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

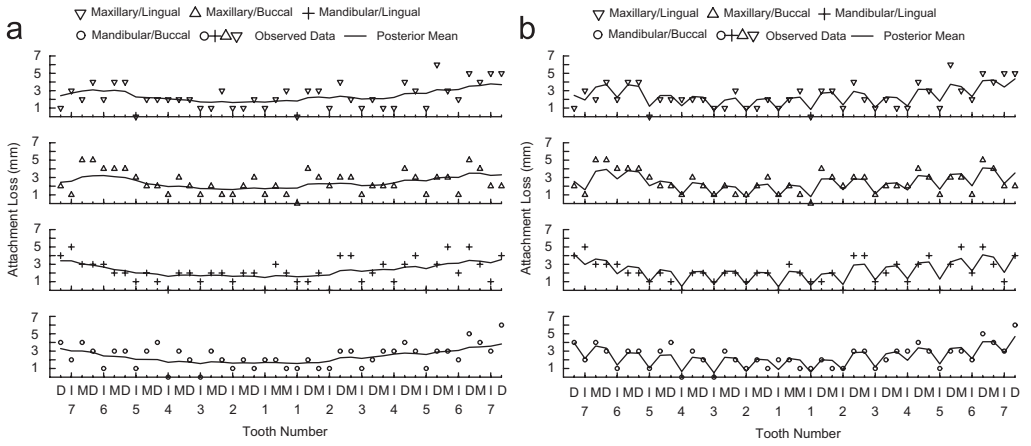

Fig. 1. Periodontal example: raw data and posterior means without (panel (a)) and with (panel (b)) terms with outlying r_i set to fixed effects. (a) Usual CAR fit. (b) Fit with two d_i set to zero.

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 $y_j|\boldsymbol{\beta}_1, \sigma_e^2$ 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

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

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 $(1/\sigma_s^2)Q$ 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 (σ_e^2, σ_s^2), with a flat prior on (σ_e^2, σ_s^2) to emphasize the data's contribution. However, this model has N observations and $N + 2$ unknowns $(\{\beta_{1j}\}, \sigma_e^2, \sigma_s^2)$, 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: 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 (σ_e^2, σ_s^2) and the ratio of variances $r = \sigma_s^2/\sigma_e^2$, 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 (σ_e^2, σ_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.