

# Factor Selection and Structural Identification in the Interaction ANOVA Model

Justin Post and Howard D. Bondell

## Abstract

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. For interpretability, when collapsing two levels of a factor it may not make sense to collapse the main effect difference to zero while still having interaction differences between those levels nonzero. 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. A group norm penalty is placed on parameters that encourages levels of factors to collapse and entire factors to be set to zero. The procedure is called GASH-ANOVA for grouping and selection using heredity in ANOVA. 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 of a modified GASH-ANOVA estimator for estimating interactions in the unreplicated case. Simulation studies show that the GASH-ANOVA 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.

# 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 their main effects also collapsed together. 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, N., Li, W. and Zhu, J. (2010) used heredity to extend the LASSO variable selection technique to include interaction terms, Yuan, M., Joseph, V. and Zou, H. (2009) used heredity in variable selection with the non-negative garrote, Yuan, M., Joseph, V. and Lin, Y. (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 also discuss the use of an unweighted version of the GASH-ANOVA estimator to estimate interaction terms in the unreplicated ANOVA model.

The rest of the article is organized as follows: In section 2, notation is introduced and the related CAS-ANOVA procedure is reviewed. In section 3.1, the GASH-ANOVA procedure is introduced. The extension to the unreplicated case is discussed 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 and section 7 shows the

method's usefulness on a real data example. Finally, section 8 gives a discussion.

## 2 Notation and CAS-ANOVA

### 2.1 Notation

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  $n = \sum_{i,j} n_{ij}$ . Note that for a balanced design we have  $n = abn_{ij}$ . 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 error terms. The parameter vector consists of the mean and three stacked vectors,  $\boldsymbol{\theta} = (\mu, \boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, (\boldsymbol{\alpha}\boldsymbol{\beta})^T)^T$ , where  $\mu$  is the intercept,  $\boldsymbol{\alpha}^T = (\alpha_1, \alpha_2, \dots, \alpha_a)$ ,  $\boldsymbol{\beta}^T = (\beta_1, \beta_2, \dots, \beta_b)$ , and  $(\boldsymbol{\alpha}\boldsymbol{\beta})^T = ((\alpha\beta)_{11}, (\alpha\beta)_{12}, \dots, (\alpha\beta)_{1b}, (\alpha\beta)_{21}, \dots, (\alpha\beta)_{2b}, \dots, (\alpha\beta)_{ab})$ . Here,  $\alpha_i$  corresponds to the main effect of level  $i$  of factor A,  $\beta_j$  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. The ordinary least squares (OLS) solution can be written as follows:

$$\hat{\boldsymbol{\theta}}_{OLS} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 \quad (1)$$

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

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.

### 2.2 CAS-ANOVA

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

$$\sum_{1 \leq k < m \leq a} w_{\alpha, CAS}^{(km)} |\alpha_k - \alpha_m| + \sum_{1 \leq k < m \leq b} w_{\beta, CAS}^{(km)} |\beta_k - \beta_m| \leq t,$$

where  $t > 0$  is a tuning constant and  $w_{\alpha, CAS}^{(km)}$  and  $w_{\beta, CAS}^{(km)}$  are scaled adaptive LASSO (Zou, 2006) type weights for the pair of levels  $k$  and  $m$  of each factor.

### 2.3 No Heredity Method

We denote the extension of the CAS-ANOVA procedure to the interaction ANOVA model as the ‘No Heredity’ (NH) method. Let  $\hat{\alpha}_{i, OLS}$ ,  $\hat{\beta}_{j, OLS}$ , and  $(\widehat{\alpha\beta})_{ij, OLS}$  be the OLS estimates of the corresponding parameters found using equation (1). The NH method’s penalty is given by

$$\begin{aligned} & \sum_{1 \leq k < m \leq a} w_{\alpha, NH}^{(km)} |\alpha_k - \alpha_m| + \sum_{1 \leq k < m \leq b} w_{\beta, NH}^{(km)} |\beta_k - \beta_m| \\ & + \sum_{1 \leq m \leq b} \sum_{1 \leq k < l \leq a} w_{\alpha\beta, NH}^{(km, lm)} |(\alpha\beta)_{km} - (\alpha\beta)_{lm}| \\ & + \sum_{1 \leq k \leq a} \sum_{1 \leq m \leq l < b} w_{\alpha\beta, NH}^{(km, kl)} |(\alpha\beta)_{km} - (\alpha\beta)_{kl}| \leq t, \end{aligned}$$

where the weights on the main effect differences are given by

$$w_{\alpha, NH}^{(km)} = |\hat{\alpha}_{k, OLS} - \hat{\alpha}_{m, OLS}|^{-1} \quad \text{and} \quad w_{\beta, NH}^{(km)} = |\hat{\beta}_{k, OLS} - \hat{\beta}_{m, OLS}|^{-1}$$

and the weights placed on the interaction differences are given by

$$w_{\alpha\beta, NH}^{(km, kl)} = |(\widehat{\alpha\beta})_{km, OLS} - (\widehat{\alpha\beta})_{kl, OLS}|^{-1}$$

and

$$w_{\alpha\beta, NH}^{(km, lm)} = |(\widehat{\alpha\beta})_{km, OLS} - (\widehat{\alpha\beta})_{lm, OLS}|^{-1}.$$

Asymptotically the NH method may 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 discussed in section 1. In problems where the factors have a large number of levels, the NH 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,  $(\alpha\beta)_{m1} - (\alpha\beta)_{k1}, (\alpha\beta)_{m2} - (\alpha\beta)_{k2}, \dots, (\alpha\beta)_{mb} - (\alpha\beta)_{kb}$ , have been set to zero. There is a chance for the NH 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, this extension of the CAS-ANOVA procedure is not ideal with interactions present. In section 6 and 7, the NH method is compared to the GASH-ANOVA procedure.

### 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 reparametrize to a full rank design matrix using a reference level as a baseline. This lessens the number of parameters and constraints needed. 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  $\boldsymbol{\theta}^* = (\mu^*, \boldsymbol{\alpha}^{*T}, \boldsymbol{\beta}^{*T}, (\boldsymbol{\alpha}\boldsymbol{\beta})^{*T})^T$ , where  $\mu^* = \mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11}$ ,

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

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

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

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  $\phi_{\alpha,ij}, 1 \leq i < j \leq a$  or  $\phi_{\beta,ij}, 1 \leq i < j \leq b$ , where  $\phi_{\alpha,1j} = (\alpha_j^*, (\alpha\beta)_{j2}^*, (\alpha\beta)_{j3}^*, \dots, (\alpha\beta)_{jb}^*)$  for  $2 \leq j \leq a$  and  $\phi_{\alpha,ij} = (\alpha_j^* - \alpha_i^*, (\alpha\beta)_{j2}^* - (\alpha\beta)_{i2}^*, (\alpha\beta)_{j3}^* - (\alpha\beta)_{i3}^*, \dots, (\alpha\beta)_{jb}^* - (\alpha\beta)_{ib}^*)$  for  $1 < i < j \leq a$  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, P., Rocha, G. and Yu, B., 2009). However, the CAP treats coefficients themselves as groups, not the differences of coefficients as groups.

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

$$\hat{\boldsymbol{\theta}}^* = \underset{\boldsymbol{\theta}^*}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}^* \boldsymbol{\theta}^*\|^2 \quad (2)$$

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

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

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

for  $2 \leq j \leq a$  and

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

for  $1 < i < j \leq a$ , and  $|\phi_{\beta,ij}|$  is similarly defined. Using equation (1) to obtain the OLS solution, the weight  $w_{\alpha}^{(ij)}$  is given by  $(\max \{|\hat{\phi}_{\alpha,ij,OLS}|\})^{-1}$ , where  $\hat{\phi}_{\alpha,ij,OLS}$  denotes the use of the OLS estimate for the differences and 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.

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. If all interaction differences for all factors are set to zero then we are left with the additive model. To accomplish the

collapsing of individual interaction differences we can explicitly add these terms to the penalty. The new penalty would be given by

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

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. With more than two factors 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  $\phi$  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  $\phi$  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  $\phi$  vector needs to include all higher order interaction differences between the levels.

### 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  $i, j$ . This is common case when using a Randomized Complete Block Design (RCBD). Using OLS for this case one usually assumes there is no interaction between the factors (or the factor and the block) as there are not enough degrees of freedom to investigate and test interaction effects. A number of solutions have been proposed to investigate the interaction when no replication is present (see Franck, C., Osborne, J., and Nielsen, D. (2011) for a detailed review and comparison of methods).



We can use a modified GASH-ANOVA procedure to investigate interactions in the unreplicated case. This modified version uses an unweighted penalty enabling us to estimate our model fits. The optional penalty that includes explicit penalization of the interaction differences discussed in the previous section could also be applied here.

One issue of note for the unweighted GASH-ANOVA procedure is that of scaling. In penalized regression procedures, having the effects on a comparable scale is important to ensure the penalization is done equally. In most penalization methods, the design matrix is scaled as to have unit  $L_2$  norm. However, since we have differences of parameters in our penalty, the way to standardize the variables is not clear. The usual GASH-ANOVA procedure remedies this issue by using adaptive weights, essentially placing the penalized terms on the same scale.

## 4 Asymptotic Properties

When investigating the asymptotic properties of the GASH-ANOVA estimator we assume that each  $\phi$  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  $\phi$  group is truly nonzero then, provided the other factor has at least two distinct levels, the corresponding interaction differences in that  $\phi$  group must all be 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  $A_{\beta, n} = \{(i, j) : \hat{\beta}_i \neq \hat{\beta}_j\}$  be defined as the set of indices for each factor whose main effect differences are estimated as nonzero. For the pairwise differences indexed by  $A_\alpha$  and  $A_\beta$ , let  $\boldsymbol{\eta}_{A_\alpha, A_\beta}$  be the vector of those pairwise differences along with their corresponding interaction differences. Notice that the sets  $A_\alpha$  and  $A_\beta$  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  $\tilde{\boldsymbol{\eta}}_{A_\alpha, A_\beta}$  as this so called ‘oracle’ estimator of  $\boldsymbol{\eta}_{A_\alpha, A_\beta}$ . It is well known that under standard conditions  $n^{-1/2} \left( \tilde{\boldsymbol{\eta}}_{A_\alpha, A_\beta} - \boldsymbol{\eta}_{A_\alpha, A_\beta} \right) \rightarrow N(0, \Sigma)$ . Let  $\hat{\boldsymbol{\eta}}_{A_\alpha, A_\beta}$  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:

$$\hat{\boldsymbol{\theta}}^* = \underset{\boldsymbol{\theta}^*}{\operatorname{argmin}} \left\{ \begin{aligned} & \|\mathbf{y} - \mathbf{X}^* \boldsymbol{\theta}^*\|^2 + \lambda_n \sum_{1 \leq i < j \leq a} \frac{w_\alpha^{(ij)}}{\sqrt{n}} \max\{|\phi_{\alpha,ij}|\} \\ & + \lambda_n \sum_{1 \leq i < j \leq b} \frac{w_\beta^{(ij)}}{\sqrt{n}} \max\{|\phi_{\beta,ij}|\} \end{aligned} \right\}.$$

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

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

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

The proof of Theorem 1 is given in the 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.

## 5 Computation and Tuning

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

Define

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

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  $\mu^*$ . The second block corresponds to factor A,  $\mathbf{M}_1 = [\mathbf{I}_{a-1} \mathbf{D}_1^T]^T$ , and consists of an identity matrix of size  $a - 1$  and a matrix  $\mathbf{D}_1$  of  $\pm 1$  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  $\pm 1$  needed to obtain  $\xi_{\alpha\beta,A}$  and define  $\mathbf{D}_4$  as the matrix of  $\pm 1$  needed to obtain  $\xi_{\alpha\beta,B}$ , then  $\mathbf{M}_3 = [\mathbf{I}_{(a-1)(b-1)} \quad \mathbf{D}_3^T \quad \mathbf{D}_4^T]^T$ .

Next, we set  $\alpha^* = \alpha^{*+} - \alpha^{*-}$  with both  $\alpha^{*+}$  and  $\alpha^{*-}$  being nonnegative (referred to respectively as the positive and negative parts of  $\alpha^*$ ). We also perform this action for all other parts of the  $\zeta$  vector except  $\mu^*$ . Define the parameter vector that includes the positive and negative parts by  $\tau$ . 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  $\phi_{\alpha,1j}^+ = (\alpha_j^{*+}, (\alpha\beta)_{j2}^{*+}, (\alpha\beta)_{j3}^{*+}, \dots, (\alpha\beta)_{jb}^{*+})^T$  for  $2 \leq j \leq a$  and

$$\phi_{\alpha,ij}^+ = ((\alpha_j^* - \alpha_i^*)^+, ((\alpha\beta)_{j2}^* - (\alpha\beta)_{i2}^*)^+, \dots, ((\alpha\beta)_{jb}^* - (\alpha\beta)_{ib}^*)^+)^T$$

for  $1 < i < j \leq a$ . We create a new design matrix corresponding to the main effects of factor A by  $\mathbf{Z}_\alpha = [\mathbf{X}_\alpha^* \quad -\mathbf{X}_\alpha^* \quad \mathbf{0}_{n \times 2 \binom{a-1}{2}}]$ , 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,  $\mathbf{Z}_{\alpha\beta} = [\mathbf{X}_{\alpha\beta}^* \quad -\mathbf{X}_{\alpha\beta}^* \quad \mathbf{0}_{n \times 2r_1} \quad \mathbf{0}_{n \times 2r_2}]$ , 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  $\mathbf{Z} = [\mathbf{Z}_\alpha \quad \mathbf{Z}_\beta \quad \mathbf{Z}_{\alpha\beta}]$  be the new full design matrix, implying  $\mathbf{Z}\tau = \mathbf{X}^*\theta^*$ . The optimization problem can be written as follows:

$$\hat{\tau} = \underset{\tau}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{Z}\tau\|^2 \quad (3)$$

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

$$\begin{aligned} (\phi_{\alpha,ij}^+ + \phi_{\alpha,ij}^-) &\leq s_{\alpha,ij}, \text{ for all } 1 \leq i < j \leq a, \\ (\phi_{\beta,ij}^+ + \phi_{\beta,ij}^-) &\leq s_{\beta,ij}, \text{ for all } 1 \leq i < j \leq b, \\ \sum_{1 \leq i < j \leq a} w_\alpha^{(ij)} s_{\alpha,ij} + \sum_{1 \leq i < j \leq b} w_\beta^{(ij)} s_{\beta,ij} &\leq t, \\ \text{and } \xi_\alpha^+, \xi_\beta^+, \xi_{\alpha\beta}^+, \xi_\alpha^-, \xi_\beta^-, \xi_{\alpha\beta}^- &, s_\alpha, s_\beta \geq 0, \end{aligned}$$

where  $s_{\alpha,ij}$  and  $s_{\beta,ij}$  are slack variables,  $s_\alpha$  and  $s_\beta$  represent the set of  $\alpha$  and  $\beta$  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,

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

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

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

Note that  $\phi_{\alpha,ij}^+$  and  $\phi_{\alpha,ij}^-$  are vectors of length  $b$  and for any given  $ij$  pair  $s_{\alpha,ij}$  is a constant. Hence, by the inequality  $(\phi_{\alpha,ij}^+ + \phi_{\alpha,ij}^-) \leq s_{\alpha,ij}$  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 use with the GASH-ANOVA procedure is 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. The logical estimate for df in the two factor case is to add the number of unique parameter estimates in each parameter group, such as  $\alpha^*$ . 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.

## 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,  $\theta_1$  and  $\theta_2$  were used and the table of cell means corresponding to each vector are given in tables 1 and 2.

\*\*\*\*\*FIGURE 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  $\theta_1$  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  $\theta_2$  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 was evaluated against four competitors. The first competitor was the No Heredity method described in section 2.2. 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  $A_{\alpha,n} = \{(i, j) : F(\hat{\phi}_{\alpha,ij}) \neq 0\}$ , where

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

and  $\mathbb{I}_{\hat{\phi}_{\alpha,ij}}$  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, NH, 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  $\frac{|A_{\alpha,n}^c \cap A_\alpha^c| + |A_{\beta,n}^c \cap A_\beta^c|}{|A_\alpha^c| + |A_\beta^c|}$ , 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.  $\frac{|A_{\alpha,n}^c \cap A_\alpha| + |A_{\beta,n}^c \cap A_\beta|}{|A_{\alpha,n}^c| + |A_{\beta,n}^c|}$ . 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.  $\frac{|A_{\alpha,n} \cap A_\alpha^c| + |A_{\beta,n} \cap A_\beta^c|}{|A_{\alpha,n}| + |A_{\beta,n}|}$ . Likewise, we de-

fine 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 3 and 4.

Table 5 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

Looking at the ‘1-Type 1’ and ‘Power’ columns of tables 3 and 4, 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 NH procedure in terms of the ‘1-Type 1’ criterion for the levels for both effect vectors and for pairwise differences for effect vector  $\theta_2$ . 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.

\*\*\*\*\*FIGURE 3 AND 4 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  $\theta_1$ , especially compared to the NH procedure, but performs very well in that respect for effect vector  $\theta_2$ . 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.

\*\*\*\*\*FIGURE 5 GOES HERE\*\*\*\*\*

We are also interested in how each method does in terms of controlling the FWER. Table 5 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 NH method performs better for this criterion as the sample size grows, but the GASH-ANOVA procedure performs extremely well in all cases and that the FWER approaches one as the sample size grows. Thus, we can see that not only does the GASH-ANOVA method tend to have the best performance in terms of power, its control of the family-wise error rate is extremely good as well.

## **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 6. The GASH-ANOVA procedure, the NH method, and the



BH and Bonferonni p-value correction methods were applied to the data and evaluated in the same manner that was done in the simulation studies.

\*\*\*\*\*FIGURE 6 AND 7 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 NH method did not collapse any levels of either factor. It did collapse the main effects 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. The NH 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 Appendix

### Proof of Theorem 1:

**Proof of a):** Let  $B_{\alpha,n} = A_{\alpha,n} \cap A_{\alpha}^c$  and  $B_{\beta,n} = A_{\beta,n} \cap A_{\beta}^c$  be the indices of the main effect differences for factor A and B, respectively, that should be estimated at zero but were incorrectly estimated as nonzero. We need to show that the true zeros will be set to zero with probability tending to one or, equivalently, that both  $P(B_{\alpha,n} \neq \emptyset) \rightarrow 0$  and  $P(B_{\beta,n} \neq \emptyset) \rightarrow 0$ . Then we must show that  $P(A_{\alpha,n}^c \cap A_{\alpha} \neq \emptyset) \rightarrow 0$  and  $P(A_{\beta,n}^c \cap A_{\beta} \neq \emptyset) \rightarrow 0$ , i.e. that none of the nonzero differences are mistakenly set to zero. The second item to show will follow directly from the  $\sqrt{n}$ -consistency of the estimators for the differences in  $A_{\alpha}$  and  $A_{\beta}$ , which will be proved in part b of the theorem.

For the first item, we show that  $P(B_{\alpha,n} \neq \emptyset) \rightarrow 0$  and the proof for  $P(B_{\beta,n} \neq \emptyset) \rightarrow 0$  is done similarly. Assuming  $B_{\alpha,n}$  is nonempty, there are two cases to consider: Case (i) A pair of indices in  $B_{\alpha,n}$  has its main effect difference as the maximum of its corresponding  $\phi_{\alpha,ij}$  group. Case (ii) No pair of indices in  $B_{\alpha,n}$  has its main effect difference as the maximum of its corresponding  $\phi_{\alpha,ij}$  group.

**Case (i):** Because we have categorical factors we can sort the levels of factor A so that  $\hat{\alpha}_1 \leq \hat{\alpha}_2 \leq \dots \leq \hat{\alpha}_a$ . Let  $m$  be the largest index of any index pair in  $B_{\alpha,n}$  that is also the maximum of its  $\phi_{\alpha,ij}$  group, i.e.

$$m = \max \{j : (i, j) \in B_{\alpha,n} \text{ for some } i, \alpha_j - \alpha_i = \max |\phi_{\alpha,ij}|\}.$$

Let  $q$  be the smallest index such that the pair  $(q, m) \in B_{\alpha,n}$ , so  $q < m$ . Now we reparameterize to the full rank design matrix using level  $q$  of factor A and some arbitrary level, say  $b$ , of factor B as our baseline. Thus, we define  $\gamma = \mu + \alpha_q + \beta_b + (\alpha\beta)_{qb}$ . We define  $\gamma_k^{\alpha}$  as  $\alpha_k - \alpha_q$  for  $k \neq q$ , define  $\gamma_j^{\beta}$  as  $\beta_j - \beta_b$  for  $j \neq b$ , and define  $\gamma_{kj}^{\alpha\beta}$  to be zero if and only if both  $(\alpha\beta)_{kj} - (\alpha\beta)_{kb} = 0$  and  $(\alpha\beta)_{kj} - (\alpha\beta)_{qj} = 0$  for  $k \neq q, j \neq b$ . We create the new full rank parameter vector,  $\gamma$  by stacking  $\gamma$ , the  $\gamma^{\alpha}$ , the  $\gamma^{\beta}$ , and the  $\gamma^{\alpha\beta}$ . By assumption  $\gamma_m^{\alpha} = 0$  and  $\hat{\gamma}_m^{\alpha} \neq 0$ . Also, by construction we have  $\hat{\gamma}_m^{\alpha} - \hat{\gamma}_k^{\alpha} \geq 0$  for all  $(k, m) \in B_{\alpha,n}$

such that the main effect difference corresponding to  $(k, m)$  is the maximum of its  $\phi_{\alpha, km}$  group. Due to the ordering of the levels chosen, for  $k < l$  at the GASH-ANOVA solution we have

$$|\alpha_l - \alpha_k| = \begin{cases} \gamma_l^\alpha - \gamma_k^\alpha & k \neq q, l \neq q \\ \gamma_l^\alpha & k = q, l > q \\ -\gamma_k^\alpha & k < q, l = q \end{cases}$$

Hence, we can rewrite the solution as

$$\hat{\gamma} = \operatorname{argmin}_{\gamma} \left[ \|\mathbf{y} - Z\gamma\|^2 + \lambda_n J(\gamma) \right], \quad (4)$$

where

$$J(\gamma) = \left\{ \begin{array}{l} \sum_{1 \leq k < q-1} \frac{w_\alpha^{(kq)}}{\sqrt{n}} \max \left\{ (-\gamma_k^\alpha, |\mathbf{\Gamma}_{k+}^{\alpha\beta}|)^T \right\} + \\ \sum_{q < k \leq a} \frac{w_\alpha^{(qk)}}{\sqrt{n}} \max \left\{ (\gamma_k^\alpha, |\mathbf{\Gamma}_{k+}^{\alpha\beta}|)^T \right\} + \\ \sum_{1 \leq k < l \leq a, k \neq q, l \neq q} \frac{w_\alpha^{(lk)}}{\sqrt{n}} \max \left\{ ((\gamma_l^\alpha - \gamma_k^\alpha), |\mathbf{\Gamma}_{l+,k+}^{\alpha\beta}|)^T \right\} + \\ \sum_{1 \leq k < b-1} \frac{w_\beta^{(kb)}}{\sqrt{n}} \max \left\{ (|\gamma_k^\beta|, |\mathbf{\Gamma}_{+k}^{\alpha\beta}|)^T \right\} + \\ \sum_{1 \leq k < l \leq b-1} \frac{w_\beta^{(kl)}}{\sqrt{n}} \max \left\{ (|\gamma_l^\beta - \gamma_k^\beta|, |\mathbf{\Gamma}_{+,k+}^{\alpha\beta}|)^T \right\} \end{array} \right\}, \quad (5)$$

$Z$  is the typical design matrix for the parametrization that treats level  $q$  of factor A and level  $b$  of factor B as the baseline and

$$\begin{aligned} |\mathbf{\Gamma}_{k+}^{\alpha\beta}| &= (|\gamma_{k1}^{\alpha\beta}|, |\gamma_{k2}^{\alpha\beta}|, \dots, |\gamma_{k(b-1)}^{\alpha\beta}|), \\ |\mathbf{\Gamma}_{l+,k+}^{\alpha\beta}| &= (|\gamma_{l1}^{\alpha\beta} - \gamma_{k1}^{\alpha\beta}|, |\gamma_{l2}^{\alpha\beta} - \gamma_{k2}^{\alpha\beta}|, \dots, |\gamma_{l(b-1)}^{\alpha\beta} - \gamma_{k(b-1)}^{\alpha\beta}|), \\ |\mathbf{\Gamma}_{+k}^{\alpha\beta}| &= (|\gamma_{1k}^{\alpha\beta}|, |\gamma_{2k}^{\alpha\beta}|, \dots, |\gamma_{(a-1)k}^{\alpha\beta}|), \\ |\mathbf{\Gamma}_{+,k+}^{\alpha\beta}| &= (|\gamma_{1l}^{\alpha\beta} - \gamma_{1k}^{\alpha\beta}|, |\gamma_{2l}^{\alpha\beta} - \gamma_{2k}^{\alpha\beta}|, \dots, |\gamma_{(a-1)l}^{\alpha\beta} - \gamma_{(a-1)k}^{\alpha\beta}|). \end{aligned}$$

To complete this part of the proof we will obtain a contradiction on a neighborhood of our solution  $\hat{\gamma}_m$ . At the solution the optimization criterion above is differentiable with respect to  $\gamma_m^\alpha$  because  $\hat{\gamma}_m^\alpha \neq 0$ . We investigate this derivative on a neighborhood of the solution on which the differences that are estimated at zero remain at zero. On this neighborhood, the terms involving  $(k, m) \in A_{\alpha, n}^c \cap A_\alpha^c$  can be omitted since they will vanish in the objective function. Because our criterion is differentiable on the neighborhood, our solution  $\hat{\gamma}$  must satisfy

$$\frac{2}{\sqrt{n}} z_m^T (\mathbf{y} - Z\hat{\gamma}) = \left\{ \begin{array}{l} \frac{\lambda_n}{\sqrt{n}} \sum_{k \neq m, (k, m) \in A_\alpha} (-1)^{I[m < k]} \frac{w_\alpha^{(km)}}{\sqrt{n}} I[\hat{\gamma}_m^\alpha] \\ + \frac{\lambda_n}{\sqrt{n}} \sum_{k \neq m, (k, m) \in B_{\alpha, n}} \frac{w_\alpha^{(km)}}{\sqrt{n}} I[\hat{\gamma}_m^\alpha] \end{array} \right\}, \quad (6)$$

where  $z_m^T$  denotes the  $m^{\text{th}}$  column of  $Z$  and  $I[\widehat{\gamma}_m^\alpha]$  is an indicator that is one if the maximum of  $|\phi_{\alpha,km}|$  was a main effect and zero otherwise.

By construction, all of the terms in the second sum on the right hand side are positive and, for this case, the second sum is nonempty. Note that for all  $(k, m) \in B_{\alpha,n}$  we have  $w_\alpha^{(km)} = |\widehat{\alpha}_{m,OLS} - \widehat{\alpha}_{k,OLS}|^{-1}$ . Also for any  $(k, m) \in A_\alpha$  the weight is of the form  $w_\alpha^{(km)} = |\widehat{\alpha}_{m,OLS} - \widehat{\alpha}_{k,OLS}|^{-1}$  or  $w_\alpha^{(km)} = \left| \widehat{(\alpha\beta)}_{ml,OLS} - \widehat{(\alpha\beta)}_{kl,OLS} \right|^{-1}$  where  $l$  is some level of factor B. Therefore, for all  $k$  such that  $(k, m) \in A_\alpha$  we have  $w_\alpha^{(km)} = O_p(1)$ , while for  $(k, m) \in A_\alpha^c$ ,  $n^{-1/2}w_\alpha^{(km)} = O_p(1)$  since the initial OLS estimator is  $\sqrt{n}$ -consistent. Thus, the first sum on the right hand side is  $O_p(\lambda_n n^{-1/2})$  and the terms in the second sum are  $O_p(\lambda_n)$ . Since the second sum is nonempty, the entire right hand side must be  $O_p(\lambda_n)$  since at least one term must be.

However, the left hand side is  $O_p(1)$  and by assumption  $\lambda_n \rightarrow \infty$ . This is a contradiction, thus it must be that  $P(B_{\alpha,n} \neq \emptyset) \rightarrow 0$ .

**Case (ii)** The proof proceeds much like that for case (i). Find the pair  $(m, q)$  such that for all  $j = 1, 2, \dots, b$ ,  $\widehat{(\alpha\beta)}_{mj} - \widehat{(\alpha\beta)}_{qj}$  is the largest estimated difference of any pair in  $B_{\alpha,n}$ . Thus,  $\left| \widehat{(\alpha\beta)}_{mj} - \widehat{(\alpha\beta)}_{qj} \right|$  is also the maximum of its corresponding  $\phi_{\alpha,ij}$  group. Without loss of generality sort the levels of factor A so that  $\widehat{(\alpha\beta)}_{1j} \leq \widehat{(\alpha\beta)}_{2j} \leq \dots \leq \widehat{(\alpha\beta)}_{aj}$ , let  $q < m$ , and assume  $j \neq b$ . As with case (i), we reparameterize to the full rank design matrix using level  $q$  of factor A and level  $b$  of factor B and form the new parameter vector  $\gamma$ . Thus,  $(\alpha\beta)_{mj} - (\alpha\beta)_{qj} = 0$  only if  $\gamma_{mj}^{\alpha\beta} = 0$  and  $\widehat{\gamma}_{mj}^{\alpha\beta}$  is positive and nonzero.

We will find a contradiction by taking the derivative of the full rank optimization criterion with respect to  $\gamma_{mj}^{\alpha\beta}$  on a neighborhood of the solution where the differences that are estimated at zero remain at zero. We can rewrite the optimization criterion with terms in the penalty not involving  $\gamma_{mj}^{\alpha\beta}$  omitted as follows:

$$\widehat{\gamma} = \underset{\gamma}{\operatorname{argmin}} \left[ \|\mathbf{y} - Z\gamma\|^2 + \lambda_n (Q_1(\gamma) + \dots + Q_6(\gamma)) \right], \quad (7)$$

where  $Q_1(\gamma), \dots, Q_6(\gamma)$  are given by

$$\begin{aligned} & \frac{w_\alpha^{(kq)}}{\sqrt{n}} \max \left\{ (|\gamma_m^\alpha|, |\gamma_{m1}^{\alpha\beta}|, \dots, |\gamma_{mj}^{\alpha\beta}|, \dots, |\gamma_{m,b-1}^{\alpha\beta}|)^T \right\}, \\ & \sum_{1 \leq k < q-1} \frac{w_\alpha^{(mk)}}{\sqrt{n}} \max \left\{ (|\gamma_m^\alpha - \gamma_k^\alpha|, |\gamma_{m1}^{\alpha\beta} - \gamma_{k1}^{\alpha\beta}|, \dots, |\gamma_{mj}^{\alpha\beta} - \gamma_{kj}^{\alpha\beta}|, \dots, |\gamma_{m,b-1}^{\alpha\beta} - \gamma_{k,b-1}^{\alpha\beta}|)^T \right\}, \\ & \sum_{q+1 < k \leq a} \frac{w_\alpha^{(mk)}}{\sqrt{n}} \max \left\{ (|\gamma_k^\alpha - \gamma_m^\alpha|, |\gamma_{k1}^{\alpha\beta} - \gamma_{m1}^{\alpha\beta}|, \dots, |\gamma_{kj}^{\alpha\beta} - \gamma_{mj}^{\alpha\beta}|, \dots, |\gamma_{k,b-1}^{\alpha\beta} - \gamma_{m,b-1}^{\alpha\beta}|)^T \right\}, \end{aligned}$$

$$\frac{w_\beta^{(jb)}}{\sqrt{n}} \max \left\{ \left( \left| \gamma_j^\beta \right|, \left| \gamma_{1j}^{\alpha\beta} \right|, \dots, \left| \gamma_{mj}^{\alpha\beta} \right|, \dots, \left| \gamma_{aj}^{\alpha\beta} \right| \right)^T \right\},$$

$$\sum_{1 \leq k < j} \frac{w_\beta^{(jk)}}{\sqrt{n}} \max \left\{ \left( \left| \gamma_j^\beta - \gamma_k^\beta \right|, \left| \gamma_{1j}^{\alpha\beta} - \gamma_{1k}^{\alpha\beta} \right|, \dots, \left| \gamma_{mj}^{\alpha\beta} - \gamma_{mk}^{\alpha\beta} \right|, \dots, \left| \gamma_{aj}^{\alpha\beta} - \gamma_{ak}^{\alpha\beta} \right| \right)^T \right\},$$

and

$$\sum_{j < k < b} \frac{w_\beta^{(jk)}}{\sqrt{n}} \max \left\{ \left( \left| \gamma_k^\beta - \gamma_j^\beta \right|, \left| \gamma_{1k}^{\alpha\beta} - \gamma_{1j}^{\alpha\beta} \right|, \dots, \left| \gamma_{mk}^{\alpha\beta} - \gamma_{mj}^{\alpha\beta} \right|, \dots, \left| \gamma_{ak}^{\alpha\beta} - \gamma_{aj}^{\alpha\beta} \right| \right)^T \right\},$$

respectively. Note that by choice of the baseline, the interaction parameters in the groups corresponding to factor B do not contain any parameters where  $q$  is the first index (i.e.  $\gamma_{qj}^{\alpha\beta}$  does not appear in the groups for all  $j$ ).

At the GASH-ANOVA solution we have  $\left| \gamma_{mj}^{\alpha\beta} \right| = \gamma_{mj}^{\alpha\beta}$  and we know that  $\gamma_{mj}^{\alpha\beta}$  is the maximum of its corresponding  $\phi_{\alpha,m,j}$  group (the first group above). For the other groups in the penalty that correspond to factor A, if the term with  $\gamma_{mj}^{\alpha\beta}$  is the maximum of the group we have that for all  $k < q$  or  $k > m$ ,  $(m, k) \notin B_{\alpha,n}$ , else a difference larger than  $(\widehat{\alpha\beta})_{mj} - (\widehat{\alpha\beta})_{qj}$  could be found and for all  $q < k < m$ ,  $\left| \gamma_{mj}^{\alpha\beta} - \gamma_{kj}^{\alpha\beta} \right| = \gamma_{mj}^{\alpha\beta} - \gamma_{kj}^{\alpha\beta}$ . Note that  $\gamma_{mj}^{\alpha\beta} \geq \left| \gamma_{mk}^{\alpha\beta} \right|$  for all  $k, k = 1, 2, \dots, b - 1$ . Thus, for the groups in the penalty that involve differences of  $\gamma_{mj}^{\alpha\beta}$  corresponding to factor B, if that difference is the maximum of the group we have that  $\left| \gamma_{mj}^{\alpha\beta} - \gamma_{mk}^{\alpha\beta} \right| = (\gamma_{mj}^{\alpha\beta} - \gamma_{mk}^{\alpha\beta})$  or  $(\gamma_{mj}^{\alpha\beta} + \gamma_{mk}^{\alpha\beta})$ .

On the neighborhood described we can differentiate our criterion to get an equation similar to equation 6. In doing so we can use a similar argument as used in case (i) to show our contradiction. The sums that involved indices in  $B_{\alpha,n}$  consist of only positive values, are nonempty, and are of order  $O_p(\lambda_n)$ . Likewise, indices in  $B_{\beta,n}$  consist only of positive values and are of order  $O_p(\lambda_n)$ . All other terms on the right hand side of the equation are  $O_p(\lambda_n n^{-1/2})$  implying the right hand side must be of order  $O_p(\lambda_n)$ . However, the left hand side is  $O_p(1)$  and by assumption  $\lambda_n \rightarrow \infty$ . This again is our contradiction, thus it must be that  $P(B_{\alpha,n} \neq \emptyset) \rightarrow 0$ .

**Proof of b):** As with the proof of the CAS-ANOVA asymptotic normality, this proof will closely follow that of Zou (2006). The proof given below is just a sketch of how the proof is adapted to this setting, for full details please see Zou (2006). Let  $\gamma_0$  be the true parameter vector for the full rank reparametrization and let  $\hat{u} = \sqrt{n}(\hat{\gamma} - \gamma_0)$ . Now,  $\hat{u} = \arg \min_u V_n(u)$ , where

$$V_n(u) = u^T \left( \frac{1}{n} Z^T Z \right) u - 2 \frac{\epsilon^T Z}{\sqrt{n}} u + \frac{\lambda_n}{\sqrt{n}} P(u),$$

and

$$\begin{aligned}
P(u) = & \sum_{k \neq q} (-1)^{I[k < q]} \frac{w_\alpha^{(kq)}}{\sqrt{n}} \sqrt{n} \max \left\{ \left( \left| \gamma_{k0}^\alpha + \frac{u_k}{\sqrt{n}} \right| - |\gamma_{k0}^\alpha| \right), \left( \left| \Gamma_{kj0}^{\alpha\beta} + \frac{u_{kj}}{\sqrt{n}} \right| - \left| \Gamma_{kj0}^{\alpha\beta} \right| \right) \right\}^T \\
& + \sum_{1 \leq k < l \leq a, k \neq q, l \neq q} \frac{w_\alpha^{(kl)}}{\sqrt{n}} \sqrt{n} \max \left\{ \left( \left| \gamma_{l0}^\alpha - \gamma_{k0}^\alpha + \frac{u_l - u_k}{\sqrt{n}} \right| - |\gamma_{l0}^\alpha - \gamma_{k0}^\alpha| \right), \right. \\
& \left. \left( \left| \Gamma_{lj0, kj0}^{\alpha\beta} + \frac{u_{lj0} - u_{kj0}}{\sqrt{n}} \right| - \left| \Gamma_{lj0, kj0}^{\alpha\beta} \right| \right) \right\}^T \\
& + \sum_{1 \leq j < b-1} \frac{w_\beta^{(jb)}}{\sqrt{n}} \sqrt{n} \max \left\{ \left( \left| \gamma_{j0}^\beta + \frac{u_j}{\sqrt{n}} \right| - |\gamma_{j0}^\beta| \right), \left( \left| \Gamma_{ij0}^{\alpha\beta} + \frac{u_{ij}}{\sqrt{n}} \right| - \left| \Gamma_{ij0}^{\alpha\beta} \right| \right) \right\}^T \\
& + \sum_{1 \leq j < m \leq b-1} \frac{w_\beta^{(jm)}}{\sqrt{n}} \sqrt{n} \max \left\{ \left( \left| \gamma_{l0}^\beta - \gamma_{j0}^\beta + \frac{u_l - u_j}{\sqrt{n}} \right| \right), \right. \\
& \left. \left( \left| \Gamma_{im0, ij0}^{\alpha\beta} + \frac{u_{im0} - u_{ij0}}{\sqrt{n}} \right| - \left| \Gamma_{im0, ij0}^{\alpha\beta} \right| \right) \right\}^T.
\end{aligned}$$

By the argument in Zou (2006),  $\frac{\lambda_n}{\sqrt{n}} P(u)$  will go to zero for the correct model structure and diverge under the incorrect model structure. Let  $V_n^O(u_O)$  be the value of the objective function obtained using the ‘oracle’ structure determined by  $A_\alpha$  and  $A_\beta$ . This implies we collapse  $Z$  to  $Z_O$  by combining the columns of the columns of each pair in  $A_\alpha^c$  and  $A_\beta^c$ , in the process forming a new  $\gamma_O$ . If  $\hat{\eta}_{A_\alpha^c, A_\beta^c} = 0$ , then  $V_n(u) = V_n^O(u_O)$ .

Assuming constant variance,  $\sigma^2$ , for our model we get  $\frac{1}{n} Z_O^T Z_O \rightarrow C$ , where  $C$  is a positive definite matrix. Also we have that  $\frac{\epsilon^T Z_O}{\sqrt{n}} \rightarrow W = N(0, \sigma^2 C)$ . As in Zou (2006), we get  $V_n(u) \rightarrow V(u)$ , where

$$V(u) = \begin{cases} u_O^T C u_O - 2u_O^T W & \hat{\eta}_{A_\alpha^c, A_\beta^c} = 0 \\ \infty & \text{otherwise} \end{cases}.$$

Since  $V_n(u)$  is convex and the unique minimizer of  $V(u)$  is  $(C^{-1}W, 0)^T$ , the asymptotic normality follows. Therefore,  $u_{A_\alpha^c, A_\beta^c}^T \rightarrow N(0, \sigma^2 C^{-1})$ . The result for all pairwise differences,  $\hat{\eta}_{A_\alpha^c, A_\beta^c}$ , follows after the (singular) transformation.

## References

- 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*.
- Hamada, M. and Wu, C. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, 24:130–137.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288.
- 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(476):1418–1429.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67(Part 2):301–320.
- Zou, H. and Yuan, M. (2008). The  $f_{\text{inf}}$ -norm support vector machine. *Statistica Sinica*, 18:379–398.

Table 1: Table of Cell Means for  $\theta_1$

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

Table 2: Table of Cell Means for  $\theta_2$

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



Table 3: Simulation Results for Effect Vector  $\theta_1$ . 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	
<i>Reps</i> = 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
	NH	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
<i>Reps</i> = 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
	NH	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
<i>Reps</i> = 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
	NH	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 4: Simulation Results for  $\theta_2$ . 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.

	I-Type1		Power		Avg Sig		Avg FSR		Avg Col		Avg FCR		
	Levels	Diffs	Levels	Diffs	Levels	Diffs	Levels	Diffs	Levels	Diffs	Levels	Diffs	
<i>Reps</i> = 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
	NH	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
<i>Reps</i> = 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
	NH	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
<i>Reps</i> = 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
	NH	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 5: Null Model Simulation Results

		Oracle	Avg Sig Levels	Avg Sig Pairwise
<i>Reps = 2</i>	GASH	0.95	0.16	0.16
	NH	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
<i>Reps = 6</i>	GASH	0.95	0.18	0.18
	NH	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
<i>Reps = 10</i>	GASH	0.98	0.08	0.08
	NH	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 6: Treatment Combination Means and Distinct Levels

Age	Group	Mean	SD	GASH	NH	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

Table 7: Distinct Levels within Factors

Age	GASH/NH/BH/Bon	Group	GASH	NH	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