Smoothing Balanced Single-Error-Term Analysis of Variance

James S. HODGES

Division of Biostatistics
University of Minnesota
Minneapolis, MN 55414
(*hodges@ccbr.umn.edu*)

Yue CUI

Division of Biostatistics
University of Minnesota
Minneapolis, MN 55414
(*yuecui@biostat.umn.edu*)

Daniel J. SARGENT

Mayo Clinic Cancer Center
Mayo Clinic
Rochester, MN 55905
(*sargent@mayo.edu*)

Bradley P. CARLIN

Division of Biostatistics
University of Minnesota
Minneapolis, MN 55414
(*carli002@umn.edu*)

We present an approach to smoothing balanced, single–error term analysis of variance (ANOVA), descended from Smith, that also allows spatial, temporal, or spatiotemporal smoothing. The approach addresses unreplicated designs, masked contrasts in effects with many degrees of freedom, and subgroup analysis, demonstrated using a study of denture-lining materials. Our approach is Bayesian but can be viewed as a way to generate frequentist procedures. A simulation experiment compares four priors, unsmoothed ANOVA, and dropping nonsignificant interactions. Three priors have advantages when some interactions are absent; dropping nonsignificant interactions has serious flaws. We contrast our approach with the approaches of Nobile–Green and Gelman.

KEY WORDS: Bayesian analysis; Degrees of freedom; Masking; Prior distribution; Shrinkage; Subgroup analysis.

1. INTRODUCTION AND MOTIVATION

Analysis of variance (ANOVA) attributes variation in a response to individual factors and to combinations of these factors, that is, interactions. Interactions are often modeled in a stepwise way, with significance tests used to delete or retain effects; but in other linear models, stepwise methods are outperformed by model-averaging and smoothing methods (Leamer 1978; Freedman 1983; Derksen and Keselman 1992; Raftery, Madigan, and Hoeting 1993). This article presents a way to smooth ANOVA that neither includes nor excludes interactions but does something intermediate, like shrinkage.

What should a smoothed ANOVA do? Statistical folklore and practical experience suggest that for a dependent variable on the proper scale, interactions are often absent or small. However, it is unwise to assume that any *specific* interaction is absent. Thus a smoothed ANOVA should mostly remove small effects, mostly retain large effects, and partly retain middling effects. This simplifies interpretation and aids estimation by reducing standard errors.

We present a smoothed ANOVA addressing this general goal as well as three specific concerns, which we introduce using an example, an unreplicated $2 \times 4 \times 8$ study of soft denture-lining materials (Pesun, Hodges, and Lai 2002). Soft denture liners are fabricated on a hard denture base. The soft liner is then polished, which can create or widen a gap between the liner and base. Such gaps harbor *Candida* and other oral pathogens. This study compared gaps, measured in microns, for a new and a standard soft-liner material (factor M) under all 32 combinations of four polishing methods (factor P) and eight finishing methods (factor F). The primary interest is how much the materials differ in gap size. Based on standard diagnostics, we analyze the common logarithm, $\log_{10}$gap.

Three issues arise in this analysis. The standard analysis of unreplicated designs uses the highest-order interaction as the error term (Scheffé 1959, sec. 4.2). Table 1 is the ANOVA table for this analysis; the dataset is given in Appendix B. This dataset has an egregious outlier on the raw scale that also fails outlier tests on the log scale, although not by much. It thus would be desirable to keep part of the three-way interaction—the outlier—in the fitted model while also deriving an error measure from this and perhaps other interactions. Second, the P $\times$ F and M$\times$P $\times$ F interactions have 21 degrees of freedom (df) each. At most, a few of these contrasts are truly present, and they may be masked by the many null contrasts. Finally, the interactions of special interest are M$\times$P and M$\times$F, with 3 and 7 df. Shrinking (smoothing) the material comparison across the four polishing methods and eight finishing methods would reduce clutter. Smoothed ANOVA addresses all of these issues: unreplicated designs, masking in effects with many df, and "subgroup analysis," where treatment-by-subgroup interactions capture subgroup treatment effects (Dixon and Simon 1991).

Our approach in this article smooths balanced, single–error term ANOVAs using hierarchical models. We focus on interactions, but exactly the same tools can be used to smooth main effects. Section 2 gives notation and results, including a smoothed ANOVA (SANOVA) table. We use a Bayesian analysis with Markov chain Monte Carlo (MCMC) draws from the posterior. This can be done in many different ways; Appendix A

TECHNOMETRICS, FEBRUARY 2007, VOL. 49, NO. 1
DOI 10.1198/004017006000000408

Table 1. Soft-Material Polishability Data: The Usual ANOVA Table

	df	*SS*	*MS*
Main effects			
Material	1	1.12	1.12
Polishing	3	.38	.13
Finishing	7	1.92	.27
Interactions			
M × P	3	.65	.22
M × F	7	1.40	.20
P × F	21	3.28	.16
M × P × F	21	2.05	.098
Error			
Replicates	0	—	—

gives our algorithm. Readers averse to Bayesian interpretations can view our approach as a way to generate smoothing procedures in which a new prior distribution specifies a new procedure. The priors do matter; Section 3 compares several in a simulation experiment. Section 4 then analyzes the polishability data. Section 5 contrasts our approach with two other approaches to "smoothed ANOVA," those of Nobile and Green (2000) and Gelman (2005). Our approach shows a clear family resemblance to the approach of Smith (1973), nourished by modern Bayesian computing.

2. THEORETICAL MACHINERY

This section illustrates the notation using a 2^3 design with six replicates per cell.

2.1 Notation

Suppose that a balanced design has c cells and $n \geq 1$ observations per cell, for cn in total. In our 2^3 example, $c = 2^3 = 8$ and $cn = 48$. Write the ANOVA as a linear model with each effect having design matrix columns orthogonal to each other and to columns for other effects, that is, an orthogonal parameterization. Effects with >2 df have infinitely many such design matrices related by orthogonal transformation. Group the M columns for the grand mean and main effects into a $cn \times M$ matrix A_1, and group the N columns for interactions into a $cn \times N$ matrix A_2. Scale the columns of A_1 and A_2 so that $A_1'A_1 = cnI_M$ and $A_2'A_2 = cnI_N$, where I_M is the M-dimensional identity.

To simplify bookkeeping, we use artificial "cases" to formulate smoothed ANOVA as a hierarchical model (Lee and Nelder 1996; Hodges 1998). The hierarchical model's data level (the "data cases") is the usual ANOVA in linear model form, $\mathbf{y} = X_1\Theta + \epsilon$, where $\mathbf{y}$ is the cn vector of data, $X_1 = [A_1|A_2]$, $\Theta = (\theta_1, \ldots, \theta_{M+N})$ is the vector of unknown mean-structure parameters, and ϵ is cn-dimensional normal with mean 0 and covariance $\frac{1}{\eta_0}I_{cn}$, with the error precision η_0 unknown. In the 2^3 example, $X_1 = H \otimes 1_6$, where $\otimes$ is the Kronecker product and

$$H = \left[\begin{array}{cccc|cccc} +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & +1 & +1 & -1 & +1 & -1 & -1 & -1 \\ +1 & +1 & -1 & +1 & -1 & +1 & -1 & -1 \\ +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \\ +1 & -1 & +1 & +1 & -1 & -1 & +1 & -1 \\ +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\ +1 & -1 & -1 & -1 & +1 & +1 & +1 & -1 \end{array}\right]. \quad (1)$$

The first four columns of H are for the grand mean and main effects, whereas the last four columns are for the interactions. The interactions θ_k are smoothed by the hierarchical model's second layer. Partition Θ as $(\Theta_1^T, \Theta_2^T)^T$ conforming to A_1 (main effects) and A_2 (interactions). Θ_2 is modeled as $\theta_{M+k}|\phi_k \sim \mathrm{N}(0, 1/\phi_k), k = 1, \ldots, N$. The unknown precisions ϕ_k control shrinkage of their respective θ_k in a manner to be elaborated; they are the key to our approach. Rewrite Θ_2's model as $0_N = Z_1\Theta + \delta$, where $Z_1 = [0_{N\times M}|I_N]$, 0_N and $0_{N\times M}$ are arrays of 0's with the given dimensions, and δ is an N-variate normal with mean 0 and diagonal covariance matrix $\mathrm{diag}(\phi_1, \ldots, \phi_N)^{-1}$. The model for Θ can be treated as artificial cases, called "constraint cases" by Hodges (1998) because they constrain Θ_2, and then combined with the data cases to give

$$\left[\frac{y}{0_N}\right] = \left[\begin{array}{c|c} A_1 & A_2 \\ \hline 0_{N\times M} & I_N \end{array}\right]\left[\frac{\Theta_1}{\Theta_2}\right] + \left[\frac{\epsilon}{\delta}\right] \quad (2)$$

or, more concisely, $Y = X\Theta + e$, where Y, X, and e have obvious definitions referring to (2).

Equation (2) has the form of a linear model; Y and X are known, Θ is unknown, and e contains unknown errors. This is precisely—and merely—an accounting identity (Whittaker 1998) that eases derivations because $\Sigma = \mathrm{cov}(e)$ is diagonal with blocks $\Sigma_1 = \frac{1}{\eta_0}I_{cn}$ for the data cases and $\Sigma_2 = [\mathrm{diag}\{\phi_1, \phi_2, \ldots, \phi_N\}]^{-1}$ for the constraint cases.

The precisions ϕ_k need not be distinct. The modeler specifies a set of distinct constraint-case precisions $\{\eta_1, \ldots, \eta_s\}, s \leq N$, and a fixed, deterministic assignment function $j(k)$ such that $\phi_k = \eta_{j(k)}$. This groups the θ_k's and their associated columns in A_2 with each group's θ_k smoothed using its own η_j. Define $\eta = (\eta_0, \eta_1, \ldots, \eta_s)$, the vector consisting of the error precision η_0 and the distinct smoothing precisions. Let n_j be the number of ϕ_k's mapping to η_j; $\sum_{j=1}^{s} n_j = N$. In the 2^3 design, A_2 has $N = 4$ columns, one three-way interaction and three two-way interactions. If each interaction is smoothed separately, then each ϕ_k is distinct, so $s = 4$ and $n_j = 1, j = 1, 2, 3, 4$. Alternatively, the two-way interactions can be smoothed using a single η_j, whereas the three-way interaction keeps its own η_j. Then $s = 2$, and, referring to (1), $j(1) = j(2) = j(3) = 1$ and $j(4) = 2$, so $n_1 = 3$ and $n_2 = 1$.

We have implicitly put a flat prior on the θ_k for the grand mean and main effects. This is not required. The main effects can be smoothed as we have smoothed the interactions, or informative priors can be added as constraint cases with fully specified covariance (Hodges 1998, sec. 2).

2.2 Conditional and Marginal Posterior Distributions

Our MCMC algorithm (Sec. A.3) draws $\log(r_j) = \log(\eta_j/\eta_0)$ and uses Rao–Blackwellization (Gelfand and Smith 1990) to estimate the marginal posteriors of Θ and η_0. This section derives the needed distributions; Section A.1 gives simplified forms that avoid matrix inversions.

In the usual notation, the joint posterior for Θ and η is

$$f(\Theta, \eta|Y) \propto \pi(\eta)|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(Y - X\Theta)'\Sigma^{-1}(Y - X\Theta)\right), \quad (3)$$

where Σ is a function of η and $\pi(\eta)$ is η's prior. Completing the square in Θ makes it possible to integrate Θ out of (3), giving

$$f(\eta|\mathbf{Y}) \propto \pi(\eta)|\mathbf{X}'\Sigma^{-1}\mathbf{X}|^{-1/2}|\Sigma|^{-1/2} \times \exp\left(-\frac{1}{2}\{\mathbf{Y}'\Sigma^{-1}\mathbf{Y} - \mathbf{Y}'\Sigma^{-1}\mathbf{X}(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma^{-1}\mathbf{Y}\}\right). \quad (4)$$

By straightforward calculation, $X'\Sigma^{-1}X$ is

$$\left[\begin{array}{c|c} cn\eta_0 I_M & 0_{M\times N} \\ \hline 0_{N\times M} & cn\eta_0 I_N + \mathrm{diag}(\phi_1, \ldots, \phi_N) \end{array}\right], \quad (5)$$

which, with further simple algebra, gives

$$|\mathbf{X}'\Sigma^{-1}\mathbf{X}|^{-1/2}|\Sigma|^{-1/2} = (cn)^{-M/2}\eta_0^{(cn-M)/2}\prod_{j=1}^{s} r_j^{n_j/2}(cn + r_j)^{-n_j/2}. \quad (6)$$

Now change variables from η to $(\eta_0, \mathbf{r})$, where $r_j = \eta_j/\eta_0, j = 1, \ldots, s$, and replace the prior $\pi(\eta)$ with $\pi(\eta_0, \mathbf{r})$, which contains the Jacobian. The joint marginal posterior of $(\eta_0, \mathbf{r})$ is

$$f(\eta_0, \mathbf{r}|\mathbf{Y}) \propto \pi(\eta_0, \mathbf{r})\eta_0^{(cn-M)/2} \times \exp\left(-\frac{1}{2}\eta_0 W(\mathbf{r})\right)\prod_{j=1}^{s} r_j^{n_j/2}(cn + r_j)^{-n_j/2}, \quad (7)$$

where $W(\mathbf{r}) = \mathbf{Y}'Q^{-1}\mathbf{Y} - \mathbf{Y}'Q^{-1}\mathbf{X}(\mathbf{X}'Q^{-1}\mathbf{X})^{-1}\mathbf{X}'Q^{-1}\mathbf{Y}$ and $Q^{-1} = \Sigma^{-1}/\eta_0$ is a function of $\mathbf{r}$ but not η_0. Equation (7) is a gamma density in η_0; integrating out η_0 gives $\mathbf{r}$'s marginal posterior.

If η_0 has a gamma prior independently of $\mathbf{r}$, $\pi(\eta_0, \mathbf{r}) \propto \pi(\mathbf{r})\eta_0^{\alpha-1}\exp(-\lambda\eta_0)$, then

$$f(\mathbf{r}|\mathbf{Y}) \propto \pi(\mathbf{r})(2\lambda + W(\mathbf{r}))^{-[(cn-M)/2+\alpha]} \times \prod_{j=1}^{s} r_j^{n_j/2}(cn + r_j)^{-n_j/2}. \quad (8)$$

From (7), the posterior of η_0 given $\mathbf{r}$ is also a gamma distribution,

$$f(\eta_0|\mathbf{r}, \mathbf{Y}) = \frac{(\lambda + W(\mathbf{r})/2)^{\xi}}{\Gamma(\xi)} \times \eta_0^{\xi-1}\exp\left(-\eta_0\left(\lambda + \frac{W(\mathbf{r})}{2}\right)\right), \quad (9)$$

where $\Gamma(\xi)$ is the gamma function evaluated at $\xi = \frac{cn-M}{2} + \alpha$. To obtain the posterior of Θ given $\mathbf{r}$, begin with (3), change variables from η to $(\eta_0, \mathbf{r})$, and integrate out the gamma variate η_0 to give a multivariate-t on $\nu = cn - M + 2\alpha$ df with center $\hat{\Theta} = (\mathbf{X}'Q^{-1}\mathbf{X})^{-1}\mathbf{X}'Q^{-1}\mathbf{Y}$ and dispersion matrix $\frac{2\lambda+W(\mathbf{r})}{\nu}(\mathbf{X}'Q^{-1}\mathbf{X})^{-1}$.

If instead all of the $\eta_j, j = 0, \ldots, s$, have independent gamma priors with shape and scale parameters (α_j, λ_j), then, following the same sequence of steps, $\mathbf{r}$'s marginal posterior is

$$f(\mathbf{r}|Y) \propto \left(W(\mathbf{r}) + 2\lambda_0 + 2\sum_{j=1}^{s} r_j\lambda_j\right)^{-((cn-M)/2+\sum_{j=0}^{s}\alpha_j)} \times \prod_{j=1}^{s} r_j^{n_j/2+\alpha_j-1}(cn + r_j)^{-n_j/2}. \quad (10)$$

Given $\mathbf{r}$, η_0's conditional posterior is gamma with shape parameter $\frac{cn-M}{2} + \sum_{j=0}^{s}\alpha_j$ and scale parameter $W(\mathbf{r})/2 + \lambda_0 + \sum_{j=1}^{s} r_j\lambda_j$, and Θ's conditional posterior is multivariate-t on $\nu = cn - M + 2\sum_{j=0}^{s}\alpha_j$ df, with center $\hat{\Theta} = (\mathbf{X}'Q^{-1}\mathbf{X})^{-1}\mathbf{X}' \times Q^{-1}\mathbf{Y}$ and dispersion matrix $(W(\mathbf{r}) + 2\lambda_0 + 2\sum_{j=1}^{s} r_j\lambda_j)/\nu \times (\mathbf{X}'Q^{-1}\mathbf{X})^{-1}$.

The MCMC algorithm in Section A.3 draws $z_i = \log(r_i)$ using (8) or (10) transformed to $\mathbf{z}$.

2.3 Degrees of Freedom

Standard ANOVA uses df in F-tests. Smoothed ANOVA emphasizes estimation, and a thoroughly Bayesian approach eschews F-tests. Nonetheless the usual notion of df can be extended to smoothed ANOVA and used to specify priors on $\mathbf{r}$ and to describe the extent of smoothing.

For given Σ, the df in the fit are $\mathrm{df} = \mathrm{trace}(X_1(X'Q^{-1}X)^{-1} \times X_1')$ (Hodges and Sargent 2001), where $Q^{-1} = \Sigma^{-1}/\eta_0$ is, as noted, a function of $\mathbf{r}$ but not of η_0. Using (5), straightforward algebra gives

$$\mathrm{df} = M + \sum_{k=1}^{N}\frac{cn\eta_0}{cn\eta_0 + \phi_k} = M + \sum_{j=1}^{s}\frac{n_j cn\eta_0}{cn\eta_0 + \eta_j} = M + \sum_{j=1}^{s}\frac{n_j cn}{cn + r_j} = M + \sum_{j=1}^{s} q_j, \quad (11)$$

where $q_j = \frac{n_j cn}{cn+r_j} \in [0, n_j]$ is the df controlled by r_j. Note that q_j depends on the precisions η only through the ratio $r_j = \eta_j/\eta_0$, so a prior on r_j induces a prior on q_j and vice versa.

2.4 Prior Distributions on the Smoothing Structure

A prior distribution on η completes a Bayesian specification. In non-Bayesian terms, Sections 2.1 and 2.2 plus a prior define a procedure; different priors define different procedures, which can be assessed in a frequentist way, such as by the mean squared error (MSE) of point estimates. We consider two kinds of priors: unconditional priors, and conditional priors that fix the fit's smoothness at a certain number of df. Other priors may be advantageous; Section 5 briefly considers two of these.

2.4.1 Unconditional Priors. An obvious prior is a gamma distribution on each element of η. Section 3's simulation experiment considers gamma(.001, .001) because with mean 1 and variance 1,000 it has become a conventional "vague" prior, although with 95th percentile 3×10^{-20}, it is far from vague.

Alternatively, a prior can be placed on each q_j, inducing a prior on each r_j and thus on η_j. The sample space of q_j is $[0, n_j]$, so a flat prior on q_j is proper and may be viewed as expressing prior indifference about the degree of smoothing. (It is also

equivalent to the uniform-shrinkage prior; cf. Daniels 1999.) Alternatively, a scaled beta prior for q_j can be used to prefer some degrees of smoothing; for example, a scaled beta(.5, .5) prefers no smoothing ($q_j = n_j$) and complete smoothing ($q_j = 0$) over intermediate smoothing. At the extreme, such a preference implies a two-point prior where $q_j = 0$ or n_j, each with probability .5. The two-point prior is easily executed by putting probability .5 on each of $q_j = \epsilon$ and $q_j = n_j - \epsilon$ for a small ϵ. As long as ϵ is quite small, its specific value is unimportant; Section 3's simulation used $\epsilon = .001$.

If $q_j \sim \mathrm{U}(0, n_j)$, then $r_j > 0$ has density $cn/(cn + r_j)^2$. As the per-cell sample size n increases, this prior moves probability to larger r_j; that is, the prior favors larger η_j to offset the data and maintain indifference about the degree of shrinkage. This is also true for any beta prior on q_j.

2.4.2 Priors Conditioned on Degrees of Freedom. The preceding priors are unconditional in that they do not fix any q_j or group of q_j's. It may be desirable to condition on $q_j = K$ or $\sum_{j \in S} q_j = K$ for some set S of indices, which is analogous to fixing a linear smoother's df. Conditioning on an inequality, say $q_j \leq K$, raises no distinct issues, so we consider only equality conditions.

If q_j's prior is conditioned on $q_j = K$, then this fixes $\eta_j/\eta_0 = r_j = cn(n_j - K)/K$, a common practice with dynamic linear models. If the condition is $\sum_{j \in S} q_j = K$ for S containing more than one j, then conditioning does not fix any q_j. Rather, it fixes the total complexity of $\{\theta_k | j(k) \in S\}$ at K df, and the affected θ_k's compete for the K df. Such a condition does not completely specify the prior on the affected q_j's; this can be done by specifying unconditional priors on the affected q_j, then imposing the condition on their sum. If $n_j = 1$ for each affected q_j, and each q_j receives an iid prior F, then conditioning on $\sum q_j = K$ makes the q_j's exchangeable, although no longer independent. Section 4 illustrates such a prior.

2.5 The SANOVA Table

The usual ANOVA table analyzes—in that word's literal sense—the sum of squares into pieces for each effect. In smoothed ANOVA as specified here, the smoothed θ_k are shrunk toward 0, that is, partly removed from the fitted model and counted as error. A SANOVA table records this division of each effect's df and sum of squares (SS) into a part retained in the fit and a part considered error. SANOVA emphasizes estimation over testing, but a SANOVA table is still a useful bookkeeping device, particularly for showing how information about error variation is derived from replication and from variation smoothed out of shrunken effects.

Section 2.3 derived the portion of an effect's df retained in the fit for a given Σ; the rest of the effect's df is deemed to be error. Section A.2 derives analogous portions of an effect's SS allotted to the fit and to error. These elements suffice to construct a SANOVA table accounting for a dataset's df and SS. The partition of each effect's SS and df is a function of the precisions η, but single-number summaries are convenient. Section A.2 argues for using the posterior expected df and SS. We defer further discussion of this topic to Section 4, which gives SANOVA tables for the polishability data.

3. A SIMULATION EXPERIMENT COMPARING PRIORS (PROCEDURES)

This experiment had two aims: to compare priors and to compare the resulting SANOVA procedures to familiar ANOVA procedures.

3.1 Design, Procedures, Outcome Measures

3.1.1 Design of the Simulation Experiment. An observation in this simulation experiment was a dataset simulated from a 2^3 design with $n = 6$, like the dataset analyzed by Hodges and Sargent (2001, sec. 6). The simulation experiment itself was a repeated-measures (split-plot) design, in which a "subject" was a set of 48 standard normal errors. We generated 1,000 such "subjects." The simulation experiment had three factors, all within-subject: (a) the true 2^3 mean structure, with levels being the number of truly present interactions (0, 1, 2, 3, or 4); (b) the 2^3 design's true error precision η_0, with three levels (1, .25, or .0625, i.e., error standard deviation 1, 2, or 4); and (c) the analysis procedure applied to the 2^3 design, with six levels as described in Table 2. Table 2 gives each procedure a brief name, which we use henceforth. We included the two-point procedure because it is a step in the Bayesian direction from the familiar "drop-non-sig" procedure.

Because we specified the true 2^3 mean structure using the orthogonal parameterization (1), we set the true θ_1, θ_2, θ_3, and θ_4 (grand mean and main effects) to 0 without loss of generality. When an interaction was absent, we set its θ_k to 0; when an interaction was present, we set θ_k to 1. We considered exchangeable priors for the interaction θ_k's with each θ_k smoothed by its own η_j, so we lose no generality by considering only how many, not which, interactions were truly present. Section A.4 gives more details on simulating the 2^3 datasets and about Monte Carlo error.

Table 2. Procedures Considered in the Simulation Experiment

Short name for procedure	*Bayesian?*	*The procedure*
Gamma	Yes	$\eta_0, \eta_i \sim$ gamma(.001, .001), $i = 1, 2, 3, 4$
Flat	Yes	$q_i \sim$ unif(0, 1), $i = 1, 2, 3, 4$; $\eta_0 \sim$ flat
Beta	Yes	$q_i \sim$ beta(.5, .5), $i = 1, 2, 3, 4$; $\eta_0 \sim$ flat
Two-point	Yes	$q_i = .999$ or .001 each with probability .5; $\eta_0 \sim$ flat
No-smoothing	No	Ordinary least squares ANOVA
Drop-non-sig	No	Two-step ANOVA Step 1: Fit with all interactions. Step 2: Refit with only interactions significant in step 1.

3.1.2 Outcome Measures. We compared the six procedures using three groups of measures, one group for the means of the eight cells in the 2^3 design, a second group for the θ_k's, and a third group for the error precision η_0. With one exception, each group of measures included the bias and MSE of the estimates (posterior means, for the Bayesian procedures) and coverage of the 95% equal-tail posterior or confidence interval. We did not consider bias of the cell means, which are simple linear functions of Θ_2's bias.

When interactions are truly present in the simulated data, the θ_k's are of three types: the "target," or truly present interactions; the "null," or truly absent interactions; and the grand mean and main effects. The simulation experiment's design implies that for any given number of truly present interactions, η_0, and procedure, the targets all have the same true bias, MSE, and coverage, as do the nulls and main effects. Thus results for target and null θ_k are described separately, with bias, MSE, and coverage averaged within each type of θ_k. For the grand mean and main effects, bias, and MSE are identical for all procedures and are not considered further.

For the θ_k's, we scaled bias by the true error standard deviation ($\frac{1}{\sqrt{\eta_0}}$); for the cell means and θ_k's, we scaled MSE by the true error variance ($\frac{1}{\eta_0}$). This gives the no-smoothing procedure constant performance, removing trends in the design factor η_0 that obscure comparisons. Similarly, for the error precision η_0, we report the bias and square root of MSE as percents of the true η_0.

3.2 Results

3.2.1 Cell Means. When no interactions are present, all procedures are unbiased for cell means. In this null case, scaled MSE is largely unaffected by the true η_0 (Table 3, "Cell means" column); all Bayesian procedures outperform no-smoothing, as does drop-non-sig. Figures 1 and 2 show scaled MSE as a function of η_0 for one and four truly present interactions. All other procedures outperformed no-smoothing when one interaction was present (Fig. 1), and even when four interactions were present (Fig. 2), gamma, flat, and beta were almost as good as no-smoothing. Intermediate numbers of interactions gave results intermediate between Figures 1 and 2 (not shown).

Coverage of 95% intervals turns out to be nearly nominal for all procedures and all numbers of interactions when $\eta_0 = 1$ or .25. Figure 3 shows coverage as a function of the number of truly present interactions when $\eta_0 = .0625$. All procedures but no-smoothing lose coverage as more interactions are added. However, gamma, flat, and beta are still near nominal, whereas two-point and drop-non-sig fall to about 90% and 80% respectively.

Table 3. Null Case: Scaled MSE for Cell Means and Parameters θ_k

Procedure	*Cell means*	θ_k
Gamma[a]	11.5–10.3%	.85–.56%
Flat	10.4%	.58%
Beta	10.2%	.54%
Two-point	8.7%	.16%
No-smoothing	16.6%	2.13%
Drop-non-sig	10.2%	.54%

[a]Scaled bias decreases as η_0 decreases.

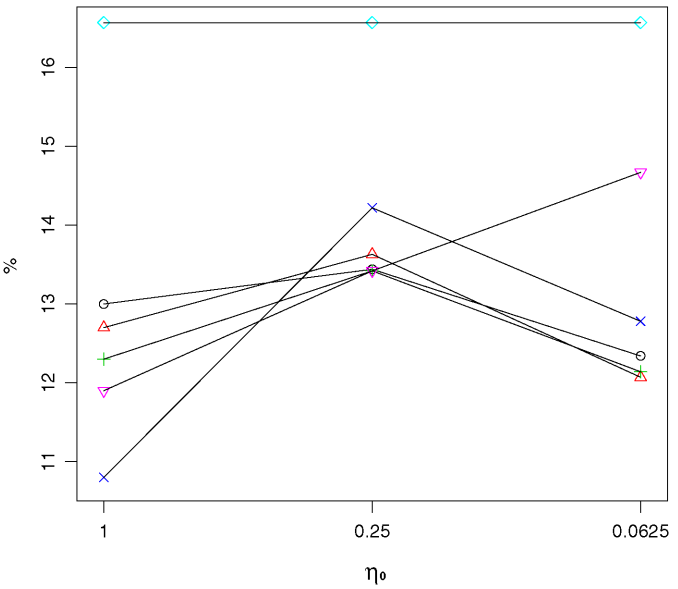

Figure 1. Cell Mean MSE (as a percentage of $1/\eta_0$) versus η_0, for the Case of One Interaction. (○, gamma; △, flat; +, beta; ×, two-point; ◇, no-smoothing; ▽, drop-non-sig.)

3.2.2 The parameters θ_k. All procedures are unbiased for null θ_k. Figure 4 plots scaled bias for target θ_k as a function of the error precision η_0 for the case of four truly present interactions. The plots for one, two, and three interactions are visually identical; only the vertical scale changes, with bias increasing in magnitude as interactions are added. Two-point and drop-non-sig degrade sharply as η_0 decreases. Gamma degrades more gracefully; flat and beta degrade when η_0 declines from 1 to .25, but actually improve as η_0 declines further to .0625.

Scaled MSE for null θ_k is again largely unaffected by η_0 (Table 3, "θ_k" column); all Bayesian procedures perform better than no-smoothing. In nonnull cases, a plot of scaled MSE

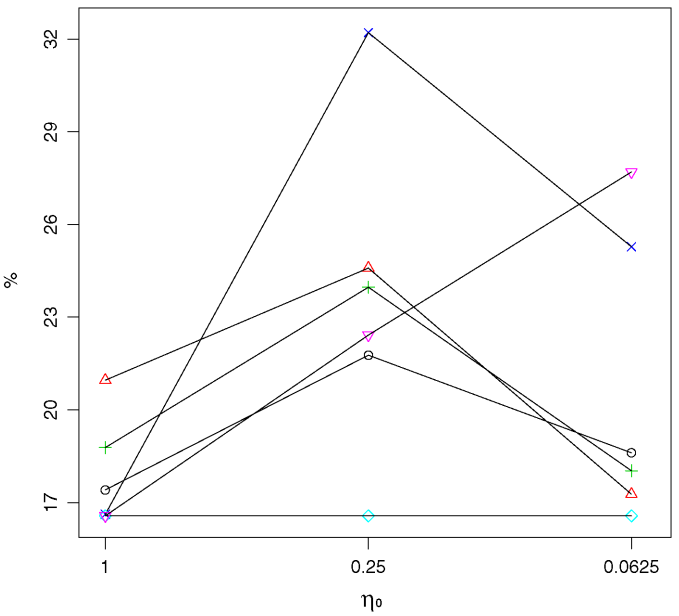

Figure 2. Cell Mean MSE (as a percentage of $1/\eta_0$) versus η_0, for the Case of Four Interactions. (○, gamma; △, flat; +, beta; ×, two-point; ◇, no-smoothing; ▽, drop-non-sig.)

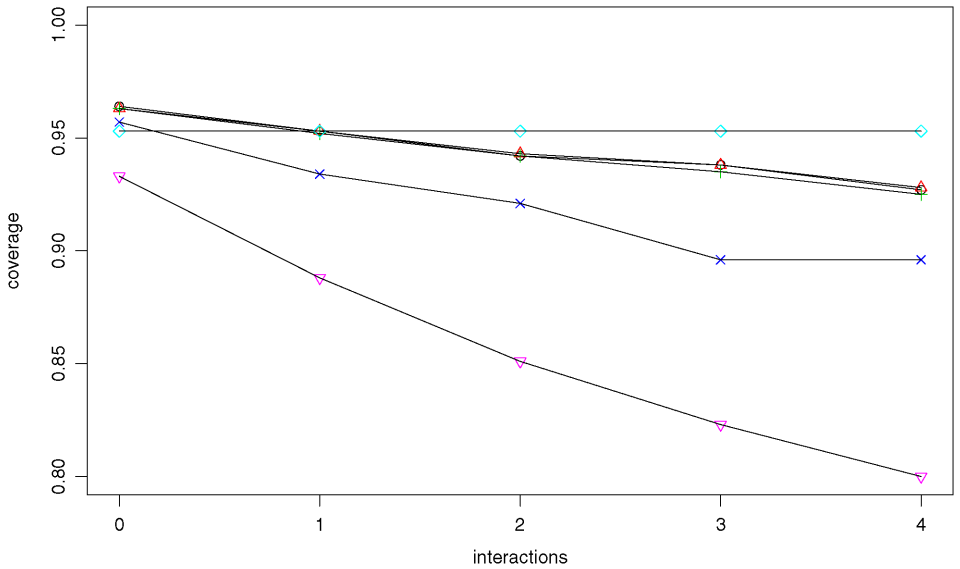

Figure 3. Cell Mean Coverage versus Number of Present Interactions, for the Case of $1/\eta_0 = .0625$. (○, gamma; △, flat; +, beta; ×, two-point; ◇, no-smoothing; ▽, drop-non-sig.)

for target θ_k looks just like Figure 2. When four interactions are truly present, no-smoothing's scaled MSE of target θ_k is about 2%. For gamma and beta, it is well under 3% for $\eta_0 = 1$ and .0625; flat reaches up to 3% for $\eta_0 = 1$ and to 4% for $\eta_0 = .25$. Two-point reaches 6% for $\eta = 0.25$ and drop-non-sig reaches 5% for $\eta_0 = .0625$.

Under all conditions, intervals for null θ_k have nearly nominal coverage for drop-non-sig but 99% or higher for the Bayesian procedures. Figure 5 shows coverage of 95% intervals for target interactions as a function of η_0. Except for no-smoothing, all procedures lose coverage as η_0 decreases. Gamma, flat, and beta degrade to about 90% coverage with little change as η_0 declines to .0625, whereas coverages of two-point and drop-non-sig plunge when $\eta_0 = .0625$.

3.2.3 The Error Precision η_0. Scaled MSE is 20–30% and coverage close to 95% for all procedures in all conditions. Drop-non-sig is nearly unbiased for η_0 in all conditions. Figure 6 plots $E(\eta_0|Y)$'s bias for the Bayesian procedures against the absolute bias of θ_k for target interactions. Generally, as η_0's bias becomes more positive, overstating the data's precision, the targets shrink less and their bias decreases. Also, although adding interactions has little effect on the bias of target θ_k, more true interactions make η_0's bias more negative, except for two-point when $\eta_0 = 1$, where the opposite is true. This bias may be specific to η_0's posterior mean; preliminary results suggest that the posterior median is much less biased.

3.3 Summary

The results support two conclusions. First, results for smooth priors (gamma, flat, and beta) degrade gracefully as error precision decreases, whereas procedures that guess the right in-

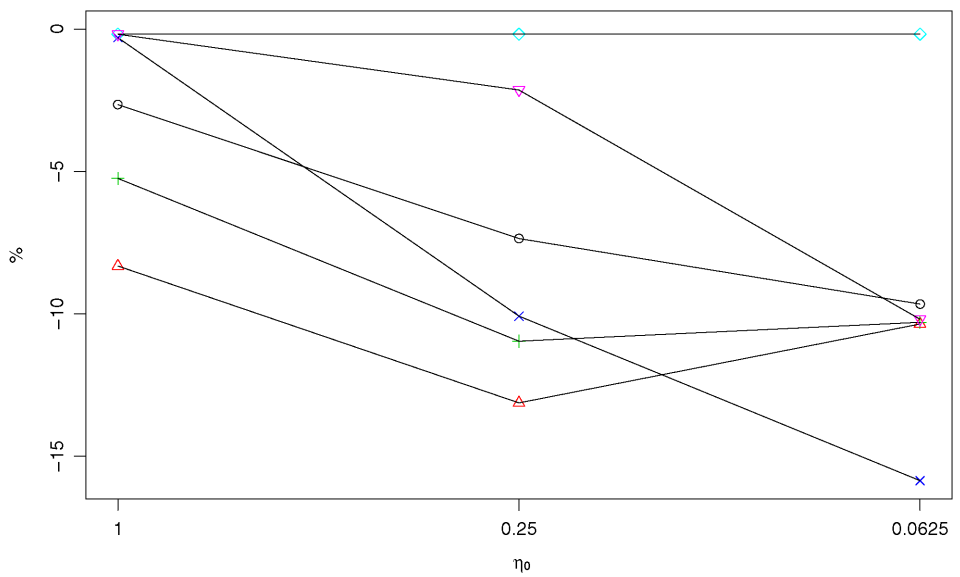

Figure 4. Average Bias for Target θ_k (as a percentage of $1/\sqrt{\eta_0}$). (○, gamma; △, flat; +, beta; ×, two-point; ◇, no-smoothing; ▽, drop-non-sig.)

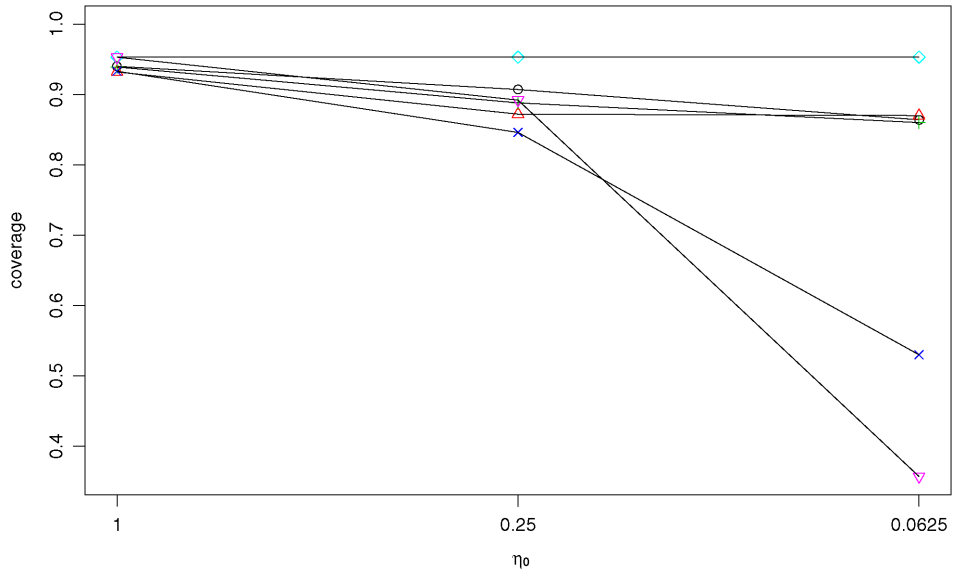

Figure 5. Target θ_k Coverage versus η_0, for the Case of Four Interactions. (○, gamma; △, flat; +, beta; ×, two-point; ◇, no-smoothing; ▽, drop-non-sig.)

teractions (two-point and drop-non-sig) degrade sharply. Second, for zero, one, or two interactions, the smooth priors give notable performance gains while giving up little or nothing to unsmoothed ANOVA in MSE and coverage. Some procedures showed performance gains even when three interactions were present. Given the common presumption—dare we say prior belief?—that real data usually have few truly present interactions, the gamma, flat, and beta priors seem to be reasonable

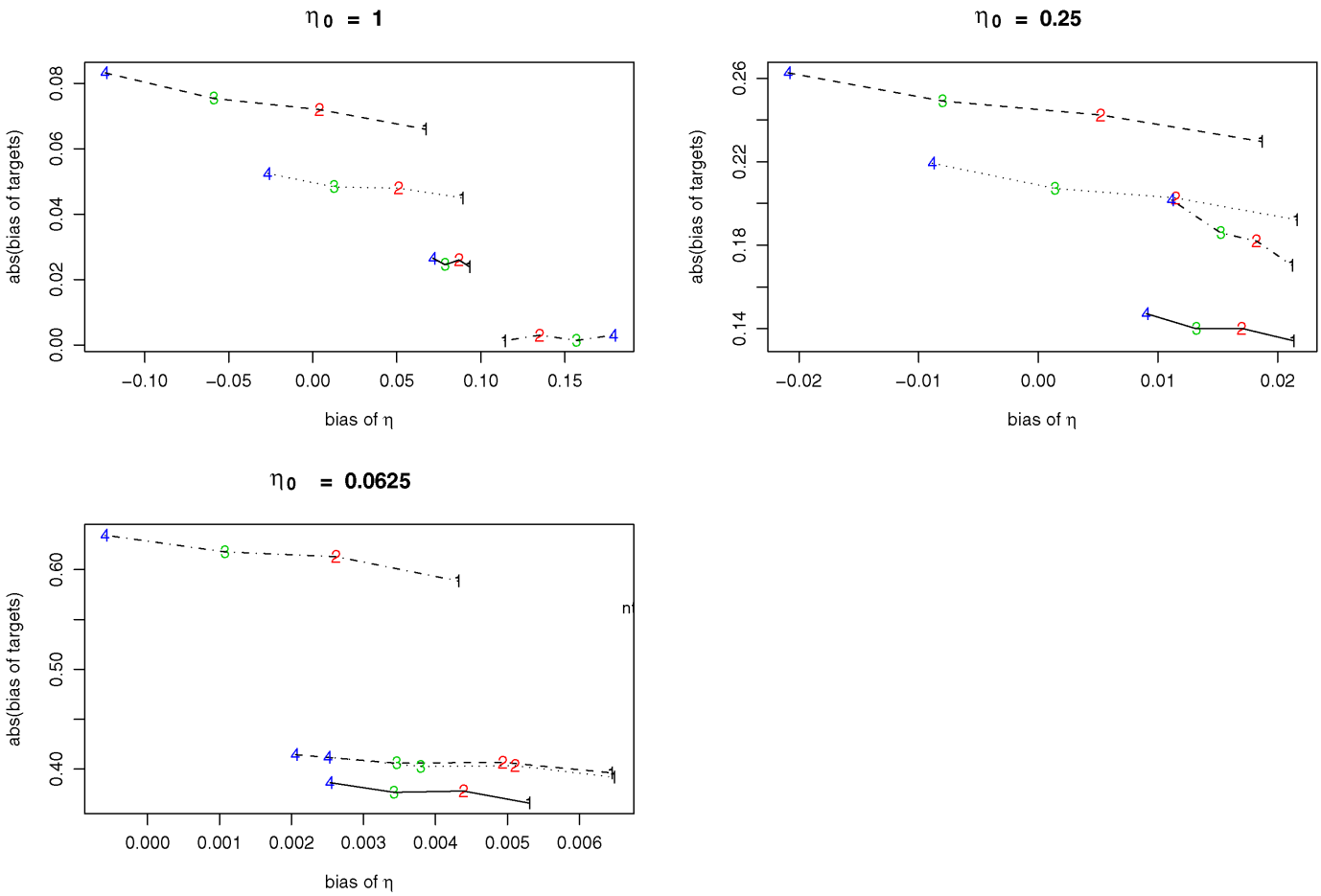

Figure 6. Average Absolute Bias of Target θ_k versus Bias of η_0. For each procedure, the plotting character (1, 2, 3, or 4) indicates the number of truly present interactions. (— gamma; - - flat; ····· beta; - · - two-point.)

candidates for general use. However, the two-point and drop-non-sig procedures are hazardous when error precision is small.

4. EXAMPLE: THE SOFT–MATERIAL POLISHABILITY DATA

We now return to the polishability dataset introduced in Section 1. A standard ANOVA of $\log_{10}$gap using the three-way interaction as the error term gives p values of .12, .097, and .15 for the material-by-polishing (M × P), material-by-finishing (M × F), and polishing-by-finishing (P × F) interactions. Without the one outlier, the p values drop to .096, .004, and .016. The investigators had no reason to consider the outlying measurement defective, so deleting it is hard to justify. Smoothed ANOVA is a smooth alternative to a binary (include/exclude) choice. This section presents three smoothed analyses; Appendix > gives the design matrix X_1 that we used.

In all three smoothed analyses, the M × P and M × F interactions were smoothed by giving each θ_k its own η_j, that is, smoothing each contrast separately. The P × F interaction is of secondary interest because it does not involve the materials, so its 21 contrasts were smoothed using a single η_j.

The three smoothed analyses differ in their handling of the M × P × F interaction, which is most directly affected by the outlier. Analysis 1 gave each of M × P × F's 21 contrasts its own η_j, that is, smoothed them separately. Analysis 2 smoothed all 21 contrasts using a single η_j, giving a posterior expected df of 6.75. To distinguish the effect of using a single η_j from the effect of having few total df, Analysis 3 fixed M × P × F's total df at the same 6.75 but allowed each contrast to be smoothed by its own η_j. (We also did an analysis like Analysis 3 but fixing M × P × F's total df at 9.83, Analysis 1's posterior mean; the results were nearly identical to those of Analysis 1.) Finally, we put a flat prior on η_0 (i.e., gamma with $\alpha = 1$ and $\lambda = 0$) and flat priors on each interaction's q_j, subject to conditioning as described.

Figure 7 summarizes the results for M × P × F. In both panels, M × P × F's contrasts are sorted from left to right in increasing order of their unsmoothed estimates' absolute values. Figure 7(a) shows, for each contrast, the posterior mean df in the fit. In Analysis 1, where each contrast has its own η_j, posterior mean df increased as the unsmoothed contrast increased, for a posterior mean of 9.83 total df. The rightmost contrast retained .8 df in the fit, reflecting the outlier's effect. In Analysis 2, all 21 contrasts were smoothed with the same η_j and had the same posterior mean df, .32. All 21 contrasts were smoothed more than in Analysis 1 (some much more), leaving 6.75 total df. Analysis 3's prior fixed M × P × F's df at the same 6.75 but allowed each contrast to be smoothed by its own η_j. It smoothed the four largest contrasts less than Analysis 2, and the other 17 contrasts were smoothed a bit more. The rightmost contrast, reflecting the outlier, changed the most, from 0.32 df in Analysis 2 to .68 in Analysis 3. Thus, although Analyses 2 and 3 gave M × P × F 6.75 df, Analysis 3 "unmasked" the outlier by allowing different contrasts to be smoothed differently and forcing them to compete for those 6.75 df.

Figure 7(b) shows absolute values of the unsmoothed contrast estimates and posterior means from Analyses 1, 2, and 3.

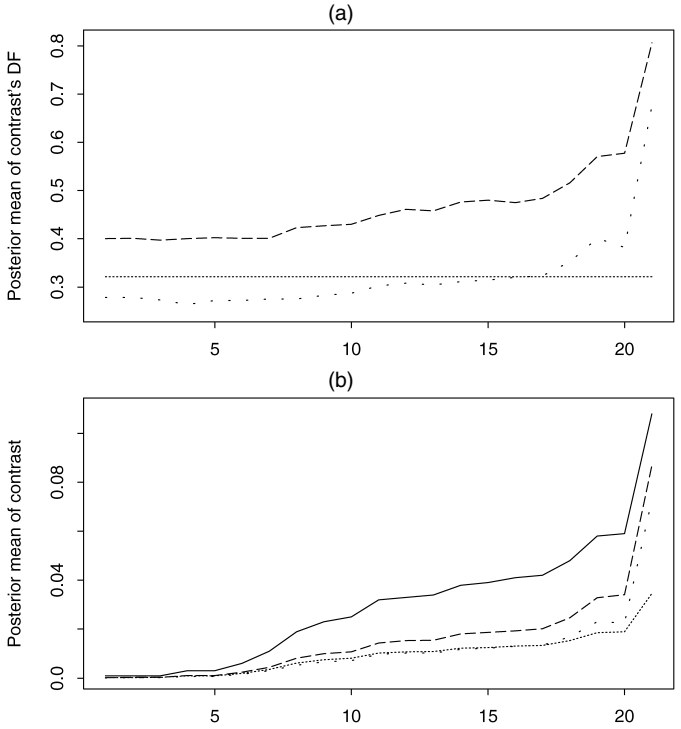

Figure 7. Polishability Data: Posterior Summaries for the 21 Contrasts in the M × P × F interaction. (a) E(DF|Y) for the three smoothed analyses. (b) Absolute values of contrast estimates for the unsmoothed and smoothed analyses [the latter being E(θ_k|Y)]. (— unsmoothed; - - Analysis 1; ····· Analysis 2; · · · Analysis 3.)

These largely reflect the posterior mean df of Figure 7(a). However, although small contrasts have fewer df in Analysis 3 compared with Analysis 2, their smoothed estimates hardly change, whereas the largest contrast's estimate increases substantially.

Table 4 gives the SANOVA tables for Analysis 1, 2, and 3. The unsmoothed section, for the grand mean and main effects, is the same as in the usual ANOVA table. The Smoothed section shows the partition of each effect's df and SS between the model fit ("Model" column) and error ("Error" column). As noted, these are posterior expectations, with respect to **r**, of the respective SS and df. The Model and Error halves of the table include a column for mean squares, which, as usual, are SS divided by df. A smoothed effect's error MS describes the effect's contribution of information about error variation. Pure Error's SS is 0 because the design is unreplicated. A replicated design would have df and SS for error from replication as well as smoothing, with the total error SS and df being the sums of the two sources. Unlike in the standard analysis, in which only the three-way interaction is deemed error, in Analysis 1 about half of the 23.08 df for error comes from variation smoothed out of two-way interactions.

The SANOVA tables for Analyses 2 and 3 have the same unsmoothed section as for Analysis 1, so Table 4 gives only their Smoothed and Error sections. Comparing Analyses 1 and 2, a single η_j for all 21 M × P × F contrasts (Analysis 2) inflates the MSE from .11 to .13 and forces M × P × F to shrink more (posterior mean df, 6.75 vs. 9.83). This occurs because the contrast reflecting the outlier is smoothed as much as the other 20 contrasts [Fig. 7(a)]. Pushing this variation into error inflates the estimate of error variance and induces further smoothing,

Table 4. SANOVA Tables for Analyses 1, 2, and 3

	Model			Error		
	SS	*DF*	*MS*	*SS*	*DF*	*MS*
Analysis 1						
Unsmoothed						
Grand mean	75.54	1.00	75.54			
M	1.12	1.00	1.12			
P	.38	3.00	.13			
F	1.92	7.00	.27			
Smoothed						
M × P	.48	1.59	.30	.17	1.41	.12
M × F	.88	3.85	.23	.52	3.15	.17
P × F	2.13	13.65	.16	1.15	7.35	.16
M × P × F	1.26	9.83	.13	.79	11.17	.07
Error						
Pure				0	0	
Smoothing				2.63	23.08	.11
Total				2.63	23.08	.11
Analysis 2						
Smoothed						
M × P	.43	1.52	.28	.22	1.48	.15
M × F	.78	3.53	.22	.62	3.47	.18
P × F	1.64	10.50	.16	1.64	10.50	.16
M × P × F	.66	6.75	.10	1.39	14.25	.10
Error						
Pure				0	0	
Smoothing				3.87	29.70	.13
Total				3.87	29.70	.13
Analysis 3						
Smoothed						
M × P	.45	1.54	.29	.20	1.46	.13
M × F	.82	3.63	.22	.58	3.37	.17
P × F	1.81	11.64	.16	1.46	9.36	.16
M × P × F	.95	6.75	.14	1.10	14.25	.08
Error						
Pure				0	0	
Smoothing				3.34	28.44	.12
Total				3.34	28.44	.12

resulting in about 3 more df smoothed out of M×P×F. Smoothing all M×P×F contrasts with a single η_j indirectly forces P×F to shrink more, with a posterior mean df in the fit of only 10.50 in Analysis 2 and 13.65 in Analysis 1. (Recall that P×F was smoothed using a single η_j.)

We illustrate a subgroup analysis using the M×F interaction, which addresses whether the material difference, standard minus new, varies between levels of F (finishing methods). Figure 8 shows the unsmoothed estimates and 95% confidence intervals for standard minus new (using M × P × F as the error term), and Analysis 2's smoothed estimates and intervals. The three smoothed analyses give nearly identical posterior means for standard minus new despite their differences for M × P × F. The SANOVA tables reflect this: M × F's line is similar in all three analyses. In Figure 8, the smoothed subgroup-specific differences are shrunk toward $-.26 \log_{10} \mu$m, the material main effect. M × F's seven contrasts were smoothed using different η_j, so F's levels 1 and 3 were shrunk toward the material main effect by larger fractions than was level 2. The intervals from the Unsmoothed analysis are wider, at $.92 \log_{10} \mu$m, than Analysis 2's intervals, which range in width from .55 to .75 (median, .65; Analyses 1's and 3's are narrower by about .1). Based on Section 3's results for cell means, we conjecture that these intervals have nearly nominal coverage despite being narrower. Finally, smoothing simplifies interpretation; while the Unsmoothed analysis has a scattered group of estimates, Analysis 2 suggests that F's levels 4 and 7 have no treatment effect, level 2 has a standard-to-new ratio of about $1/4$ ($\log_{10} .25 = -.6$), and F's other levels cluster around a ratio of about $1/2$ ($\log_{10} .5 = -0.3$).

5. DISCUSSION

We have presented a way to smooth balanced ANOVAs with one error term. We focused on interactions, but smoothing main effects is a trivial extension. Section 4 showed how

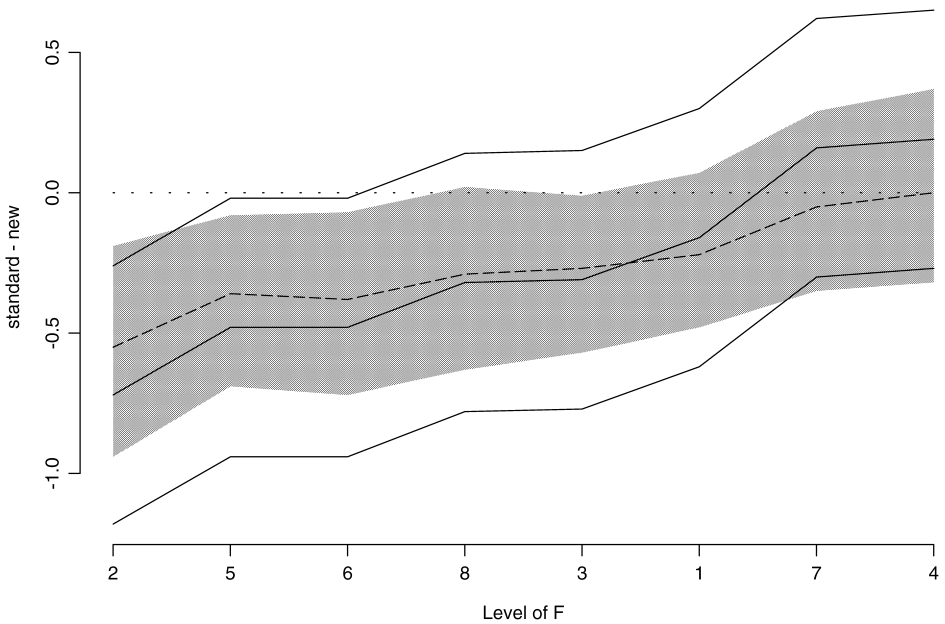

Figure 8. Polishability Data: Subgroup Analysis, Standard Material Minus New Material for the Levels of Factor F. Unsmoothed analysis: solid lines, point estimates and 95% confidence interval. Smoothed Analysis 2: dashed line, posterior mean; gray region, 95% posterior intervals. The dotted line indicates no difference between standard and new.

smoothed ANOVA addresses three common problems: unreplicated designs, masking of large contrasts, and subgroup analysis. Our approach allows flexibility in grouping contrasts to be smoothed by the same η_j and in choosing priors to control smoothing. Section 3's simulation experiment compared four priors, unsmoothed ANOVA, and a strategy of dropping nonsignificant interactions. The three smooth priors (gamma, flat, and beta) had performance advantages over nonsmoothed ANOVA and few significant disadvantages for zero, one, and two truly present interactions. The two-point prior and the strategy of dropping nonsignificant interactions entailed serious problems when error variation was large.

We considered only iid priors on the q_j or η_j, but other priors are possible. Spatial smoothing for lattice data fits easily into our framework using, for example, an improper conditional autoregressive (CAR) prior reexpressed as a proper prior as done by Hodges, Carlin, and Fan (2003). This is a case in which smoothing a spatial main effect is clearly indicated and smoothing considerably simplifies the effects. Similarly, the response surface priors considered by Smith (1973) fit into our framework, allowing an effect to be smoothed toward, but not forced to fit, a polynomial.

Nobile and Green (2000) presented a Bayesian ANOVA for a two-way design in which they modeled each of the row and column factors and the interaction as draws from finite mixtures of normally distributed components. The prior structure for (say) the row factor enforces small variation between levels from the same mixture component and large variation between components, so that levels from the same component are "practically indistinguishable." Differences between levels are tested using the posterior probabilities of the allocations of levels to components. The theory handles unbalanced or unreplicated designs as well as different error variances in each design cell. This bounty comes at the price of a prior more complex than ours, which we find daunting despite Nobile and Green's ingenious interpretation. Also, although how to place an order constraint on (say) the row factor in this approach seems clear, how to smooth a spatial effect is unclear.

Gelman (2005) stated that his fertile treatment of ANOVA derived from an early version of the present article, but his goal was more ambitious than and rather different from ours. He aimed to reinterpret ANOVA, to treat all effects as random effects and replace F-tests with variance-component estimation, and to make error-term selection implicit and automatic. In contrast, we consider only single–error term ANOVAs; in ongoing work on multiple–error term ANOVAs, we observe the usual distinction between fixed and random effects. However, having labeled each effect as fixed or random, we focus on estimation so we too let our machinery select an error term automatically, as in the approach of Gelman (2005) and the SAS MIXED procedure, among others. Moreover, Gelman (2005) smoothed each effect's θ_k using one η_j. He defended this off-the-shelf ANOVA as "a tool for data exploration . . . used to construct useful models," and in this sense, our machinery may be viewed as providing some useful models.

In our approach to smoothed ANOVA, the prior distribution on the smoothing structure is important. The three smooth priors (gamma, flat, and beta) had similar performance in the simulation experiment, giving little basis for a preference. Many hierarchical model applications have meaningful data about second-level precisions, but in smoothed ANOVA, these precisions are purely a device to induce smoothing. This suggests that priors on df or other scale-invariant quantities will ultimately prevail as reference priors.

The scope of our simulation experiment is limited in that it considers only effects with 1 df, each with its own η_j. But the prior should have its greatest effect in this case because the datum for each η_j is a single contrast in the cell averages. Of course, the information in this contrast depends symmetrically on the error precision η_0, a design factor in our experiment, and on the per-cell sample size, which we fixed at $n = 6$.

A prior on q_j induces a prior on η_j and vice versa. For a 2^3 ANOVA with $n = 6$, giving the η_j a gamma(.001, .001) prior induces a prior on each q_j marginally that is well approximated by a beta(.001, .001) distribution, which is in effect a two-point prior. This gamma prior on the η_j also induces a prior on any pair of q_j with probability near 1 on the unit square's perimeter. This happens because gamma(.001, .001) puts nearly all its probability extremely close to 0, with a tiny bit on huge values. Thus for independently drawn η_0 and η_j, with rare exceptions both are microscopically small but differ by orders of magnitude, so their ratio r_j is effectively 0 or infinite. In the simulation experiment, gamma dominated two-point for small η_0 apparently because two-point placed *independent* two-point priors on the q_j, whereas gamma induced *correlated* two-point priors on the q_j.

As presented here, smoothed ANOVA has some potentially undesirable features. Section 2's tidy theory requires an orthogonal parameterization, and in general smoothing depends on the specific orthonormal parameterization. We know of one exception, when an effect's contrasts are treated as iid with a single smoothing precision (e.g., P × F in Sec. 4). Otherwise, a different design matrix implies different smoothing, although we conjecture that this effect is minor for smooth priors. Certainly it is advisable to use contrasts motivated by subject matter considerations whenever possible.

We considered the balanced case here. Unbalanced cases have no universally accepted ANOVA even without smoothing. Of course, one could order the effects, orthogonalize columns with respect to preceding columns, and apply our approach, which amounts to smoothing SAS's type I analysis. Smoothing unbalanced ANOVA more thoughtfully will require a different treatment of variation in the dependent variable $\mathbf{y}$ than is "claimed" by collinear effects.

Finally, our approach could be extended to a Lasso-like procedure (Tibshirani 1996) in which the θ_k's have double-exponential priors, and the posterior is maximized. As with the Lasso, most contrasts would be shrunk to 0, whereas shrinkage of the other contrasts would depend on the prior on the double-exponential's scale. A final extension is to balanced ANOVAs with two or more error terms; preliminary work indicates that results as explicit as those in Section 2 are possible.

ACKNOWLEDGMENTS

This work was supported in part by National Institute of Allergy and Infectious Diseases (NIAID) contract NO1-AI05073 and NIAID Graduate Training in Biostatistics grant 1-T32-

A107432-01A1. The authors thank Tom Louis, the RAND Statistics group, and two referees for helpful comments and Dr. Igor Pesun of the School of Dentistry, University of Minnesota, for permission to use the polishability data.

APPENDIX A: DETAILS FOR SECTIONS 2 AND 3

A.1 Computational Forms for Expressions in Section 2.2

Simplified forms of Section 2.2's expressions are easily derived; we omit the derivations. First,

$$(\mathbf{X}'Q^{-1}\mathbf{X})^{-1} = \left[\begin{array}{c|c} (cn)^{-1}I_M & 0_{M\times N} \\ \hline 0_{N\times M} & \operatorname{diag}(cn + r_{j(k)})^{-1} \end{array}\right]; \quad \text{(A.1)}$$

therefore,

$$\hat{\Theta} = \left[\begin{array}{c} (cn)^{-1}A_1'y \\ \hline \operatorname{diag}(cn + r_{j(k)})^{-1}A_2'y \end{array}\right]. \quad \text{(A.2)}$$

The upper part of $\hat{\Theta}$—the main effects' conditional posterior mean—does not depend on $\mathbf{r}$. In the lower part, the diagonal matrix depends on $\mathbf{r}$, whereas $A_2'y$ does not. Similarly,

$$W(\mathbf{r}) = y'y - (cn)^{-1}y'A_1A_1'y - y'A_2\operatorname{diag}\big(cn + r_{j(k)}\big)^{-1}A_2'y. \quad \text{(A.3)}$$

The first two terms of $W(\mathbf{r})$ do not depend on $\mathbf{r}$ and are the usual residual sum of squares in a main effects–only ANOVA. The Bayesian analysis gives this because of the flat prior on Θ_1. The last term in (A.3) involves $\mathbf{r}$ and can be written as

$$\sum_k \frac{1}{cn + r_{j(k)}}(y'A_2)_k^2, \quad \text{(A.4)}$$

where again $(y'A_2)_k$ does not depend on $\mathbf{r}$.

The conditional posterior of Θ given $\mathbf{r}$ has center $\hat{\Theta}$, given in (A.2). Its dispersion matrix is

$$\frac{2\lambda + W(\mathbf{r})}{\nu}(\mathbf{X}'Q^{-1}\mathbf{X})^{-1}, \quad \text{(A.5)}$$

which simplifies using (A.3) and (A.1). Posterior variances for individual θ_k can be derived using $\operatorname{var}(\theta_k) = E(\operatorname{var}(\theta_k|\mathbf{r})) + \operatorname{var}(E(\theta_k|\mathbf{r}))$, where the outer expectation and variance are with respect to $\mathbf{r}$. For main effect θ_k, $E(\theta_k|\mathbf{r})$ is independent of $\mathbf{r}$, so its variance is 0. For these θ_k's,

$$\operatorname{var}(\theta_k|\mathbf{r}) = \frac{2\lambda + W(\mathbf{r})}{(\nu - 2)cn}, \quad \text{(A.6)}$$

from (A.5); this is the same for each main effect θ_k. $E(\operatorname{var}(\theta_k|\mathbf{r}))$ can be estimated by the average of (A.6) evaluated at the MCMC draws of $\mathbf{r}$. For interaction θ_k,

$$E(\theta_k|\mathbf{r}) = \frac{1}{cn + r_{j(k)}}\{A_2'y\}_k, \quad \text{(A.7)}$$

so $\operatorname{var}(E(\theta_k|\mathbf{r}))$ is

$$\left(E\left[\frac{1}{(cn + r_{j(k)})^2}\right] - E^2\left[\frac{1}{cn + r_{j(k)}}\right]\right)\{A_2'y\}_k^2; \quad \text{(A.8)}$$

both expectations can be estimated by the obvious averages. From (A.5),

$$\operatorname{var}(\theta_k|\mathbf{r}) = \frac{2\lambda + W(\mathbf{r})}{(\nu - 2)(cn + r_{j(k)})}, \quad \text{(A.9)}$$

the expectation of which can be estimated by the obvious average.

A.2 The SANOVA Table

In an ordinary one–error term ANOVA, the usual projection theory analyzes $\mathbf{y}$ as

$$\mathbf{y} = P_u\mathbf{y} + P_s\mathbf{y} + Q_A\mathbf{y}, \quad \text{(A.10)}$$

where $P_u = \frac{1}{cn}A_1A_1'$ and $P_s = \frac{1}{cn}A_2A_2'$ are orthogonal projections onto the columns of A_1 and A_2 and $Q_A = I_{cn} - \frac{1}{cn}(A_1A_1' + A_2A_2')$ is the residual projection. P_u, P_s, and Q_A have dimension cn. Smoothing of interactions is captured by splitting $P_s\mathbf{y}$ into two pieces, $P_s\mathbf{y} = P_{ss}\mathbf{y} + P_{se}\mathbf{y}$, the smoothed part in the fit and the error part.

To do this, extend (A.10) to Section 2.1's constraint-case formulation as

$$\left[\begin{array}{c} \mathbf{y} \\ 0_{N\times 1} \end{array}\right] = Y = P_uY + P_sY + Q_AY, \quad \text{(A.11)}$$

where each projection matrix now has dimension $cn + N$. Now P_u is partitioned into four parts; the upper left part is the original P_u in (A.10), whereas the other parts are appropriate-sized zero matrices. P_s and Q_A are related to their counterparts in (A.10) in the same way. Split P_s, treating Σ as fixed for now. Define $X_s' = [A_2'|I_N]$, the columns of X corresponding to the smoothed contrasts. The projection matrix for the smoothed contrasts is

$$P_{ss} = X_s(X_s'\Sigma^{-1}X_s)^{-1}X_s'\Sigma^{-1} \quad \text{(A.12)}$$

$$= \left[\begin{array}{cc} A_2\operatorname{diag}(cn + \frac{\eta_{j(k)}}{\eta_0})^{-1}A_2' & A_2\operatorname{diag}(\frac{\eta_{j(k)}/\eta_0}{cn+\eta_{j(k)}/\eta_0}) \\ \operatorname{diag}(cn + \frac{\eta_{j(k)}}{\eta_0})^{-1}A_2' & \operatorname{diag}(\frac{\eta_{j(k)}/\eta_0}{cn+\eta_{j(k)}/\eta_0}) \end{array}\right] \quad \text{(A.13)}$$

(Hodges and Sargent 2001). Thus the projection for the error part of the smoothed contrasts is

$$P_{se} = P_s - P_{ss} \quad \text{(A.14)}$$

$$= \left[\begin{array}{cc} A_2\operatorname{diag}(\frac{\eta_{j(k)}/\eta_0}{cn(cn+\eta_{j(k)}/\eta_0)})A_2' & -A_2\operatorname{diag}(\frac{\eta_{j(k)}/\eta_0}{cn+\eta_{j(k)}/\eta_0}) \\ -\operatorname{diag}(cn + \frac{\eta_{j(k)}}{\eta_0})^{-1}A_2' & -\operatorname{diag}(\frac{\eta_{j(k)}/\eta_0}{cn+\eta_{j(k)}/\eta_0}) \end{array}\right]. \quad \text{(A.15)}$$

Neither P_{ss} nor P_{se} is an orthogonal projection, although their sum, P_s, is.

The SS $\mathbf{y}'\mathbf{y} = Y'Y$ partitions as

$$\mathbf{y}'\mathbf{y} = Y'Y = Y'(P_u + P_{ss} + P_{se} + Q_A)Y$$

$$= \frac{1}{cn}\mathbf{y}'A_1A_1'\mathbf{y} + \mathbf{y}'A_2\operatorname{diag}\left(cn + \frac{\eta_{j(k)}}{\eta_0}\right)^{-1}A_2'\mathbf{y} \quad \text{(A.16)}$$

$$+ \mathbf{y}'A_2\operatorname{diag}\left(\frac{\eta_{j(k)}/\eta_0}{cn(cn + \eta_{j(k)}/\eta_0)}\right)A_2'\mathbf{y} \quad \text{(A.17)}$$

$$+ \mathbf{y}'\left(I_{cn} - \frac{1}{cn}(A_1A_1' + A_2A_2')\right)\mathbf{y}, \quad \text{(A.18)}$$

where the four terms in the sum correspond to unsmoothed effects, the part of smoothed effects in the fit, the part of smoothed effects considered error, and pure error.

The first three terms in (A.16)–(A.18) have the form $\mathbf{y}'A_iG \times A_i'\mathbf{y}$, for diagonal G. Because $A_i'\mathbf{y}$ is a vector, these terms can be written as $\sum_k G_k(A_{ik}'\mathbf{y})^2$, where G_k is G's kth diagonal entry and A_{ik} is A_i's kth column. Thus each SS decomposes into

summands for the individual columns of A_i, which a SANOVA table can group in any convenient way.

The traces of P_u and P_{ss} give each effect's df in the fit (Hodges and Sargent 2001). For unsmoothed effects and pure error, the mean squares are as usual. For any grouping of smoothed contrasts, the mean square for the part in the fit is the part of SS in the fit divided by the part of df in the fit, with the obvious analog for the part of the effect smoothed into error.

For smoothed effects, the df in the fit and error are functions of Σ, as is the partition of the SS. Because Σ is unknown and has a posterior distribution, so do each effect's df and SS in the fit and in error. In a SANOVA table, it is much simpler to use single-number summaries of df and SS instead of a distribution. Here are three candidate summaries, all easily computed by MCMC:

a. Compute the DF and SS at the posterior median of each r_j.
b. Compute the DF and SS at the posterior mean of each r_j.
c. Use the posterior means of the DF and SS themselves.

The r_j tend to have posteriors with long upper tails, because in hierarchical models the data usually provide little information about higher-level variances. Thus r_j's posterior mean is likely to be much larger than its median and less representative of the distribution's "middle." Option (c) is less subject to this problem; each smoothed effect's df and SS lie in closed intervals, so their median and mean cannot differ as much as r_j's. In this sense, (a) and (c) seem to dominate (b). Because expectations of sums are sums of expectations, the posterior mean df and SS always add properly for model and error parts and across effects. Given the ANOVA table's bookkeeping function, (c) thus seems to dominate (a). With (c) as the single-number summary, if a smoothed contrast is not grouped with other contrasts, then its model and error parts have mean squares identical to each other and to the contrast's conventional mean square.

A.3 MCMC Algorithms for Bayesian Analyses

We used the following procedures but claim no optimality. Our algorithms draw samples from the marginal posterior of $\mathbf{r}$, the vector of smoothing ratios, which are then used with $f(\Theta|Y, \mathbf{r})$ and $f(\eta_0|Y, \mathbf{r})$ to Rao–Blackwellize. First, we give an algorithm suitable for unconditional priors, then a modification that works better with priors conditioned on df.

Case 1. Unconditional Priors. The chain is on $z_j = \log(r_j)$. To start, set z_j so $q_j = .5n_j$. For each cycle, update each z_j conditional on the other z_j using this Metropolis–Hastings step:

1. $z_{j,old} \leftarrow$ current value of z_j.
2. For $t = 1, \ldots, T$,
   a. $z_{j,t} \leftarrow z_{j,old} + C \times \kappa$, where $\kappa \sim \mathrm{N}(0, 1)$
   b. With probability $\min(1, \text{MH ratio for } z_{j,t})$, $z_{j,old} \leftarrow z_{j,t}$.
3. Return $z_{j,old}$.

In Section 3's simulation experiment, each analysis used this algorithm for 12,000 cycles with $T = 2$ and $C = 3$, discarding the first 2,000 draws as burn-in.

Case 2. Prior Conditioned on df. For r_j not affected by a prior conditioned on df, use Case 1's algorithm. The algorithm that follows describes draws for a group of r_j's whose q_j's add up to K. If several groups of r_j's have such constraints, then apply this algorithm separately to each group. An inequality condition on df raises no distinct issues.

Again, the chain is on $z_j = \log(r_j)$. In each MCMC cycle, the condition on $\mathbf{z}$ is handled by randomly selecting one z_j, $z_{j'}$, to use as a "pivot" in drawing $z_j, j \neq j'$. Specifically, draw z_l given $\{z_j | j \neq l, j'\}$ and adjust the pivot $z_{j'}$ according to the drawn z_l, so that $q_l + q_{j'} = K - \sum_{j \neq l,j'} q_j$. The df condition limits z_l's sample space. The specific algorithm follows.

Index the affected r_j by $j = 1, \ldots, B$. To start, set each z_j so that q_j is K/B. Then each cycle through the z_j includes the following steps:

1. Randomly select a pivot index j' from $\{1, \ldots, B\}$, with the probability of index j proportional to $q_j^*(1 - q_j^*)$, q_j^* being the current value of q_j (a function of the current z_j).
2. Randomly permute the remaining index values $\{1, 2, \ldots, j' - 1, j' + 1, \ldots, B\}$.
3. Sample z_l given $\{z_j | j \neq l, j'\}$ in the order selected in the previous step. Use Case 1's Metropolis–Hasting step, with two changes:
   a. z_l's conditional sample space is bounded so at least one $z_{j'}$ satisfies $q_l + q_{j'} = K - \sum_{j \neq l,j'} q_j$.
   b. The pivot $z_{j'}$ is updated using the df constraint and returned along with $z_{j,old}$.

Originally, we used z_B as the pivot. For datasets strongly indicating more than K df, each q_j's samples quickly moved close to 0 or 1 and could not move away. Some might take this to mean that they should use a different prior, but for overpowered experiments, one might prefer to avoid trivial detail in the fit by imposing more smoothing than the data would suggest.

A.4 Some Details About Simulated Datasets and Monte Carlo Error

In the simulation experiment, we created artificial datasets as follows. Consider simulating datasets with two truly present interactions and $\eta_0 = .25$. We set two interaction θ_k's to 1 and all other θ_k's to 0. We multiplied the 1,000 "subjects" (sets of 48 errors) by 2 to give error precision .25 and added them to the true means to give 1,000 artificial datasets. We then applied the six procedures to each artificial dataset. For the Bayesian procedures, we used the MCMC algorithm in Section A.3 (Case 1). We used true parameter values as starting values, sampled 10,000 $\mathbf{r}$'s from their marginal joint posterior (8), and then Rao–Blackwellized θ_k and η_0 to estimate posterior means and interval coverages. Preliminary testing indicated that this was sufficient.

The contribution of MCMC error to overall Monte Carlo error was quite small, because it was averaged over 1,000 artificial datasets. For the comparisons in Figures 1 and 2, when comparing two methods according to scaled MSE (i.e., as a percent of true error variance), the contribution of MCMC error is at most .02 percentage points per dataset (i.e., before averaging). The contribution of variation between simulated datasets is somewhat larger, but still small compared with differences

between design cells, because of the simulation experiment's design. For example, in Figure 1 (one interaction truly present), for the gamma method, comparing $\eta_0 = .25$ versus $\eta_0 = .0625$, the two design cells differ in scaled MSE by 1.09 percentage points, and the relevant Monte Carlo standard error is .08 percentage points.

APPENDIX B: THE POLISHABILITY DATASET

Pesun et al. (2002) described this study in detail. Table B.1 gives raw gap measurements (leftmost column) and the design matrix used in Section 4's analyses. Design matrix columns are given for the main effects; columns for interactions (A_2) are

Table B.1. The Polishability Data and Design Matrix for Main Effects

Gap (μm)	*Material*	*Polishing*			*Finishing*						
5.02	1	3	0	0	7	0	0	0	0	0	0
8.84	1	3	0	0	−1	6	0	0	0	0	0
3.61	1	3	0	0	−1	−1	5	0	0	0	0
10.55	1	3	0	0	−1	−1	−1	4	0	0	0
3.90	1	3	0	0	−1	−1	−1	−1	3	0	0
5.64	1	3	0	0	−1	−1	−1	−1	−1	2	0
98.95	1	3	0	0	−1	−1	−1	−1	−1	−1	1
10.75	1	3	0	0	−1	−1	−1	−1	−1	−1	−1
2.91	1	−1	2	0	7	0	0	0	0	0	0
3.00	1	−1	2	0	−1	6	0	0	0	0	0
5.94	1	−1	2	0	−1	−1	5	0	0	0	0
8.64	1	−1	2	0	−1	−1	−1	4	0	0	0
16.33	1	−1	2	0	−1	−1	−1	−1	3	0	0
7.44	1	−1	2	0	−1	−1	−1	−1	−1	2	0
11.26	1	−1	2	0	−1	−1	−1	−1	−1	−1	1
16.35	1	−1	2	0	−1	−1	−1	−1	−1	−1	−1
4.75	1	−1	−1	1	7	0	0	0	0	0	0
3.93	1	−1	−1	1	−1	6	0	0	0	0	0
4.90	1	−1	−1	1	−1	−1	5	0	0	0	0
13.44	1	−1	−1	1	−1	−1	−1	4	0	0	0
2.82	1	−1	−1	1	−1	−1	−1	−1	3	0	0
6.44	1	−1	−1	1	−1	−1	−1	−1	−1	2	0
20.88	1	−1	−1	1	−1	−1	−1	−1	−1	−1	1
9.30	1	−1	−1	1	−1	−1	−1	−1	−1	−1	−1
178.22	1	−1	−1	−1	7	0	0	0	0	0	0
1.95	1	−1	−1	−1	−1	6	0	0	0	0	0
3.70	1	−1	−1	−1	−1	−1	5	0	0	0	0
18.11	1	−1	−1	−1	−1	−1	−1	4	0	0	0
16.40	1	−1	−1	−1	−1	−1	−1	−1	3	0	0
9.61	1	−1	−1	−1	−1	−1	−1	−1	−1	2	0
36.52	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1
14.88	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
18.68	−1	3	0	0	7	0	0	0	0	0	0
49.02	−1	3	0	0	−1	6	0	0	0	0	0
4.55	−1	3	0	0	−1	−1	5	0	0	0	0
10.85	−1	3	0	0	−1	−1	−1	4	0	0	0
20.04	−1	3	0	0	−1	−1	−1	−1	3	0	0
5.65	−1	3	0	0	−1	−1	−1	−1	−1	2	0
47.00	−1	3	0	0	−1	−1	−1	−1	−1	−1	1
34.14	−1	3	0	0	−1	−1	−1	−1	−1	−1	−1
8.12	−1	−1	2	0	7	0	0	0	0	0	0
10.30	−1	−1	2	0	−1	6	0	0	0	0	0
10.10	−1	−1	2	0	−1	−1	5	0	0	0	0
1.11	−1	−1	2	0	−1	−1	−1	4	0	0	0
21.49	−1	−1	2	0	−1	−1	−1	−1	3	0	0
19.02	−1	−1	2	0	−1	−1	−1	−1	−1	2	0
13.49	−1	−1	2	0	−1	−1	−1	−1	−1	−1	1
48.75	−1	−1	2	0	−1	−1	−1	−1	−1	−1	−1
37.62	−1	−1	−1	1	7	0	0	0	0	0	0
36.22	−1	−1	−1	1	−1	6	0	0	0	0	0
10.58	−1	−1	−1	1	−1	−1	5	0	0	0	0
11.60	−1	−1	−1	1	−1	−1	−1	4	0	0	0
33.44	−1	−1	−1	1	−1	−1	−1	−1	3	0	0
51.28	−1	−1	−1	1	−1	−1	−1	−1	−1	2	0
24.19	−1	−1	−1	1	−1	−1	−1	−1	−1	−1	1
23.25	−1	−1	−1	1	−1	−1	−1	−1	−1	−1	−1
9.75	−1	−1	−1	−1	7	0	0	0	0	0	0
8.38	−1	−1	−1	−1	−1	6	0	0	0	0	0
14.23	−1	−1	−1	−1	−1	−1	5	0	0	0	0
27.90	−1	−1	−1	−1	−1	−1	−1	4	0	0	0
16.72	−1	−1	−1	−1	−1	−1	−1	−1	3	0	0
37.83	−1	−1	−1	−1	−1	−1	−1	−1	−1	2	0
12.51	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1
11.51	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1

constructed by multiplying main-effect columns. For the Material factor, $+1$ corresponds to standard and -1 corresponds to new. The columns in Table B.1 are *not* standardized so that $A_i'A_i = 64I$, as the theory in Section 2.1 requires.

[Received September 2005. Revised May 2006.]

REFERENCES

Daniels, M. (1999), "A Prior for the Variance in Hierarchical Models," *Canadian Journal of Statistics*, 27, 567–578.

Derksen, S., and Keselman, H. J. (1992), "Backward, Forward and Stepwise Automated Subset Selection Algorithms: Frequency of Obtaining Authentic and Noise Variables," *British Journal of Mathematics and Statistics in Psychology*, 45, 265–282.

Dixon, D. O., and Simon, R. (1991), "Bayesian Subset Analysis," *Biometrics*, 47, 871–881.

Freedman, D. A. (1983), "A Note on Screening Regression Equations," *The American Statistician*, 37, 152–155.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A. (2005), "Analysis of Variance—Why It Is More Important Than Ever" (with discussion), *The Annals of Statistics*, 33, 1–53.

Hodges, J. S. (1998), "Some Algebra and Geometry for Hierarchical Models, Applied to Diagnostics" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 60, 497–536.

Hodges, J. S., and Sargent, D. J. (2001), "Counting Degrees of Freedom in Hierarchical and Other Richly Parameterised Models," *Biometrika*, 88, 367–379.

Hodges, J. S., Carlin, B. P., and Fan, Q. (2003), "On the Precision of the Conditionally Autoregressive Prior in Spatial Models," *Biometrics*, 59, 317–322.

Leamer, E. E. (1978), *Specification Searches*, New York: Wiley.

Lee, Y., and Nelder, J. A. (1996), "Hierarchical Generalized Linear Models" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 58, 619–678.

Nobile, A., and Green, P. J. (2000), "Bayesian Analysis of Factorial Experiments by Mixture Modelling," *Biometrika*, 87, 15–35.

Pesun, I. J., Hodges, J. S., and Lai, J. H. (2002), "Effect of Finishing and Polishing Procedures on the Gap Width Between a Denture Base Resin and Two Long-Term Resilient Denture Liners," *Journal of Prosthodontic Dentistry*, 87, 311–318.

Raftery, A. E., Madigan, D., and Hoeting, J. (1993), "Model Selection and Accounting for Model Uncertainty in Linear Regression Models," Technical Report 262, University of Washington, Dept. of Statistics.

Scheffé, H. (1959), *The Analysis of Variance*, New York: Wiley.

Smith, A. F. M. (1973), "Bayes Estimates in the One-Way and Two-Way Models," *Biometrika*, 60, 319–329.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

Whittaker, J. (1998), Discussion of "Some Algebra and Geometry for Hierarchical Models, Applied to Diagnostics," by J. S. Hodges, *Journal of the Royal Statistical Society*, Ser. B, 60, 533.