Smoothed ANOVA with Application to Subgroup Analysis

By DANIEL J. SARGENT[1] and JAMES S. HODGES[2]

[1]*Department of Biostatistics, Mayo Clinic*

[2]*Division of Biostatistics, School of Public Health, University of Minnesota*

Correspondence Author: Daniel J. Sargent
Department of Biostatistics, Mayo Clinic
200 First Street SW, Rochester, Minnesota, 55905, U.S.A.
telephone: (507) 284-5380
fax: (507) 284-1902
email: sargent.daniel@mayo.edu

February 14, 1997

Smoothed ANOVA with Application to Subgroup Analysis

ABSTRACT

A major issue in the analysis of variance (ANOVA) and associated modeling is how to handle interactions. One way to fit models with many interaction terms is to use a hierarchical model, where terms remain in the model if they are important predictors, are removed if they are not, and are partly smoothed away if the data are indecisive. Recent advances in computing make it straightforward to fit such models. In this paper, we propose a method for specifying and fitting ANOVA-related models that has as its goal using all of the information provided by the data, while avoiding assumptions about the presence or absence of certain terms such as interactions. We investigate this new method's properties through simulation and examples.

Key Words: Hierarchical model, Gibbs sampling, Smoothing

1 Introduction

Techniques for analyzing differential responses due to controlled experimental predictors have been a central technique in applied statistics in this century. A particular technique, the analysis of variance (ANOVA; Scheffé 1959), has been studied extensively. ANOVA is concerned with assessing whether variations in a response can be attributed to one or more factors or to combinations of these factors. A major issue in the analysis of variance is how to handle the many possible combinations of the factors, that is, the interactions of the factors with one another.

For example, consider a 2^3 factorial experimental design, with factors x_1, x_2, and x_3 taking on the values 0 or 1; the resulting design cells are show in Table 1. An ANOVA of such a design considers, among other things, the saturated linear model

$$y_l = \sum_{i,j,k} \theta_{ijk} w_{ijkl} + \epsilon_l \tag{1}$$

where i, j, and k take the values 0 or 1 and jointly designate the design cell, w_{ijkl} is 1 if observation l is in design cell ijk and 0 otherwise, and $\epsilon_l \sim N(0, \sigma^2)$. If the main effects are present but no interaction is present, the θ_{ijk} are related in simple ways:

$$\begin{aligned} &\theta_{111} - \theta_{011} = \theta_{110} - \theta_{010} = \theta_{101} - \theta_{001} = \theta_{100} - \theta_{000}, \\ &\theta_{111} - \theta_{101} = \theta_{110} - \theta_{100} = \theta_{011} - \theta_{001} = \theta_{010} - \theta_{000}, \quad \text{and} \\ &\theta_{111} - \theta_{110} = \theta_{101} - \theta_{100} = \theta_{011} - \theta_{010} = \theta_{001} - \theta_{000}, \end{aligned} \tag{2}$$

corresponding to the main effects of x_1, x_2, and x_3, respectively. The presence of an interaction breaks some of these equalities, thus complicating interpretation. This leads to the common question: which interactions should be in the model?

In this paper, we are especially concerned with one particular instance of this question: subgroup

design cell	x_1	x_2	x_3
111	1	1	1
110	1	1	0
101	1	0	1
100	1	0	0
011	0	1	1
010	0	1	0
001	0	0	1
000	0	0	0

Table 1: Definition of design cells based on binary covariates x_1, x_2, and x_3

or subset analysis (Louis 1984; Oxman and Guyatt 1992; Pocock, Hughes, and Lee 1987; Thomas et al. 1985). If x_1 indicates a treatment assignment, subgroup analysis is concerned with whether the treatment effect – the effect of x_1 – is the same in the $2^2 = 4$ groups of units defined by the combinations of x_2 and x_3. This is precisely the question of which interactions are present.

A common approach to this problem is a stepwise method, where terms enter the model based on significance tests. However, a large literature (Derksen and Keselman 1992; Freedman 1983; Leamer 1978; Raftery, Madigan, and Hoeting 1993) shows that stepwise procedures are beaten by model-averaging and smoothing methods. In this paper, we investigate a smoothing method for ANOVA and subgroup analysis that neither includes or excludes the interactions, but does something intermediate, as shrinkage estimators do. Section 2 motivates our method; Section 3 describes it and gives a method for fitting it to data. Section 4 simulates the new method's performance and compares it to some competing methods. Section 5 provides an example, and Section 6 describes extensions of the method to more complicated cases. Finally, Section 7 discusses the results and gives areas for future research.

2 Background and motivation

Our method extends an approach with roots in the 1950's. The common thread in this approach is models with many parameters – in our case, one per cell of the design – where the parameters are in turn treated as realizations of a linear model. Such models are a special case of hierarchical linear models (Bryk and Raudenbush 1992). Various methods have been used to fit these models, but all such methods induce shrinkage.

2.1 James-Stein estimators

Stein (1955) showed that for estimating the θ_{ijk} in (1), the cell averages are inadmissible for total squared error risk. The James-Stein estimator (James and Stein 1961) shrinks the individual cell averages toward zero; subsequent elaborations (e.g., Efron and Morris 1975) shrink toward the center of the cell averages. The model implicit in this approach is

$$\theta_{ijk} \sim N(\mu, \tau^2) \tag{3}$$

where the assumption about μ varies depending on the method. Many shrinkage estimators can be viewed as ways to fit this model.

While celebrated in theory, these methods have not been widely used. Shrinking each θ_{ijk} toward zero has little appeal. In many problems, particularly subgroup analysis, shrinking the cell averages toward a common center is hardly better, because if a single interaction is in fact present it will often be obscured.

2.2 Morris's method

Morris (1983) outlined an empirical Bayes (henceforth EB) approach to this problem, appropriate when one of the explanatory variables, say x_1, is a treatment indicator. The goal of Morris's method is estimating treatment effects in subgroups defined by the levels of x_2 and x_3. If we define the subgroup-specific treatment effects $\beta_{11} = \theta_{111} - \theta_{011}, \beta_{10} = \theta_{110} - \theta_{010}, \beta_{01} = \theta_{101} - \theta_{001}$, and $\beta_{00} = \theta_{100} - \theta_{000}$, then the model behind this method is

$$\beta_{jk} \sim N(B, \Lambda). \tag{4}$$

The fitting method is EB: estimate B and Λ and then estimate the β_{jk} as if the estimates for B and Λ are true. Specifically, call the unshrunken subgroup-specific treatment effects Y_{jk} and assume that Y_{jk} is normal with mean β_{jk} and known variance σ^2_{jk}. Then the EB estimate of β_{jk} is

$$\begin{aligned}
\hat{\beta}_{jk} &= Y_{jk} - s_{jk}(Y_{jk} - m), \\
\text{where} \quad s_{jk} &= \tfrac{p-3}{p-1} \frac{\sigma^2_{jk}}{\sigma^2_{jk} + \hat{\Lambda}} \\
m &= \frac{\sum_{j,k} v_{jk} Y_{jk}}{\sum_{j,k} v_{jk}}, \\
v_{jk} &= (\sigma^2_{jk} + \hat{\Lambda})^{-1}, \\
\text{and} \quad \hat{\Lambda} &= \max\left\{0, \left[\textstyle\sum_{j,k} v_{jk} [\frac{p}{p-1}] (Y_{jk} - \hat{\mu})^2 - \sigma^2_{jk}\right] / \textstyle\sum_{j,k} v_{jk}\right\}.
\end{aligned}$$

In this method, the subgroup treatment effects (the β_{jk}) are shrunk toward a single mean. If a non-treatment factor and the treatment interact, then this modeling assumption – that the β_{jk} are draws from the same distribution – is incorrect, and in fact different β_{jk} should be shrunk towards different centers. The EB method of Louis (1984) also suffers from this limitation.

2.3 Dixon and Simon's method

Dixon and Simon (1991; henceforth DS) proposed a method for subgroup analysis; Simon, Dixon, and Friedlin (1996) extended it. DS propose using a model with all of the main effects but only certain interactions, specifically, the interactions of each non-treatment covariate with the treatment indicator. To induce shrinkage, DS build a hierarchical model on the interactions. For the 2^3 factorial design, their model is

$$\theta_{ijk} = \mu + \delta_i\zeta_1 + \delta_j\zeta_2 + \delta_k\zeta_3 + \delta_i\delta_j\xi_{12} + \delta_i\delta_k\xi_{13}, \tag{5}$$

where δ_a takes the value 1 if covariate $a = 1$ and 0 otherwise, ζ_s is the main effect of covariate s, and the $\xi_r, r = 12, 13$, are the treatment-by-covariate interactions. DS induce shrinkage by assuming that $\xi_r \sim N(0, \eta)$, where η is unknown, and using a fully Bayesian method to fit their model.

A limiting assumption of this model is that the ξ_r are drawn from the same distribution. If, for example, x_1 and x_2 interact but x_1 and x_3 do not, then ξ_{12} is non-zero while ξ_{13} is zero. Requiring both ξ_r to be draws from the same normal distribution will cause their estimates to be shrunk toward their average and away from the truth. Additionally, DS's method does not contain a three-way interaction term or a term for an interaction between x_2 and x_3. If these effects are present, the DS method will be unable to adjust to them.

2.4 Motivating our methods

The methods just described have two common features. First, implicitly or explicitly, each treats some group of parameters as draws from a single distribution: for the 2^3 factorial, the model behind the James-Stein method draws eight θs from a single normal distribution, while Morris's method draws four βs and DS's method draws two ξs. Second, each method uses a problem-specific fitting

method. James-Stein and descendants like Efron and Morris (1975) are *ad hoc*, though intended to have specific properties; Morris's method uses a general approach (maximum likelihood) but a problem-specific algorithm; and DS, though they used Bayesian methods, were constrained by the computing methods available to them. Our method departs from both of these common features by more flexible modeling and a generic fitting method.

3 The constraint method of ANOVA

In this section, we describe our method for the 2^3 factorial design; extensions to more general designs are considered in Section 6.

3.1 Hierarchical models

Consider the features one might want in an procedure that fits ANOVA models. We seek information about how experimental units respond to factors. We expect that this response will tend to be similar in cells differing on few of their defining factors, and perhaps more different in cells differing on many of their defining factors. We would like a procedure that uses this expectation without ignoring major signals in the data. We do not want to assume that any specific effect does not exist; nonetheless, most of the interactions are probably not present. Thus, we want a model and fitting procedure that remove unnecessary terms from the model while retaining "important" terms and partly retaining terms of intermediate or indeterminate importance. Such model reduction aids in parameter estimation, by giving smaller standard errors, and in model interpretation, because models without interaction terms are easier to interpret than models containing them.

One way to approach this goal is by using a hierarchical model. To see this, consider the method of DS. Setting η to infinity removes the hierarchical component of their model by leaving the ξ_r unconstrained, in which case DS's model has six degrees of freedom, as $\mu, \zeta_1, \zeta_2, \zeta_3, \xi_{12}, \xi_{13}$ can vary

freely. On the other hand, fixing η at zero means the model has four degrees of freedom, the ξ_r being fixed at zero. Estimating η in the hierarchical model corresponds to smoothly including the interaction terms in the model, and in this sense, η controls the amount of shrinkage in estimating the cell means or interactions. More complicated models could have several variance components, each controlling the shrinkage of a specific aspect of the model.

3.2 The model

Consider the ANOVA model (1). The constraint method makes this a hierarchical model by modeling the interactions as follows:

$$\begin{aligned}
(\theta_{110} - \theta_{010}) - (\theta_{100} - \theta_{000}) &\sim N(0, \sigma^2_{12}) \\
(\theta_{101} - \theta_{001}) - (\theta_{100} - \theta_{000}) &\sim N(0, \sigma^2_{13}) \\
(\theta_{011} - \theta_{001}) - (\theta_{010} - \theta_{000}) &\sim N(0, \sigma^2_{23}) \\
(\theta_{111} - \theta_{011}) - (\theta_{110} - \theta_{010}) - (\theta_{101} - \theta_{001}) + (\theta_{100} - \theta_{000}) &\sim N(0, \sigma^2_{123}),
\end{aligned} \tag{6}$$

where the first equation corresponds to the (x_1, x_2) interaction, the second equation to the (x_1, x_3) interaction, and so on, and the variances σ^2_r are unknown. Each interaction has its own variance component; this is discussed below. This model corresponds to the same kind of parameterization of the main effects and interactions used in DS's method, namely:

$$\theta_{ijk} = \mu + \delta_i\zeta_1 + \delta_j\zeta_2 + \delta_k\zeta_3 + \delta_i\delta_j\xi_{12} + \delta_i\delta_k\xi_{13} + \delta_j\delta_k\xi_{23} + \delta_i\delta_j\delta_k\xi_{123}. \tag{7}$$

Expressed in this way, the model in (6) is identical to the model (7) with $\xi_r \sim N(0, \sigma^2_r), r = 12, 13, 23, 123$. The orthogonal parameterization for ANOVA models, in which the main effects and interactions sum to zero across any single subscript i, j, or k, leads by obvious analogy to a slightly

different model.

Each of (6) and (7) illuminates an aspect of the constraint method. Model (7) is straightforward: as any of the σ_r^2 becomes smaller, its associated interaction is smoothly removed from the model, and as that variance approaches infinity, its interaction smoothly becomes unconstrained. In this sense, appropriate prior distributions on the variance components act as "accelerators" on the interaction estimates. If an interaction looks small, our procedure will tend to shrink it towards zero. If an interaction looks large, our procedure will largely leave it alone, and if an interaction is somewhere in the middle, it will be shrunk by an intermediate amount.

Model (6) gives insight into how the shrinkage affects estimation of the θ_{ijk}, the cell means. If all the interactions are shrunk away, all of the subgroups formed using x_2 and x_3 have the same treatment effect, given by the first equation in (2). But if, for example, only the (x_1, x_2) interaction is absent – that is, if the treatment effect does not depend on x_2 – only the simple effects $\theta_{110} - \theta_{010}$ and $\theta_{100} - \theta_{000}$ are forced to be equal.

The hierarchical model proposed here has one observation per variance component, but the variances are identified because the normal distributions in (6) have mean zero. We will define our inferential procedure using the Bayesian formalism, which requires a prior distribution for each variance component; thus, we have prior information for each variance component as well. The prior distribution strongly influences the amount of shrinkage; this is discussed in Sections 3.4, 4, and 7.

An important feature of the constraint method is that it treats each of the covariates equally, that is, there is no need to define one of them to be the "treatment". Therefore, the constraint method can be used to estimate any contrast, including subgroup-specific treatment effects.

3.3 Model fitting

Fitting this model to the data is straightforward once we re-express it as in Hodges (1996). The base for the hierarchical model is (1), which can be written as

$$\mathbf{y} = X_1\Theta + E_1, \tag{8}$$

where $\mathbf{y}$ is the n-vector containing the y_l, X_1 is the $n \times 8$ matrix with l^{th} row $(w_{111l}, \ldots, w_{000l})$, $\Theta = (\theta_{111}, \ldots, \theta_{000})^T$, and E_1 contains the error terms ϵ_l. The hierarchical constraints are given in (6) and, following Hodges (1996), can be re-written in matrix notation as

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 & 0 & 1 & 0 & 1 & 0 & -1 \\ 0 & 0 & -1 & 1 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \end{bmatrix} \Theta + E_2 \tag{9}$$

where E_2 is a 4-dimensional normal vector with mean 0 and covariance matrix Γ_2 having diagonal elements $\sigma^2_{12}, \sigma^2_{13}$, σ^2_{23}, and σ^2_{123} and zeroes off the diagonal.

We can express (9) as

$$\mathbf{0} = Z\Theta + E_2 \tag{10}$$

and then combine equations (8) and (10) into the linear model

$$\begin{bmatrix} \mathbf{y} \\ \hline \mathbf{0} \end{bmatrix} = \begin{bmatrix} X_1 \\ \hline Z \end{bmatrix} \Theta + \begin{bmatrix} E_1 \\ \hline E_2 \end{bmatrix} \qquad \text{or} \qquad Y = X\Theta + E, \tag{11}$$

where X and Y are known, Θ is unknown, E is the error term, and we define $\mathrm{cov}(E) = \Gamma$,

which contains unknown parameters. Using the terminology of Hodges (1996), rows of X and Y corresponding to X_1 in (11) are called "data cases"; they comprise the terms in the joint posterior into which the data enter directly. Rows of X and Y corresponding to Z are called "constraint cases" because they place restrictions – stochastic constraints – on Θ. Equation (9) defines the variance components $\sigma^2_{123}, \sigma^2_{12}, \sigma^2_{13}$, and σ^2_{23} as constraint case variances. Note that DS's method simply specifies a different set of constraint cases, while Morris's method specifies a third set.

To use the Bayesian formalism, we need prior distributions for $\sigma^2, \sigma^2_{12}, \sigma^2_{13}, \sigma^2_{23}$, and σ^2_{123}, and for the four main-effect contrasts in Θ. We use improper flat priors for the latter and for σ^2, although this is not necessary. (Informative normal priors for the main effects can be added to (11) as constraint cases, as in Hodges 1996.) For computational convenience, we use gamma priors for the constraint case precisions $h_r = \frac{1}{\sigma^2_r}$ with parameters (α_r, ϕ_r). Selection of α_r and ϕ_r is considered in Section 3.4. (A prior for σ^2 of the same form adds no complication.)

Given the foregoing, Θ has full conditional posterior density:

$$\Theta|\ Y, X, \Gamma \sim N((X'\Gamma^{-1}X)^{-1}X'\Gamma^{-1}Y, (X'\Gamma^{-1}X)^{-1}). \tag{12}$$

The gamma priors for the h_r imply that their full conditional posteriors, $g(h_r|\mathbf{y}, X, \Theta)$, are also gamma distributions. With these specifications, the model is easily fit using a Gibbs sampler (Gelfand and Smith 1990) that iterates between drawing Θ given Γ and drawing Γ given Θ. In all cases discussed below, we ran multiple chains and monitored for convergence using the Gelman and Rubin diagnostic (Gelman and Rubin 1992) and within-chain autocorrelations.

3.4 Prior distributions

To fit the model described in Section 3.2 we must specify prior distributions for the constraint case precisions. Prior distributions are typically considered in two ways: either they are viewed as a nuisance and made as uninformative as possible, or they are intended to represent some particular belief. By contrast, we view the prior distribution as providing knobs by which a procedure can be tuned, as one tunes a radio or an engine. The prior influences the method's performance, and we choose a prior to give good performance.

We have found it helpful to consider the mass that a prior allocates to various intervals. Somewhat arbitrarily (but after extensive experimentation), we define precision values in the interval $I_1 = [0, 10]$ to be indicative of little or no shrinkage, precision values in the interval $I_2 = (10, 1000]$ to be indicative of moderate shrinkage, and precision values in the interval $I_3 = [1000, \infty)$ to be indicative of major shrinkage. The choice of these intervals depends on the size and variability of the anticipated interaction effects, but not on σ^2. We use these intervals for anticipated effect sizes on the order of 1, as was the case in our simulations. In Section 4 we investigate the performance of the method under priors that allocate mass in various ways to these intervals. Anticipated effects of different orders would require our priors to be transformed accordingly.

This paper does not fully explore the choice of prior distributions. The priors that we do explore present contrasting benefits, as will be seen in Section 4, and provide a proof of the principle of smoothing ANOVA. Future work will explore more fully the choice of prior distributions.

4 Comparing the methods by simulation

We used simulations to evaluate the various methods according to several criteria, the most important of which is mean squared error of the parameter estimates. For DS's method and

	Effects present			Effect sizes						
Truth	Main effect	Two-way	Three-way	ζ_1	ζ_2	ζ_3	ξ_{12}	ξ_{13}	ξ_{23}	ξ_{123}
1	Yes	None	No	1	0	0	0	0	0	0
2	No	One	No	0	0	0	1	0	0	0
3	No	None	Yes	0	0	0	0	0	0	1
4	No	One	Yes	0	0	0	1	0	0	1
5	No	Two	No	0	0	0	1	1	0	0
6	No	One	No	0	0	0	0	0	1	0

Table 2: Definition of the six truths

the constraint method, we used posterior means as point estimates. An important goal of these simulations is to find robustly-performing prior distributions for the precisions in the constraint method.

The first set of simulations consists of datasets following the structure of Section 1, that is, with eight design cells defined by three binary factors. We use the first covariate x_1 as a treatment indicator so we can compare the constraint method to the DS and EB methods. For each of six "truths" we generated 1000 datasets and applied the various methods to each dataset. Table 2 defines the six truths.

Truths 1, 2, and 3 were chosen for obvious reasons, truth 4 is an example of a more complicated data structure, truth 5 was chosen because DS's method should handle it well, and truth 6 was chosen because it involves the interaction not present in the DS model. In all cases the data were generated using $\sigma^2 = 1$, with 10 observations per subgroup.

The DS and constraint methods require prior distributions for the precision parameters; DS's method has one such precision, while the constraint method has four. In the constraint method, we used the same prior distribution for all four constraint-case precisions, reflecting our desire not to discriminate among the interactions. Other choices are possible; for example, those wishing to shrink the three-way interaction more than the two-way interactions would use a prior for h_{123}

putting more weight on high precisions (low variances) than the priors for h_{12}, h_{13}, and h_{23}.

Both DS's and the constraint method are sensitive to the prior, so we present three priors for each method. The following discussion labels the method and prior using the notation M(κ, π), where M refers to the method (DS for Dixon and Simon, C for constraint), and κ and π refer to the mean and standard deviation, respectively, of the prior for the constraint-case precisions. We considered three priors: $\kappa = \pi = 1$, $\kappa = \pi = 1000$, and $\kappa = 1000$, $\pi = 10000$. The (1,1) prior allocates 99.995% of its mass to $I_1 = [0, 10]$ and nearly all the rest to $I_2 = (10, 1000]$. This prior induces some shrinkage but less than the other two priors. The (1000,1000) prior allocates 0.995% of its mass to I_1 and 62.22% of its mass to I_2; it induces the most shrinkage of our three priors. Finally, the (1000,10000) prior allocates 91.72% of its mass to I_1 and 4.31% of its mass to I_2, which seems to permit both no shrinkage and great shrinkage.

We fit the DS and constraint methods using Gibbs samplers. Test runs for both methods showed rapid convergence, with Gelman and Rubin statistics very close to one after 50 iterations in all cases. Thus for the DS and constraint methods we use iterations 50 to 200 from a run of 200 iterations.

Table 3 presents the average squared error for the four subgroup-specific treatment effects $\beta_{11} = \theta_{111} - \theta_{011}, \beta_{10} = \theta_{110} - \theta_{010}, \beta_{01} = \theta_{101} - \theta_{001}$, and $\beta_{00} = \theta_{100} - \theta_{000}$. Monte Carlo standard errors were small relative to the average squared error; for each method, the maximum Monte Carlo standard error occurred for β_{11} in Truth 4, with the maxima ranging from 0.009 to 0.021. The no-shrinkage results are consistent across truths, as would be expected, with the average squared error hovering around the true mean squared error of 0.2. The EB method provides some benefit compared to no-shrinkage in truths 1 and 6, but no benefit – or apparent penalty – when the treatment and subgrouping variables interact. In the absence of a three-way interaction (truths 1, 2, 5, and 6) the DS method performs better than no-shrinkage, but it suffers when a three-way

interaction is present. Viewed across the six truths, the (1,1) and (1000,10000) priors are clearly superior to the (1000,1000) prior for the DS method.

Turning to the constraint method, C(1,1) is superior to no-shrinkage for every contrast and truth we have examined. However, in many cases C(1,1) gives a smaller gain over no-shrinkage than C(1000,1000) or C(1000,10000): the (1,1) prior rarely induces too much shrinkage but shrinks less aggressively than the other priors in truths 1 and 6. C(1000,1000) performs well for truths 1 and 6, but poorly for the other truths. C(1000,10000) performs as well as or better than C(1,1) with a few exceptions. The exceptions are the contrasts most strongly affected by the interaction(s) in truths 3, 4, and 5, where C(1000,10000) tends to shrink too much.

Of the three constraint methods, we prefer C(1000,10000) slightly. First, C(1000,10000) is not beaten badly by no-shrinkage for any contrast and truth. The main concern here is truth 5, where C(1000,10000) does not do well for β_{00}. Second, when DS's method outperforms no-shrinkage, the constraint method does at least as well as DS's method and often better. The exception is truth 5, for which DS is designed. Third, C(1000,10000) accommodates a three-way interaction better than DS's method. Finally, in the situation that is probably most common (truth 1), C(1000,10000) gives substantially better performance than C(1,1). These arguments notwithstanding, a reasonable case can be made for C(1,1), as noted above.

We considered other properties of the constraint method in addition to average squared error. All of the shrinkage methods produced estimates for the θ_i that tested significant for bias (i.e., had P-values < 0.05). Despite this, 95% highest posterior density (HPD) regions produced by the constraint methods usually had coverage at or above the nominal 95% level (Table 4). Coverage for C(1,1) was always reasonably calibrated, with some indication of conservatism. Coverage was low for the C(1000,10000) method for those contrasts with average squared error greater than 0.2, the mean squared error for no-shrinkage. Coverage for the DS and EB methods (data not shown)

Table 3: Simulation results: Average squared error for various methods under various truths

β	No shrinkage	EB	DS(1,1)	DS(1000,1000)	DS(1000,10000)	C(1,1)	C(1000,1000)	C(1000,10000)
Truth 1: No interactions								
11	0.199	0.118	0.121	0.097	0.108	0.155	0.051	0.082
10	0.185	0.114	0.116	0.092	0.103	0.123	0.051	0.066
01	0.203	0.128	0.119	0.097	0.106	0.132	0.050	0.067
00	0.184	0.110	0.110	0.087	0.097	0.106	0.050	0.064
Truth 2: One two-way interaction								
11	0.204	0.199	0.135	0.176	0.152	0.162	0.243	0.148
10	0.196	0.193	0.137	0.182	0.153	0.150	0.256	0.168
01	0.209	0.209	0.144	0.200	0.165	0.143	0.274	0.153
00	0.183	0.188	0.135	0.189	0.156	0.129	0.294	0.156
Truth 3: Three-way interaction								
11	0.187	0.250	0.233	0.295	0.259	0.164	0.531	0.232
10	0.188	0.154	0.169	0.111	0.142	0.132	0.099	0.098
01	0.210	0.178	0.188	0.211	0.195	0.143	0.099	0.102
00	0.212	0.173	0.167	0.132	0.156	0.129	0.091	0.082
Truth 4: One two-way and the three way interaction								
11	0.191	0.212	0.290	0.468	0.325	0.175	0.970	0.247
10	0.203	0.186	0.161	0.128	0.163	0.150	0.166	0.204
01	0.200	0.195	0.237	0.365	0.258	0.145	0.401	0.148
00	0.202	0.195	0.144	0.164	0.159	0.127	0.390	0.142
Truth 5: Two equal two-way interactions								
11	0.191	0.219	0.147	0.201	0.166	0.156	0.846	0.251
10	0.203	0.178	0.122	0.115	0.120	0.146	0.070	0.162
01	0.204	0.178	0.126	0.116	0.123	0.146	0.064	0.156
00	0.201	0.226	0.159	0.212	0.179	0.175	0.886	0.334
Truth 6: Two-way non-treatment interactions								
11	0.199	0.122	0.121	0.096	0.106	0.159	0.049	0.098
10	0.214	0.131	0.121	0.096	0.107	0.141	0.048	0.072
01	0.206	0.126	0.118	0.094	0.105	0.123	0.048	0.071
00	0.188	0.113	0.120	0.096	0.107	0.112	0.048	0.069

	C(1,1)						C(1000,10000)					
Truth	1	2	3	4	5	6	1	2	3	4	5	6
β_{11}	96	96	96	96	96	97	99	93	88	91	88	98
β_{10}	97	95	97	96	96	96	98	89	97	90	94	98
β_{01}	97	97	96	96	97	97	98	90	97	93	95	98
β_{00}	98	97	97	97	94	98	98	90	98	93	75	98

Table 4: Simulation results: Percent coverage of 95% HPD regions for the C(1,1) and C(1000,10000) methods under various truths

was inversely related to average squared error; if these methods were compared to the constraint method based on coverage, the conclusions would be identical to those we drew based on average squared error.

We repeated the entire simulation using effect sizes twice as large as those used in the first simulation. The conclusions from this second simulation were consistent with those presented above. Other results (data not shown) show that the constraint methods are not sensitive to small changes in the prior distributions of the constraint case precisions.

We specifically chose not to compare the various methods on their abilities to detect interactions *per se*. The object of our procedure is not to make yes/no decisions about the presence of interactions, but to accommodate them smoothly to the extent that they appear to be present.

5 Subgroup analysis of CD4+ changes in the CPCRA 002 study

We illustrate our method on data from the ddI/ddC study (CPCRA 002) conducted by the Community Programs for Clinical Research on AIDS (CPCRA). CPCRA 002 compared didanosine (ddI) and zalcitabine (ddC) in patients with HIV infection who were intolerant to or had treatment failure on zidovudine (ZDV). Abrams et al. (1994) give further details and a standard analysis.

For this example, the response is the within-patient change in CD4+ count between baseline and two months; these two measurements were available on 367 patients. The questions of interest are

	Method		
Effect	No Shrinkage: Full model	No Shrinkage: Main effects only	C(1000,10000)
average change	0.14 (0.15)	-0.02 (0.11)	0.15 (0.12)
Main Effects			
x_1, treatment	-0.60 (0.23)	-0.26 (0.10)	-0.33 (0.15)
x_2, failed/intolerant	0.07 (0.22)	0.27 (0.11)	0.22 (0.14)
x_3, Karnofsky	-0.31 (0.19)	-0.13 (0.11)	-0.16 (0.13)
Interactions			
x_1,x_2	0.42 (0.34)	–	0.06 (0.14)
x_1,x_3	0.39 (0.28)	–	0.06 (0.14)
x_2,x_3	0.13 (0.30)	–	0.00 (0.11)
x_1,x_2,x_3	-0.24 (0.44)	–	0.02 (0.15)
Treatment effects			
β_{11}	-0.03 (0.24)	-0.26 (0.10)	-0.16 (0.17)
β_{10}	-0.18 (0.25)	-0.26 (0.10)	-0.26 (0.17)
β_{01}	-0.21 (0.16)	-0.26 (0.10)	-0.26 (0.13)
β_{00}	-0.60 (0.23)	-0.26 (0.10)	-0.33 (0.15)

Table 5: Estimated effects for the ddI/ddC data

whether either drug increased patients' CD4+ counts, and whether this increase differed between the two drugs and in different groups of patients. We consider the treatment indicator (x_1: 1 = ddC, 0 = ddI) and two binary covariates: reason for eligibility (x_2: 1 = failed ZDV, 0 = intolerant to ZDV), and baseline Karnofsky score (x_3: 1 = score > 85, 0 = score ≤ 85, where higher scores are better). We standardized each response by subtracting the overall mean response and dividing by the overall standard deviation to make the scale of the problem similar to that of our simulations. Convergence for the constraint method was immediate; we present results from multiple chains run for 500 iterations each.

Table 5 presents results from fitting three models: no-shrinkage ANOVA with all interactions (model (7)), no-shrinkage ANOVA with main effects only, and the C(1000,10000) constraint method (i.e., priors for the constraint case precisions have mean 1000 and standard deviation 10,000). We

tried several typical variable selection procedures, but they did not agree on a model. Looking first at the full-model no-shrinkage results, the estimated treatment effect varies substantially between subgroups defined by x_2 and x_3. The within-subgroup treatment effect appears to be significant for one subgroup (β_{00}) but not the others. However, only two of the interaction terms are larger in absolute value than their standard errors. Each interaction term is non-significant, and a four-degree-of-freedom F-test for the interactions has a P-value greater than 0.50. The second model, with main effects only, removes the interactions from the model completely. However, it also removes the possibly interesting interactions. Attempts to arrive at an intermediate model cannot avoid the issue of determining which terms should be removed or included.

The constraint method resolves the issue by smoothly including terms in the model as the data indicate. In this example the three-way interaction and the (x_2, x_3) interaction are shrunk almost entirely away, while the other two-way interactions are shrunk considerably. This results in an estimated treatment effect that varies slightly between the subgroups, with the two extreme subgroups (β_{11} and β_{00}) differing by about one posterior standard deviation. Based on the constraint method, we conclude that there is a no evidence of a boost in CD4+ from ddI (the intercept is not significantly different than zero), but there is evidence that CD4+ counts decrease on ddC (the treatment main effect has posterior mean and standard deviation of -0.33 and 0.15 respectively). The analysis leaves hardly any indication that the difference between ddI and ddC varies among the subgroups.

6 Extensions

6.1 2^k designs

We presented the constraint method for ANOVA using a 2^3 design. Generalizing to other 2^k designs is straightforward. For example, a model for $k = 2$ has a single interaction term, so the single constraint case is $\theta_{11} - \theta_{10} - \theta_{01} + \theta_{00} \sim N(0, \sigma_{12}^2)$, in the obvious notation. Designs with $k > 3$ are mathematically simple but notationally burdensome. If we allow each interaction its own variance component, we have $\sum_{i=2}^{k} \binom{k}{i} = 2^k - (k+1)$ variance components and constraint cases. As k increases, the ratio of the number of θs (2^k) to the number of constraint cases ($2^k - (k+1)$) approaches 1, which suggests that opportunities for shrinkage and performance of the constraint method should improve as k increases.

To test this hypothesis, we applied the C(1000,10000) method to data simulated from a 2^4 design following truths 1 through 5 from Section 4, with the additional (fourth) factor having no effect on the response. Table 6 gives the average squared error of the treatment effects (βs) in the eight subgroups defined by the three non-treatment factors, using 500 simulated datasets per truth. As in Section 4, the average squared error for the no-shrinkage method hovers around the mean squared error of 0.2, so it is omitted from Table 6. The constraint method is uniformly better than the no-shrinkage method, and its performance relative to the no-shrinkage method is better than in the 2^3 design.

As k increases, it becomes tempting to fix some high-order interactions at zero. If we set the k-way interaction to zero, we can have two constraint cases for each two-way and three-way interaction. For example, in a 2^4 design, allowing all interactions, there are 11 constraint cases and variance components. If we set the four-way interaction to zero, two linear combinations of the θ_{ijkl} estimate each of the two- and three-way interactions, so we can construct 20 constraint cases, two

β	Truth 1	Truth 2	Truth 3	Truth 4	Truth 5
111	0.083	0.101	0.131	0.143	0.151
110	0.054	0.108	0.169	0.170	0.173
101	0.053	0.093	0.062	0.136	0.115
100	0.041	0.110	0.062	0.148	0.113
011	0.052	0.074	0.074	0.082	0.119
010	0.041	0.078	0.067	0.084	0.119
001	0.039	0.084	0.046	0.073	0.165
000	0.040	0.086	0.049	0.077	0.192

Table 6: Simulation results: average squared error for the C(1000,10000) method, 2^4 design

per remaining interaction. This is related to the concept of confounding a high-order interaction in an experimental design.

To examine this possibility, we compared two constraint methods for the simulated data sets summarized in Table 6. One method used all interactions and one constraint case per interaction (1cc), and the other method set the 4-way interaction to zero and used two constraint cases per interaction (2cc). For each method, we considered two priors for the constraint-case precisions, one having mean 1000 and standard deviation 10,000 (high shrinkage) and the other having mean 0.5 and standard deviation 2.5 (low shrinkage). (The latter was selected after some experimentation.) Average squared error for the 1cc/high-shrinkage method was given in Table 6. The 2cc/low-shrinkage method had performance very similar to the 1cc/high-shrinkage method for all truths and contrasts. The 1cc/low-shrinkage method was rather worse for Truths 1 and 2 but competitive for truths 3, 4, and 5; like the C(1,1) method in the 2^3 design, it outperformed no-shrinkage for all truths and contrasts. The 2cc/high-shrinkage method did best of all on truth 1 and worst of all on the other 4 truths, with especially bad performance for truths 3, 4, and 5. The latter results – markedly worse than no-shrinkage for some contrasts – arose because the 2cc/high-shrinkage method tends to shrink interactions too much.

We do not fully understand these last results, although clearly the interplay of constraint-case

priors and multiple constraint cases is not trivial.

6.2 Factors with more than two levels

A less straightforward extension is to cases where the subgroups are defined by non-binary covariates. Suppose we have an ANOVA with k factors, where the i^{th} covariate has c_i levels. The saturated model has $\prod_{i=1}^{k} c_i$ degrees of freedom, which can be shrunk in at least two ways. An approach consistent with Section 3.2 gives each interaction its own variance component. Sections 4 and 6.1 suggest that this is feasible and may yield acceptable results.

However, if we have (say) d degrees of freedom for a two-way interaction and no *a priori* basis for distinguishing among these d contrasts, we might exploit this fact. To illustrate, consider a 3^3 design. The saturated model for this design has 27 degrees of freedom: one for the grand mean, six for main effects, 12 for two-way interactions, and eight for three-way interactions. We could assume that the eight three-way interaction terms are drawn from a single distribution with center 0 and scale σ_{123}. With this approach, the number of variance components is a function of the number of factors only, not the number of levels of each factor.

Suppose only one of the three-way interaction terms is non-zero and the others are zero; then the non-zero term will be shrunk substantially if f is a normal distribution. A possible alternative is to let f be a t distribution, which would readily allow one non-zero interaction term while shrinking the others near zero. The use of the Student's t distribution is possible in this case because we have multiple observations per variance component. The extension to t errors complicates the computing slightly; two solutions are to use the more general Metropolis-Hastings algorithm (Smith and Roberts 1993), or to use Gibbs sampling with latent gamma variables for the t errors (Carlin and Polson 1991, Evans and Swartz 1995).

Other extensions of the constraint method are possible, depending on the investigator's prior

beliefs about the interactions and on the goals of the investigation.

7 Discussion

The constraint method for 2^k designs uses equal numbers of variance components and constraint cases. The variance components are intended to allow interactions to be non-zero while providing a mechanism to shrink them in cases where their existence is unclear. Consider a specific interaction, say the interaction between x_1 and x_2 in the 2^3 design. The constraint case for this interaction is $\theta_{110} - \theta_{010} - \theta_{100} + \theta_{000} \sim N(0, \sigma^2_{12})$. If x_1 and x_2 interact, $\theta_{110} - \theta_{010} - \theta_{100} + \theta_{000}$ will tend to be fairly large. The mean of the constraint case error is zero, so the estimated magnitude of $\theta_{110} - \theta_{010} - \theta_{100} + \theta_{000}$ will suggest that σ^2_{12} is fairly large. Additionally, the prior distribution for $h_{12} = \frac{1}{\sigma^2_{12}}$ provides information to stabilize the posterior density. It appears important to have a prior with a large percentage of its mass at precisions near zero, so the posterior distribution can put mass on small precisions when an interaction is present.

An alternative approach to our problem is through model averaging (Hodges 1987; Raftery, Madigan, and Hoeting 1993; Draper 1995). A model-averaging approach begins with a parameterization such as (7), i.e., a model that includes all main effects and interactions. The method proceeds by fitting the saturated and all simpler models, then computing the posterior distribution across the models. Subgroup comparisons are estimated using a weighted average of the subgroup comparisons in each model, the weights being the posterior probabilities of the models. This approach is promising but many technical issues remain to be examined, such as setting prior distributions within each model and putting prior probabilities on the models themselves.

The constraint method can be viewed as a smooth version of model averaging: model averaging can only include or exclude terms, while the constraint method allows partial inclusion. One might

expect superior performance from the constraint method because it *is* a smooth method, but for 2^k designs in particular model averaging may be more stable and thus provide better performance.

The methods developed here suggest a relationship between the amount of shrinkage and the degrees of freedom in (or size of) the fitted model. This issue has been investigated in the smoothing literature, where it is common to read about the degrees of freedom of a smooth (e.g., Hastie and Tibshirani 1990). In a typical smooth, individual data points are down-weighted and the estimate of the curve at each data point is a combination of the individual point and some sort of local average or median. The constraint method for ANOVA is a different sort of smooth, in which individual cell averages are down-weighted and the estimate of a cell mean is a weighted average of that cell's average and the averages of cells associated through constraint cases. An explicit connection between the model, the amount of shrinkage, and the number of degrees of freedom is currently under investigation.

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 are grateful for the comments of Tom Louis and of the RAND Statistics group on earlier versions of this work.

References

Abrams, D.I., Goldman, A.I., Launer, C., Korvick, J.A., Neaton, J.D., Crane, L.R. et al. (1994), "Comparative trial of didanosine and zalcitabine in patients with human immunodeficiency virus infection who are intolerant or have failed zidovudine therapy", *New England Journal*

of Medicine, **330**, 657-662.

Bryk, A.S. and Raudenbush, S.W. (1992), *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, California: SAGE Publications.

Carlin, B.P. and Polson, N.G. (1991), "Inference for nonconjugate Bayesian models using the Gibbs sampler", *Canadian Journal of Statistics*, **19**:399-405.

Dixon, D.O. and Simon, R. (1991), "Bayesian subset analysis", *Biometrics*, **47**, 871-881.

Draper, D. (1995), "Assessment and propagation of model uncertainty", *Journal of the Royal Statistical Society, Series B*, **57**:45-97.

Efron, B. and Morris, C.N. (1975), "Data analysis using Stein's estimator and its generalizations", *Journal of the American Statistical Association*, **70**:311-319.

Evans, M. and Swartz, T. (1995), "Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems", *Statistical Science*, **10**:254-272.

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., Rubin, D.B. (1992), "Inference from iterative simulation using multiple sequences" (with discussion), *Statistical Science*, **7**:457-511.

Hastie, T.J. and Tibshirani, R.J. (1990), *Generalized Additive Models*. New York: Chapman and Hall.

Hodges, J.S. (1987), "Uncertainty, policy analysis, and statistics", *Statistical Science*, **2**:259-291.

Hodges J.S. (1996), "Some algebra and geometry for hierarchical models, applied to diagnostics", Research Report 94-009 (revised), Division of Biostatistics, University of Minnesota.

James, W., and Stein, C. (1961), "Estimation with quadratic loss", *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, (Vol 1). Berkeley: University of California Press, 361-379.

Louis, T.A. (1984), "Estimating a population of parameter values using Bayes and empirical Bayes methods", *Journal of the American Statistical Association*, **79**:393-398.

Morris, C.N. (1983), "Parametric empirical Bayes inference: theory and applications", *Journal of the American Statistical Association*, **78**:47-55.

Oxman, A.D. and Guyatt, G.H. (1992), "A consumer's guide to subgroup analyses", *Annals of Internal Medicine*, **116**:78-84.

Pocock, S.J., Hughes, M.D., and Lee, R.J. (1987), "Statistical problems in the reporting of clinical trials: a survey of three medical journals", *New England Journal of Medicine*, **317**:426-432.

Raftery, A.E., Madigan, D., Hoeting, J. (1993), "Model selection and accounting for model uncertainty in linear regression models", Technical Report 262, University of Washington.

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

Simon, R., Dixon, D.O., Freidlin, B. (1996), "Bayesian subset analysis of a clinical trial for the treatment of HIV infections", In *Bayesian Biostatistics*, eds. D. Berry, D. Stangl, New York: Marcel Dekker.

Smith, A.F.M. and Roberts, G.O. (1993), "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods", *Journal of the Royal Statistical Society*, Series B, **55**:3-23.

Stein, C. (1955), "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution", *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley: University of California Press, 197-206.

Thomas, D.C., Siemiatycki, J., Dewar, R., Robins, J., Goldberg, M., and Armstrong, B.G. (1985), "The problem of multiple inference in studies designed to generate hypotheses", *American Journal of Epidemiology*, **122**:1080-1095.