Biometrics 00, 0-0
00 0000

Factor Selection and Structural Identification in the Interaction ANOVA Model

Justin B. Post*

Department of Statistics, North Carolina State University

Raleigh, NC 27695-8203, USA

**email:* jbpost@ncsu.edu

and

Howard D. Bondell*

Department of Statistics, North Carolina State University

Raleigh, NC 27695-8203, USA

**email:* bondell@stat.ncsu.edu

Summary: When faced with categorical predictors and a continuous response, the objective of analysis often consists of two tasks: finding which factors are important and determining which levels of the factors differ significantly from one another. Often times these tasks are done separately using Analysis of Variance (ANOVA) followed by a post-hoc hypothesis testing procedure such as Tukey's Honestly Significant Difference test. When interactions between factors are included in the model the collapsing of levels of a factor becomes a more difficult problem. When testing for differences between two levels of a factor, claiming no difference would refer not only to equality of main effects, but also equality of each interaction involving those levels. This structure between the main effects and interactions in a model is similar to the idea of heredity used in regression models. This paper introduces a new method for accomplishing both of the common analysis tasks simultaneously in an interaction model while also adhering to the heredity-type constraint on the model. An appropriate penalization is constructed that encourages levels of factors to collapse and entire factors to be set to zero. It is shown that the procedure has the oracle property implying that asymptotically it performs as well as if the exact structure were known beforehand. We also discuss the application to estimating interactions in the unreplicated case. Simulation studies show the procedure outperforms post hoc hypothesis testing procedures as well as similar methods that do not include a structural constraint. The method is also illustrated using a real data example.

Key words: Interaction ANOVA model; Grouping; Multiple comparisons; Oracle propery; Shrinkage; Variable selection.

1. Introduction

Consider the common case of a continuous response variable and categorical predictors (factors). A conventional way to judge the importance of these factors is to use Analysis of Variance (ANOVA). Once factors are deemed important, to check which levels of the factors differ from one another the next step is often to do a post hoc analysis such as Tukey's honestly significantly different test, Fisher's least significant difference test, or pairwise comparisons of levels using a Bonferroni adjustment or a Benjamini-Hochberg (Benjamini and Hochberg, 1995) type adjustment. Rather than carry out these two tasks separately, a technique called CAS-ANOVA, for collapsing and shrinkage in ANOVA (Bondell and Reich, 2009), has been developed to perform these actions simultaneously. This procedure is a constrained, or penalized, regression technique. Much of the recent variable selection literature is of this form, and examples include the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), the Elastic Net (EN) (Zou and Hastie, 2005), the Smoothly Clipped Absolute Deviation penalty (SCAD) (Fan and Li, 2001), and the Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) (Bondell and Reich, 2008). The CAS-ANOVA procedure places an L_1 constraint directly on the pairwise differences in each factor allowing an entire factor to be zeroed out while also allowing levels within a factor to collapse (be set equal) into groups. Not only does the CAS-ANOVA procedure accomplish both tasks at once, the nature of the penalty requires the levels of each factor to be collapsed into non-overlapping groups. This method was shown to have the oracle property, implying that its performance is asymptotically equivalent to having the true grouping structure known beforehand and performing the standard least squares ANOVA fit to this collapsed design.

A limitation of the CAS-ANOVA method is that it was developed assuming a main effect only model and often an interaction model is more realistic. In the interaction ANOVA

model, when determining whether two levels of a factor should collapse we must look at more than just their respective main effects. It is not appropriate to claim the two levels should collapse if there is an interaction term that differs between the two levels for any given level of another factor. This leads to the idea that only levels whose interaction effects are all identical should have the possibility of their main effects also declared identical. Thus, if we can enforce this type of structure on our model we will be able to accomplish both tasks of our analysis. The notion of this model structure is similar to the idea of heredity often used in regression models when higher order terms are involved. For a regression model such as $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i1}^2$, where y_i is a continuous response with x_{i1} and x_{i2} continuous predictors, strong heredity implies x_{i1}^2 should only appear in the model if x_{i1} appears in the model and the interaction term $x_{i1}x_{i2}$ should only appear in the model if both x_{i1} and x_{i2} appear in the model. The heredity principle is important because it aids in interpretation of the model and ensures that the model makes sense structurally. This type of constraint on the structure of the predictors in a regression model has seen much use. For example, Choi, Li, and Zhu (2010) used heredity to extend the LASSO variable selection technique to include interaction terms, Yuan, Joseph, and Zou (2009) used heredity in variable selection with the non-negative garrote, Yuan, Joseph, and Lin (2007) used the constraint with the LARS algorithm, and Chipman (1996) used the constraint in the Bayesian variable selection context. However, all of these approaches deal with continuous predictors, in that an interaction is a single term derived from the product of two other predictors. In the interaction ANOVA model, interaction terms are not single terms, they arise as products of groups of variables, and thus need to be treated differently.

This paper develops a method for use in the interaction ANOVA model to simultaneously perform the two main goals of analysis. The method utilizes a weighted penalty that enforces the heredity-type structure on the model. The new method is called GASH-ANOVA for

Grouping And Selection using Heredity in ANOVA. We show that the oracle property holds for the GASH-ANOVA procedure, in that asymptotically it performs as well as if the exact structure were known beforehand. This property also implies that, even after the possible dimension reduction, asymptotic inference can be based on standard ANOVA theory.

We discuss an application of the GASH-ANOVA estimator to the unreplicated mutli-way ANOVA design. When no replication exists, standard practice yields no way to evaluate interactions among the factors and one is forced to make the assumption of no interaction effects. A number of solutions have been proposed to investigate interaction effects in the unreplicated design (see Franck, Osborne, and Nielsen, 2011 for a detailed review and comparison of methods). We propose the use of an unweighted version of the GASH-ANOVA estimator to inspect and estimate the interaction terms.

The rest of the article is organized as follows: In section 2, notation is introduced and the use of penalized regression for collapsing levels is discussed. In section 3.1, the GASH-ANOVA procedure is introduced. The extension to the unreplicated case is expanded upon in section 3.2. In section 4, the asymptotic properties of the GASH-ANOVA solution are presented. Section 5 discusses computation and tuning for the GASH-ANOVA method. To illustrate the method, section 6 gives simulation studies and their results. Section 7 shows the method's usefulness on a real data example. Finally, section 8 gives a discussion.

2. Collapsing via Penalized Regression

To simplify notation, consider the two factor ANOVA model with factor A and factor B having a and b levels respectively. The extension to more than two factors is straightforward and will be discussed shortly. We denote the number of observations at each level combination by n_{ij} and denote the total sample size by $n = \sum_{i,j} n_{ij}$. We use the matrix representation of the linear model, $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is the $n \times 1$ vector of responses, $\mathbf{X}$ is the typical $n \times p$ overparameterized ANOVA design matrix consisting of

zeros and ones corresponding to the combination of levels for each observation (where $p = 1 + a + b + ab$), $\boldsymbol{\theta}$ is a $p \times 1$ vector of parameters, and $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of constant variance error terms. The parameter vector consists of the mean and three stacked vectors, $\boldsymbol{\theta} = (\mu, \boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, (\boldsymbol{\alpha\beta})^T)^T$, where μ is the intercept, $\boldsymbol{\alpha}^T = (\alpha_1, \alpha_2, ...\alpha_a)$, $\boldsymbol{\beta}^T = (\beta_1, \beta_2, ...\beta_b)$, and $(\boldsymbol{\alpha\beta})^T = ((\alpha\beta)_{11}, (\alpha\beta)_{12}, ...(\alpha\beta)_{1b}, (\alpha\beta)_{21}, ..., (\alpha\beta)_{2b}, ..., (\alpha\beta)_{ab})$. Here, α_i corresponds to the main effect of level i of factor A, β_j corresponds to the main effect of level j of factor B, and $(\alpha\beta)_{ij}$ corresponds to the interaction effect of level i of factor A and level j of factor B. One ordinary least squares (OLS) solution can be written as follows:

$$\widehat{\boldsymbol{\theta}}_{OLS} = argmin_{\boldsymbol{\theta}} ||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||^2 \tag{1}$$

$$\text{subject to} \quad \sum_{i=1}^{a} \alpha_i = 0, \qquad \sum_{i=1}^{a} (\alpha\beta)_{ij} = 0 \text{ for all } j,$$
$$\sum_{j=1}^{b} \beta_j = 0, \qquad \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0 \text{ for all } i,$$

where the constraints on the parameters are aptly named the sum to zero constraints. Note that n_{ij} must be at least two in order to estimate the interaction terms when using OLS.

A naive penalized regression approach to collapsing levels in the interaction ANOVA model would be to individually penalize all of the differences of main effect parameters and interaction effect parameters that correspond to collapsing levels. Specifically, let $\widehat{\alpha}_{i,OLS}$, $\widehat{\beta}_{j,OLS}$, and $\widehat{(\alpha\beta)}_{ij,OLS}$ be the OLS estimates of the corresponding parameters found using equation (1). A weighted naive penalty is given by

$$\sum_{1 \leqslant k < m \leqslant a} w^{(km)}_{\alpha,Naive} |\alpha_k - \alpha_m| + \sum_{1 \leqslant k < m \leqslant b} w^{(km)}_{\beta,Naive} |\beta_k - \beta_m|$$

$$+ \sum_{1 \leqslant m \leqslant b} \sum_{1 \leqslant k < l \leqslant a} w^{(km,lm)}_{\alpha\beta,Naive} |(\alpha\beta)_{km} - (\alpha\beta)_{lm}|$$

$$+ \sum_{1 \leqslant k \leqslant a} \sum_{1 \leqslant m \leqslant l < b} w^{(km,kl)}_{\alpha\beta,Naive} |(\alpha\beta)_{km} - (\alpha\beta)_{kl}| \leqslant t,$$

where the weights are scaled adaptive LASSO (Zou, 2006) type weights and $t \geqslant 0$ is a tuning parameter. The weights on the main effect differences are given by

$$w^{(km)}_{\alpha,Naive} = |\widehat{\alpha}_{k,OLS} - \widehat{\alpha}_{m,OLS}|^{-1} \text{ and } w^{(km)}_{\beta,Naive} = \left|\widehat{\beta}_{k,OLS} - \widehat{\beta}_{m,OLS}\right|^{-1}$$

and the weights placed on the interaction differences are given by

$$w^{(km,kl)}_{\alpha\beta,Naive} = \left|\widehat{(\alpha\beta)}_{km,OLS} - \widehat{(\alpha\beta)}_{kl,OLS}\right|^{-1}$$

and

$$w^{(km,lm)}_{\alpha\beta,Naive} = \left|\widehat{(\alpha\beta)}_{km,OLS} - \widehat{(\alpha\beta)}_{lm,OLS}\right|^{-1}.$$

We refer to this penalty and its estimator as the 'Naive method'.

Note that the Naive method is the basic extension of the CAS-ANOVA procedure to the interaction ANOVA case. In order to shrink the coefficients and to perform variable selection in the additive model, the (adaptive) CAS-ANOVA (Bondell and Reich, 2009) procedure places a weighted constraint directly on the pairwise differences of the levels of each factor. In the two factor main effect only model, which has the same set up as the above model ignoring the interaction terms and constraints on the interaction terms, the CAS-ANOVA procedure adds a constraint that is of the form

$$\sum_{1 \leqslant k < m \leqslant a} w^{(km)}_{\alpha,Naive} |\alpha_k - \alpha_m| + \sum_{1 \leqslant k < m \leqslant b} w^{(km)}_{\beta,Naive} |\beta_k - \beta_m| \leqslant t,$$

where $t \geqslant 0$ is a tuning constant.

Asymptotically the Naive method will perform well, although it is unlikely to perform well at the two main tasks of choosing significant factors and collapsing levels in small samples because of the lack of important heredity-type structure previously discussed. In problems where the factors have a large number of levels, the Naive method will have difficulty collapsing levels due to the sizable number of interaction differences that need to be set to zero. Using our notation we can now describe the procedure for collapsing two levels of a factor when interactions are present in more detail. In the two factor situation considered here, our

procedure implies that level k of factor A should only be collapsed to level m of factor A if the main effect difference, $\alpha_m - \alpha_k$, and all the interaction differences between factor A and B involving level k and m of factor A, $(\alpha\beta)_{m1} - (\alpha\beta)_{k1}, (\alpha\beta)_{m2} - (\alpha\beta)_{k2}, ..., (\alpha\beta)_{mb} - (\alpha\beta)_{kb}$, have been set to zero. There is a chance for the Naive method to select factors and collapse levels but nothing forces this to be the case, a stray nonzero interaction difference can prevent the collapsing from occurring. Thus, for interpretability of the model and to accomplish our two main tasks of analysis simultaneously, the Naive method is not ideal with interactions present.

3. GASH-ANOVA

3.1 *Method*

For computation and the statement of the theoretical results of the GASH-ANOVA estimator it is more convenient to reparameterize to a full rank design matrix using a reference level as a baseline. This lessens the number of parameters and yields a full rank design. Thus, from this point forward we will use the full rank design. We choose the first level of each factor as the reference level, although this choice is arbitrary as the levels can be relabeled. Define the new design matrix by $\mathbf{X}^*$ and the new parameter vector by $\boldsymbol{\theta}^* = (\mu^*, \boldsymbol{\alpha}^{*T}, \boldsymbol{\beta}^{*T}, (\boldsymbol{\alpha\beta})^{*T})^T$, where $\mu^* = \mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11}$,

$$\boldsymbol{\alpha}^{*T} = (\alpha_2^*, ..., \alpha_a^*) = (\alpha_2 - \alpha_1, ..., \alpha_a - \alpha_1),$$

$\boldsymbol{\beta}^*$ is a $(b-1) \times 1$ vector defined similarly for factor B, and

$$(\boldsymbol{\alpha\beta})^{*T} = ((\alpha\beta)_{22}^*, (\alpha\beta)_{23}^*, ..., (\alpha\beta)_{2b}^*, (\alpha\beta)_{32}^*, ..., (\alpha\beta)_{3b}^*, ..., (\alpha\beta)_{ab}^*),$$

where $(\alpha\beta)_{ij}^* = (\alpha\beta)_{ij} - (\alpha\beta)_{11}$.

To achieve the automatic factor selection and collapsing of levels in the interaction model the GASH-ANOVA approach uses a weighted heredity-type constraint. To encourage the collapsing of levels, an infinity norm constraint is placed on (overlapping) groups of pairwise

differences belonging to different levels of each factor. In detail, we form $G = \binom{a}{2} + \binom{b}{2}$ groups where each group contains a main effect difference between two levels of a factor along with all interaction differences that involve those same two levels. We denote each group of parameters by either $\phi_{\alpha,ij}, 1 \leqslant i < j \leqslant a$ or $\phi_{\beta,ij}, 1 \leqslant i < j \leqslant b$, where $\phi_{\alpha,1j} = (\alpha_j^*, (\alpha\beta)_{j2}^*, (\alpha\beta)_{j3}^*, ..., (\alpha\beta)_{jb}^*)$ for $2 \leqslant j \leqslant a$ and $\phi_{\alpha,ij} = (\alpha_j^* - \alpha_i^*, (\alpha\beta)_{j2}^* - (\alpha\beta)_{i2}^*, (\alpha\beta)_{j3}^* - (\alpha\beta)_{i3}^*, ..., (\alpha\beta)_{jb}^* - (\alpha\beta)_{ib}^*)$ for $1 < i < j \leqslant a$ and $\phi_{\beta,ij}$ is defined similarly. Note that these groups share some interaction terms. By judicious choice of overlapping groups, two main effects of a factor can be set equal to one another only if all of the interactions for those two levels are also set equal and, with probability one, an interaction difference is only present if the corresponding main effect differences are also present. Thus, the GASH-ANOVA procedure adheres to our heredity-type structure which encourages levels of each factor to be estimated with exact equality and entire factors to be set to zero. This overlapping group penalty on the differences is related to the family of Composite Absolute Penalties (CAP) (Zhang, Rocha, and Yu, 2009). However, the CAP treats coefficients themselves as groups, not the differences of coefficients as groups.

We specify an infinity norm to achieve our goals, however other norms could also be applied to attain this grouping effect. For example, the L_2 norm could be used and would require entire groups to be set to zero. We choose the infinity norm for computational purposes as it leads to a quadratic programming problem (see section 5).

The GASH-ANOVA solution can be written in detail as follows:

$$\widehat{\boldsymbol{\theta}}^* = argmin_{\boldsymbol{\theta}^*} ||\mathbf{y} - \mathbf{X}^*\boldsymbol{\theta}^*||^2 \tag{2}$$

$$\text{subject to } \sum_{1 \leqslant i < j \leqslant a} w_\alpha^{(ij)} \max\{|\phi_{\alpha,ij}|\} + \sum_{1 \leqslant i < j \leqslant b} w_\beta^{(ij)} \max\{|\phi_{\beta,ij}|\} \leqslant t,$$

where $w_\alpha^{(ij)}$ and $w_\beta^{(ij)}$ are adaptive weights, $t \geqslant 0$ is a tuning constant,

$$|\phi_{\alpha,1j}| = (|\alpha_j^*|, |(\alpha\beta)_{j2}^*|, |(\alpha\beta)_{j3}^*|, ..., |(\alpha\beta)_{jb}^*|)^T$$

for $2 \leqslant j \leqslant a$ and

$$|\phi_{\alpha,ij}| = (|\alpha_j^* - \alpha_i^*|, |(\alpha\beta)_{j2}^* - (\alpha\beta)_{i2}^*|, |(\alpha\beta)_{j3}^* - (\alpha\beta)_{i3}^*|, ..., |(\alpha\beta)_{jb}^* - (\alpha\beta)_{ib}^*|)^T$$

for $1 < i < j \leqslant a$, and $|\phi_{\beta,ij}|$ is similarly defined. Using the OLS solution, the weight $w_\alpha^{(ij)}$ is given by $(\max\left\{\left|\widehat{\phi}_{\alpha,ij,OLS}\right|\right\})^{-1}$, where $\widehat{\phi}_{\alpha,ij,OLS}$ denotes the use of the OLS estimate for the differences. The form of the weight $w_\beta^{(ij)}$ is given similarly. The adaptive weights allow for the asymptotic properties of the GASH-ANOVA procedure given in section 4.

If all interaction differences for all factors are set to zero then we are left with the additive model. Thus, it may be of interest to not only be able to collapse entire levels of a factor, but to also be able to collapse individual interaction differences. To accomplish the collapsing of individual interaction differences we can explicitly add these terms to the penalty. The new penalty would be given by

$$\begin{aligned}
&\sum_{1\leqslant i<j\leqslant a} w_\alpha^{(ij)} \max\{|\phi_{\alpha,ij}|\} + \sum_{1\leqslant i<j\leqslant b} w_\beta^{(ij)} \max\{|\phi_{\beta,ij}|\} \\
&\quad + \sum_{1\leqslant m\leqslant b} \sum_{1\leqslant k<l\leqslant a} w_{\alpha\beta,Naive}^{(km,lm)} |(\alpha\beta)_{km} - (\alpha\beta)_{lm}| \\
&\quad + \sum_{1\leqslant k\leqslant a} \sum_{1\leqslant m\leqslant l<b} w_{\alpha\beta,Naive}^{(km,kl)} |(\alpha\beta)_{km} - (\alpha\beta)_{kl}| \leqslant t,
\end{aligned}$$

where the weights are as defined previously. The asymptotic theory for this penalty given in section 4 should still hold. If one desired even more control of the interaction differences we could leave the original GASH-ANOVA penalty alone and create a second penalty with its own tuning parameter, say t_2, that involved only the interaction differences (explicitly given by the last two sums of the previous penalty). However, in this case we would need to fit a lattice of points over our tuning parameters to find a solution. This greatly increases the number of GASH-ANOVA solutions to compute.

When more than two factors are included in the model the method follows directly because the idea of collapsing two levels of a factor remains the same. In order to collapse two levels, we need the main effect difference and any interaction differences that involve those two

levels to be set to zero. Thus, we need only augment the ϕ vectors with all interaction differences necessary for collapsing the two levels of the given factor. If we assume interactions of order three or greater are null, the ϕ vectors need only be augmented with all two-way interaction differences that involve the given levels of the factor. If we allow for all higher order interactions, the ϕ vector needs to include all higher order interaction differences between the levels as well.

The GASH-ANOVA method can also be suitably altered for use with ordinal factors. For example, suppose factor A is ordinal with the levels of the factor being ordered appropriately. To utilize the GASH-ANOVA procedure we can modify the penalty to only include differences between successive levels as opposed to all pairs, while still imbedding them within the penalized groups. Gertheiss and Tutz (2010) penalized successive differences for ordinal predictors, with the focus being only on simple effects with no interactions.

We can also apply the GASH-ANOVA method when quantitative predictors are included in the model by adding the appropriate interaction parameter differences to our penalty groups. This new quantitative factor then acts like a two level categorical predictor in terms of the penalty groups. We may also include an L_1 constraint with adaptive weight on the continuous predictor if one desires.

3.2 *Investigating Interactions in the Unreplicated Case*

Assume there is only one observation for each level combination, i.e. $n_{ij} = 1$ for all i, j. Using OLS for this case, one usually assumes there is no interaction between the factors as there are not enough degrees of freedom to investigate and test interaction effects. In fact, the error term used for testing the main effect of each factor is the interaction mean square. This is common case when using a Randomized Complete Block Design (RCBD) without replication within blocks. Many of the solutions for inspecting the interaction effects involve partitioning the interaction sum of squares into a part that corresponds to detecting

interactions and a part that is used as the error term. For example, Alin and Kurt (2006) show that Tukey's one degree of freedom test for nonadditivity (Tukey, 1949) is equivalent to looking at the model

$$y_{ij} = \mu + \alpha_i + \beta_j + \kappa\alpha_i\beta_j + \epsilon_{ij}$$

and testing $H_0 : \kappa = 0$ vs $H_A : \kappa \neq 0$. This test separates one degree of freedom from the interaction degrees of freedom (or error degrees of freedom) for estimation and testing of this parameter.

In order to estimate interactions without replication we can use a modified GASH-ANOVA procedure rather than specify a model with new parameters. Recall that the usual GASH-ANOVA penalty requires we first find the OLS solution and use these estimates in our weights. The modification we suggest is to use an unweighted version of the GASH-ANOVA estimator. With the penalty no longer being weighted we lose our asymptotic properties given in section 4, however we are now able to estimate all parameters. Thus, we can inspect the structure of the model and the interactions between the factors by looking at the chosen model fit.

If one is mainly interested in testing if the interaction effects are nonzero, penalized regression can also be used. One method is to use one of the optional penalties that include explicit penalization of the interaction differences discussed in the previous section. Another method would be to simply penalize only the interaction terms or their differences. One can say there are no interactions present if the chosen model collapses all interactions.

4. Asymptotic Properties

When investigating the asymptotic properties of the GASH-ANOVA estimator we assume that each ϕ group is either truly all zero (collapsed) or all differences in that group are truly nonzero. This implies that if a main effect difference in a ϕ group is truly nonzero

then, provided the other factor has at least two distinct levels, the corresponding interaction differences in that ϕ group are all nonzero as well.

Let $A_\alpha = \{(i,j) : \alpha_i \neq \alpha_j\}$ and $A_\beta = \{(i,j) : \beta_i \neq \beta_j\}$ be defined as the set of indices for the main effect differences of each factor that are truly nonzero and let $A_{\alpha,n} = \{(i,j) : \hat{\alpha}_i \neq \hat{\alpha}_j\}$ and $A_{\beta,n} = \left\{(i,j) : \hat{\beta}_i \neq \hat{\beta}_j\right\}$ be defined as the set of indices for each factor whose main effect differences are estimated as nonzero. For the pairwise differences indexed by A_α and A_β, let $\boldsymbol{\eta}_{A_\alpha,A_\beta}$ be the vector of those pairwise differences along with their corresponding interaction differences. Notice that the sets A_α and A_β contain the indices for the truly significant level and factor structure. If this information were known a priori, the solution would be estimated by collapsing down to this structure and then conducting the usual ANOVA analysis. Define $\widetilde{\boldsymbol{\eta}}_{A_\alpha,A_\beta}$ as this so called 'oracle' estimator of $\boldsymbol{\eta}_{A_\alpha,A_\beta}$. It is well known that under standard conditions $n^{-1/2}\left(\widetilde{\boldsymbol{\eta}}_{A_\alpha,A_\beta} - \boldsymbol{\eta}_{A_\alpha,A_\beta}\right) \rightarrow N(0,\Sigma)$, where Σ is singular due to the overparameterization to the pairwise differences. Let $\widehat{\boldsymbol{\eta}}_{A_\alpha,A_\beta}$ denote the GASH-ANOVA estimator of $\boldsymbol{\eta}_{A_\alpha,A_\beta}$. Theorem 1 given below shows that the GASH-ANOVA obtains the oracle property.

The theorem is most easily stated when we rewrite the GASH-ANOVA criterion in its corresponding Lagrangian formulation:

$$\widehat{\boldsymbol{\theta}}^* = argmin_{\boldsymbol{\theta}^*}\left\{\begin{array}{c} ||\mathbf{y} - \mathbf{X}^*\boldsymbol{\theta}^*||^2 + \lambda_n \sum_{1\leqslant i<j\leqslant a} \frac{w_\alpha^{(ij)}}{\sqrt{n}} \max\{|\phi_{\alpha,ij}|\} \\ +\lambda_n \sum_{1\leqslant i<j\leqslant b} \frac{w_\beta^{(ij)}}{\sqrt{n}} \max\{|\phi_{\beta,ij}|\} \end{array}\right\}.$$

Note that there is a one-to-one correspondence with the tuning parameter t and λ_n.

Theorem 1: Suppose that $\lambda_n \rightarrow \infty$ and $\frac{\lambda_n}{\sqrt{n}} \rightarrow 0$. The GASH-ANOVA estimator $\widehat{\theta}$ and its corresponding estimator of the pairwise differences $\widehat{\boldsymbol{\eta}}$ has the following properties:

a) $P(A_{\alpha,n} = A_\alpha) \rightarrow 1$ and $P(A_{\beta,n} = A_\beta) \rightarrow 1$

b) $n^{-1/2}(\widehat{\boldsymbol{\eta}}_{A_\alpha,A_\beta} - \boldsymbol{\eta}_{A_\alpha,A_\beta}) \rightarrow N(0,\Sigma)$

The proof of Theorem 1 is given in the web appendix.

The oracle property states that the method determines the correct structure of the model

with probability tending to one. Additionally, it tells us that one can create a new design matrix corresponding to the reduced model structure selected and conduct inference using the standard asymptotic variance obtained from OLS estimation on that design. Note that this second level of inference may not be necessary, depending on the goals of one's study.

As we mention in section 2, the asymptotic theory developed here will hold for the Naive estimator as well. In fact the proof would follow as a direct extension of Bondell and Reich (2009). Although they would be asymptotically equivalent when the model followed the appropriate heredity structure, the main advantage of the GASH-ANOVA method is its small sample performance. In particular, the structured form of the penalty allows for easier collapsing of levels, selection of factors, and interpretation of the resulting estimated model. This is similar to the relationship between the adaptive LASSO (Zou, 2006) and the adaptive group LASSO (Yuan and Lin, 2006; Wang and Leng, 2008). If the underlying assumptions about the structure of the model hold, in both cases, we would expect increased small sample performance even though the methods would be asymptotically equivalent.

5. Computation and Tuning

The GASH-ANOVA problem can be expressed as a quadratic programming problem. Define

$$\boldsymbol{\zeta} = \boldsymbol{M}\boldsymbol{\theta}^* = (\mu^*, \boldsymbol{\alpha}^{*T}, \boldsymbol{\xi}_\alpha^T, \boldsymbol{\beta}^{*T}, \boldsymbol{\xi}_\beta^T, (\boldsymbol{\alpha}\boldsymbol{\beta})^{*T}, \boldsymbol{\xi}_{\alpha\beta,A}^T, \boldsymbol{\xi}_{\alpha\beta,B}^T)^T,$$

where $\boldsymbol{\xi}_\alpha$ and $\boldsymbol{\xi}_\beta$ are vectors containing the main effect pairwise differences for each factor that do not involve the baseline level and $\boldsymbol{\xi}_{\alpha\beta,A}$ and $\boldsymbol{\xi}_{\alpha\beta,B}$ are vectors containing the interaction pairwise differences of interest for factor A and factor B, respectively, that do not involve

the baseline levels. The matrix

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \mathbf{M}_1 & 0 & 0 \\ 0 & 0 & \mathbf{M}_2 & 0 \\ 0 & 0 & 0 & \mathbf{M}_3 \end{bmatrix}$$

needed to create this new parameter vector is block diagonal. The first block (a scalar) corresponds to μ^*. The second block corresponds to factor A, $\mathbf{M}_1 = \left[\boldsymbol{I}_{a-1}\ \mathbf{D}_1^T\right]^T$, and consists of an identity matrix of size $a-1$ and a matrix $\mathbf{D}_1$ of ± 1 that creates $\boldsymbol{\xi}_\alpha$ that is of dimension $\binom{a-1}{2} \times (a-1)$. The third block, $\mathbf{M}_2$, is defined likewise for factor B. The fourth block, $\mathbf{M}_3$, is also defined similarly except that two difference matrices are needed. Define $\mathbf{D}_3$ as the matrix of ± 1 needed to obtain $\boldsymbol{\xi}_{\alpha\beta,A}$ and define $\mathbf{D}_4$ as the matrix of ± 1 needed to obtain $\boldsymbol{\xi}_{\alpha\beta,B}$, then $\mathbf{M}_3 = \left[\boldsymbol{I}_{(a-1)(b-1)}\ \mathbf{D}_3^T\ \mathbf{D}_4^T\right]^T$.

Next, we set $\boldsymbol{\alpha^*} = \boldsymbol{\alpha^{*+}} - \boldsymbol{\alpha^{*-}}$ with both $\boldsymbol{\alpha^{*+}}$ and $\boldsymbol{\alpha^{*-}}$ being nonnegative (referred to respectively as the positive and negative parts of $\boldsymbol{\alpha^*}$). We also perform this action for all other parts of the $\boldsymbol{\zeta}$ vector except μ^*. Define the parameter vector that includes the positive and negative parts by $\boldsymbol{\tau}$. We split the groups of pairwise differences of parameters into positive and negative parts, denoted by $\phi^+_{\alpha,ij}$, $\phi^+_{\beta,ij}$ and $\phi^-_{\alpha,ij}$, $\phi^-_{\beta,ij}$, respectively. In detail, examples of these groups are $\phi^+_{\alpha,1j} = (\alpha^{*+}_j, (\alpha\beta)^{*+}_{j2}, (\alpha\beta)^{*+}_{j3}, ..., (\alpha\beta)^{*+}_{jb})^T$ for $2 \leqslant j \leqslant a$ and

$$\phi^+_{\alpha,ij} = ((\alpha^*_j - \alpha^*_i)^+, ((\alpha\beta)^*_{j2} - (\alpha\beta)^*_{i2})^+, ..., ((\alpha\beta)^*_{jb} - (\alpha\beta)^*_{ib})^+)^T$$

for $1 < i < j \leqslant a$. We create a new design matrix corresponding to the main effects of factor A by $\mathbf{Z}_\alpha = \left[\mathbf{X}^*_\alpha\ -\mathbf{X}^*_\alpha\ \mathbf{0}_{n\times 2\binom{a-1}{2}}\right]$, where $\mathbf{X}^*_\alpha$ denotes the columns of the design matrix corresponding to factor A. Likewise, we create a new design matrix for the main effect of factor B, $\mathbf{Z}_\beta$. A new design matrix is created similarly for the interactions with two zero matrices appended, $\mathbf{Z}_{\alpha\beta} = \left[\mathbf{X}^*_{\alpha\beta}\ -\mathbf{X}^*_{\alpha\beta}\ \mathbf{0}_{n\times 2r_1}\ \mathbf{0}_{n\times 2r_2}\right]$, where $r_1 = (b-1)\binom{a-1}{2}$ and $r_2 = (a-1)\binom{b-1}{2}$ are the number of pairwise interaction differences corresponding to factor A and factor B, respectively. Let $\mathbf{Z} = [\mathbf{Z}_\alpha\ \mathbf{Z}_\beta\ \mathbf{Z}_{\alpha\beta}]$ be the new full design matrix, implying

$\mathbf{Z}\boldsymbol{\tau} = \mathbf{X}^*\boldsymbol{\theta}^*$. The optimization problem can be written as follows:

$$\widehat{\boldsymbol{\tau}} = argmin_{\boldsymbol{\tau}} \, ||\mathbf{y} - \mathbf{Z}\boldsymbol{\tau}||^2 \tag{3}$$

$$\text{subject to } \mathbf{L}\boldsymbol{\tau} = 0,$$

$$(\boldsymbol{\phi}^+_{\alpha,ij} + \boldsymbol{\phi}^-_{\alpha,ij}) \leqslant s_{\alpha,ij}, \text{ for all } 1 \leqslant i < j \leqslant a,$$

$$(\boldsymbol{\phi}^+_{\beta,ij} + \boldsymbol{\phi}^-_{\beta,ij}) \leqslant s_{\beta,ij}, \text{ for all } 1 \leqslant i < j \leqslant b,$$

$$\textstyle\sum_{1 \leqslant i < j \leqslant a} w^{(ij)}_{\alpha} s_{\alpha,ij} + \sum_{1 \leqslant i < j \leqslant b} w^{(ij)}_{\beta} s_{\beta,ij} \leqslant t,$$

$$\text{and } \boldsymbol{\xi}^+_{\alpha}, \boldsymbol{\xi}^+_{\beta}, \boldsymbol{\xi}^+_{\alpha\beta}, \boldsymbol{\xi}^-_{\alpha}, \boldsymbol{\xi}^-_{\beta}, \boldsymbol{\xi}^-_{\alpha\beta}, s_{\alpha}, s_{\beta} \geqslant 0,$$

where $s_{\alpha,ij}$ and $s_{\beta,ij}$ are slack variables, s_α and s_β represent the set of α and β slack variables respectively, and

$$\mathbf{L} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \mathbf{L}_1 & 0 & 0 \\ 0 & 0 & \mathbf{L}_2 & 0 \\ 0 & 0 & 0 & \mathbf{L}_3 \end{bmatrix}$$

is a block diagonal matrix with four blocks that ensures the estimated parameters maintain their relationships. The first block (a scalar) corresponds to the mean. The second block corresponds to the factor A main effect differences,

$$\mathbf{L}_1 = \begin{pmatrix} \mathbf{D}_1 & -\mathbf{D}_1 & -\boldsymbol{I}_{\binom{a-1}{2}} & \boldsymbol{I}_{\binom{a-1}{2}} \end{pmatrix}.$$

The third block corresponds to the factor B main effect difference and is defined similarly. The fourth block corresponds to the interaction differences is given by

$$\mathbf{L}_3 = \begin{pmatrix} \mathbf{D}_3 & -\mathbf{D}_3 & -\boldsymbol{I}_{r1} & \boldsymbol{I}_{r1} & \mathbf{0}_{r2} & \mathbf{0}_{r2} \\ \mathbf{D}_4 & -\mathbf{D}_4 & \mathbf{0}_{r1} & \mathbf{0}_{r1} & -\boldsymbol{I}_{r2} & \boldsymbol{I}_{r2} \end{pmatrix}.$$

Note that $\boldsymbol{\phi}^+_{\alpha,ij}$ and $\boldsymbol{\phi}^-_{\alpha,ij}$ are vectors of length b and for any given ij pair $s_{\alpha,ij}$ is a constant. Hence, by the inequality $(\boldsymbol{\phi}^+_{\alpha,ij} + \boldsymbol{\phi}^-_{\alpha,ij}) \leqslant s_{\alpha,ij}$ we really mean each element being less than the slack variable. This is now a quadratic objective function with linear constraints, and hence can be solved by standard quadratic programming methods. Note that the GASH-ANOVA

computation remains a quadratic programming problem when more than two factors are included in the model.

The tuning parameter t can be chosen in a number of standard ways such as k-fold cross-validation, generalized cross-validation, or by minimizing AIC or BIC. The method recommended for choosing the tuning parameter for the GASH-ANOVA procedure is that of minimizing BIC as it has been shown that under general conditions BIC is consistent for model selection. In order to compute BIC, an estimate of the degrees of freedom (df) of the model is needed. A logical estimate for df in the two factor case is to add the number of unique parameter estimates in each parameter group, such as $\boldsymbol{\alpha}^*$. Specifically,

$$\widehat{df} = 1 + a^* + b^* + (ab)^*,$$

where we use one df for the mean, a^* and b^* denote the number of estimated unique coefficients for factor A and B respectively, and $(ab)^*$ denotes the number of estimated unique interaction coefficients.

Although intuitively this estimate of df seems reasonable, we have no proof of its unbiasedness. It may be possible to derive an unbiased estimate using the methods in Kato (2009), however, for tuning purposes, this estimate of degrees of freedom works reasonably well.

In practice, one must choose a grid of tuning parameters to select the model from. We advocate fitting models on an equally spaced grid from zero to $G = \binom{a}{2} + \binom{b}{2}$, the total number of ϕ groups. This will approximate the entire path of solutions since, due to the adaptive weights, any value of t greater than G will yield the OLS solution. Using our estimate for degrees of freedom, the minimum BIC chosen estimate will likely fall at a point along the path where one of our groups are collapsed to zero or at the full least squares solution. Thus, we need to be sure that we fit a fine enough grid to approximately capture these points.

6. Simulation Studies

In order to assess the performance of the GASH-ANOVA procedure two Monte Carlo simulation studies were performed and analysis on a number of different criteria were compared with two different types of competitors: constrained regression with no heredity-type constraint and post hoc hypothesis testing procedures.

6.1 *Simulation Set-up*

The simulation set-up consisted of a two factor design having eight levels for factor A and four levels for factor B. A balanced design was used with sample sizes of 64, 192, and 320, corresponding to two, six, and ten replications per treatment combination respectively. The response was generated according to a normal distribution with an error variance of one. Two different effect vectors, θ_1 and θ_2 were used and the table of cell means corresponding to each vector are given in tables 1 and 2.

*****TABLES 1 AND 2 GO HERE*****

We can see that in terms of the cell means table, a $\phi_{\alpha,ij}$ group is zero only if column i and j are equal and a $\phi_{\beta,ij}$ group is zero only if row i and j are equal. The vector θ_1 consisted of four distinct levels with 18 true nonzero differences between levels for factor A and three distinct levels having five true nonzero differences between levels for factor B. Using the full rank baseline reparametrization, there were 68 truly nonzero pairwise differences of interest and 71 truly zero pairwise differences of interest. The vector θ_2 consisted of three distinct levels for both factors, with 13 and five true nonzero differences between levels for factor A and B, respectively. In terms of pairwise differences of interest, there were 61 truly nonzero differences and 78 truly zero differences. The analysis was run on 300 independent data sets at each setting of sample size and effect vector.

In order to inspect the control of the family-wise error rate (FWER), null model simulations

(all true parameter vectors set to zero) were also conducted. The simulation set-up above was used with two, four, and eight replications and an error variance of 16.

6.2 *Competitors and Methods of Evaluation*

The GASH-ANOVA procedure, fit on an equally spaced grid from zero to $G = 34$ by increments of 0.01, was evaluated against four competitors. The first competitor was the No Heredity method described in section 2. The other competitors were post hoc hypothesis testing methods that tested pairwise comparisons of interest. The p-values from these tests (HT method) along with p-values corrected for multiple comparisons using the conservative Bonferroni approach (Bon) and the false discovery rate approach (BH) of Benjamini and Hochberg (Benjamini and Hochberg, 1995) were obtained and groups of p-values that corresponded to tests of each member of $\phi_{\alpha,ij}$ and $\phi_{\beta,ij}$ were evaluated at the 0.05 level. If none of the p-values in a $\phi_{\alpha,ij}$ or $\phi_{\beta,ij}$ group were significant then levels i and j for the corresponding factor were considered collapsed. If any of the p-values in a group were significant, then the corresponding levels were considered to be significantly different. Note that these methods also do not encourage the collapsing of levels the way we desire. For the GASH-ANOVA and the No Heredity methods, if all of the differences in a $\phi_{\alpha,ij}$ or $\phi_{\beta,ij}$ group were estimated at zero then levels i and j of the corresponding factor were considered collapsed. If any of the differences in a group were nonzero then the corresponding levels were considered to be significantly different.

We use the sets $A_\alpha, A_{\alpha,n}, A_\beta$, and $A_{\beta,n}$ to define the criteria that are used for comparisons of the procedures. Due to the different procedures for deciding if we collapse two levels or deem them significantly different, we must extend our definitions of $A_{\alpha,n}$ and $A_{\beta,n}$. Define

$A_{\alpha,n} = \left\{(i,j) : F(\hat{\phi}_{\alpha,ij}) \neq 0\right\}$, where

$$F(\hat{\phi}_{\alpha,ij}) = \begin{cases} ||\hat{\phi}_{\alpha,ij}||^2 & \text{for GASH and Naive methods} \\ \mathbb{1}_{\hat{\phi}_{\alpha,ij}} & \text{for hypothesis testing methods} \end{cases}$$

and $\mathbb{1}_{\hat{\phi}_{\alpha,ij}}$ is an indicator function that is one if any p-value in the $\hat{\phi}_{\alpha,ij}$ is deemed significant and zero otherwise. The set $A_{\beta,n}$ is defined similarly.

The GASH-ANOVA, Naive, HT, Bon, and BH methods were all evaluated and compared on a number of criteria. Let us consider the null hypothesis that we collapse two levels of a factor against the alternative that those levels differ significantly. A '1-TypeI' error criterion is defined as $\frac{|A^c_{\alpha,n} \cap A^c_\alpha| + |A^c_{\beta,n} \cap A^c_\beta|}{|A^c_\alpha| + |A^c_\beta|}$, where $|A|$ is the cardinality of A. In words, this is the number of collapsed level differences found that truly should have been collapsed divided by the true total number of collapsed level differences. A 'Power' criterion was likewise defined as $\frac{|A_{\alpha,n} \cap A_\alpha| + |A_{\beta,n} \cap A_\beta|}{|A_\alpha| + |A_\beta|}$ or the number of significantly different level differences found that truly differed divided by the true total number of significantly different level differences. The number of collapsed differences between levels in each data set (Collapsed) was found along with the false collapse rate (FCR), which is the number of incorrectly collapsed differences divided by the number of collapses found, i.e. $\frac{|A^c_{\alpha,n} \cap A_\alpha| + |A^c_{\beta,n} \cap A_\beta|}{|A^c_{\alpha,n}| + |A^c_{\beta,n}|}$. The number of significant differences between levels in each data set (Sig) was also found along with the false significance rate (FSR), which is the number of incorrect significantly different level differences found divided by the total number of significantly different level differences found, i.e. $\frac{|A_{\alpha,n} \cap A^c_\alpha| + |A_{\beta,n} \cap A^c_\beta|}{|A_{\alpha,n}| + |A_{\beta,n}|}$. Likewise, we define these criteria for the pairwise differences of interest. All of the criteria were averaged across the 300 data sets. These results are given in tables 4 and 5.

Table 3 was produced from the null model simulation. This table gives the oracle percent. That is, the percent of datasets such that $A_\alpha = A_{\alpha,n}$ and $A_\beta = A_{\beta,n}$, which acts like the

FWER in this situation. The average number of significant differences found between both levels and pairwise differences of interest was also reported.

6.3 *Simulation Results*

We are interested in how each method does in terms of controlling the FWER. Table 3 shows that the corrected hypothesis testing methods do as they are designed to, hold the FWER approximately at 0.95. We can see that the Naive method performs better for this criterion as the sample size grows, but the GASH-ANOVA procedure performs extremely well in all cases and that its FWER approaches one as the sample size grows.

*****TABLE 3 GOES HERE*****

Looking at the '1-Type 1' and 'Power' columns of tables 4 and 5, we see that the GASH-ANOVA procedure is the only method that has high 'power' for finding both the significant level differences and significant pairwise differences. This is due to the heredity-type structure that the method requires its model to have. The other methods may be able to find one or more of the pairwise differences of a truly nonzero level difference significant (leading to high level power), but the GASH-ANOVA procedure's structural constraint forces all pairwise differences of interest for a level to be significant if the level difference is significant. Thus, we see the advantage and usefulness of the constraint. The GASH-ANOVA procedure also dominates the Naive procedure in terms of the '1-Type 1' criterion for the levels for both effect vectors and for pairwise differences for effect vector θ_2. The corrected hypothesis testing procedures perform very well in this aspect, but lack the power to find significant pairwise differences, especially compared to the GASH-ANOVA procedure. Thus, we can see that not only does the GASH-ANOVA method tend to have good performance in terms of controlling the FWER, it also performs very well in terms of power.

*****TABLES 4 AND 5 GO HERE*****

We also see that the average number of significant level differences the GASH-ANOVA procedure found is very close to the true number of significant level differences for both effect vectors. The procedure does tend to find too many significant pairwise differences on average for effect vector θ_1, especially compared to the Naive procedure, but performs very well in that respect for effect vector θ_2. The corrected hypothesis testing methods perform very poorly in terms of average number of significant level differences found for the more difficult sample size cases, but perform well with larger samples sizes. However, we again see the usefulness of the structural constraint when we look at the average number of significant pairwise differences found. For the corrected hypothesis testing procedures the average number of significantly different level differences is very close to the correct number, but the average number of significant pairwise differences found is far too small in every case.

7. Real Data Example

The GASH-ANOVA procedure was applied to data from a memory trial done by Eysenck (1974). The trial was designed to investigate the memory capacity of two ages of people (Young and Old) by having them recall a categorized word list. There were 50 subjects in each age group that were randomly assigned to one of five learning groups: Counting, Rhyming, Adjective, Imagery, and Intentional. The Counting group was to count and record the number of letters in each word. The Rhyming group was told to think of and say out loud a word that rhymed with each word given. The Adjective group was to find a suitable modifying adjective for each word and to say each out loud. The Imagery group was to create an image of the word in their mind. These learning groups were increasing in the level of processing required with Counting being the lowest level and Imagery being the highest. The subjects assigned to the first four learning groups were not told they would need to recall the words given, but the Intentional group was told they would be asked to recall the words.

The setup of this experiment is that of a balanced two-way ANOVA with replication (10 per treatment combination), allowing for interactions to be investigated. The standard analysis was run treating the Old age group and the Adjective learning group as the baseline levels. The analysis showed that both main effects and the interaction effect were significant at the 0.05 level. To get an idea about the data, the means for each treatment combination are given in table 7. The GASH-ANOVA procedure, the Naive method, and the BH and Bonferroni p-value correction methods were applied to the data and evaluated in the same manner that was done in the simulation studies.

*****TABLES 7 AND 6 GO HERE*****

As we can see from table 6, the GASH-ANOVA solution collapsed the Counting and Rhyming treatment groups but did not collapse any other levels from either factor. The Naive method did not collapse any levels of either factor. From table 7 we can see that the Naive method did collapse the main effect for the Rhyming and Counting groups, but the corresponding interaction difference was estimated as nonzero. This implies that Rhyming and Counting learning groups were collapsed for the Old age group only. The BH and Bonferonni procedures found that the Counting and Rhyming groups, the Adjective and Imagery groups, and the Imagery and Intentional groups were not different. These two methods did happen to collapse the interactions corresponding to those main effects.

Here we can see that the p-value correction methods form overlapping groups in that the imagery and intentional groups form one set, while the imagery group is also collapsed to the adjective group, whereas the intentional group is not. The Naive method does create non-overlapping groups, however, it seems that the lack of model structure may have prevented two levels from being collapsed. The p-value correction methods do follow the level collapsing structure in this example, but this need not be the case. Based on the simulation results, the p-value correction methods also suffer from lack of power. We see this here as the

GASH-ANOVA procedure is able to detect more significant differences between the levels of the learning group factor. Thus, we can see the advantages inherent in the GASH-ANOVA procedure. The GASH-ANOVA procedure's estimates are designed to encourage the collapsing of levels and they have the advantage of automatically creating non-overlapping groups.

8. Discussion

In this paper we have proposed a constrained regression method that enforces a structural constraint on the model using an infinity norm penalty on groups of pairwise differences of parameters. The method automatically selects important factors and forms non-overlapping groups of levels within a factor. The method is shown to enjoy the 'oracle' property. Simulation studies and a real data example show the effectiveness of the method and the benefit it gives over a similar method that does not impose a structural constraint and over post hoc hypothesis testing procedures. The simulation studies show that in terms of identifying the correct structure of the model, finding the significant pairwise differences of interest, and maintaining high family wise error rate, the GASH-ANOVA procedure performs the best of all the methods compared. The computation for the problem is shown to be a quadratic programming problem with linear constraints and is feasible in most situations.

9. Supplementary Materials

The Web Appendix referenced in Section 4 is available with this paper at the Biometrics website on Wiley Online Library.

10. Acknowledgements

The authors would like to thank the editor, associate editor and an anonymous referee for their help in improving this manuscript. Bondell's research was supported in part by NSF grant DMS-1005612 and NIH grants R01-MH-084022 and P01-CA-142538.

References

Alin, A., and Kurt, S. (2006). Testing non-additivity (interaction) in the two-way anova tables with no replication. *Statistical Methods in Medical Research*, 15:63–85.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. *Biometrics*, 64:115–123.

Bondell, H. D. and Reich, B. J. (2009). Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, 65:169–177.

Chipman, H. (1996). Bayesian variable selection with related predictors. *The Canadian Journal of Statistics*, 24:17–36.

Choi, N., Li,W. and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105:354–364.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

Franck, C., Osborne, J., and Nielsen, D. (2011). An all configurations approach for detecting hidden-additivity in two-way unreplicated experiments. *North Carolina State University Technical Report Number 2636.*

Gertheiss, J. and Tutz, G. (2010). Sparse modeling of categorical explanatory variables.

Annals of Applied Statistics, 4:2150–2180.

Hamada, M. and Wu, C. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, 24:130–137.

Kato, K. (2009). On the degrees of freedom in shrinkage estimation. *Journal of Multivariate Analysis*, 100:1338–1352.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society B*, 67:91–108.

Tukey, J. (1949). One degree of freedom for non-additivity. *Biometrics*, 5:232–242.

Wang, H. and Leng, C. (2008). A note on adaptive group lasso. *Computational Statistics and Data Analysis*, 52:5277–5286.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67.

Yuan, M., Joseph, V. and Lin, Y. (2007). An efficient variable selection approach for analyzing designed experiments. *Technometrics*, 49:430.

Yuan, M., Joseph, V. and Zou, H. (2009). Structured variable selection and estimation. *The Annals of Applied Statistics*, 3:1738–1757.

Zhang, P., Rocha, G. and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37:3468–3497.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67:301–320.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the 'degrees of freedom' of the lasso. *The*

Annals of Statistics, 35:2173–2192.

Zou, H. and Yuan, M. (2008). The f_{inf}-norm support vector machine. *Statistica Sinica*, 18:379–398.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

[Table 7 about here.]

Table 1

Table of Cell Means for θ_1

		Factor A							
	Level	1	2	3	4	5	6	7	8
Factor B	1	2	4.5	-1	0	2	2	2	2
	2	4.5	8.5	2.5	3	4.5	4.5	4.5	4.5
	3	3	5	2	0	3	3	3	3
	4	3	5	2	0	3	3	3	3

Table 2

Table of Cell Means for θ_2

		Factor A							
	Level	1	2	3	4	5	6	7	8
Factor B	1	7	-2	2	2	2	2	2	2
	2	15	8	8	8	8	8	8	8
	3	3	0	-1	-1	-1	-1	-1	-1
	4	3	0	-1	-1	-1	-1	-1	-1

Table 3

Null Model Simulation Results

		Oracle	Avg Sig Levels	Avg Sig Pairwise
$Reps = 2$	GASH	0.95	0.16	0.16
	Naive	0.66	7.24	14.86
	HT	0.18	5.26	7.07
	Bon	0.96	0.06	0.06
	BH	0.96	0.15	0.17
$Reps = 4$	GASH	0.95	0.18	0.18
	Naive	0.83	2.67	4.36
	HT	0.11	5.72	7.68
	Bon	0.97	0.07	0.07
	BH	0.96	0.13	0.16
$Reps = 8$	GASH	0.98	0.08	0.08
	Naive	0.92	1.16	1.81
	HT	0.09	5.44	7.18
	Bon	0.96	0.05	0.05
	BH	0.96	0.07	0.09

Table 4

Simulation Results for Effect Vector θ_1. There are 23 truly nonzero level differences and 11 level differences that should be collapsed. There are 68 truly nonzero pairwise differences and 71 pairwise differences that should be collapsed.

		1-Type1		Power		Avg Sig		Avg FSR		Avg Col		Avg FCR	
		Levels	Diffs	Levels	Diffs	Levels	Diffs	Levels	Diffs	Levels	Diffs	Levels	Diffs
$Reps = 2$	GASH	0.703	0.599	0.950	0.906	25.120	90.073	0.112	0.284	8.880	48.927	0.060	0.106
	Naive	0.196	0.543	0.998	0.736	31.797	82.500	0.270	0.355	2.203	56.500	0.008	0.312
	HT	0.850	0.949	0.691	0.286	17.557	23.093	0.081	0.130	16.443	115.907	0.396	0.417
	Bon	0.999	1.000	0.114	0.040	2.627	2.727	0.002	0.005	31.373	136.273	0.647	0.479
	BH	0.981	0.994	0.243	0.091	5.803	6.637	0.016	0.027	28.197	132.363	0.599	0.466
$Reps = 6$	GASH	0.777	0.642	0.999	0.988	25.437	92.603	0.089	0.264	8.563	46.397	0.002	0.017
	Naive	0.419	0.753	1.000	0.759	29.387	69.190	0.202	0.209	4.613	69.810	0.000	0.220
	HT	0.840	0.948	0.982	0.538	24.347	40.303	0.067	0.083	9.653	98.697	0.037	0.317
	Bon	0.998	1.000	0.540	0.203	12.447	13.817	0.001	0.002	21.553	125.183	0.480	0.432
	BH	0.958	0.987	0.849	0.384	19.993	27.077	0.020	0.027	14.007	111.923	0.221	0.372
$Reps = 10$	GASH	0.848	0.707	1.000	0.994	24.673	88.357	0.064	0.227	9.327	50.643	0.000	0.008
	Naive	0.514	0.809	1.000	0.816	28.347	68.997	0.172	0.162	5.653	70.003	0.000	0.173
	HT	0.863	0.952	0.998	0.653	24.467	47.807	0.058	0.067	9.533	91.193	0.004	0.257
	Bon	0.998	1.000	0.781	0.325	17.993	22.107	0.001	0.002	16.007	116.893	0.297	0.392
	BH	0.963	0.987	0.982	0.543	22.983	37.850	0.016	0.022	11.017	101.150	0.033	0.306

Table 5

Simulation Results for θ_2. There are 18 truly nonzero level differences and 16 level differences that should be collapsed. There are 61 truly nonzero pairwise differences and 78 pairwise differences that should be collapsed.

		1-Type1		Power		Avg Sig		Avg FSR		Avg Col		Avg FCR	
		Levels	Diffs	Levels	Diffs	Levels	Diffs	Levels	Diffs	Levels	Diffs	Levels	Diffs
$Reps = 2$	GASH	0.894	0.918	0.941	0.914	18.640	62.210	0.073	0.073	15.360	76.790	0.049	0.053
	Naive	0.230	0.583	0.994	0.838	30.213	83.637	0.386	0.346	3.787	55.363	0.005	0.172
	HT	0.852	0.950	0.995	0.652	20.273	43.690	0.106	0.080	13.727	95.310	0.007	0.221
	Bon	0.998	1.000	0.738	0.304	13.320	18.607	0.002	0.001	20.680	120.393	0.209	0.350
	BH	0.956	0.985	0.957	0.544	17.927	34.327	0.033	0.027	16.073	104.673	0.041	0.262
$Reps = 6$	GASH	0.941	0.961	1.000	0.992	18.950	63.580	0.045	0.040	15.050	75.420	0.000	0.006
	Naive	0.583	0.832	1.000	0.859	24.677	65.490	0.233	0.162	9.323	73.510	0.000	0.111
	HT	0.848	0.949	1.000	0.791	20.427	52.230	0.109	0.070	13.573	86.770	0.000	0.146
	Bon	0.998	0.999	0.999	0.644	18.017	39.327	0.002	0.001	15.983	99.673	0.001	0.217
	BH	0.941	0.981	1.000	0.744	18.940	46.827	0.045	0.028	15.060	92.173	0.000	0.169
$Reps = 10$	GASH	0.956	0.974	1.000	0.997	18.697	62.813	0.034	0.028	15.303	76.187	0.000	0.002
	Naive	0.709	0.896	1.000	0.862	22.653	60.683	0.170	0.105	11.347	78.317	0.000	0.102
	HT	0.848	0.944	1.000	0.838	20.440	55.477	0.110	0.073	13.560	83.523	0.000	0.117
	Bon	0.999	1.000	1.000	0.694	18.020	42.403	0.001	0.001	15.980	96.597	0.000	0.192
	BH	0.934	0.978	1.000	0.794	19.053	50.123	0.050	0.031	14.947	88.877	0.000	0.141

Table 6

Distinct Levels within Factors

Age	GASH/Naive/BH/Bon	Group	GASH	Naive	BH/Bon
Old	A	Counting	A	A	A
Young	B	Rhyming	A	B	A
		Adjective	B	C	B
		Imagery	C	D	BC
		Intentional	D	E	C

Table 7

Treatment Combination Means and Distinct Levels

Age	Group	Mean	SD	GASH	Naive	BH/Bon
Young	Counting	6.5	1.43	A	A	A
Young	Rhyming	7.6	1.96	A	B	A
Young	Adjective	14.8	3.49	B	C	B
Young	Imagery	17.6	2.59	C	D	BC
Young	Intentional	19.3	2.67	D	E	C
Old	Counting	7.0	1.83	E	F	D
Old	Rhyming	6.9	2.13	E	F	D
Old	Adjective	11.0	2.49	F	G	E
Old	Imagery	13.4	4.50	G	H	EF
Old	Intentional	12.0	3.74	H	I	F