

# Identification of the Variance Components in the General Two-Variance Linear Model

BY BRIAN J. REICH AND JAMES S. HODGES <sup>1</sup>

*Division of Biostatistics, School of Public Health, University of Minnesota,  
2221 University Ave SE, Suite 200, Minneapolis, Minnesota 55414, U.S.A.*

brianr@biostat.umn.edu

hodges@ccbr.umn.edu

Correspondence Author: James S. Hodges

telephone: (612) 626-9626

fax: (612) 626-9054

December 14, 2004

---

<sup>1</sup>The second author was supported by the University of Minnesota School of Dentistry. The authors thank Dr. Bradley Carlin for his helpful comments on the manuscript.

# Identification of the Variance Components in the General Two-Variance Linear Model

## Abstract

Bayesians frequently employ two-stage hierarchical models consisting of 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), one-way random effects, and multiple membership 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.

*Key Words:* Conditional autoregressive prior; hierarchical models; identification; mixed linear model; variance components.

# 1 Introduction

Advances in computing allows Bayesians to fit 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. 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 Figure 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 Figure 1a plot 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 to 1 mm. Figure 1a shows a severe case of periodontal disease, so measurement error with a 1 mm standard deviation is substantial.

Reich et al. (2004) 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  $\beta_{1i}$ ,  $i = 1, 2, \dots, N$  (here location  $i$ ’s true AL). Let  $y_i$  be the region  $i$ ’s observable; assume  $y_i | \beta_{1i}, \sigma_e^2$  is normal with mean  $\beta_{1i}$  and variance  $\sigma_e^2$ , independent across  $i$ . Spatial dependence is introduced through the prior (or model)

on  $\beta_1 = (\beta_{11}, \dots, \beta_{1N})'$ . The CAR model with  $L_2$  norm (also called a Gaussian Markov random field) for  $\beta_1$  has improper density

$$p(\beta_1 | \sigma_s^2) \propto (\sigma_s^2)^{-(N-G)/2} \exp\left(-\frac{1}{2\sigma_s^2} \beta_1' Q \beta_1\right), \quad (1)$$

where  $\sigma_s^2$  controls the smoothing induced by this prior, smaller values smoothing more than larger;  $G$  is the number of “islands” (disconnected groups of regions) in the spatial structure (Hodges et al., 2003); and  $Q$  is  $N \times N$  with non-diagonal entries  $q_{ij} = -1$  if regions  $i$  and  $j$  are neighbors and 0 otherwise, and diagonal entries  $q_{ii}$  equal to the number of region  $i$ ’s neighbors. This is a multivariate normal kernel, specified by its precision matrix  $\frac{1}{\sigma_s^2} Q$  instead of the usual covariance matrix.

Figure 1a plots the posterior mean of  $\beta_1$  (solid lines) for the attachment loss data described above. For this fit, both  $\sigma_e^2$  and  $\sigma_s^2$  have Inverse Gamma(0.01,0.01) priors and 30,000 samples were drawn using Gibbs sampling. The posterior distribution of  $\beta_1$  is well-identified; the  $\beta_{1i}$  have posterior standard deviations between 0.40 and 0.59 and their posterior means are smoothed considerably. The variances are also well-identified. Figure 2a is a contour plot of the log marginal posterior of  $(\sigma_e^2, \sigma_s^2)$ , with a flat prior on  $(\sigma_e^2, \sigma_s^2)$  to emphasize the data’s contribution (the shaded lines are explained later). However, this model has  $N$  observations and  $N + 2$  unknowns ( $\{\beta_{1i}\}, \sigma_e^2, \sigma_s^2$ ), 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:  $\sigma_e^2$ , which describes measurement error and  $\sigma_s^2$ , which controls smoothing. Section 2 also gives a useful decomposition of the posterior distribution and derives the marginal posteriors of  $(\sigma_e^2, \sigma_s^2)$  and  $r = \sigma_s^2/\sigma_e^2$ . 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  $(\sigma_e^2, \sigma_s^2)$ . Section 3 applies the theory of Section 2 to the periodontal example and then explores identification for several common two-variance models including CAR, one-way random effects and multiple membership models. Section 4 concludes by connecting this theory to constrained regression methods such as the Lasso and to variance components estimation.

## 2 The general two-variance model

The general two-variance model has the form

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

where  $\frac{1}{\sigma_e^2} I_N$  and  $\frac{1}{\sigma_s^2} Q$  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  $\boldsymbol{\beta}_2$ ,  $\sigma_e^2$  and  $\sigma_s^2$ . For this paper, we use a flat prior for  $\boldsymbol{\beta}_2$  and inverse gamma priors for  $\sigma_e^2$  and  $\sigma_s^2$ . (This is, of course, much like the model introduced in Lindley and Smith 1972)

This paper focuses on the marginal posterior of the variances  $(\sigma_e^2, \sigma_s^2)$  and of the smoothing parameter  $r = \sigma_s^2 / \sigma_e^2$ . The next paragraph 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  $(\sigma_e^2, \sigma_s^2)$ .

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  $\frac{1}{\sigma_s^2} (Q - QZ'(Z'QZ)^{-1}Z'Q)$ . Now reparameterize  $X_1 \boldsymbol{\beta}_1$  to give an orthogonal design matrix and diagonal prior precision matrix. Let  $\Phi = (X_1' X_1)^{-\frac{1}{2}} (Q - QZ'(Z'QZ)^{-1}Z'Q) (X_1' X_1)^{-\frac{1}{2}}$  have spectral decomposition  $\Gamma' D \Gamma$  for a  $p \times p$

orthogonal  $\Gamma$  and rank  $c$   $p \times p$  diagonal  $D$  with positive diagonal elements  $d_1 \geq \dots \geq d_c$ , and define

$$X = X_1(X_1'X_1)^{-\frac{1}{2}}\Gamma' \quad (3)$$

$$\boldsymbol{\theta} = \Gamma(X_1'X_1)^{\frac{1}{2}}\boldsymbol{\beta}_1.$$

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

$$\mathbf{y} \sim N(X\boldsymbol{\theta}, \frac{1}{\sigma_e^2}I_N) \quad (4)$$

$$\boldsymbol{\theta} \sim N(0, \frac{1}{\sigma_s^2}D), \quad (5)$$

where  $\frac{1}{\sigma_e^2}I_N$  and  $\frac{1}{\sigma_s^2}D$  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 the examples in Section 3 show that with suitable displays they are often highly interpretable.

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

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

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,

$$p(\boldsymbol{\theta}, \sigma_e^2, \sigma_s^2 | \mathbf{y}) \propto (\sigma_e^2)^{-(N-p)/2} \exp\left(-\frac{SSE}{2\sigma_e^2}\right) \quad (7)$$

$$\times \prod_{i=1}^p \left[ (\sigma_e^2 \gamma_i)^{-1/2} \exp\left(-\frac{(\theta_i - \gamma_i \hat{\theta}_i)^2}{2\sigma_e^2 \gamma_i}\right) \right] \quad (8)$$

$$\times \prod_{i=1}^c \left[ (\sigma_e^2 \delta_i)^{-1/2} \exp\left(-\frac{\hat{\theta}_i^2}{2\sigma_e^2 \delta_i}\right) \right] p(\sigma_e^2, \sigma_s^2), \quad (9)$$

where  $\hat{\boldsymbol{\theta}} = X'\mathbf{y}$  is the maximum likelihood estimator of  $\boldsymbol{\theta}$  ignoring the higher-level structure,  $SSE = (\mathbf{y} - X\hat{\boldsymbol{\theta}})'(\mathbf{y} - X\hat{\boldsymbol{\theta}})$  is the usual sum of squared errors,  $r = \sigma_s^2/\sigma_e^2$ ,  $\gamma_i = \sigma_s^2/(\sigma_s^2 + d_i\sigma_e^2) = r/(r + d_i)$ , and  $\delta_i = (\sigma_s^2 + d_i\sigma_e^2)/(d_i\sigma_e^2) = (1 - \gamma_i)^{-1} = 1 + r/d_i$ .

Equation (7) involves only  $\sigma_e^2$  and the usual residual sum of squares of linear model theory. Equation (8) is  $\boldsymbol{\theta}$ 's conditional posterior,  $p(\boldsymbol{\theta}|\sigma_e^2, \sigma_s^2, \mathbf{y})$ . Conditional on  $(\mathbf{y}, \sigma_e^2, \sigma_s^2)$ , the  $\theta_i$  have independent normal posteriors with mean  $\gamma_i \hat{\theta}_i$  and variance  $\sigma_e^2 \gamma_i$ . As  $r$  decreases toward zero, the factor  $\gamma_i = r/(r + d_i)$  shrinks the posterior mean of  $\theta_i$  towards zero and reduces its posterior variance, reducing the effective dimension of  $\boldsymbol{\theta}$ . Hodges and Sargent (2001) defined  $\rho = \sum_{i=1}^c \gamma_i$ , 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  $\gamma_i \in [0, 1]$  being the fraction of a degree of freedom for the contrast  $\theta_i$ . This  $\rho$  is an overall measure of the complexity of the fitted model for fixed  $r$ , and grows as  $d_i$  decreases.

Equation (9) is  $\hat{\theta}$ 's marginal density given the variances,  $p(\hat{\theta}|\sigma_e^2, \sigma_s^2)$ , a function familiar from likelihood analysis. The  $G = p - c$  terms with  $d_i = 0$  and thus  $\gamma_i = 1$  contribute to (8) but not to (9). For the  $c$  terms having  $d_i > 0$ , the  $\hat{\theta}_i$  given  $(\sigma_e^2, r)$  (but *not*  $\theta_i$ ) have independent normal distributions with mean zero and variance  $\sigma_e^2(1 + r/d_i) = \sigma_e^2 \delta_i \geq \sigma_e^2$ . If the  $\theta_i$  do not vary, i.e.,  $\sigma_s^2 = 0$  so  $r = 0$ ,  $\hat{\theta}_i$  has variance  $\sigma_e^2$ . Integrating out the unknown  $\theta_i$  instead of conditioning on it inflates the variance of  $\hat{\theta}_i$  by a factor  $\delta_i$ .

## 2.1 The marginal posterior of $(\sigma_e^2, \sigma_s^2)$

Integrating  $\boldsymbol{\theta}$  out of the posterior  $p(\boldsymbol{\theta}, \sigma_e^2, \sigma_s^2|\mathbf{y})$  simply removes (8), leaving

$$p(\sigma_e^2, \sigma_s^2|\mathbf{y}) \propto p(\sigma_e^2, \sigma_s^2) (\sigma_e^2)^{-(N-p)/2} \exp\left(-\frac{SSE}{2\sigma_e^2}\right) \quad (10)$$

$$\times \prod_{i=1}^c \left(\frac{d_i}{\sigma_s^2 + d_i \sigma_e^2}\right)^{1/2} \exp\left(-\frac{\hat{\theta}_i^2}{2} \cdot \frac{d_i}{\sigma_s^2 + d_i \sigma_e^2}\right). \quad (11)$$

Leaving aside  $p(\sigma_e^2, \sigma_s^2)$ , the  $N - p$  terms in (10) do not involve  $\sigma_s^2$ ; we define these to be *free terms* for  $\sigma_e^2$ . The terms in (11) involve both  $\sigma_e^2$  and  $\sigma_s^2$  and we define them to be *mixed terms* for  $\sigma_e^2$  and  $\sigma_s^2$ . (Reich et al. (2004) 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  $\sigma_e^2$ ,  $c$  are mixed terms for  $\sigma_e^2$  and  $\sigma_s^2$  and  $G = p - c$  have  $d_i = 0$ , that is, their  $\theta_i$  has a flat prior and thus provides no marginal information about the variance parameters. The decomposition in (10), (11) shows that the sufficient statistic for the variances is  $(SSE, \hat{\theta}_1, \dots, \hat{\theta}_c)$  and the *design* information that determines variance identification is  $(N - p, d_1, \dots, d_c)$ .

The  $i^{th}$  mixed term in (11) has the form of a gamma density with variate  $d_i/(\sigma_s^2 + d_i\sigma_e^2)$ , shape parameter  $3/2$ , rate parameter  $\hat{\theta}_i/2$  and mode  $1/\hat{\theta}_i^2$ . Since the  $i^{th}$  mixed term is a function of  $\sigma_e^2$  and  $\sigma_s^2$  only through  $d_i/(\sigma_s^2 + d_i\sigma_e^2)$ , each mixed term is constant for pairs of  $(\sigma_e^2, \sigma_s^2)$  that give the same value of  $d_i/(\sigma_s^2 + d_i\sigma_e^2)$ , so a mixed term cannot identify both variances. The variances are both identified if and only if there are both free and mixed terms or there is more than one distinct  $d_i$ .

Define the set of  $(\sigma_e^2, \sigma_s^2)$  that satisfy  $d_i/(\sigma_s^2 + d_i\sigma_e^2) = C_i$  for some constant  $C_i$  as an *unidentified line*. Unidentified lines can be used to graphically examine identification of  $\sigma_e^2$  and  $\sigma_s^2$ . For example, Figure 2b plots these unidentified lines for the periodontal CAR model of Section 1, specifically, for the free terms the line is  $\sigma_e^2 = SSE/(N - p)$ , while for the  $i^{th}$  mixed term  $d_i/(\sigma_s^2 + d_i\sigma_e^2) = 1/\hat{\theta}_i^2$ , the posterior mode for term  $i$ . Note that two unidentified lines from mixed terms are discrepant; we examine these further in Section 3.1. The shaded lines in Figure 2a are the unidentified lines with  $C_i$  determined by the overall posterior median of  $(\sigma_e^2, \sigma_s^2)$ . In the  $(\sigma_e^2, \sigma_s^2)$  quarter-plane with  $\sigma_e^2$  on the horizontal axis, these sets of points form the vertical line  $\sigma_e^2 = SSE/(N - p)$  for free terms (although there are no free terms for this model) and lines with slopes  $-d_i$  and intercepts  $d_i\hat{\theta}_i^2$  for mixed terms; 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 variances are identified if and only if the lines corresponding to the free and mixed terms do not all have the same slope.

The  $d_i$  also determine whether the  $i^{th}$  mixed term is more informative for  $\sigma_e^2$  or  $\sigma_s^2$ . Mixed

terms with large  $d_i$  give nearly vertical lines, similar to lines arising from free terms for  $\sigma_e^2$ . These terms are more informative about  $\sigma_e^2$  than  $\sigma_s^2$  because any segment of an unidentified line spans a wider range of values for  $\sigma_s^2$  than  $\sigma_e^2$ . By contrast, mixed terms with  $d_i$  near zero have nearly horizontal slopes, similar to lines arising from free terms for  $\sigma_s^2$  (which never actually exist), so they are more informative about  $\sigma_s^2$ . Models that give free terms for  $\sigma_e^2$  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  $(\sigma_e^2, r)$  where  $r = \sigma_s^2/\sigma_e^2$ . If  $\sigma_e^2$  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  $r$ ,

$$p(r|\mathbf{y}) \propto \left[ \prod_{i=1}^c \frac{1}{\hat{\sigma}_e^2 \delta_i} \right]^{\frac{1}{2}} \left[ 1 + \frac{1}{\nu} \sum_{i=1}^c \frac{\hat{\theta}_i^2}{\hat{\sigma}_e^2 \delta_i} \right]^{-\frac{c+\nu}{2}} p(r) \quad (12)$$

where  $\nu = N - p + a_e$  and  $\hat{\sigma}_e^2 = (SSE + b_e)/\nu$  and recalling  $\delta_i = 1 + r/d_i$ . 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  $\sigma_e^2$ 's prior. The likelihood piece of (12) 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  $(\hat{\theta}_1, \dots, \hat{\theta}_c)$ ,  $\nu$  degrees of freedom, location vector zero and a diagonal scale matrix with diagonal entries  $\hat{\sigma}_e^2 \delta_i$ .

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  $\sigma_e^2$ 's free terms, this suggests  $\delta_i$  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  $r > 0$ . Conversely, each  $d_i$  controls the sensitivity of its  $\delta_i$  to changes in  $r$  and thus controls contrast  $i$ 's contribution of information to  $p(r|\mathbf{y})$ . If  $d_i$  is large,  $\delta_i \approx 1$  for a large range of  $r$ , so the  $i^{th}$  term of (12) is flat in  $r$  except for very large  $r$ . As  $d_i$  is

reduced  $\delta_i$  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 (9) that  $E(\hat{\theta}_i^2 | \sigma_e^2, r) = \sigma_e^2 \delta_i = \sigma_e^2(1 + r/d_i)$ , define  $\hat{r}_i$  by solving  $\hat{\theta}_i^2 = \hat{\sigma}_e^2(1 + \hat{r}_i/d_i)$ , i.e.,

$$\hat{r}_i = d_i \left( \frac{\hat{\theta}_i^2}{\hat{\sigma}_e^2} - 1 \right)^+ . \quad (13)$$

This measures the data's smoothness in the  $i^{\text{th}}$  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  $(\sigma_e^2, r)$ ,  $\hat{\theta}_i/\sqrt{\sigma_e^2 \delta_i} = \hat{\theta}_i/\sqrt{\sigma_e^2(1 + r/d_i)}$  have independent standard normal distributions. Although  $(\sigma_e^2, r)$  is not actually known, to gauge the magnitude of the  $\hat{r}_i$ ,  $\sigma_e^2$  and  $r$  could be replaced by the posterior medians,  $\tilde{\sigma}_e^2$  and  $\tilde{r}$ , and smoothing in the  $i^{\text{th}}$  direction could be identified as outlying if  $|\hat{\theta}_i| > 3\sqrt{\tilde{\sigma}_e^2(1 + \tilde{r}/d_i)}$  or equivalently

$$\hat{r}_i = d_i \left( \frac{\hat{\theta}_i^2}{\hat{\sigma}_e^2} - 1 \right)^+ > d_i \left( \frac{9\tilde{\sigma}_e^2(1 + \tilde{r}/d_i)}{\hat{\sigma}_e^2} - 1 \right)^+ . \quad (14)$$

If an  $\hat{r}_i$  exceeds this threshold, the corresponding  $\theta_i$  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 3.1 using the periodontal data of Section 1.

### 3 The periodontal data and other 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  $\sigma_e^2$  is straightforward: it is the degrees of freedom for error in the usual linear model (ignoring the higher level structure on  $\beta_1$ ).

However, the  $d_i$  are less obvious. This section describes model features that give large and small  $d_i$ , first for the periodontal example of Section 1 and then more generally for CAR models (Section 3.2), one-way random effects models (Sections 3.3) and multiple membership models (Section 3.4).

### 3.1 Analysis of periodontal data

As mentioned earlier, Figure 2a shows a contour plot of the joint log posterior of  $(\sigma_e^2, \sigma_s^2)$  modelling one subject’s attachment loss data with a CAR prior on true AL and a flat prior on  $(\sigma_e^2, \sigma_s^2)$  to emphasize the data’s contribution to the posterior. This model has no free terms for  $\sigma_e^2$  and  $N - G$  mixed terms for  $\sigma_e^2$  and  $\sigma_s^2$  where  $G$  is the number of islands in the spatial structure; the  $G$  coordinates of  $\theta$  corresponding to the average AL of each island have prior precision zero and do not contribute to the marginal distribution of  $(\sigma_e^2, \sigma_s^2)$ . The  $c = N - G$  superimposed gray lines in Figure 2a are the unidentified lines arising from mixed terms. The  $i^{th}$  line is the set of points for which  $\sigma_s^2/d_i + \sigma_e^2 = 0.25/d_i + 1.25$ , where 0.25 and 1.25 are the posterior medians of  $\sigma_s^2$  and  $\sigma_e^2$  respectively when the variances have InvGamma(0.01,0.01) priors.

Both variances are fairly well-identified, but  $\sigma_e^2$  is better identified than  $\sigma_s^2$  in the sense that the posterior mass is more concentrated. 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  $\sigma_e^2$ . Some lines have small  $d_i$  and thus nearly horizontal unidentified lines, resembling free terms for  $\sigma_s^2$ .

In the periodontal grid, terms with large  $d_i$  are more informative for  $\sigma_e^2$  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. Figure 3a shows one of the 12 corresponding eigenvectors (i.e., contrasts in  $\beta_1$ , 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

twelve 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  $\sigma_s^2$  and are more informative about  $\sigma_s^2$  because they measure low-frequency trends in the data. Figures 3c and 3d 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 Figure 3, the upper and lower sides) of the jaw. The eigenvector in Figure 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. Figure 1a plots the raw data and posterior mean of  $\beta_1$  for the attachment loss data set. The posterior mean of  $\beta_1$  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  $\rho$  (Hodges and Sargent 2001) is 23.5 of a maximum possible 168 coordinates of  $\beta_1$ , and the effective model size,  $p_D$ , is 26.0 (Spiegelhalter et al., 2002). Figure 4 plots the  $\hat{r}_i$  statistics (13) for this model fit. Since this model has no free terms for  $\sigma_e^2$ ,  $\hat{\sigma}_e^2 = b_e/a_e = 1$ . Most of the  $\hat{r}_i$  are near zero, suggesting that these contrasts are smooth. Although a few other terms have  $\hat{r}_i$  above ten, the two terms with  $d_i = 3$  have  $\hat{r}_i > 100$  and appear to be outliers. The solid line in Figure 4 represents the threshold (14) evaluated at the posterior median of  $(\sigma_e^2, \sigma_s^2)$  from the fit of Figure 1a, (1.25, 0.25). Only the two terms with  $d_i = 3$  exceed the threshold.

These outlying  $\hat{r}_i$  suggest a spatially varying covariate (fixed effect) 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 Figure 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).

Figure 1b shows the posterior mean of  $\beta_1$  from modelling 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  $\rho$  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 to 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  $\sigma_e^2$  and increased  $\sigma_s^2$ : the posterior median of  $(\sigma_e^2, \sigma_s^2)$  changed from (1.25, 0.25) to (0.63, 0.41). Section 3.2 explores this further.

### 3.2 General CAR model

The CAR model is a special case of (2) with  $X_1 = I_N$  and  $Z\theta_2 = 0$ , i.e.,  $Z$  is null. Each observation in this model has its own mean parameter, that is,  $p = N$ , leaving no free terms for  $\sigma_e^2$ , but both variances are identified unless each island is the same size and its neighborhood structure is saturated (Appendix A.1). That is, there is no one-dimensional function  $g(\sigma_e^2, \sigma_s^2)$  such that  $p(\mathbf{y}|\sigma_e^2, \sigma_s^2) = p(\mathbf{y}|g(\sigma_e^2, \sigma_s^2))$  for all  $(\mathbf{y}, \sigma_e^2, \sigma_s^2)$ . However, identification is poor for some spatial structures.

In the periodontal grid of Section 3.1, the orthogonal contrasts in Figure 3 having large and small  $d_i$  measured high and low frequency trends, respectively, providing mainly information about  $\sigma_e^2$  and  $\sigma_s^2$ , respectively. Similarly interpretable contrasts are present in relatively unstructured spatial lattices, like the map of Minnesota's counties in Figure 5. As Figure 5a shows, the eigenvalue associated with the largest  $d_i$  measures the difference between the counties with largest number of neighbors and their neighbors (high-frequency contrast), while Figure 5b 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  $\sigma_e^2$  or  $\sigma_s^2$ . High-frequency contrasts are always present in a spatial lattice, so there will always be some terms with large  $d_i$ , which favor  $\sigma_e^2$ . 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  $i^{th}$  site and  $m_{max} = \max(m_i)$ , so  $d_1 \geq 3$  except for spatial structures consisting only of “islands” with 2 connected regions.

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

$$\lambda_i(A + B) \leq \lambda_i(A) + \lambda_1(B), \quad (15)$$

where  $\lambda_i(A)$  is the  $i^{th}$  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

$$\lambda_i(Q) \geq \lambda_i(Q_{SP}) \quad (16)$$

i.e., a lower bound for  $d_i$  is the  $i^{th}$  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 less than one, so the same is true for any grid of under 1000 locations containing a simple path. This suggests  $\sigma_e^2$  will generally be better identified than  $\sigma_s^2$  for CAR models.

In the class of size  $N$  connected grids,  $\sum_{i=1}^c d_i = \text{trace}(Q) = \sum_{j=1}^n m_j$  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  $\sigma_e^2$  at the expense of  $\sigma_s^2$ .

In Section 3.1, the posterior of  $r$  changed in a non-intuitive way when the contrasts  $\theta_i$  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  $d_i/(\sigma_s^2 + d_i\sigma_e^2)$ , shape parameter  $3/2$ , rate parameter  $\hat{\theta}_i^2/2$  and mode  $1/\hat{\theta}_i^2$ . Figure 2b shows the unidentified modal line of each mixed term for the periodontal example, i.e., the set of points  $(\sigma_e^2, \sigma_s^2)$  such that  $d_i/(\sigma_s^2 + d_i\sigma_e^2) = 1/\hat{\theta}_i^2$ . The two dashed lines represent the unidentified modal lines of the two terms with outlying  $\hat{r}_i$ . These 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  $\sigma_e^2$  than  $\sigma_s^2$ ; the posterior median of  $(\sigma_e^2, \sigma_s^2)$  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  $\sigma_e^2$ , 50 mixed terms for  $\sigma_e^2$  and  $\sigma_s^2$  and  $d = \{0.1, 0.2, \dots, 5\}$ . Figure 6a plots the 50 unidentified modal lines and the log marginal posterior of  $(\sigma_e^2, \sigma_s^2)$  assuming  $\hat{r} = \{10, 1.02, 1.04, \dots, 2\}$  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  $d_i = 0.1 < 1$  (Figure 6b) affects  $\sigma_s^2$  more than  $\sigma_e^2$  and shifts the posterior down and to the right; the posterior median of  $(\sigma_e^2, \sigma_s^2)$  changes from (0.82, 2.25) to (1.23, 1.02) and the posterior median of  $r$  decreases from 2.86 to 0.83.

### 3.3 One-way random effects model

The one-way random effects model with  $p$  groups and  $n_i$  observations in group  $i$ ,  $n_1 \geq \dots \geq n_p$ , is

$$\begin{aligned} y_{kj} | \mu_k, \sigma_e^2 &\sim N(\mu_k, \sigma_e^2), \quad j = 1, \dots, n_k \\ \mu_k | \mu, \sigma_s^2 &\sim N(\mu, \sigma_s^2), \quad k = 1, \dots, p \end{aligned} \quad (17)$$

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

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

Both variances are identified if and only if  $p > 1$  and  $n_1 > 1$ . There are  $N - p$  free terms for  $\sigma_e^2$  and  $p - 1$  mixed terms for  $\sigma_e^2$  and  $\sigma_s^2$ . The free terms give information about  $\sigma_e^2$  through the within-group sample variances, while the mixed terms give information about both variances through the  $\hat{\theta}_i$ , which are linear combinations of the group means,  $\bar{y}_k$ ,  $k = 1, \dots, p$ .

The  $d_i$  are the eigenvalues of  $\Phi = \text{diag}(n_k)^{-1/2} (I_p - \frac{1}{p} J_p) \text{diag}(n_k)^{-1/2}$ , where  $J_p$  is the  $p \times p$  matrix of ones. Wang and Zhang (1992) show that for any positive semi-definite  $K \times K$  matrices  $A$  and  $B$ ,

$$\lambda_i(A) \lambda_K(B) \leq \lambda_i(AB), \quad (18)$$

where  $\lambda_i(A)$  is the  $i^{\text{th}}$  largest eigenvalue of  $A$ . Applying this theorem gives  $d_i \leq \lambda_i(I_p - \frac{1}{p} J_p) / n_p$ , with equality occurring in the balanced design with  $n_1 = \dots = n_p = n$ . Since  $I_p - \frac{1}{p} J_p$  has  $p - 1$  eigenvalues of one and one eigenvalue of zero, each  $d_i \leq 1/n_p$ , where  $n_p$  is the smallest  $n_k$ . The slope of the unidentified line arising from each mixed term (Section 2.1) is less than one, i.e., each mixed term is at least as informative for  $\sigma_s^2$  as  $\sigma_e^2$ . As  $n_p$  increases, each  $d_i$  goes to zero and the mixed terms resemble free terms for  $\sigma_s^2$ . In this case,  $\theta$  is estimated with high precision and there

are essentially a full  $p - 1$  degrees of freedom for estimating  $\sigma_s^2$ .

For the balanced case,  $d_1 = \dots = d_{p-1} = 1/n$  and  $\theta_1, \dots, \theta_{p-1}$  can be any orthogonal contrasts in the  $\mu_k$  that are also orthogonal to the  $p$ -vector of ones. The term with  $d_i = 0$  has  $\theta_i = \frac{1}{\sqrt{p}} \sum_{j=1}^p \mu_j$ . Since each  $d_i$  equals  $1/n$ , the mixed terms can be combined, and the marginal distribution of the variances, (10), (11), reduces to

$$p(\sigma_e^2, \sigma_s^2 | \mathbf{y}) \propto p(\sigma_e^2, \sigma_s^2) (\sigma_e^2)^{-(N-p)/2} \exp\left(-\frac{SSE}{2\sigma_e^2}\right) \quad (19)$$

$$\times (\lambda)^{-(p-1)/2} \exp\left(-\frac{SSM}{2\lambda}\right), \quad (20)$$

where  $\lambda = \sigma_s^2 + \sigma_e^2/n$  and  $SSM = \sum_{i=1}^p (\bar{y}_i - \bar{y})^2$ . The marginal posterior of the variances can be factored into two pieces: the free terms, which are a function of only  $\sigma_e^2$ , and the mixed terms, which are a function of only  $\lambda$ . This suggests a reparameterization from  $(\sigma_e^2, \sigma_s^2)$  to  $(\sigma_e^2, \lambda)$ . Apart from the constraint that  $\lambda > \sigma_e^2/n$ , the likelihoods for these two parameters factor.

As shown by (12), the data provide information about  $r$  by comparing the  $\hat{\theta}_i^2$  to  $\hat{\sigma}_e^2$ . For the balanced design, each  $\delta_i$  is  $1 + nr$  and  $\sum_{i=1}^{p-1} \hat{\theta}_i^2 = n \sum_{i=1}^p (\bar{y}_i - \bar{y})^2 = n(p-1)\hat{\sigma}_s^2$  for any orthogonal basis of  $\Phi$ . The marginal posterior of  $r$  thus reduces to

$$p(r | \mathbf{y}) \propto \left(\frac{1}{1 + rn}\right)^{(p-1)/2} \left(1 + \frac{\sum_{i=1}^{p-1} \hat{\theta}_i^2}{\nu(1 + nr)\hat{\sigma}_e^2}\right)^{-(p-1+\nu)/2} p(r). \quad (21)$$

The  $\hat{r}_i$  statistics in (13) measure the data's smoothness in the  $i^{\text{th}}$  direction. Because  $\theta_1, \dots, \theta_{p-1}$  can be any orthogonal basis for the space of contrasts, there is a lot of freedom in defining the  $\{\hat{r}_i\}$ .

Because the balanced model has only one unique  $d_i$ , the terms can be combined to give one  $\hat{r}_i$ ,

$$\hat{r} = \frac{1}{n} \left(\frac{\sum_{i=1}^{p-1} \hat{\theta}_i^2 / (p-1)}{\hat{\sigma}_e^2} - 1\right)^+ = \left(\frac{\hat{\sigma}_s^2}{\hat{\sigma}_e^2} - \frac{1}{n}\right)^+, \quad (22)$$

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

### 3.4 Multiple membership model (MMM)

This extension of the one-way random effects model, common in education research (Browne et al. 2001, McCaffrey et al. 2004), allows observations to be “members” of more than one classification level (group). The membership weights of the  $j^{th}$  observation are given by  $w_{jk} \geq 0$ , with  $\sum_{k=1}^p w_{jk} \leq 1$  for  $j = 1, \dots, N$  and the  $w_{jk}$  assumed known. The model is

$$\begin{aligned} y_j | \mu_k, \sigma_e^2 &\sim N\left(\sum_{k=1}^p w_{jk} \mu_k, \sigma_e^2\right), \quad j = 1, \dots, N \\ \mu_k | \mu, \sigma_s^2 &\sim N(\mu, \sigma_s^2), \quad k = 1, \dots, p. \end{aligned} \quad (23)$$

This model becomes a special case of the general two-variance model by setting  $\theta'_1 = (\mu_1, \dots, \mu_p)$ ,  $\theta'_2 = \mu$ ,  $Q = I_p$ ,  $Z = \mathbf{1}_p$  and the  $(j, k)$  element of  $X_1$  to  $w_{jk}$ .

To see how membership in multiple groups affects variance identification, consider the simple MMM where observations are equally weighted between at most two groups. Let  $n_{kk} = \phi n$  be the number of observations that are members of group  $k$  only and  $n_{kl} = 2(1 - \phi)n/(p - 1)$  be the number of observations that are equally weighted between groups  $k$  and  $l$ , for a total of  $N = np$  observations. If  $\phi = 1$ , this reduces to the balanced one-way random effects model.

This model, like the balanced one-way random effects model,  $N - p$  free terms for  $\sigma_e^2$  and  $p - 1$  mixed terms, each with the same  $d_i$ . Each  $d_i = \omega/n$ , where  $\omega = (p - 1)/(\frac{\phi+1}{2}p - 1) \geq 1$ . Since the counts of free and mixed terms are the same as a comparably-sized balanced one-way random effects model, but the  $d_i$  are larger, this simple MMM gives less information about  $\sigma_s^2$  and thus  $r$ , with  $\omega$  serving as an index of the information “cost” of multiple membership.

## 4 Discussion

This paper explored identification of variance parameters in two-variance linear models. In Section 2.1, the marginal posterior of  $(\sigma_e^2, \sigma_s^2)$  is decomposed (10, 11) into the product of  $N - p$  free terms

for  $\sigma_e^2$  and  $c = p - G$  mixed terms for  $\sigma_e^2$  and  $\sigma_s^2$ . This decomposition illuminates the sufficient statistic relevant to variances,  $(SSE, \hat{\theta}_1, \dots, \hat{\theta}_c)$ , and the design information that determines how well the variances are identified,  $(N - p, d_1, \dots, d_c)$ . The statistics  $\hat{r}_i$  (13) were used in Section 2.2 to summarize each mixed term's contribution to the identification of  $r$  and a method for identifying outlying  $\hat{r}_i$  was proposed (14). Section 3 used these ideas to explain variance identification in the periodontal CAR example as well as other models. 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  $\frac{1}{\sigma_s^2}D$ . The  $d_i$  represent the strength of the prior for  $\theta_i$  relative to the prior on the other  $\theta_i$ ; many models have some  $d_i = 0$ . A given prior constraint structure  $(Z, Q)$  risks bias in contrasts  $\theta_i$  with large  $d_i$  (much smoothing) for the sake of reducing posterior variance in those directions. However the prior precisions are tuned by  $\sigma_s^2$ , so even models with large  $d_i$  permit unsmooth fits if the data indicate a large  $\sigma_s^2$ .

#### 4.2 The $d_i$ in constrained regression

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

$$\frac{1}{\sigma_e^2}X'(\mathbf{y} - X\boldsymbol{\theta}) = 0. \quad (24)$$

To achieve better prediction, penalized regression techniques have been proposed that choose  $\hat{\boldsymbol{\theta}}$  to maximize (24) 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), (8) and

(9), 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 (24) under this constraint solves

$$\frac{1}{\sigma_e^2} X'(\mathbf{y} - X(I + \frac{1}{r}D)\boldsymbol{\theta}) = 0. \quad (25)$$

The solution  $\theta_i = \gamma_i \hat{\theta}_i$ , recalling that  $\gamma_i = r/(r + d_i)$ , is the mean of the conditional posterior of  $\theta_i$  in (8). This weighted ridge estimate smooths each  $\theta_i$  towards zero by its own factor  $\gamma_i$ . Terms with large  $d_i$  are smoothed more than terms with small  $d_i$  and the degree of smoothing is controlled by  $r = \sigma_s^2/\sigma_e^2$ . 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  $\theta_i$  independent double exponential priors and using the posterior mode as the estimator.

## References

- Besag J, York JC, Mollié A (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.
- Browne W, Goldstein H, Rasbash J (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling*. **1**:103-124.
- Fulton W (2000). Eigenvalues, invariant factors, highest weights, and Schubert calculus. *Bulletin of the American Mathematical Society*. **37**:209:249.
- Hodges JS, Carlin BP, Fan Q (2003). On the precision of the conditionally autoregressive prior in spatial models. *Biometrics*, **59**, 317-322.
- Hodges JS, Sargent DJ (2001). Counting degrees of freedom in hierarchical and other richly parameterized models. *Biometrika*, **88**, 367-379.
- Lindley DV, Smith AFM (1972) Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, **34**, 1-41.
- McCaffrey DF, Lockwood JR, Koretz D, Louis TA, Hamilton L (2004). Models for value-added modeling of teacher effects. To appear in *J. Behavioral and Educational Statistics*.
- Reich BJ, Hodges JS, and Carlin BP (2004). Spatial analyses of periodontal data using conditionally autoregressive priors having two types of neighbor relations. Research Report 2004-004, Division of Biostatistics, University of Minnesota, 2004. Submitted to J. Amer. Statist. Assoc.

- 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. MS Thesis, School of Dentistry, University of Minnesota, Minneapolis, MN.
- Spiegelhalter DJ, Best NG, Carlin BP, 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.
- Wang B, Zhang F (1992). Some inequalities for the eigenvalues of the product of positive semidefinite hermitian matrices. *Linear Algebra and its Application* **160**, 113-118.

## Appendix

### A.1 Proof that $\sigma_e^2$ and $\sigma_s^2$ 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  $p(y|\sigma_e^2, \sigma_s^2)$  can be written as a function of  $\sigma_s^2 + n\sigma_e^2$ , so the variance parameters are not individually identified.

Assume the grid is connected and that  $\sigma_e^2$  and  $\sigma_s^2$  are unidentified, i.e., all the  $d_i$  equal a common value, say  $d$ . Using the facts that  $\text{trace}(Q) = \sum_{i=1}^n m_i = \sum_{i=1}^{n-1} d_i$  and  $\sum_{i=1}^n m_i^2 + m_i = \text{trace}(QQ) = \text{trace}(DD) = \sum_{i=1}^{n-1} d_i^2$ , where  $m_i$  is the number of regions neighboring the  $i^{\text{th}}$  site, one can show that  $n \cdot \text{var}(m_i) = \sum_{i=1}^n m_i^2 - \frac{1}{n}(\sum_{i=1}^n m_i)^2 > 0$  implies  $d > n$ . This implies  $\sum_{i=1}^n m_i = \sum_{i=1}^{n-1} d_i > n(n-1)$ , which is a contradiction because each of the  $n$   $m_i$  are at most  $n-1$ . So, if the  $d_i$  are equal, the  $m_i$  are equal. The solution of the equations  $\sum_{i=1}^n m_i = \sum_{i=1}^{n-1} d_i$  and  $\sum_{i=1}^n m_i^2 + m_i = \sum_{i=1}^{n-1} d_i^2$  assuming common  $m_i$  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  $i^{\text{th}}$  island has  $n_i > 1$  locations, the  $d_i$  corresponding to that island will be  $n_i$ . Therefore, if all the  $d_i$  are equal, all the islands are saturated and the same size.

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

Say the  $j^{th}$  site has  $m_{max} = \max(m_i)$  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  $d_1 = \max_y \{y'Qy/y'y\}$ ,  $d_1 \geq m_{max} + 1$ . Also,  $d_i = \max_y \{y'Qy/y'y\} \leq 2m_{max}$  because  $y'Qy/y'y > 2m_{max}$  implies  $\sum_{i=1}^N m_i y_i^2 > \sum_{i=1}^N m_{max} y_i^2$ , a contradiction.



Figure 3: Eigenvectors 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.

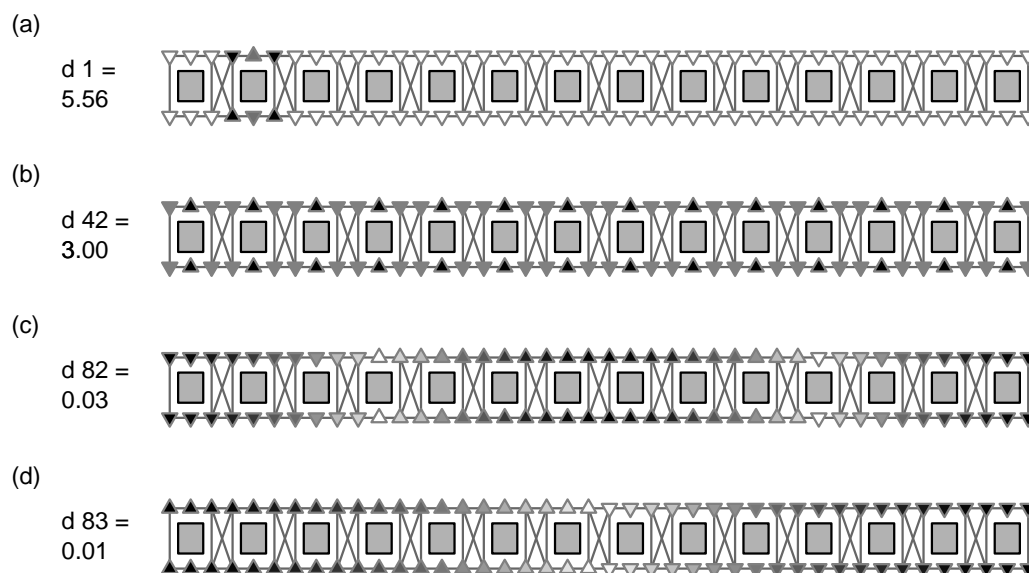


Figure 4: Plot of the  $\hat{r}_i$  for the attachment-loss data set. The solid line is the threshold (14) for declaring an  $\hat{r}_i$  outlying.

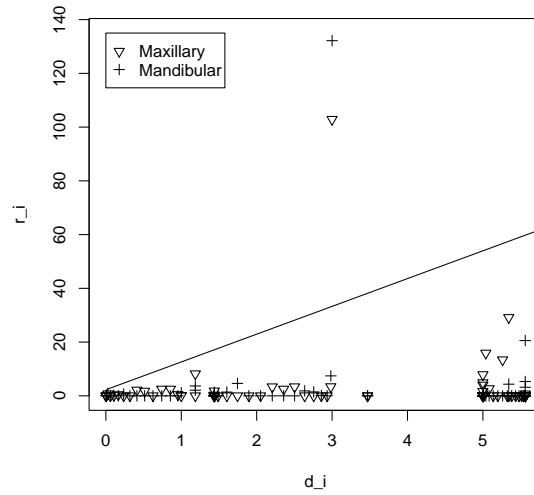


Figure 5: Eigenvectors associated with the largest and smallest eigenvalues of the counties of Minnesota adjacency matrix. Panel (a) shows the contrast in county means  $\beta_{1i}$  associated with the largest  $d_i$ , while panel (b) shows the contrast associated with the smallest  $d_i$ .

(a)

(b)

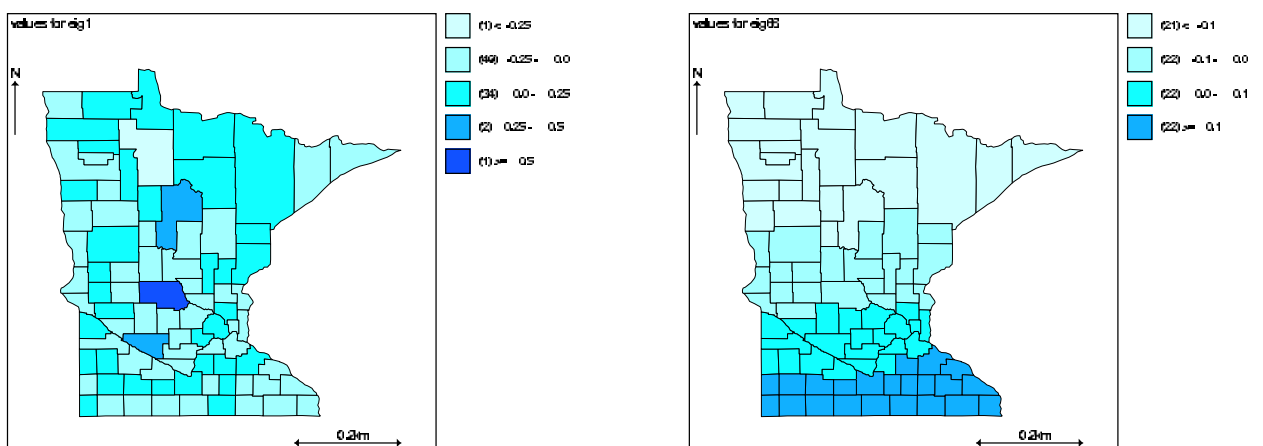
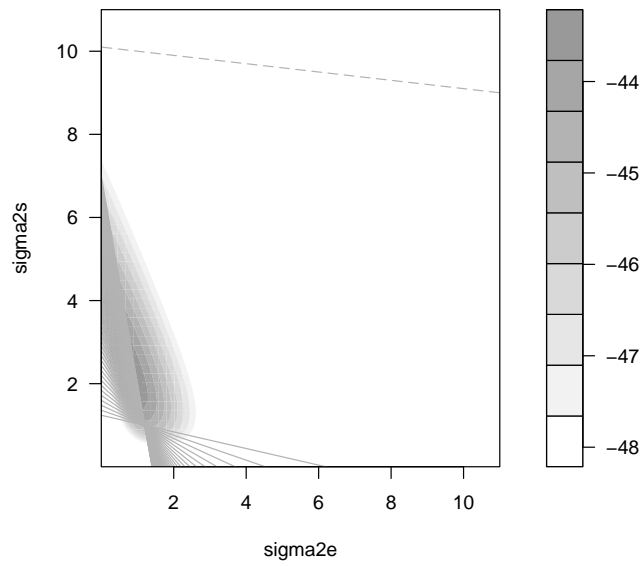


Figure 6: Joint log posterior of  $(\sigma_e^2, \sigma_s^2)$  assuming flat priors for the variances with and without term with outlying  $\hat{r}_i$ .

(a)  $d = \{0.1, 0.2, \dots, 5\}$ ,  $\hat{r} = \{10, 1.02, 1.04, \dots, 2\}$



(b)  $d = \{0.2, 0.3, \dots, 5\}$ ,  $\hat{r} = \{1.02, 1.04, \dots, 2\}$

