Available online at www.sciencedirect.com

ScienceDirect

Journal of Statistical Planning and Inference 138 (2008) 1592–1604

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Identification of the variance components in the general two-variance linear model[☆]

Brian J. Reich[a,*], James S. Hodges[b]

[a]Department of Statistics, North Carolina State University, 2501 Founders Drive, Box 8203, Raleigh, NC 27695, USA
[b]Division of Biostatistics, School of Public Health, University of Minnesota, 2221 University Ave. SE, Suite 200, Minneapolis, MN 55414, USA

Received 26 February 2006; received in revised form 20 April 2007; accepted 6 May 2007
Available online 12 August 2007

Abstract

Bayesian analyses frequently employ two-stage hierarchical models involving two-variance parameters: one controlling measurement error and the other controlling the degree of smoothing implied by the model's higher level. These analyses can be hampered by poorly identified variances which may lead to difficulty in computing and in choosing reference priors for these parameters. In this paper, we introduce the class of two-variance hierarchical linear models and characterize the aspects of these models that lead to well-identified or poorly identified variances. These ideas are illustrated with a spatial analysis of a periodontal data set and examined in some generality for specific two-variance models including the conditionally autoregressive (CAR) and one-way random effect models. We also connect this theory with other constrained regression methods and suggest a diagnostic that can be used to search for missing spatially varying fixed effects in the CAR model.

Keywords: Conditional autoregressive prior; Hierarchical models; Identification; Mixed linear model; Variance components

1. Introduction

Advances in computing allow Bayesian analyses of complicated hierarchical models with relative ease. However, these powerful tools must be used cautiously; the posterior for, say, a richly parameterized model may be weakly identified, particularly for variance parameters. This may lead to computational problems and highlights the difficulty of choosing reference priors for these parameters (Gelman, 2005). The present paper develops some theory and tools for analyzing identification for the simplest interesting class of such models, those with two unknown variances. This includes scatterplot and lattice smoothers and random-intercept models, among others.

To motivate this problem, consider the periodontal data in Fig. 1a from one subject in a clinical trial of a new periodontitis treatment, conducted at the University of Minnesota's Dental School (Shievitz, 1997). One of the trial's outcome measures was attachment loss (AL), the distance down a tooth's root (in millimeters) that is no longer attached to the surrounding bone by periodontal ligament. AL is measured at six locations on each tooth, for a total of $N = 168$ locations, and is used to quantify cumulative damage to a subject's periodontium. The first two rows of Fig. 1a plot

[☆] The research conducted by Reich has been supported by National Science Foundation Grant DMS 0354189.

[*] Corresponding author. Tel.: +919 513 7686; fax: +919 515 7591.
E-mail address: reich@stat.ncsu.edu (B.J. Reich).

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.05.046

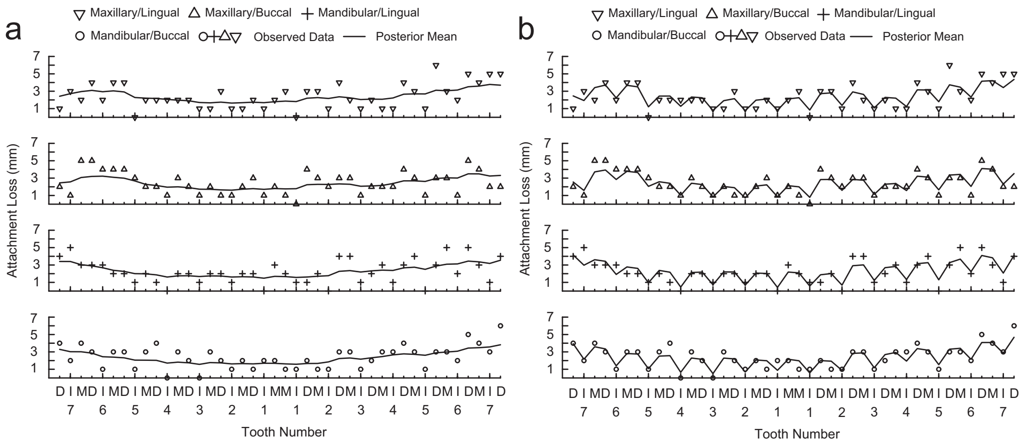

Fig. 1. Periodontal example: raw data and posterior means without (panel (a)) and with (panel (b)) terms with outlying r_i set to fixed effects. (a) Usual CAR fit. (b) Fit with two d_i set to zero.

AL measured along the lingual (cheek side) and buccal (tongue side) strips of locations, respectively, of the maxilla (upper jaw), while the final two rows plot the AL measured at mandibular (lower jaw) locations. Calibration studies commonly show that a single AL measurement has an error with standard deviation of roughly 0.75–1 mm. Fig. 1a shows a severe case of periodontal disease, so measurement error with a 1 mm standard deviation is substantial.

Reich et al. (2007) analyzed AL data using a conditionally autoregressive (CAR) distribution, popularized for Bayesian disease mapping by Besag et al. (1991). In a map with N regions, suppose each region is associated with an unknown quantity β_{1j}, $j = 1, 2, \ldots, N$ (here location j's true AL). Let y_j be the region j's observable (measured AL); assume $y_j|\boldsymbol{\beta}_1, \sigma_{\mathrm{e}}^2$ is normal with mean β_{1j} and variance σ_{e}^2, independent across j. Spatial dependence is introduced through the prior (or model) on $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1N})'$. The CAR model with L_2 norm (also called a Gaussian Markov random field) for $\boldsymbol{\beta}_1$ has improper density

$$p(\boldsymbol{\beta}_1|\sigma_{\mathrm{s}}^2) \propto (\sigma_{\mathrm{s}}^2)^{-(N-G)/2} \exp\left(-\frac{1}{2\sigma_{\mathrm{s}}^2}\boldsymbol{\beta}_1' Q \boldsymbol{\beta}_1\right), \tag{1}$$

where σ_{s}^2 controls the smoothing induced by this prior, smaller values smoothing more than larger; G is the number of "islands" in the spatial structure ($G = 2$ for the periodontal grid since the two jaws are disconnected; Hodges et al., 2003); and Q is $N \times N$ with non-diagonal entries $q_{lj} = -1$ if regions l and j are neighbors and 0 otherwise, and diagonal entries q_{jj} equal to the number of region j's neighbors. This is a multivariate normal kernel, specified by its precision matrix $(1/\sigma_{\mathrm{s}}^2)Q$ instead of the usual covariance matrix.

Fig. 1a plots the posterior mean of $\boldsymbol{\beta}_1$ (solid lines) for the AL data described above. For this fit, both σ_{e}^2 and σ_{s}^2 have Inverse Gamma(0.01, 0.01) priors and 30,000 samples were drawn using Gibbs sampling. The posterior distribution of $\boldsymbol{\beta}_1$ is well identified; the β_{1j} have posterior standard deviations between 0.40 and 0.59 and their posterior means are smoothed considerably. The variances are also well identified. Fig. 2a is a contour plot of the log marginal posterior of $(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$, with a flat prior on $(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$ to emphasize the data's contribution. However, this model has N observations and $N + 2$ unknowns $(\{\beta_{1j}\}, \sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$, so it is far from clear why the variances are identified, how the data are informative about the variances, and how this depends on the spatial structure.

This paper's objectives are to explain how, in problems like this, the data are informative about the variances and to determine which features of a model lead to well-identified variances. Section 2 introduces a class of models with two variances as above: σ_{e}^2, which describes measurement error and σ_{s}^2, which controls smoothing. Section 2 also gives a useful decomposition of the posterior distribution and derives the marginal posteriors of $(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$ and the ratio of variances $r = \sigma_{\mathrm{s}}^2/\sigma_{\mathrm{e}}^2$, which controls the degree of smoothing. The marginal posterior of r suggests a diagnostic that can be used to search for contrasts in the data that are outlying with regard to the information they provide about $(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$. Section 3 explores identification for two common two-variance models, the one-way random effects and CAR models, and applies the theory of Section 2 to the periodontal example. Section 4 concludes by connecting this theory to constrained regression methods such as the Lasso, among other things.

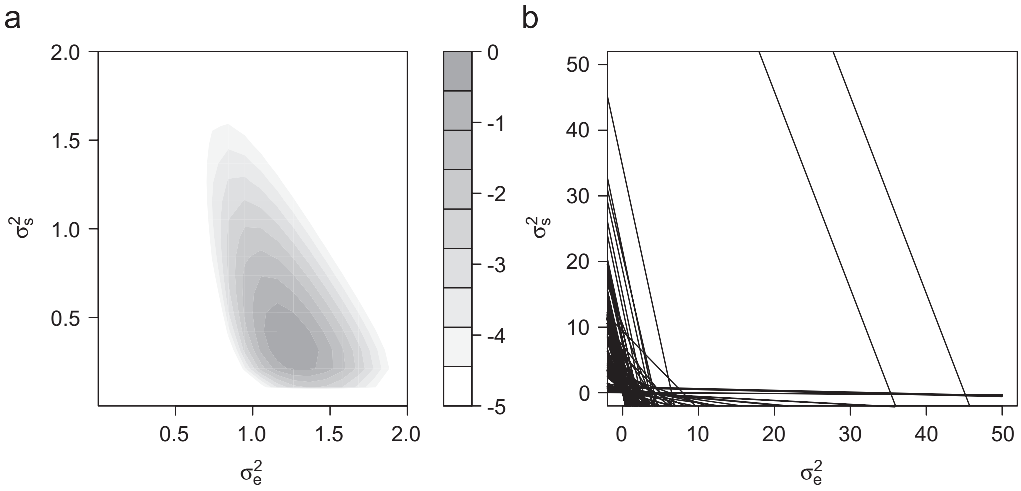

Fig. 2. Joint log posterior (scaled to be zero at the posterior mode) of (σ_e^2, σ_s^2) assuming flat priors for the variances. (a) $\log[p(\sigma_e^2, \sigma|\mathbf{y})]$. (b) Unidentified modal lines of the mixed terms, each evaluated at its own posterior mode $\sigma_s^2 = d_i\hat{\theta}_i^2 - d_i\sigma_e^2$.

2. The general two-variance model

The general two-variance model has the form

$$\begin{aligned} \mathbf{y}|\boldsymbol{\beta}_1, \sigma_e^2 &\sim \mathrm{N}\left(X_1\boldsymbol{\beta}_1, \frac{1}{\sigma_e^2} I_N\right), \\ \boldsymbol{\beta}_1|\boldsymbol{\beta}_2, \sigma_s^2 &\sim \mathrm{N}\left(Z\boldsymbol{\beta}_2, \frac{1}{\sigma_s^2} Q\right), \end{aligned} \tag{2}$$

where $(1/\sigma_e^2)I_N$ and $(1/\sigma_s^2)Q$ are precision matrices with dimensions N and p, respectively, p is $\boldsymbol{\beta}_1$'s length and X_1's rank, Q is assumed known, Z is $p \times l$ and $\boldsymbol{\beta}_2$ is $l \times 1$. A complete Bayesian specification adds priors for $\boldsymbol{\beta}_2$, σ_e^2 and σ_s^2. In this section we assume a flat prior for $\boldsymbol{\beta}_2$, although in some applications $\boldsymbol{\beta}_2$ may be fixed, e.g., the CAR model takes $\boldsymbol{\beta}_2 = 0$. The variances σ_e^2 and σ_s^2 have inverse gamma priors. (This is, of course, much like the model introduced in Lindley and Smith, 1972.)

This paper focuses on the marginal posterior of the variances (σ_e^2, σ_s^2) and of the smoothing parameter $r = \sigma_s^2/\sigma_e^2$. The next section gives a not-too-intuitive reparameterization of the data-level mean structure $X_1\boldsymbol{\beta}_1$, which simplifies derivations and is examined at length in Sections 3 and 4. The reparameterization simplifies but does not change the marginal posterior of (σ_e^2, σ_s^2).

Everything novel in this problem arises from mean-structure effects in $X_1\boldsymbol{\beta}_1$ that are smoothed or shrunk, so integrating out the fixed effects $\boldsymbol{\beta}_2$ simplifies the posteriors of interest. The prior for $\boldsymbol{\beta}_1$ after integrating out $\boldsymbol{\beta}_2$ has mean zero and precision $(1/\sigma_s^2)(Q - QZ'(Z'QZ)^{-1}Z'Q)$. Now reparameterize $X_1\boldsymbol{\beta}_1$ to give an orthogonal design matrix and diagonal prior precision matrix. To do this, let $\Phi = (X_1'X_1)^{-1/2}(Q - QZ'(Z'QZ)^{-1}Z'Q)(X_1'X_1)^{-1/2}$ have spectral decomposition $\Gamma' D \Gamma$ for a $p \times p$ orthogonal Γ and rank c $p \times p$ diagonal D with positive diagonal elements $d_1 \geqslant \cdots \geqslant d_c$, and define

$$\begin{aligned} X &= X_1(X_1'X_1)^{-1/2}\Gamma', \\ \boldsymbol{\theta} &= \Gamma(X_1'X_1)^{1/2}\boldsymbol{\beta}_1. \end{aligned} \tag{3}$$

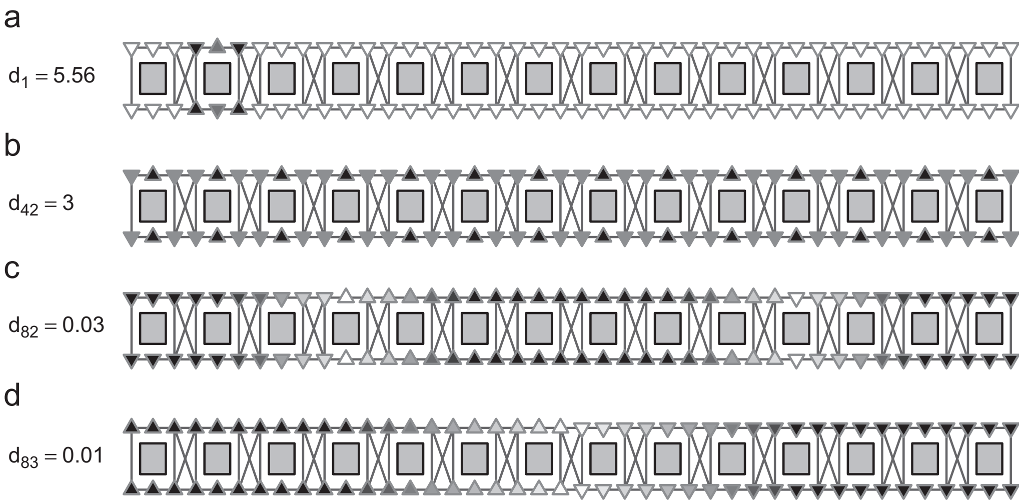

Fig. 3. Eigenvectors (columns of X) associated with large, medium, and small eigenvalues of a 14-tooth periodontal grid. Arrows pointing up (down) represent positive (negative) values; dark shades represent large magnitude and light shades small magnitudes.

This reparameterization depends only on known items. The standardized general two-variance model can then be written as

$$\mathbf{y}|\boldsymbol{\theta}, \sigma_e^2 \sim \mathrm{N}\left(X\boldsymbol{\theta}, \frac{1}{\sigma_e^2} I_N\right), \tag{4}$$

$$\boldsymbol{\theta}|\sigma_s^2 \sim \mathrm{N}\left(0, \frac{1}{\sigma_s^2} D\right), \tag{5}$$

where $(1/\sigma_e^2)I_N$ and $(1/\sigma_s^2)D$ are precisions, $X'X$ is the $p \times p$ identity, and D is known. Note that the last $G = p - c$ coordinates of $\boldsymbol{\theta}$ have prior precision zero.

At this level of generality, X's columns and $\boldsymbol{\theta}$'s coordinates are rather obscure, but with suitable displays they are often highly interpretable. For example, Fig. 3 shows four columns of X for a 14-tooth periodontal grid. In order from Fig. 3a–d, the transformed predictors measure local differences on tooth 2, the effect for a site in a gap between teeth, and the quadratic and linear trends (averaged across the sides of the jaw), respectively.

The posterior arising from this standardized model, $p(\boldsymbol{\theta}, \sigma_e^2, \sigma_s^2|\mathbf{y})$, is proportional to

$$(\sigma_e^2)^{-N/2} \exp\left(-\frac{(\mathbf{y} - X\boldsymbol{\theta})'(\mathbf{y} - X\boldsymbol{\theta})}{2\sigma_e^2}\right) (\sigma_s^2)^{-c/2} \exp\left(-\frac{\boldsymbol{\theta}' D \boldsymbol{\theta}}{2\sigma_s^2}\right) p(\sigma_e^2, \sigma_s^2), \tag{6}$$

where $p(\sigma_e^2, \sigma_s^2)$ is the prior on the variances. The posterior can be decomposed into a product of $p(\sigma_e^2, \sigma_s^2)$ and three terms,

$$\begin{aligned} p(\boldsymbol{\theta}, \sigma_e^2, \sigma_s^2|\mathbf{y}) \propto\ & (\sigma_e^2)^{-(N-p)/2} \exp\left(-\frac{\mathrm{SSE}}{2\sigma_e^2}\right) \\ & \times \prod_{i=1}^{p} \left[(\sigma_e^2[r/(r+d_i)])^{-1/2} \exp\left(-\frac{(\theta_i - [r/(r+d_i)]\hat{\theta}_i)^2}{2\sigma_e^2[r/(r+d_i)]}\right)\right] \\ & \times \prod_{i=1}^{c} \left[(\sigma_e^2[1 + r/d_i])^{-1/2} \exp\left(-\frac{\hat{\theta}_i^2}{2\sigma_e^2[1 + r/d_i]}\right)\right] p(\sigma_e^2, \sigma_s^2), \end{aligned} \tag{7}$$

where $\hat{\theta} = X'\mathbf{y}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$ ignoring the higher-level structure, $\mathrm{SSE} = (\mathbf{y} - X\hat{\theta})'(\mathbf{y} - X\hat{\theta})$ is the usual sum of squared errors, and $r = \sigma_s^2/\sigma_e^2$.

The first term in (7) involves only σ_{e}^2 and the usual residual sum of squares of linear model theory. The second term in (7) is $\boldsymbol{\theta}$'s conditional posterior, $p(\boldsymbol{\theta}|\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2, \mathbf{y})$. Conditional on $(\mathbf{y}, \sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$, the θ_i have independent normal posteriors with mean $\hat{\theta}_i r/(r+d_i)$ and variance $\sigma_{\mathrm{e}}^2 r/(r+d_i)$. As r decreases toward zero, the factor $r/(r+d_i)$ shrinks the posterior mean of θ_i towards zero and reduces its posterior variance, reducing the effective dimension of $\boldsymbol{\theta}$. Hodges and Sargent (2001) defined $\rho = \sum_{i=1}^{c} r/(r+d_i)$, a scalar between zero and p, as the degrees of freedom in the fitted model's mean structure, as implied by the smoothing constant r, with $r/(r+d_i) \in [0, 1]$ being the fraction of a degree of freedom for the contrast θ_i. This ρ is an overall measure of the complexity of the fitted model for fixed r, and grows as $\sum_{i=1}^{c} d_i$ decreases.

The final term in (7) is $\hat{\theta}$'s marginal density given the variances, $p(\hat{\theta}|\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$, a function familiar from likelihood analysis. The $G = p - c$ terms with $d_i = 0$ and thus $r/(r+d_i) = 1$ contribute to second row, but not to the third row of (7). For the c terms having $d_i > 0$, the $\hat{\theta}_i$ given $(\sigma_{\mathrm{e}}^2, r)$ (but *not* θ_i) have independent normal distributions with mean zero and variance $\sigma_{\mathrm{e}}^2(1 + r/d_i) \geqslant \sigma_{\mathrm{e}}^2$. If the θ_i do not vary, i.e., $\sigma_{\mathrm{s}}^2 = 0$ so $r = 0$, $\hat{\theta}_i$ has variance σ_{e}^2. Integrating out the unknown θ_i instead of conditioning on it inflates the variance of $\hat{\theta}_i$ by a factor $(1 + r/d_i)$.

2.1. The marginal posterior of $(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$

Integrating $\boldsymbol{\theta}$ out of the posterior $p(\boldsymbol{\theta}, \sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2|\mathbf{y})$ simply removes the second row of (7), leaving

$$p(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2|\mathbf{y}) \propto p(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)(\sigma_{\mathrm{e}}^2)^{-(N-p)/2} \exp\left(-\frac{\mathrm{SSE}}{2\sigma_{\mathrm{e}}^2}\right) \prod_{i=1}^{c} \left(\frac{d_i}{\sigma_{\mathrm{s}}^2 + d_i\sigma_{\mathrm{e}}^2}\right)^{1/2} \exp\left(-\frac{\hat{\theta}_i^2}{2} \cdot \frac{d_i}{\sigma_{\mathrm{s}}^2 + d_i\sigma_{\mathrm{e}}^2}\right). \tag{8}$$

Leaving aside $p(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$, the first row of (8) does not involve σ_{s}^2; we define these to be *free terms for* σ_{e}^2. The terms in the second row of (8) involve both σ_{e}^2 and σ_{s}^2; we define them to be *mixed terms for* σ_{e}^2 *and* σ_{s}^2. (Reich et al., 2007 used "free terms" and "mixed terms" to refer to analogous free and mixed terms conditional on $\boldsymbol{\theta}$.) Of the data set's original N observations or "degrees of freedom", $N - p$ are free terms for σ_{e}^2, c are mixed terms for σ_{e}^2 and σ_{s}^2, and $G = p - c$ have $d_i = 0$, that is, their θ_i has a flat prior and thus provides no marginal information (after integrating out θ_i) about the variance parameters. The decomposition in (8) shows that the sufficient statistic for the variances is $(\mathrm{SSE}, \hat{\theta}_1, \ldots, \hat{\theta}_c)$ and the *design* information that determines variance identification is $(N - p, d_1, \ldots, d_c)$.

The ith mixed term in (8) has the form of a gamma density with variate $d_i/(\sigma_{\mathrm{s}}^2 + d_i\sigma_{\mathrm{e}}^2)$, shape parameter $\frac{3}{2}$, rate parameter $\hat{\theta}_i/2$ and mode $1/\hat{\theta}_i^2$. Since the ith mixed term is a function of σ_{e}^2 and σ_{s}^2 only through $d_i/(\sigma_{\mathrm{s}}^2 + d_i\sigma_{\mathrm{e}}^2)$, each mixed term is constant for pairs of $(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$ that give the same value of $d_i/(\sigma_{\mathrm{s}}^2 + d_i\sigma_{\mathrm{e}}^2)$, so a mixed term cannot identify both variances. The variances are both identified—that is there is no one-dimensional function $g(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$ such that $p(\mathbf{y}|\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2) = p(\mathbf{y}|g(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2))$ for all $(\mathbf{y}, \sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$—if and only if there are both free and mixed terms or there is more than one distinct d_i.

The ith mixed term is maximized by any $(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$ on the line $\sigma_{\mathrm{s}}^2 = d_i\hat{\theta}_i^2 - d_i\sigma_{\mathrm{e}}^2$. The mixed term favors these $(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$ but cannot differentiate between any two points on the line. Therefore, we define this as an *unidentified modal line* for the ith mixed term. Unidentified modal lines can be used to graphically examine identification of σ_{e}^2 and σ_{s}^2. For example, Fig. 2b plots these unidentified lines for the periodontal CAR model of Section 1. In the $(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$ quarter-plane with σ_{e}^2 on the horizontal axis, these sets of points form the vertical line $\sigma_{\mathrm{e}}^2 = \mathrm{SSE}/(N - p)$ for free terms (although there are no free terms for this model), and for the mixed terms lines with slopes $-d_i$ and intercepts $d_i\hat{\theta}_i^2$. Note that only the intercepts, not the slopes, of the unidentified lines depend on $\hat{\theta}_i$. The location of these modal lines determines the location of the posterior.

The d_i also determine whether the ith mixed term is more informative for σ_{e}^2 or σ_{s}^2. Mixed terms with large d_i give nearly vertical lines, similar to lines arising from free terms for σ_{e}^2. These terms are more informative about σ_{e}^2 than σ_{s}^2 because any segment of an unidentified line spans a wider range of values for σ_{s}^2 than σ_{e}^2. By contrast, mixed terms with d_i near zero have nearly horizontal slopes, similar to lines arising from free terms for σ_{s}^2 (which never actually exist), so they are more informative about σ_{s}^2. Models that give free terms for σ_{e}^2 or mixed terms with large d_i, and also give some terms with small d_i, provide information about both variances and generally lead to well-identified posteriors.

2.2. *Marginal posterior of r*

The variances can also be parameterized as (σ_e^2, r) where $r = \sigma_s^2/\sigma_e^2$. If σ_e^2 is given an Inverse Gamma$(a_e/2, b_e/2)$ prior parameterized so $E(\sigma_e^2) = b_e/(a_e - 2)$, it can be integrated out of $p(\sigma_e^2, r|\mathbf{y})$ leaving the marginal posterior of r,

$$p(r|\mathbf{y}) \propto \left[\prod_{i=1}^{c} \frac{1}{\hat{\sigma}_e^2(1+r/d_i)}\right]^{1/2} \left[1 + \frac{1}{v}\sum_{i=1}^{c} \frac{\hat{\theta}_i^2}{\hat{\sigma}_e^2(1+r/d_i)}\right]^{-(c+v)/2} p(r), \tag{9}$$

where $v = N - p + a_e$ and $\hat{\sigma}_e^2 = (\mathrm{SSE} + b_e)/v$. For models with $N = p$, such as the CAR model of Section 1, $\hat{\sigma}_e^2 = b_e/a_e$ is a function only of σ_e^2's prior. The likelihood piece of (9) is the marginal density of $\hat{\boldsymbol{\theta}}$ given r and SSE, i.e., integrating out $(\boldsymbol{\theta}, \sigma_e^2)$. It has the form of a multivariate t density with c-dimensional variate $(\hat{\theta}_1, \ldots, \hat{\theta}_c)$, v degrees of freedom, location vector zero and a diagonal scale matrix with diagonal entries $\hat{\sigma}_e^2(1 + r/d_i)$.

Information about r arises, loosely speaking, by comparing the $\hat{\theta}_i^2$ to $\hat{\sigma}_e^2$. If $\hat{\theta}_i^2$ is near $\hat{\sigma}_e^2$, the estimated error variance from σ_e^2's free terms, this suggests $1 + r/d_i$ is near one and r is near zero. If $\hat{\theta}_i^2$ is considerably greater than $\hat{\sigma}_e^2$, this is evidence of variability in the data that cannot be explained by measurement error, and indicates $r \gg 0$. Conversely, each d_i controls the sensitivity of its $1 + r/d_i$ to changes in r and thus controls contrast i's contribution of information to $p(r|\mathbf{y})$. If d_i is large, $1 + r/d_i \approx 1$ for a large range of r, so the ith term of (9) is flat in r except for very large r. As d_i is reduced, $1 + r/d_i$ becomes more sensitive to changes in r and thus $\hat{\theta}_i$ becomes informative over a wider range of r. Note that this and other interpretations we give for d_i are invariant to changes in the data's scale because d_i refers to $r = \sigma_s^2/\sigma_e^2$.

A two-variance model assumes that smoothing in each orthogonal direction is controlled by the same smoothing parameter, r. Verifying this assumption by visual inspection of the data may be difficult due to the often obscure nature of the canonical directions. Recalling from (7) that $\mathrm{E}(\hat{\theta}_i^2|\sigma_e^2, r) = \sigma_e^2(1 + r/d_i)$, define $\hat{r}_i$ by solving $\hat{\theta}_i^2 = \hat{\sigma}_e^2(1 + \hat{r}_i/d_i)$, i.e.,

$$\hat{r}_i = d_i \left(\frac{\hat{\theta}_i^2}{\hat{\sigma}_e^2} - 1\right)^{+}. \tag{10}$$

This measures the data's smoothness in the ith direction. Plotting the $\hat{r}_i$ is an exploratory tool that can be used for checking that smoothing in each direction is similar. For known (σ_e^2, r), $\hat{\theta}_i/\sqrt{\sigma_e^2(1 + r/d_i)}$ have independent standard normal distributions. Although (σ_e^2, r) is not actually known, to gauge the magnitude of the $\hat{r}_i$, σ_e^2 and r could be replaced by the posterior medians, $\tilde{\sigma}_e^2$ and $\tilde{r}$, and smoothing in the ith direction could be identified as outlying if $|\hat{\theta}_i| > 3\sqrt{\tilde{\sigma}_e^2(1 + \tilde{r}/d_i)}$ or equivalently

$$\hat{r}_i = d_i \left(\frac{\hat{\theta}_i^2}{\hat{\sigma}_e^2} - 1\right)^{+} > d_i \left(\frac{9\tilde{\sigma}_e^2(1 + \tilde{r}/d_i)}{\hat{\sigma}_e^2} - 1\right)^{+}. \tag{11}$$

If an $\hat{r}_i$ exceeds this threshold, the corresponding θ_i could be given a separate smoothing parameter or smoothing in that direction could be removed altogether by setting the d_i to zero. This diagnostic is illustrated in Section 3.2 using the periodontal data of Section 1.

3. Illustrative special cases

Section 2 showed that the number of free terms and the d_i capture the *design* information that determines variance identification. The number of free terms for σ_e^2 is straightforward: it is the degrees of freedom for error in the usual linear model (ignoring the higher level structure on $\boldsymbol{\beta}_1$ or $\boldsymbol{\theta}$). However, the d_i are less obvious. This section describes model features that give large and small d_i for the one-way random effect models, the periodontal example of Section 1, and then for CAR models more generally.

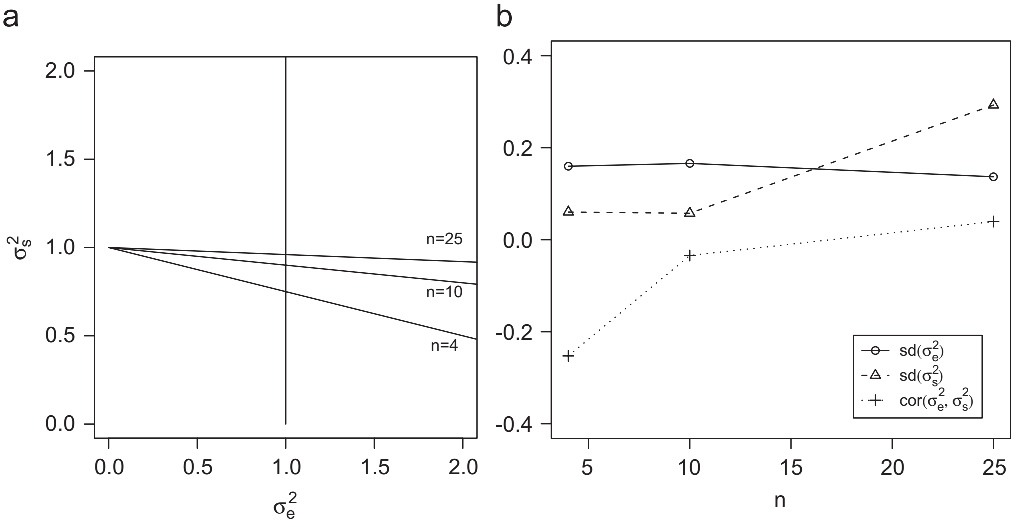

Fig. 4. Identification of the variance components in the one-way random effect model. Panel (a) gives the unidentified modal lines for various within-group samples sizes n assuming the overall sample size is $N = 100$, $\hat{\sigma}_e^2 = \text{SSE}/(N-p) = 1$ and $\hat{\sigma}_s^2 = \text{SSM}/(p-1) = 1$. Panel (b) plots the medians across 1000 data sets, generated assuming $N = 100$ and $\sigma_e^2 = \sigma_s^2 = 1$, of the posterior standard deviations of σ_e^2 and σ_s^2 and the posterior correlation between σ_e^2 and σ_s^2. (a) Unidentified model lines. (b) Medians of the posterior standard deviations of σ_e^2 and σ_s^2.

3.1. One-way random effect model

The one-way random effect model with p groups and n_i observations in group i, $n_1 \geqslant \cdots \geqslant n_p$, is

$$\begin{aligned} &y_{kj}|\mu_k, \sigma_e^2 \sim \text{N}(\mu_k, \sigma_e^2), \quad j = 1, \ldots, n_k, \\ &\mu_k|\mu, \sigma_s^2 \sim \text{N}(\mu, \sigma_s^2), \quad k = 1, \ldots, p. \end{aligned} \tag{12}$$

This is a special case of the general two-variance model obtained by setting $N = \sum_{k=1}^p n_k$, $\boldsymbol{\beta}_1' = (\mu_1, \ldots, \mu_p)$, $\boldsymbol{\beta}_2' = \mu$, $Z = \mathbf{1}_p$, $Q = I_p$ and

$$X_1 = \begin{pmatrix} \mathbf{1}_{n_1} & 0 & 0 & \ldots & 0 \\ 0 & \mathbf{1}_{n_2} & 0 & \ldots & 0 \\ & & \vdots & & \\ 0 & \ldots & & 0 & \mathbf{1}_{n_p} \end{pmatrix}.$$

Both variances are identified if and only if $p > 1$ and $n_1 > 1$. There are $N - p$ free terms for σ_e^2 and $p - 1$ mixed terms for σ_e^2 and σ_s^2.

The balanced case with $n_1 = \cdots = n_p \equiv n$ and $N = np$ has $d_1 = \cdots = d_{p-1} = 1/n$ and $\theta_1, \ldots, \theta_{p-1}$ can be any orthogonal contrasts in the μ_k that sum to zero. The term with $d_i = 0$ has $\theta_i = (1/\sqrt{p})\sum_{j=1}^p \mu_j$, which represents the overall mean and provides no information about the variances. Since each d_i equals $1/n$, the mixed terms can be combined, and the marginal distribution of the variances (8) reduces to

$$p(\sigma_e^2, \sigma_s^2|\mathbf{y}) \propto p(\sigma_e^2, \sigma_s^2)\ (\sigma_e^2)^{-(N-p)/2} \exp\left(-\frac{\text{SSE}}{2\sigma_e^2}\right) (\sigma_s^2 + \sigma_e^2/n)^{-(p-1)/2} \exp\left(-\frac{\text{SSM}}{2(\sigma_s^2 + \sigma_e^2/n)}\right), \tag{13}$$

where $\text{SSM} = \sum_{i-1}^p (\bar{y}_i - \bar{y})^2$. The marginal posterior of the variances can be factored into two pieces: the free terms, which provide information for only σ_e^2 via SSE, and the mixed terms, which provide information for both σ_s^2 and σ_e^2 via SSM.

Fig. 4a plots the unidentified modal lines for different n assuming $N = 100$, $\hat{\sigma}_e^2 = \text{SSE}/(N-p) = 1$, and $\hat{\sigma}_s^2 = \text{SSM}/(p-1) = 1$. The vertical line $\sigma_e^2 = 1$ is the free term's unidentified modal line and is the same for all n. The mixed terms' unidentifed modal line is $\sigma_s^2 = \hat{\sigma}_s^2 - \sigma_e^2/n$. The slope is $-1/n$, so as n increases the mixed terms more closely resemble free terms for σ_s^2 and there are essentially a full $p-1$ degrees of freedom for estimating σ_s^2.

To show how the unidentified modal lines relate to the posterior of the variance components, we generate 1000 data sets assuming $N=100$ and $\sigma_e^2=\sigma_s^2=1$ for various n. Fig. 4b plots the median (over the 1000 data sets) of the posterior standard deviations of σ_e^2 and σ_s^2 and the correlation between σ_e^2 and σ_s^2. For $n=4$ ($N=100$, so $p=25$ groups) the slope of the mixed term is $-\frac{1}{4}$ and there is generally a negative posterior correlation between σ_e^2 and σ_s^2. For large n, the slope of the mixed terms is near zero and the correlation decreases. Also, for $n=25$ the posterior standard deviation of σ_s^2 is large because there are only $p=4$ groups and $p-1=3$ mixed terms involving σ_s^2.

As (9) shows, the data provide information about r by comparing the $\hat{\theta}_i^2$ to $\hat{\sigma}_e^2$. For the balanced design, each $1+r/d_i = 1+rn$ and $\sum_{i=1}^{p-1}\hat{\theta}_i^2 = n(p-1)\hat{\sigma}_s^2$. The marginal posterior of r thus reduces to

$$p(r|\mathbf{y}) \propto \left(\frac{1}{1+rn}\right)^{(p-1)/2}\left(1+\frac{n(p-1)\hat{\sigma}_s^2}{v(1+rn)\hat{\sigma}_e^2}\right)^{-(p-1+v)/2} p(r). \tag{14}$$

The $\hat{r}_i$ statistics in (10) measure the data's smoothness in the ith direction. Because the balanced model has only one unique d_i, the terms can be combined to give one $\hat{r}_i$,

$$\hat{r} = \frac{1}{n}\left(\frac{\sum_{i=1}^{p-1}\hat{\theta}_i^2/(p-1)}{\hat{\sigma}_e^2} - 1\right)^+ = \left(\frac{\hat{\sigma}_s^2}{\hat{\sigma}_e^2} - \frac{1}{n}\right)^+, \tag{15}$$

which is also r's posterior mode (ignoring its prior).

3.2. Analysis of periodontal data

As mentioned earlier, Fig. 2a shows a contour plot of the joint log posterior of (σ_e^2, σ_s^2) modeling one subject's AL data with a CAR prior on true AL and a flat prior on (σ_e^2, σ_s^2). This model has no free terms for σ_e^2 and $N-G$ mixed terms for σ_e^2 and σ_s^2, where G is the number of islands in the spatial structure; the $G=2$ coordinates of $\boldsymbol{\theta}$ corresponding to the average AL of each jaw have prior precision zero and do not contribute to the marginal distribution of (σ_e^2, σ_s^2). Both variances are fairly well identified, but σ_e^2 is better identified than σ_s^2 in the sense that the posterior mass is more concentrated. Fig. 2b plots the unidentified modal lines for the periodontal data. While there are no free terms for either variance, several terms have large d_i and thus nearly vertical unidentified lines, resembling free terms for σ_e^2. Some lines have small d_i and thus nearly horizontal unidentified lines, resembling free terms for σ_s^2.

In the periodontal grid, terms with large d_i are more informative for σ_e^2 because they measure high-frequency trends in the data which are implicitly presumed to be "noise". For a "half-mouth", a 14-tooth periodontal grid, the largest distinct eigenvalue, $d_1=5.56$, has multiplicity 12. Fig. 3a shows one of the 12 corresponding eigenvectors (i.e., contrasts in $\boldsymbol{\beta}_1$, the true AL). It is non-zero only at locations around the second tooth from the left and contrasts the two sides of the tooth according to lag-one differences. The eigenvectors having the analogous pattern at each of the 12 interior teeth are an orthogonal basis for the span of eigenvectors with $d_1=5.56$.

Several terms have d_i near zero. These resemble free terms for σ_s^2 and are more informative about σ_s^2 because they measure low-frequency trends in the data. Fig. 3c and d show the eigenvectors associated with the two smallest distinct eigenvalues of a 14-tooth periodontal grid. These two d_i are associated with the linear and quadratic trends, averaged over the inner and outer sides (in Fig. 3, the upper and lower sides) of the jaw. The eigenvector in Fig. 3b is associated with a medium-sized d_i and will be discussed later.

This data set also gives a nice example of the sometimes non-intuitive way that data determine smoothing. Fig. 1a plots the raw data and posterior mean of $\boldsymbol{\beta}_1$ for the AL data set. The posterior mean of $\boldsymbol{\beta}_1$ is smoothed considerably; the posterior median of $r=\sigma_s^2/\sigma_e^2$ is 0.20. The posterior median of the degrees of freedom ρ (Hodges and Sargent, 2001) is 23.5 of a maximum possible 168 coordinates of $\boldsymbol{\beta}_1$, and the effective model size, p_D, is 26.0 (Spiegelhalter et al., 2002). Fig. 5 plots the $\hat{r}_i$ statistics (10) for this model fit. Most of the $\hat{r}_i$ are near zero, suggesting that these contrasts are smooth. A few other terms have $\hat{r}_i$ above 10: the two terms with $d_i=3$ have $\hat{r}_i>100$ and appear to be outliers. The solid line in Fig. 5 represents the threshold (11) evaluated at the posterior median of (σ_e^2, σ_s^2) from the fit of Fig. 1a, (1.25, 0.25). Only the two terms with $d_i=3$ exceed the threshold.

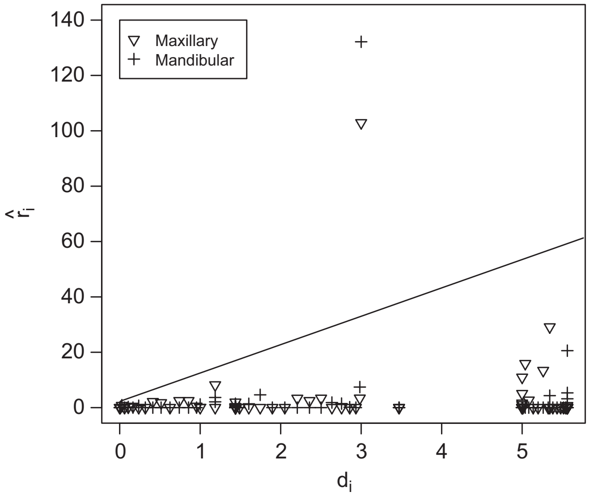

Fig. 5. Plot of the $\hat{r}_i$ for the attachment loss data set. The solid line is the threshold (11) for declaring an $\hat{r}_i$ to be outlying.

These outlying $\hat{r}_i$ suggest a spatially varying covariate that is not included in the model. Each jaw has one outlying $\hat{r}_i$ and both are associated with the contrast in site means shown in Fig. 3b. The corresponding $\hat{\theta}_i$ are proportional to the difference between the average observed AL at direct sites (those that do not border an interproximal region) and the average observed AL at non-direct sites (those that border the interproximal region) for each jaw. This pattern has also been found in more conventional mixed model analyses (e.g., Roberts, 1999).

Fig. 1b shows the posterior mean of $\boldsymbol{\beta}_1$ from modeling the two terms with outlying $\hat{r}_i$ as fixed effects, by setting their d_i to zero in (5). Since the prior in this model is less restrictive, it is not surprising that the measures of effective model size increase: the posterior median of ρ increases from 23.5 to 45.6 and the posterior median of p_D from 26.0 to 46.8. Despite the increased complexity, the *DIC* statistic (Spiegelhalter et al., 2002) prefers this model 139.4–231.6.

It is somewhat surprising that the posterior median of r actually increases from 0.25 to 0.63 after removing the two terms with the largest $\hat{r}_i$. It seems clear that this happens because removing the $\hat{r}_i$ outliers reduced σ_e^2 and increased σ_s^2: the posterior median of (σ_e^2, σ_s^2) changed from (1.25,0.25) to (0.63,0.41). Section 3.3 explores this further.

3.3. General CAR model

The CAR model is a special case of (2) with $X_1 = I_N$ and $Z\boldsymbol{\theta}_2 = 0$, i.e., Z is null. Each observation in this model has its own mean parameter, that is, $p = N$, leaving no free terms for σ_e^2, but both variances are identified unless each island is the same size and its neighborhood structure is saturated (Appendix A.1). However, identification is poor for some spatial structures.

In the periodontal grid of Section 3.2, the orthogonal contrasts in Fig. 3 having large and small d_i measured high and low frequency trends, respectively, providing mainly information about σ_e^2 and σ_s^2, respectively. Similarly interpretable contrasts are present in relatively unstructured spatial lattices, like the map of Minnesota's counties in Fig. 6. Although we do not analyze data using the Minnesota map, inspecting the eigenstructure of the Minnesota county adjacency matrix illustrates the types of contrasts generally associated with large and small d_i. As Fig. 6a shows, the eigenvector associated with the largest d_i measures the difference between the counties with largest number of neighbors and their neighbors (high-frequency contrast), while Fig. 6b shows that the smallest d_i corresponds to Minnesota's north/south gradient (low-frequency contrast).

Because there are no free terms for the CAR model, the d_i alone determine whether a grid favors identification of σ_e^2 or σ_s^2. High-frequency contrasts are always present in a spatial lattice, so there will always be some terms with large d_i, which favor σ_e^2. Appendix A.2 shows that for any CAR grid, d_1, the largest d_i, is in the interval $[m_{\max} + 1, 2 * m_{\max}]$, where m_i is the number of sites neighboring the ith site and $m_{\max} = \max(m_i)$, so $d_1 \geqslant 3$ except for spatial structures consisting only of "islands" with two connected regions.

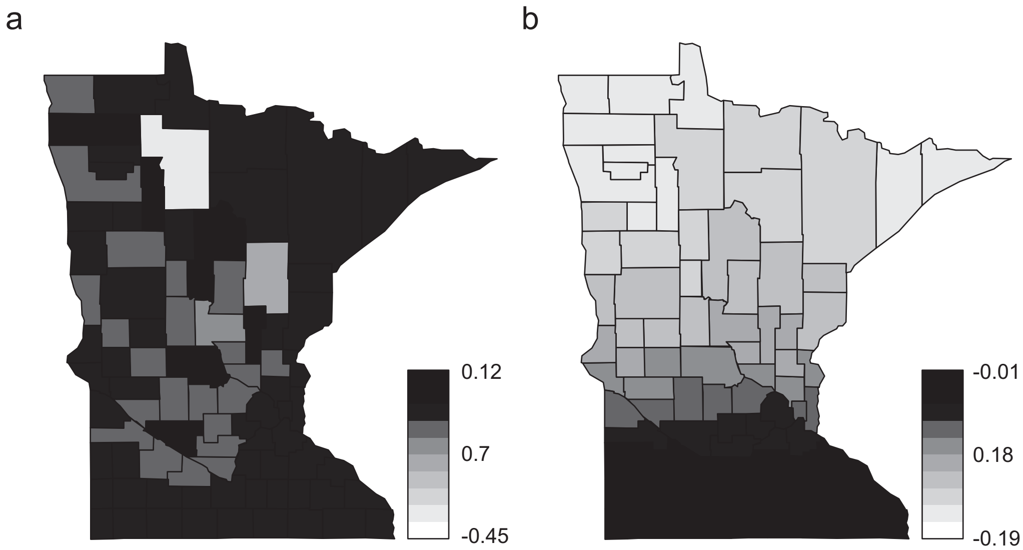

Fig. 6. Eigenvectors associated with the largest and smallest eigenvalues of the counties of Minnesota adjacency matrix. Panel (a) shows the contrast in county means β_{1i} associated with the largest d_i, while panel (b) shows the contrast associated with the smallest d_i.

For a large class of spatial maps, a CAR grid has more $d_i > 1$ (good for σ_{e}^2) than $d_i < 1$ (good for σ_{s}^2). This can be shown using Horn's theorem (Fulton, 2000), which states that for any positive semi-definite matrices A and B,

$$\lambda_i(A+B) \leqslant \lambda_i(A) + \lambda_1(B), \tag{16}$$

where $\lambda_i(A)$ is the ith largest eigenvalue of A. A simple path through a spatial lattice is a path that uses each location exactly once. If Q is the adjacency matrix for a spatial grid with a simple path, let Q_{SP} be the adjacency matrix of the neighbor pairs that make up a simple path. Then the adjacency matrix Q can be written as the sum of Q_{SP} (using the edges that form the simple path) and a residual adjacency matrix, Q_{rest} (using the edges not in the simple path). Horn's theorem then gives

$$\lambda_i(Q) \geqslant \lambda_i(Q_{\mathrm{SP}}), \tag{17}$$

i.e., a lower bound for d_i is the ith largest eigenvalue of Q_{SP}. Although no proof is available, inspection of each Q_{SP} grid with $N < 1000$ shows that each of these grids has more d_i greater than one than d_i less than one, so the same is true for any grid of under 1000 locations containing a simple path. This suggests σ_{e}^2 will generally be better identified than σ_{s}^2 for CAR models.

In the class of size N connected grids, $\sum_{i=1}^{c} d_i = \mathrm{trace}(Q) = \sum_{j=1}^{n} m_j$ is maximized by the saturated grid where each pair of locations are neighbors, and it is minimized by the simple path grid. Generally speaking, it seems that more densely connected graphs have larger d_i, favoring identification of σ_{e}^2 at the expense of σ_{s}^2.

In Section 3.2, the posterior of r changed in a non-intuitive way when the contrasts θ_i with the two largest $\hat{r}_i$ were changed to fixed effects. We conjecture that this can be explained in general in the following way. As discussed in Section 2.1, each mixed term has the form of a gamma density with variate $d_i/(\sigma_{\mathrm{s}}^2 + d_i\sigma_{\mathrm{e}}^2)$, shape parameter $\frac{3}{2}$, rate parameter $\hat{\theta}_i^2/2$ and mode $1/\hat{\theta}_i^2$. Fig. 2b shows the unidentified modal line of each mixed term for the periodontal example, i.e., the set of points $(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$ such that $d_i/(\sigma_{\mathrm{s}}^2 + d_i\sigma_{\mathrm{e}}^2) = 1/\hat{\theta}_i^2$. The two outlying terms have large $\hat{\theta}_i^2$ and thus large intercepts, suggesting that both variances are large, but since $d_i = 3 > 1$, it seems that these $\hat{\theta}_i^2$ have more effect on σ_{e}^2 than σ_{s}^2; the posterior median of $(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$ changed from (1.25, 0.25) to (0.63, 0.41) and the posterior median of r increased from 0.25 to 0.63 after removing these two terms.

Removing terms with outlying $\hat{r}_i$ has the opposite effect on r when the outlying terms have small d_i. Consider the hypothetical CAR model with no free terms for σ_{e}^2, 50 mixed terms for σ_{e}^2 and σ_{s}^2 and $d = \{0.1, 0.2, \ldots, 5\}$. Fig. 7a plots the 50 unidentified modal lines and the log marginal posterior of $(\sigma_{\mathrm{e}}^2, \sigma_{\mathrm{s}}^2)$ assuming $\hat{r} = \{10, 1.02, 1.04, \ldots, 2\}$

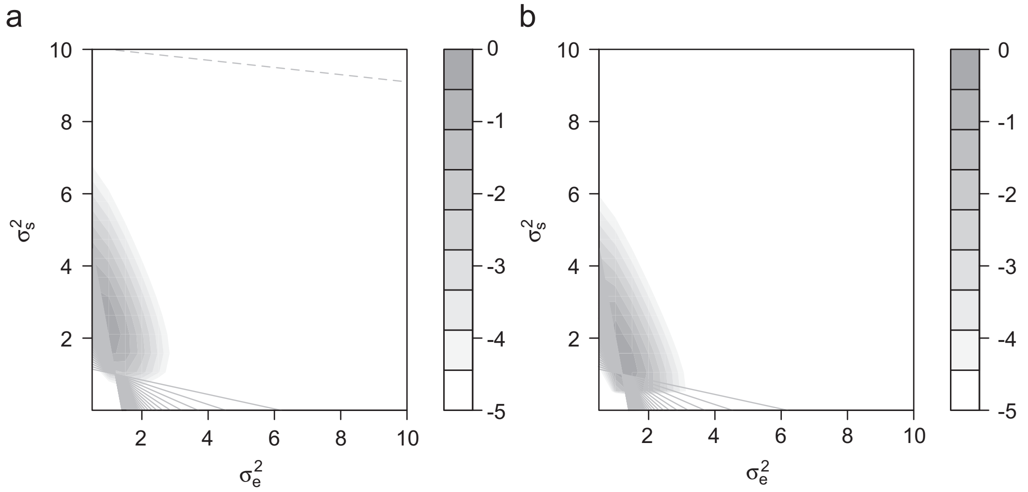

Fig. 7. Joint log posterior of (σ_e^2, σ_s^2) (scaled to be zero at the posterior mode) assuming flat priors for the variances with and without the term with outlying $\hat{r}_i$. (a) $d = \{0.1, 0.2, \ldots, 5\}$, $\hat{r} = \{10, 1.02, 1.04, \ldots, 2\}$. (b) $d = \{0.2, 0.3, \ldots, 5\}$, $\hat{r} = \{1.02, 1.04, \ldots, 2\}$.

was observed, so the outlying $\hat{r}_i$ has the smallest d_i, 0.1. The modal lines all intersect near (1,1) except the dashed line representing the term with $d_1 = 0.1$ and $\hat{r}_1 = 10$. The presence of this term pulls the center of the posterior up and to the left, away from (1,1). Removing the term with $d_i = 0.1$ (Fig. 7b) affects σ_s^2 more than σ_e^2 and shifts the posterior down and to the right; the posterior median of (σ_e^2, σ_s^2) changes from (0.82, 2.25) to (1.23, 1.02) and the posterior median of r decreases from 2.86 to 0.83.

4. Discussion

This paper explored identification of variance parameters in two-variance linear models. In Section 2.1, the marginal posterior of (σ_e^2, σ_s^2) is decomposed (Eq. (8)) into the product of $N - p$ free terms for σ_e^2 and $c = p - G$ mixed terms for σ_e^2 and σ_s^2. This decomposition illuminates the sufficient statistic relevant to variances, (SSE, $\hat{\theta}_1, \ldots, \hat{\theta}_c$), and the design information that determines how well the variances are identified, $(N - p, d_1, \ldots, d_c)$. The statistics $\hat{r}_i$ (10) were used in Section 2.2 to summarize each mixed term's contribution to the information about r and a method for identifying outlying $\hat{r}_i$ was proposed (11). Section 3 used these ideas to explain variance identification in the periodontal CAR example as well as other models.

The methods proposed here also apply to more elaborate random effect models, such as the multiple membership model (Browne et al., 2001; McCaffrey et al., 2004) which allows observations to be "members" of more than one classification level (group). Also, the main ideas presented here apply to the standard mixed effect model (Laird and Ware, 1982)

$$\begin{aligned} \mathbf{y}|b_1, b_2, \sigma_e^2 &\sim \mathrm{N}\left(Z_1 b_1 + Z_2 b_2, \frac{1}{\sigma_e^2} I_N\right), \\ b_2|\sigma_s^2 &\sim \mathrm{N}\left(0, \frac{1}{\sigma_s^2} Q\right), \end{aligned} \tag{18}$$

where b_1 has a flat prior. This mixed effect model differs from (2) because there are fixed effects in the likelihood. However, the present paper proceeded by integrating out $\mathbf{y}$'s fixed effects; after integrating out b_1, the mixed effect model (18) can be written as a special case of (2), i.e.,

$$\begin{aligned} \mathbf{y}_2|b_2, \sigma_e^2 &\sim \mathrm{N}\left(X_2 b_2, \frac{1}{\sigma_e^2} I_N\right), \\ b_2|\sigma_s^2 &\sim \mathrm{N}\left(0, \frac{1}{\sigma_s^2} Q\right), \end{aligned} \tag{19}$$

where $\mathbf{y}_2 = \Gamma_z \mathbf{y}$, $X_2 = \Gamma_z Z_2$, and Γ_z is the matrix with columns equal to the eigenvectors of $I - Z_1(Z_1'Z_1)^{-1}Z_1'$ having non-zero eigenvalues.

Our data analysis has been done from a Bayesian perspective; we now show connections to other areas of statistics.

4.1. Shrinkage: d_i controls the bias vs. variance tradeoff

In the standardized model (5), $\boldsymbol{\theta}$ has prior precision $(1/\sigma_\text{s}^2)D$. The d_i represent the strength of the prior for θ_i relative to the prior on the other θ_i; many models have some $d_i = 0$. A given prior constraint structure (Z, Q) risks bias in contrasts θ_i with large d_i (much smoothing) for the sake of reducing posterior variance in those directions. However, the prior precisions are tuned by σ_s^2, so even models with large d_i permit unsmooth fits if the data indicate a large σ_s^2.

4.2. The d_i in constrained regression

The maximum likelihood estimate of $\boldsymbol{\theta}$ (ignoring $\boldsymbol{\theta}$'s prior), $\hat{\theta} = X'\mathbf{y}$, solves the normal equation

$$\frac{1}{\sigma_\text{e}^2} X'(\mathbf{y} - X\boldsymbol{\theta}) = 0. \tag{20}$$

To achieve better prediction, penalized regression techniques have been proposed that choose $\hat{\theta}$ to maximize (20) subject to constraints on $\boldsymbol{\theta}$. For example, ridge and Lasso regression impose the constraints $\sum_{i=1}^p \theta_i^2 < t$ and $\sum_{i=1}^p |\theta_i| < t$, respectively (Tibshirani, 1996). Although the resulting estimates are biased, they are attractive because they have smaller variance and often smaller mean squared error (MSE) compared to the maximum likelihood estimate.

Implicit in the standardized general two-variance model, specifically the posterior (7), is the weighted ridge regression constraint $\sum_{i=1}^p d_i\theta_i^2 < t$. To see this, note that for fixed r, the solution to (20) under this constraint solves

$$\frac{1}{\sigma_\text{e}^2} X' \left(\mathbf{y} - X \left(I + \frac{1}{r} D \right) \boldsymbol{\theta} \right) = 0. \tag{21}$$

The solution $\theta_i = \hat{\theta}_i r/(r + d_i)$ is the mean of the conditional posterior of $\boldsymbol{\theta}_i$ in the second row of (7). This weighted ridge estimate smooths each θ_i towards zero by its own factor $r/(r + d_i)$. Terms with large d_i are smoothed more than terms with small d_i and the degree of smoothing is controlled by $r = \sigma_\text{s}^2/\sigma_\text{e}^2$. In ridge regression and related methods, r is fixed to optimize a criterion that penalizes model complexity, while in a Bayesian analysis r is just another unknown parameter. This suggests a way to apply the Lasso and related methods to any two-variance model, e.g., to scatter-plot or lattice smoothers, by imposing the constraint $\sum_{i=1}^p d_i|\theta_i| < t$, which is equivalent to giving the θ_i independent double exponential priors and using the posterior mode as the estimator.

Acknowledgment

The authors thank Dr. Bradley Carlin for his helpful comments on the manuscript.

Appendix A

A.1. Proof that σ_e^2 and σ_s^2 are unidentified in the CAR model if and only if each island is the same size and saturated

If each island has n locations and is saturated, each positive eigenvalue of Q is n and $p(y|\sigma_\text{e}^2, \sigma_\text{s}^2)$ can be written as a function of $\sigma_\text{s}^2 + n\sigma_\text{e}^2$, so the variance parameters are not individually identified.

Assume the grid is connected and that σ_e^2 and σ_s^2 are unidentified, i.e., all the d_i equal a common value, say d. Using the facts that $\text{trace}(Q) = \sum_{j=1}^n m_j = \sum_{i=1}^n d_i$ and $\sum_{j=1}^n (m_j^2 + m_j) = \text{trace}(QQ) = \text{trace}(DD) = \sum_{i=1}^{n-1} d_i^2$, where m_j is the number of regions neighboring the jth site, one can show that $n \cdot var(m_j) = \sum_{j=1}^n m_j^2 - (1/n)(\sum_{j=1}^n m_j)^2 > 0$ implies $d > n$. This implies $\sum_{j=1}^n m_j = \sum_{i=1}^{n-1} d_i > n(n-1)$, which is a contradiction because each of the n m_j are at most $n-1$. So, if the d_i are equal, the m_j are equal. The solution of the $\sum_{j=1}^n m_j = \sum_{i=1}^{n-1} d_i$ and $\sum_{j=1}^n m_j^2 + m_j = \sum_{i=1}^{n-1} d_i^2$ assuming common m_j and d_i are $d = n$ and $m = n - 1$, i.e., the grid is saturated.

If there are multiple islands and all the d_i equal a common value, we have shown above that each island must be saturated. If the kth island has $n_k > 1$ locations, the d_i corresponding to that island will be n_k. Therefore, if all the d_i are equal, all the islands are saturated and the same size.

A.2. *Proof that* $m_{\max} + 1 \leqslant d_1 \leqslant 2m_{\max}$ *for any spatial grid*

Say the jth site has $m_{\max} = \max(m_j)$ neighbors. If z is the N-vector with $z_j = 1$ and $z_k = -1/m_{\max}$ if $k \sim j$, 0 otherwise, then $z'Qz/z'z = m_{\max} + 1$. Since $d_1 = \max_y\{y'Qy/y'y\}$, $d_1 \geqslant m_{\max} + 1$. Also, $d_i = \max_y\{y'Qy/y'y\} \leqslant 2m_{\max}$ because $y'Qy/y'y > 2m_{\max}$ implies $\sum_{j=1}^{N} m_j y_j^2 > \sum_{j=1}^{N} m_{\max} y_j^2$, a contradiction.

References

Besag, J., York, J.C., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics (with discussion). Ann. Inst. Statist. Math. 43, 1–59.

Browne, W., Goldstein, H., Rasbash, J., 2001. Multiple membership multiple classification (MMMC) models. Statist. Model. 1, 103–124.

Fulton, W., 2000. Eigenvalues, invariant factors, highest weights, and Shubert calculus. Bull. Amer. Math. Soc. 37, 209–249.

Gelman, A., 2005. Prior distributions for variance parameters in hierarchical models. Bayesian Anal. 1, 1–19.

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

Hodges, J.S., Carlin, B.P., Fan, Q., 2003. On the precision of the conditionally autoregressive prior in spatial models. Biometrics 59, 317–322.

Laird, N., Ware, J., 1982. Random effects models for longitudinal data. Biometrics 38, 963–974.

Lindley, D.V., Smith, A.F.M., 1972. Bayes estimates for the linear model. J. Roy. Statist. Soc. Ser. B 34, 1–41.

McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T.A., Hamilton, L., 2004. Models for value-added modeling of teacher effects. J. Behavioral Ed. Statist. 29, 67–101.

Reich, B.J., Hodges, J.S., Carlin, B.P., 2007. Spatial analyses of periodontal data using conditionally autoregressive priors having two types of neighbor relations. J. Amer. Statist. Assoc. 102, 44–55.

Roberts, T., 1999. Examining mean structure, covariance structure and correlation of interproximal sites in a large periodontal data set. Master of Science Plan B Paper, Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN.

Shievitz, P., 1997. The effect of a non-steroidal anti-inflammatory drug on periodontal clinical parameters after scaling. M.S. Thesis, School of Dentistry, University of Minnesota, Minneapolis, MN.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit (with discussion and rejoinder). J. Roy. Statist. Soc. Ser. B 64, 583–639.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B 58, 267–288.