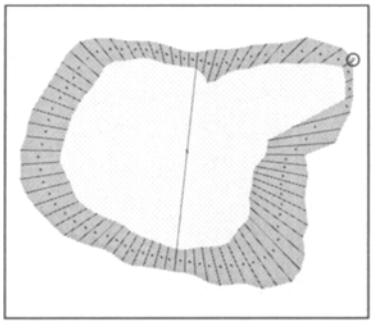

Fig. 4. A digitized image of the cross section of human bone through a slice taken at the lesser trochanter.

the hip bones and connective tissue. Patients in the OTH category tend to have acute problems, such as injuries or bone cancer; prior to their evaluation they can be regarded as nearly normal.

The roughly circular character of the cortical shell thickness suggested the following form for the Stage I model. The i-th curve was summarized by the curve-specific coefficients $(\mu_i, \alpha_{i1}, \alpha_{i2}, \alpha_{i3}, \beta_{i1}, \beta_{i2}, \beta_{i3})$ from the Fourier fit

$$Y_{i,j} = \mu_i + \sum_{f=1}^{3} [\alpha_{if} \cos(2\pi f \times j/72) + \beta_{if} \sin(2\pi f \times j/72)] + \eta_{ij} \tag{5.1}$$

for $i = 1, \ldots, 41$ and $j = 1, \ldots, 72$. Independence of the deviations from the Fourier series model was not tenable and we based the within-curve errors on the autoregressive model

$$\eta_{ij} = \phi_1 \eta_{i,j-1} + \phi_2 \eta_{i,j-2} + \varepsilon_{ij}, \tag{5.2}$$

subject to the initialization conditions described in Section 2.

In Stage II, the vector of the 41 Fourier intercepts $(\mu_1, \ldots, \mu_{41})$ was regressed on

$$\begin{aligned} &\gamma_{\mu,1} + \gamma_{\mu,2}\mathrm{I}[\mathrm{G} = \mathrm{MALE}] + \gamma_{\mu,3}\mathrm{I}[\mathrm{DX} = \mathrm{OA}] + \gamma_{\mu,4}\mathrm{I}[\mathrm{DX} = \mathrm{JRA}] \\ &\quad + \gamma_{\mu,5}\mathrm{AGE} + \gamma_{\mu,6}\mathrm{AGE}^2 \end{aligned} \tag{5.3}$$

where the Stage II regression coefficients $\boldsymbol{\gamma}_\mu = (\gamma_{\mu,1}, \gamma_{\mu,2}, \gamma_{\mu,3}, \gamma_{\mu,4}, \gamma_{\mu,5}, \gamma_{\mu,6})$ are unknown and $\mathrm{I}[E]$ is the 0/1 indicator function of the event E. Equation (5.3) describes each subject's Stage I intercept μ_i by a quadratic equation in that subject's AGE with an intercept that is additive in the subject's gender and diagnostic-group. For example, $\gamma_{\mu,1}$ is the intercept for a 45 year-old female in the OTH diagnosis group while $\gamma_{\mu,1} + \gamma_{\mu,3}$ is the intercept for a 45 year-old female with OA. The same regression variables (with harmonic-specific regression coefficients) were used to summarize the variability in *each* of the 41×1 vectors corresponding to the six remaining harmonic coefficients. Peruggia *et al.* (1994) contains additional detail.

We determined outliers relative to the null model (5.1)–(5.3). The extended model was specified as follows. In Stage II, the components of the thickness location shift, $\boldsymbol{c}_i$, were assigned the common mean $\mu_c = 0$ with fairly large variance, $\sigma_c^2 = 4.0$. The prior

variances of the autoregressive parameters were set to be $\sigma^2_\phi[1] = 4 = \sigma^2_\phi[2]$. The shape and scale parameters of the inverse gamma distribution of the first and second innovations ε_{i1} and ε_{i2} were set equal to 1. As was done in Section 4, the shape parameters and the scale parameters of all remaining inverse gamma distributions were set equal to 3.0 and to 1.0, respectively.

In Stage III, the prior means of the six Stage II regression coefficients γ_μ for the parameters $(\mu_1, \ldots, \mu_{41})$ of equation (5.3) were all set to zero and their prior variances were set to be $(\sigma^2_\gamma[\mu, 1], \ldots, \sigma^2_\gamma[\mu, 6]) = (25.0, 1.0, 1.0, 1.0, 1.0, 1.0)$. The fairly large variance for the intercept of the regression equation of $(\mu_1, \ldots, \mu_{41})$ which is $\gamma_{\mu,1}$, the mean thickness for a 45 year old female with OTH, allowed adequate support for values far from zero; the bulk of the prior support for the other regression coefficients was restricted to smaller values. Each vector of regression coefficients of the six sine and cosine harmonics was given zero prior mean and the prior variances $(\sigma^2_\gamma[l, 1], \ldots, \sigma^2_\gamma[l, 6])$ were set equal to $(4.0, 1.0, 1.0, 1.0, 1.0, 1.0)$, for $l = 2, \ldots, 7$. The variances of the intercept terms were set differently than the variances of the coefficients of the remaining regressors for the same reasons that suggested a similar assumption about the $\sigma^2_\gamma[\mu, k]$. The means of all Stage II deviation magnitudes, d_{il}, were set equal to zero (i.e., $\mu_d[l] = 0$) and their prior variances were set equal to 4.0, i.e., $\sigma^2_d[l] = 4.0$, for all l.

In Stage IV, for the reasons explained in Section 4, the parameters of the beta distributions for the probabilities p_y and p_β of a Stage I and Stage II deviation were set equal to $u_y = 1.5 = u_\beta$ and $v_y = 8.5 = v_\beta$.

Draws from the posterior distribution of the model parameters given the data were generated using our BUGS program. The first 2,000 iterations were used as a warm-up for the algorithm. The estimates stated below were based on subsequent cycles of the MCMC sampler spaced 10 apart, for a total of 2,500 samples.

The estimated posterior probabilities p^y_i are all zero, indicating that there are no bone sections containing measurement error outliers. Thus, the 7-term Fourier series expansion in equation (5.1) appears to approximate the thickness curve well for all 41 bone sections. The largest estimated posterior probabilities of a Stage II outlier (standard errors) are $p^\beta_2 = p^\beta_6 = p^\beta_8 = 1.00(0.00)$, $p^\beta_{39} = 0.92(0.03)$, and $p^\beta_{17} = 0.41(0.08)$. The standard errors were also computed using the overlapping batch means estimator of Chen and Schmeiser (1993). All other estimated posterior probabilities p^β_i are less than 0.07. Thus, our diagnostic identifies four bone sections (2, 6, 8, and 39) as Stage II outliers, and suggests that bone section 17 might also have Stage II anomalies. We expect these bone sections to have one or more Fourier coefficients that differ from the pattern of the corresponding coefficients for sections having similar covariates.

There are several methods that can be used to provide more detailed information about the bone sections identified as Stage II outliers. First, we assess how Stage II violations are reflected in the Stage I coefficients of the individual curves. We do this by examining the size of the estimated shifts $\widehat{\delta_{\beta_i} d_{i\mu}} = 2500^{-1} \sum_{m=1}^{2500} \delta^{(m)}_{\beta i} d^{(m)}_{i\mu}$, where $\{\delta^{(m)}_{\beta_i}, d^{(m)}_{i\mu}\}^{2500}_{m=1}$, for $i = 1, \ldots, 41$, are the 2,500 draws from the Gibbs sampler for those coefficients (based on the expanded model). Similar estimates can be constructed for each harmonic coefficient α_{if} and β_{if}. Figure 5 displays an index plot, over the 41 bone sections, of the estimated shifts for each of the seven Stage I Fourier coefficients. These estimated shifts should be large in the presence of a Stage II violation. For example, Fig. 5 shows that the Fourier intercepts are ill-fit by the Stage II regression for bone sections 2, 6 and 39. Similarly all but one of the harmonic coefficients are poorly fit for

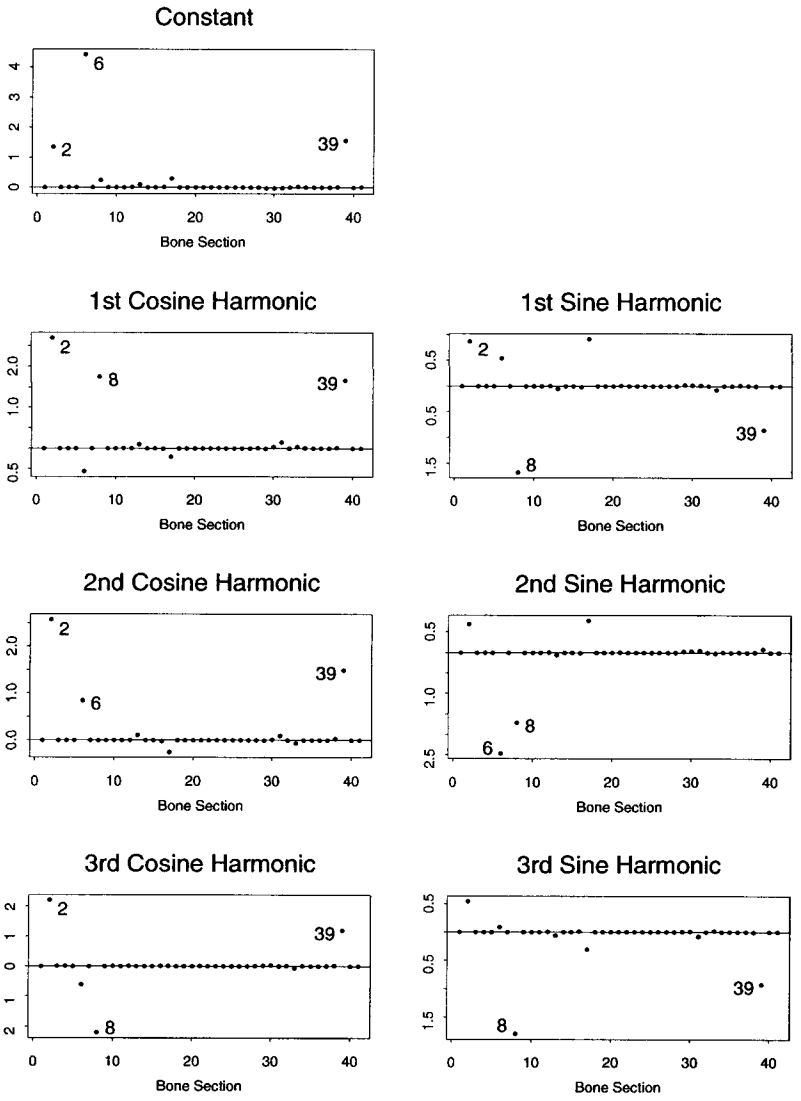

Fig. 5. Index plots of the estimated posterior mean shifts for the intercept and the six harmonic coefficients in the Fourier expansion of equation (5.1).

bone sections 8 and 39.

A second diagnostic gives an overall view of how the presence of deviations in the Stage II regression model affects the fit of the data. This diagnostic compares the predicted thicknesses based on the null and expanded models. Each of the seven coefficients in the Fourier model for each subject is estimated in two ways based on the MCMC draws from the corresponding model. The first estimate of μ_i is $\widehat{\mu}_i^{\text{null}} = \boldsymbol{z}_{i\mu}^{\top}\widehat{\boldsymbol{\gamma}}_{\mu}^{\text{null}}$, where $\widehat{\boldsymbol{\gamma}}_{\mu}^{\text{null}}$ is the estimated posterior mean of the Stage II regression coefficients based on the null model. The second estimate of μ_i is $\widehat{\mu}_i^{\text{exp}} = \boldsymbol{z}_{i\mu}^{\top}\widehat{\boldsymbol{\gamma}}_{\mu}^{\text{exp}} + \delta_{\beta_i}\widehat{d_{i\mu}}$, where $\widehat{\boldsymbol{\gamma}}_{\mu}^{\text{exp}}$ is the estimated posterior mean of the Stage II regression coefficients based on the expanded model and $\delta_{\beta_i}\widehat{d_{i\mu}}$ is defined in the previous paragraph. Similar pairs of estimates can be

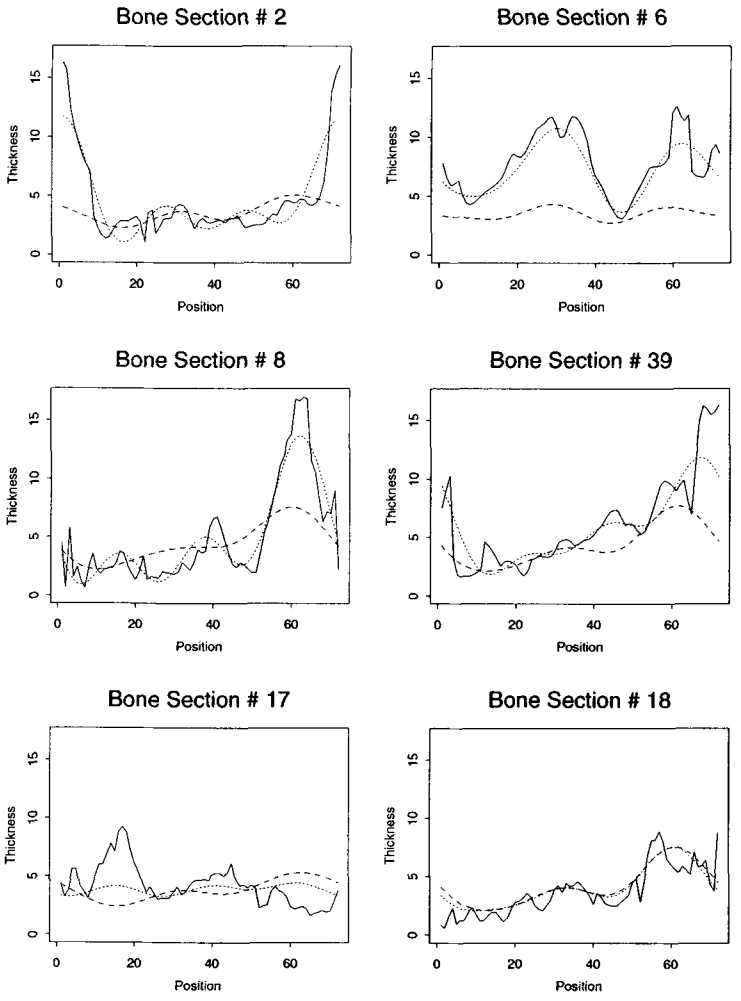

Fig. 6. Plots of the raw thickness $\boldsymbol{Y}_i$ (solid curve) and the two estimates given by equation (5.4). The dashed and dotted curves are based on the null and expanded model, respectively. Plots are provided for the four bone sections strongly indicated as outliers (2, 6, 8, and 39), the bone section marginally indicated as an outlier (17), and, for comparison, a bone section showing no indication of being an outlier (18).

constructed for the other six Fourier coefficients.

This second diagnostic is presented in Fig. 6. For each bone section identified as an outlier, we display index plots of the raw thickness curve (solid line) overlaid with two estimates of the mean curve

$$\widehat{\mu}_i^{[k]} + \sum_{f=1}^{3} [\widehat{\alpha}_{if}^{[k]} \cos(2\pi f \times j/72) + \widehat{\beta}_{if}^{[k]} \sin(2\pi f \times j/72)], \tag{5.4}$$

$$\text{for} \quad k \in \{\text{null}, \text{exp}\}.$$

The lack of fit of the null Fourier model without shift adjustment (dashed line) is most

pronounced for bone section 6. Lack of fit is also very clear for bone sections 2, 8, and 39 (in the region of maximum thickness).

Sensitivity analyses can be conducted in a way similar to that described in Section 4. In this more highly computational setting, we used importance sampling to investigate the effect of the inverse gamma prior distributions for the variance parameters σ^2_ϵ, σ^2_μ, $\sigma^2_{\alpha_1}$, $\sigma^2_{\alpha_2}$, $\sigma^2_{\alpha_3}$, $\sigma^2_{\beta_1}$, $\sigma^2_{\beta_2}$, and $\sigma^2_{\beta_3}$. For example, by varying the scale parameters of the above prior distributions over the lattice $\{0.67, 1.0, 1.33\}^8$, the estimated value of p^β_{39} ranged from 0.86 to 0.96. Qualitatively, the dependence of p^β_{39} on the scale parameters is similar to the one described for the 19-th observation in the example of Section 4. In particular larger values of the scale parameters lead to smaller estimated values of p^β_{39}.

6. Discussion

We mention but a few of the many methods to extend the basic model (2.1)–(2.4). If the within-curve measurements are not equally spaced, then an autoregressive error structure could still be specified as in Jones and Boadi-Boateng (1991). The case of innovation outliers can be modeled using shift indicators at the innovation level, in which case the impact of the deviations would be filtered through the AR structure.

Given the relevant prior information, one could assume a more complicated (and possibly random) covariance structure for the regression parameters β_{il}. A discussion of various possible specifications that would also apply to our setting is given in Section 2.3 of George and McCulloch (1993). The basic model can also be expanded to allow outlier detection in n-stage hierarchical Bayesian models with $n > 3$. To check for assumption violations at Stage III or higher, deviation indicators and deviation magnitudes can be added to appropriate parameters of the prior distributions for which it is desired to detect deviations. Of course, this extension would introduce a tremendous amount of complexity to the problem.

As mentioned in Section 3, Rao-Blackwellized estimates can be computed for any of the posterior diagnostics that we recommend (Gelfand and Smith (1990)). We initially calculated Rao-Blackwellized estimates for p^y_i and p^β_i but found that, in our examples, they differed little from the tabulation estimates reported by our BUGS program. For this reason, we used the more automatic BUGS estimates of all posterior quantities in the examples of Sections 4 and 5.

The diagnostics proposed in this paper are model-based—they assume a multi-stage, null model for repeated measures data; location shift extensions are added to the null model at various levels of the hierarchy. Alternative model elaboration steps to accommodate outliers (e.g., replacement of the normal priors by t-like priors) could also be entertained as in Wakefield *et al.* (1994) and, more recently, in Spiegelhalter and Marshall (1999).

For data sets where the distinction between usual and unusual observations is clear-cut (as judged by the modeling assumptions and prior specifications) the Gibbs sampler for our extended model converges rapidly and the estimated probabilities that individual observations are outliers are highly reliable. However, for data sets where either the number of "outliers" is large relative to the prior assumption or other anomalies occur, it is possible for the posterior distribution to assign large probability to several nearly disjoint regions of the parameter space. In such instances, the Gibbs sampler does not mix well over these regions, successive draws exhibit strong dependencies, and longer runs are required to obtain reliable estimates. A situation of this sort occurs in

the example of Section 5, where the estimate of p_{17}^{β} is only accurate to within ± 0.08 despite being based on 2,500 subsamples from a total of 25,000 iterations of our BUGS code. A similar difficulty occurs in problems of Bayesian variable selection (George and McCulloch (1993)). Called by another name, such situations are related to the masking phenomenon described in Justel and Peña (1996).

Appropriate Metropolis-Hastings steps can be introduced to facilitate communication between multiple regions (see, for example, George and McCulloch (1997)). The effectiveness of such a strategy, however, is highly dependent on having some knowledge of the structure of the posterior distribution. We are developing such algorithms for outlier detection models.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. SES-0214574. The authors would like to thank the Hospital for Special Surgery, New York, NY for allowing use of the cortical bone data analyzed in Section 5. They would also like to thank an Associate Editor and two referees for suggestions that improved the paper.

References

Becker, R. A., Cleveland, W. S. and Shyu, M.-J. (1996). The visual design and control of trellis display, *Journal of Computational and Graphical Statistics*, **5**, 123–155.

Berger, J. O. and Hui, S. L. (1983). Empirical Bayes estimation of rates in longitudinal studies, *Journal of the American Statistical Association*, **78**, 753–760.

Berlin, J. A., Santanna, J., Schmid, C. H., Szczech, L. A. and Feldman, H. I. (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head, *Statistics in Medicine*, **21**, 371–387.

Carota, C., Parmigiani, G. and Polson, N. G. (1996). Diagnostic measures for model criticism, *Journal of the American Statistical Association*, **91**, 753–762.

Chaloner, K. (1994). Residual analysis and outliers in Bayesian hierarchical models, *Aspects of Uncertainty. A Tribute to D. V. Lindley* (eds. P. R. Freeman and A. F. M. Smith), 149–157, Wiley, Chichester.

Chaloner, K. and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis, *Biometrika*, **75**, 651–659.

Chen, M.-H. and Schmeiser, B. (1993). Performance of the Gibbs, hit-and-run, and metropolis samplers, *Journal of Computational and Graphical Statistics*, **2**, 251–272.

Cnaan, A., Laird, N. M. and Slasor, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data, *Statistics in Medicine*, **16**, 2349–2380.

Crowder, M. J. and Hand, D. J. (1990). *Analysis of Repeated Measures*, Chapman & Hall, New York.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398–409.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling, *Journal of the American Statistical Association*, **88**, 881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection, *Statistica Sinica*, **7**, 339–374.

Ho, Y.-Y., Peruggia, M. and Santner, T. J. (1995). Diagnostics for hierarchical Bayesian repeated measures models, *27th Symposium of the Interface: Computing Science and Statistics* (eds. M. M. Meyer and J. L. Rosenberger), 387–391, Interface Foundation of North America, Fairfax Station, Virginia.

Hodges, J. S. (1998). Some algebra and geometry for hierarchical models, applied to diagnostics (with discussion), *Journal of the Royal Statistical Society, Series B*, **60**, 497–536.

Johnson, M. E., Moore, L. M. and Ylvisaker, D. (1990). Minimax and maximin distance designs, *Journal of Statistical Planning and Inference*, **26**, 131–148.

Jones, M. C. and Rice, J. A. (1992). Displaying the important features of large collections of similar curves, *The American Statistician*, **46**, 140–145.

Jones, R. H. and Boadi-Boateng, F. (1991). Unequally spaced longitudinal data with AR(1) serial correlation, *Biometrics*, **47**, 161–175.

Joseph, L., Wolfson, D. B., Belisle, P., Brooks, J. O. 3rd, Mortimer, J. A., Tinklenberg, J. R. and Yesavage, J. A. (1999). Taking account of between-patient variability when modeling decline in Alzheimer's disease, *American Journal of Epidemiology*, **149**, 963–973.

Justel, A. and Peña, D. (1996). Gibbs sampling will fail in outlier problems with strong masking, *Journal of Computational and Graphical Statistics*, **5**, 176–189.

Koehler, J. R. and Owen, A. B. (1996). Computer experiments, *Handbook of Statistics* (eds. S. Ghosh and C. R. Rao), 261–308, North Holland, Elsevier, Amsterdam.

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

Lambert, P. C., Abrams, K. R., Jones, D. R., Halligan, A. W. F. and Shennan, A. (2001). Analysis of ambulatory blood pressure monitor data using a hierarchical model incorporating restricted cubic splines and heterogeneous within-subject variances, *Statistics in Medicine*, **20**, 3789–3805.

Langford, I. H. and Lewis, T. (1998). Outliers in multilevel data (Disc: P153-160), *Journal of the Royal Statistical Society, Series A, General*, **161**, 121–153.

Lindsey, J. K. (1993). *Models for Repeated Measurements*, Clarendon Press, Oxford.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data, *Biometrics*, **46**, 673–687.

McKay, M. D., Beckman, R. J. and Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics*, **21**, 223–245.

Palmer, J. L. and Müller, P. (1998). Bayesian optimal design in population models for haematologic data, *Statistics in Medicine*, **17**, 1613–1622.

Pauler, D. K. and Laird, N. M. (2000). A mixture model for longitudinal data with application to assessment of noncompliance, *Biometrics*, **56**, 464–472.

Pauler, D. K. and Laird, N. M. (2002). Non-linear hierarchical models for monitoring compliance, *Statistics in Medicine*, **21**, 219–229.

Peruggia, M., Santner, T. J., Ho, Y. Y. and Macmillan, N. J. (1994). A hierarchical Bayesian analysis of circular data with autoregressive errors: Modeling the mechanical properties of cortical bone, *Statistical Decision Theory and Related Topics V* (eds. S. S. Gupta and J. O. Berger), 201–220, Springer-Verlag, New York.

Pettit, L. I. and Smith, A. F. M. (1985). Outliers and influential observations in linear models, *Bayesian Statistics II* (eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), 473–494, North Holland, Elsevier, Amsterdam.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*, Springer-Verlag, New York.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989). Design and analysis of computer experiments, *Statistical Sciences*, **4**, 409–423.

Segal, M. R. (1994). Representative curves for longitudinal data via regression trees, *Journal of Computational and Graphical Statistics*, **3**, 214–233.

Sharples, L. D. (1990). Identification and accommodation of outliers in general hierarchical models, *Biometrika*, **77**, 445–453.

Spiegelhalter, D. J. and Marshall, E. C. (1999). Inference-robust institutional comparisons: A case study of school examination results, *Bayesian Statistics 6, Proceedings of the Sixth Valencia International Meeting* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 613–630, Clarendon Press, Oxford.

Spiegelhalter, D. J., Thomas, A., Best, N. G. and Gilks, W. R. (1996). *BUGS Bayesian Inference Using Gibbs Sampling, Version 0.5, (version ii)*, MRC Biostatistics Unit, Cambridge, U.K.

Tan, M., Qu, Y., Mascha, E. and Schubert, A. (1999). A Bayesian hierarchical model for multi-level repeated ordinal data: Analysis of oral practice examinations in a large anaesthesiology training

programme, *Statistics in Medicine*, **18**, 1983–1992.

Verdinelli, I. and Wasserman, L. (1991). Bayesian analysis of outlier problems using the Gibbs sampler, *Statistics and Computing*, **1**, 105–117.

Wakefield, J. C., Smith, A. F. M., Racine-Poon, A. and Gelfand, A. E. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler, *Applied Statistics*, **43**, 201–221.

Weisberg, S. (1983). Comment on "Developments in linear regression methodology: 1959–1982", *Technometrics*, **25**, 240–244.

Weiss, R. E. (1995). Residuals and outliers in repeated measures random effects models, Tech. Report, Department of Biostatistics, UCLA School of Public Health.

Welch, W. J. (1985). ACED: Algorithms for the construction of experimental designs, *The American Statistician*, **39**, p. 146.

Zellner, A. (1975). Bayesian analysis of regression error terms, *Journal of the American Statistical Association*, **70**, 138–144.