DEPARTMENT OF STATISTICS
North Carolina State University
2501 Founders Drive, Campus Box 8203
Raleigh, NC 27695-8203

Institute of Statistics Mimeo Series No. 2605

Simultaneous factor selection and collapsing levels in ANOVA

Howard D. Bondell and Brian J. Reich

Department of Statistics, North Carolina State University

Raleigh, NC 27695-8203, U.S.A.

(email: bondell@stat.ncsu.edu)

Simultaneous factor selection and collapsing levels in ANOVA

Howard D. Bondell and Brian J. Reich

Department of Statistics, North Carolina State University,

Raleigh, NC 27695-8203, U.S.A.

October 1, 2007

SUMMARY. When performing an Analysis of Variance, the investigator often has two main goals: to determine which of the factors have a significant effect on the response, and to detect differences among the levels of the significant factors. Level comparisons are done via a post-hoc analysis based on pairwise differences. One common complaint about this approach is that the levels are not necessarily collapsed into non-overlapping groups. This paper proposes a novel constrained regression approach to simultaneously accomplish both goals via shrinkage within a single automated procedure. The form of this shrinkage has the ability to collapse levels within a factor by setting their effects to be equal, while also achieving factor selection by zeroing out entire factors. Using this approach also leads to the identification of a structure within each factor, as levels can be automatically collapsed to form groups. In contrast to the traditional pairwise comparison methods, these groups are necessarily non-overlapping so that the results are interpretable in terms of distinct subsets of levels. The proposed procedure is shown to perform well in simulations and in a real data example.

KEY WORDS: ANOVA; Grouping; Multiple comparisons; Shrinkage; Variable selection.

---

email: bondell@stat.ncsu.edu

1. Introduction

Analysis of Variance (ANOVA) is a commonly used statistical technique to determine a relationship between a continuous response and categorical predictors. ANOVA is particularly applicable to designed experiments where the focus is on discovering the relevant factors that affect the response. An initial goal in ANOVA is to judge the overall importance of these categorical factors. If a factor is deemed important, a secondary analysis is performed to determine which levels of an important factor really differ from one another and which do not. This second question is typically answered via a post-hoc analysis involving pairwise comparisons within factors that are found to be important. Some common approaches to judge significance when the pairwise comparisons are considered include Tukey's Honestly Significantly Different (HSD) test, Fisher's Least Significant Difference (LSD) procedure, Bonferroni or Scheffe multiple comparison adjustments, and more recently, procedures based on the False Discovery Rate (Benjamini and Hochberg, 1995; Storey, 2002). One common complaint of applied scientists in this type of analysis is that it is often not possible to collapse the levels into non-overlapping groups based on the pairwise difference results.

Accomplishing the goal of judging importance of the factor levels is a variable selection problem that has received a great deal of attention in the literature. In particular, penalized, or constrained, regression has emerged as a highly-successful technique for variable selection. For example, the LASSO (Tibshirani, 1996) imposes a bound on the L_1 norm of the coefficients. This results in both shrinkage and variable selection due to the nature of the constraint region which often results in several coefficients becoming identically zero. Alternative choices of constraints, or penalties, have also been proposed for shrinkage and selection in regression (Frank and Friedman, 1993; Fan and Li, 2001; Tibshirani et al., 2005; Zou and Hastie, 2005; Bondell and Reich, 2006; Zou, 2006).

However, in ANOVA, a variable, or factor, actually corresponds to a group of coefficients,

typically coded in terms of dummy variables, as opposed to a single coefficient as in the case of regression using continuous predictors. Naive use of a penalization technique does not achieve variable selection properly in this situation, as it can set some of the dummy variables corresponding to a factor to zero, while others are not. Yuan and Lin (2006) propose the Group LASSO in order to perform variable selection for entire factors by appropriately treating these dummy variables as a group and constraining the sum of group norms as opposed to a single overall norm. This approach accomplishes the first goal in ANOVA of selecting the important factors. However, the post-hoc analysis of pairwise comparisons must still be performed, as it is not built in to this procedure. For each significant factor, each of the dummy variables will be given a distinct coefficient in the estimated model.

A novel constrained regression approach called CAS-ANOVA (for Collapsing And Shrinkage in ANOVA) is proposed in this paper to simultaneously perform the two main goals of ANOVA within the estimation procedure. The form of the constraint encourages similar effect sizes within a factor to be estimated with exact equality, thus collapsing the corresponding levels. In addition, via the typical ANOVA design parametrization combined with this constraint, entire factors can be collapsed to a zero effect, hence providing the desired factor selection. In this manner, the proposed procedure allows the investigator to conduct a complete analysis on both the factors and the individual levels in a single step, thus eliminating the need for a secondary analysis.

In addition to performing the collapsing of levels within the estimation procedure itself, another important distinction between this CAS-ANOVA approach and the typical post-hoc pairwise comparison approach is that the typical approaches in many instances will not actually yield a feasible collapsing of levels, whereas in the proposed procedure, this is automatic. As a simple example, in a typical ANOVA analysis for a factor with 3 levels, one may find that level 1 is significantly different from level 2, but neither level 1 nor 2 are

significantly different from level 3. This result would not correspond to any possible way of grouping the three levels, whereas in the proposed procedure, this infeasible grouping cannot occur, as the structure is incorporated into the estimation procedure.

This procedure requires a single tuning parameter to control the degree of shrinkage. Various methods to choose this parameter are discussed including a new technique to tune the procedure based on a notion of False Selection rate. The intuitive idea is to add false differences to the data. This is accomplished by artificially splitting cells of the design and monitoring how well the procedure fuses them back together.

The remainder of the paper proceeds as follows. The CAS-ANOVA procedure is introduced in §2, while computation and a data adaptive version of the procedure are discussed in §3. Techniques for tuning the procedure including a novel procedure based on estimating the False Selection Rate appear in §4. Finally, §5 gives a simulation study and a real data example.

2. Collapsing And Shrinkage in ANOVA

2.1 *The procedure*

To establish notation, consider the additive ANOVA model with J factors and a total sample size n. Factor j has p_j levels, and denote $p = \sum_{j=1}^{J} p_j$. Let $n_j^{(k)}$ be the total number of observations with factor j at level k, for $k = 1, ..., p_j$ and $j = 1, ..., J$. Under a balanced design this simplifies to $n_j^{(k)} = n_j = n/p_j$ for all k.

Assume that the responses have been centered to have mean zero, so that the intercept can be omitted in the case of a balanced design. Let the $n \times p$ design matrix X be the typical over-parameterized ANOVA design matrix composed of zeros and ones denoting the combination of levels for each of the n observations. This parametrization is useful for this approach as equality of coefficients correspond exactly to collapsing levels.

The standard least squares estimation procedure for the ANOVA parameterization esti-

mates the p coefficients, which can be interpreted as the effect sizes for each level, by the solution to

$$\begin{gathered}\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||^2 \\ \text{subject to} \\ \textstyle\sum_{k=1}^{p_j} \beta_{jk} = 0 \text{ for all } j = 1, ..., J,\end{gathered} \tag{1}$$

where the constraint forces the effects for the levels within each factor to sum to zero for identifiability.

An additional constraint can then be added to this optimization problem to shrink the coefficients in order to perform variable selection, such as the LASSO, or Group LASSO. To perform the collapsing of levels, a desirable constraint would not only have the ability to set entire groups to zero, it would additionally be able to collapse the factor levels by setting subsets of coefficients within a group to be equal. For the linear regression setting, the OSCAR (Bondell and Reich, 2006) uses a mixture of L_1 and pairwise L_∞ penalty terms to simultaneously perform variable selection and find clusters among the important predictors. The nature of this penalty naturally handles both positive and negative correlations among continuous predictors, which does not come into play in the ANOVA design.

With the ANOVA goals in mind, the proposed CAS-ANOVA procedure places a constraint directly on the pairwise differences as

$$\begin{gathered}\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||^2 \\ \text{subject to} \\ \textstyle\sum_{k=1}^{p_j} \beta_{jk} = 0 \text{ for all } j = 1, ..., J \text{ and } \sum_{j=1}^{J} \sum_{1 \le k < m \le p_j} w_j^{(km)} |\beta_{jk} - \beta_{jm}| \le t,\end{gathered} \tag{2}$$

where $t > 0$ is a tuning constant and $w_j^{(km)}$ is a weight for the pair of levels k and m of factor j. These weights are needed to account for the fact that the number of levels for each factor are not necessarily the same and that the design may also be unbalanced. In the balanced design case, the weights for all pairs of levels within a given factor would be equal, so in that case $w_j^{(km)} = w_j$. An appropriate choice of weights is discussed in the next section.

The first constraint is again the sum-to-zero constraint on the effects for the levels within

each factor used in the standard ANOVA. The second constraint is a generalized version of the Fused LASSO-type constraint (Tibshirani et al., 2005) taken on the pairwise differences within each factor. The Fused LASSO itself constrains consecutive differences in the regression setting to yield a smoother coefficient profile for ordered coefficients. Here the pairwise constraints smooth the within factor differences towards one another. This second constraint enables the automated collapsing of levels within the factor. In addition, combining the two constraints implies that entire factors will be set to zero once their levels are collapsed.

As opposed to the typical pairwise comparison methods, this approach will automatically yield non-overlapping groups so that a feasible group structure among the levels of a factor will always emerge. This additional benefit alleviates the common problem faced by applied scientists in the interpretation of the post-hoc analysis.

2.2 *Choosing the weights*

In any penalized regression, having each predictor on a comparable scale is essential so that the predictors are penalized equally. This is typically done by standardization, so that each column of the design matrix has its 2-norm equal unity. This could instead be done by weighting each term of the penalty as in (2). Clearly, when the penalty is based on a norm, rescaling a column of the design matrix is equivalent to instead weighting the corresponding coefficient in the penalty term by the inverse of this rescaling factor. In standard situations such as the LASSO, each column contributes equally to the penalty, so it is clear how to standardize the columns (or, equivalently, weight the terms in the penalty). However, standardizing the predictors in the CAS-ANOVA procedure is not as straightforward, as the penalization involves the pairwise differences. Although the design matrix consists of only zeros and ones, the number of terms in the penalty changes dramatically with the number of levels per factor, while the amount of information in the data to estimate each coefficient also varies across factors and levels.

It is now shown that an appropriate set of weights to use in the CAS-ANOVA procedure (2) are

$$w_j^{(km)} = (p_j + 1)^{-1}\sqrt{n_j^{(k)} + n_j^{(m)}}, \tag{3}$$

which for the special case of the balanced design would be $w_j = (p_j + 1)^{-1}\sqrt{2n_j}$. The idea behind these weights is based on the notion of 'standardized predictors' as follows.

Let $\boldsymbol{\theta}$ denote the vector of the pairwise differences taken within each factor. Hence $\boldsymbol{\theta}$ is a vector of length $d = \sum_{j=1}^{J} d_j = \sum_{j=1}^{J} p_j(p_j - 1)/2$. Let the over-parameterized model with $q = p + d$ parameters arising from the p coefficients plus the d pairwise differences have parameter vector denoted by

$$\boldsymbol{\gamma} = M\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\beta}_J^T, \boldsymbol{\theta}_J^T\right)^T,$$

where $\boldsymbol{\beta}_j$ and $\boldsymbol{\theta}_j$ are the coefficients and pairwise differences corresponding to factor j. The matrix M is block diagonal with j^{th} block given by $M_j = [I_{p_j \times p_j}\ D_j^T]^T$, with D_j the $d_j \times p_j$ matrix of ± 1 that creates the vector $\boldsymbol{\theta}_j$ from $\boldsymbol{\beta}_j$ for a given factor j by picking off each of the pairwise differences from that factor.

The corresponding design matrix Z for this overparameterized design is then an $n \times q$ matrix such that $Z\boldsymbol{\gamma} = X\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$. Hence it is desired to solve $ZM = X$. Clearly Z is not uniquely defined. An uninteresting choice would be $Z = [X\ 0_{n \times d}]$. Also, $Z = XM^*$, with M^* as any left inverse of M, would suffice. We choose

$$Z = XM^-,$$

where M^- denotes the Moore-Penrose generalized inverse of M. This resulting matrix Z is an appropriate design matrix for the overparamterized space. One could directly work with this new design matrix and the additional set of constraints defined via $\boldsymbol{\gamma} = M\boldsymbol{\beta}$ and then standardize the columns and perform the estimation. However, the resulting parameters

after standardization would no longer have the same interpretation as differences between effects, so that collapsing levels is no longer accomplished. Hence the corresponding approach to weighting the terms in the penalty is advocated. The appropriate weights are then just the Euclidean norm of the columns of Z that correspond to each of the differences. It can be shown that the proposed weights are exactly these norms.

PROPOSITION. The Moore-Penrose generalized inverse of the matrix M is block diagonal with the j^{th} diagonal block corresponding to the j^{th} factor and of the form $(M^-)_j = (p_j + 1)^{-1}[(I_{p_j \times p_j} + \mathbf{1}_{p_j}\mathbf{1}_{p_j}^T)\ D_j^T]$. Furthermore, under the ANOVA design, the column of $Z = XM^-$ that corresponds to the difference in effect for levels k and m of factor j has Euclidean norm $(p_j + 1)^{-1}\sqrt{n_j^{(k)} + n_j^{(m)}}$.

The above proposition, whose proof is given in the appendix, allows the determination of the weights via the standardization in this new design space. Under the balanced design, the weights only depend on the factor regardless of the levels considered, whereas that is not the case in the unbalanced design.

Note that standardizing the columns of Z and imposing an L_1 penalty on $\boldsymbol{\theta}$ as in the typical LASSO approach (Tibshirani, 1996), is strictly equivalent to choosing the weights to be the Euclidean norm of the corresponding column only when the columns have mean zero. Now for the matrix Z, the columns corresponding to a difference consist of elements given by $\{0, \pm(p_j + 1)^{-1}\}$, and the mean of each of these columns is given by $\{n(p_j+1)\}^{-1}\{n_j^{(k)} - n_j^{(m)}\}$, so that under the balanced design the columns have mean zero. Meanwhile, in an unbalanced design, the weighting is at least approximately equivalent to standardization, as the mean of the column would typically be negligible due to the factor of n^{-1}.

3. Computation

The CAS-ANOVA optimization problem can be expressed as a quadratic programming problem as follows. For each $k = 1, ..., d$, set $\theta_k = \theta_k^+ - \theta_k^-$ with both θ_k^+ and θ_k^- being non-negative, and only one is nonzero. Then $|\theta_k| = \theta_k^+ + \theta_k^-$. Now let the full $p + 2d$ dimensional parameter vector be denoted by η. For each factor j, let $X_j^* = [X_j \ 0_{n \times 2d_j}]$, where X_j denotes the columns of the design matrix corresponding to factor j and $d_j = p_j(p_j - 1)/2$ is the number of differences for factor j. Then let $X^* = [X_1^* ... X_J^*]$ be the $n \times (p + 2d)$ dimensional design matrix formed by combining all of the X_j^*. Then $X^*\eta = X\beta$. The optimization problem can be written as

$$\begin{gathered} \tilde{\eta} = \arg\min_\eta ||\mathbf{y} - X^*\eta||^2 \\ \text{subject to} \\ L\eta = 0, \ \textstyle\sum_{k=1}^d w^{(k)}(\theta_k^+ + \theta_k^-) \leq t \text{ and } \theta_k^+, \theta_k^- \geq 0 \text{ for all } k = 1, ..., d, \end{gathered} \tag{4}$$

where $w^{(k)}$ denotes the weight for pairwise difference k and the matrix L is a block diagonal matrix with j^{th} block L_j givenby

$$L_j = \begin{bmatrix} D_j & I_{d_j} & -I_{d_j} \\ \mathbf{1}_{p_j}^T & \mathbf{0}_{d_j}^T & \mathbf{0}_{d_j}^T \end{bmatrix}$$

for each $j = 1, ..., J$.

The optimization problem given by (4) is just a standard quadratic programming problem as all constraints are linear. This quadratic programming problem has $p + 2d$ parameters and $p + 3d$ linear constraints. Note that both the number of parameters and constraints grow with p, but the growth is not quadratic in p, it is only quadratic in the number of levels within a factor, which is typically not too large. Hence this direct computational algorithm is feasible for the majority of practical problems.

4. An adaptive CAS-ANOVA

It has been shown both theoretically and in practice that a weighted version of the LASSO with data-adaptive weights exhibits better performance than the LASSO itself in the sense

of selecting the correct model and not overshrinking large coefficients (Zou, 2006). The intuition is to weight each coefficient in the penalty so that estimates of larger coefficients are penalized less and in the limiting case, coefficients that are truly zero will be infinitely penalized unless the estimate is identically zero. Specifically the optimization problem for the adaptive LASSO is given by

$$\begin{gathered}\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||^2 \\ \text{subject to} \\ \textstyle\sum_{k=1}^{p} |\hat{\beta}_k|^{-1}|\beta_k| \leq t,\end{gathered} \tag{5}$$

where $\hat{\beta}_k$ denotes the ordinary least squares estimate for β_k. The idea is that the weights for the non-zero components will converge to constants, whereas the weights for the zero components diverge to infinity. Zou (2006) has shown that if the bound is chosen appropriately, the resulting estimator has the oracle property in that it obtains consistency in variable selection along with asymptotic efficiency for the non-zero components.

This approach can also be used here directly by modifying the CAS-ANOVA optimization problem given by (2) as

$$\begin{gathered}\tilde{\boldsymbol{\beta}}^A = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||^2 \\ \text{subject to} \\ \textstyle\sum_{k=1}^{p_j} \beta_{jk} = 0 \text{ for all } j = 1, ..., J \text{ and } \sum_{j=1}^{J} \sum_{1 \leq k < m \leq p_j} w_j^{(km)*} |\beta_{jk} - \beta_{jm}| \leq t,\end{gathered} \tag{6}$$

where

$$w_j^{(km)*} = w_j^{(km)} |\hat{\beta}_{jk} - \hat{\beta}_{jm}|^{-1},$$

with $w_j^{(km)}$ as in (3), and the vector $\hat{\boldsymbol{\beta}}$ denotes the estimate obtained from the typical least squares ANOVA with the sum to zero constraint as given by (1). This is exactly the adaptive LASSO on the space of pairwise differences.

Note that computationally, a single ANOVA fit is done initially and the weight for each difference is adjusted based on this fit as above. Then the computation in the adaptive CAS-ANOVA procedure is exactly the same as in the original version using the above weights.

5. Tuning the procedure

5.1 *Cross-validation and degrees of freedom*

In general penalized, or constrained, regression techniques, one must choose the value of the tuning parameter, the bound t in this case. The choice of t yields a trade-off of fit to the data with model complexity, or sparsity. This choice can be accomplished via minimizing any of the standard techniques to estimate out-of-sample prediction error, such as AIC, BIC, Generalized Cross-Validation (GCV), or directly via k-fold Cross-Validation. To use any of the criteria such as AIC, BIC, or GCV, an estimate of the number of degrees of freedom is needed. For the LASSO, it is known that the number of non-zero coefficients is an unbiased estimate of the degrees of freedom (Zou et al., 2004). For the Fused LASSO (Tibshirani et al., 2005), the number of non-zero blocks of coefficients is used as an estimate of degrees of freedom. In this ANOVA design case, the natural estimate of degrees of freedom for each factor is the number of unique coefficients minus one for the constraint, this represents the number of resulting levels minus one. Specifically, one has

$$\hat{\mathrm{df}} = \sum_{j=1}^{J} \left(p_j^* - 1\right), \tag{7}$$

where p_j^* denotes the number of estimated unique coefficients for factor j.

5.2 *False selection rate*

Using prediction error to judge the fit in order to choose the tuning parameter is not necessarily the most appropriate measure when the main goal in the ANOVA analysis is to determine the important factors and levels. In keeping with this goal, an alternative tuning procedure is now introduced for the CAS-ANOVA procedure based on the idea of False Selection Rate (FSR) as proposed in the context of variable selection by Wu et al. (2006). FSR in variable selection is the analogue of False Discovery Rate (FDR) in that it measures the proportion of unimportant variables selected, $U(X, Y)$, out of the total number selected, $S(X, Y)$, i.e., $\gamma = U(X, Y)/\{1 + S(X, Y)\}$, where the dependence on the particular data is

made explicit, and the 1 in the denominator reflects the inclusion of an intercept and also avoids division by zero in the case that no variables are selected.

However, which of the variables are truly important or unimportant is, of course, unknown, so that the number of unimportant variables selected by a procedure, $U(X, Y)$, is not known and must somehow be estimated. The approach of Wu et al. (2006) is to generate new data sets by appending known irrelevant variables to the original data and, for a given tuning parameter, monitor the proportion of these variables that are selected out of the total number of selected variables, on average. Under some simplifying assumptions, a multiplicative factor is derived to obtain a direct estimate of the FSR for the original data set corresponding to that tuning parameter. One then chooses the tuning parameter that selects the largest model whose estimated FSR, $\hat{\gamma}$, remains below a desired level, typically $\gamma_0 = 0.05$. This idea is intuitively appealing for this situation as it is more in line with the goals of ANOVA, as opposed to a criterion based on prediction error.

Specifically, the method of Wu et al. (2006) proposes to estimate the FSR for a given tuning parameter t as

$$\hat{\gamma} = \frac{\hat{k}_U \bar{U}_P^* / k_P}{1 + S(X, Y, t)}, \tag{8}$$

where $\hat{k}_U$ is the estimated number of truly unimportant variables in the original data set, $\bar{U}_P^*$ is the average number of known irrelevant variables that were selected from the appended data sets, k_P denotes the number of irrelevant variables that were appended to create each new data set, and the dependence on the tuning parameter t is made explicit.

However, simply adding extra variables to the ANOVA design matrix is not directly applicable for tuning this procedure, as just adding additional factors is not sufficient to also judge the collapsing of levels. Instead, a novel method is proposed to implicitly add new differences that are known to be zero and monitor the FSR on the full set of pairwise differences. This is exactly in line with the idea of controlling an error rate for pairwise

comparisons with the currently used methods. However, with the proposed CAS-ANOVA technique, this is accomplished simultaneously within the estimation procedure where the pairwise differences from all factors, including those that are set to zero, is taken into account by the FSR.

To fix ideas, consider the replicated balanced ANOVA design so that there are $c = \prod_{j=1}^{J} p_j$ total combinations (cells), each having $r > 1$ observations. To effectively add these known irrelevant differences, the proposed method splits each cell of the ANOVA table as follows. Let $\boldsymbol{\nu}$ be a $c \times 1$ vector of parameters, then a derived new model is

$$\mathbf{y} = X\boldsymbol{\beta} + A\boldsymbol{\nu} + \boldsymbol{\epsilon},$$

with the matrix A containing elements given by $\{0, \pm 1/2\}$ that randomly splits the observations for each cell into two groups of size $r/2$.

Remark: This idea naturally generalizes to unequally splitting of cells if r is odd or if the design is unbalanced, including splitting of only some cells if there are replications for some combinations but not for others. Additionally, splitting could be into multiple groups if there were a large number of replications thus obtaining more irrelevant differences. This approach to tuning the procedure does require that there are replications of at least some combinations so that it is possible to divide at least some of the cells.

As parameterized, the vector $\boldsymbol{\nu}$ represents the differences in predicted response between each half of the split cells, which plays exactly the role that the pairwise differences play. Hence for estimation in this derived model, each of the $|\nu_j|$ is penalized along with $\boldsymbol{\theta}$, the vector of pairwise differences, in (2) and thus the matrix A plays the role of adding the irrelevant difference variables to the full vector of pairwise differences. Using the standardization idea, assuming a balanced design and equal splitting of each cell into two groups, the weight for $|\nu_j|$ in the penalty term is given by $\sqrt{r}/2$, as this is the Euclidean norm of each column

of A.

Now for a given tuning parameter t, one estimates the full vector $(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\nu})$ based on the matrix A corresponding to a possible splitting of the cells. The number of total nonzero estimates of elements of $\boldsymbol{\nu}$ then plays the role of U_P^*, the number of chosen irrelevant differences. Note that if $r = 2$ there is only one split possible so that the term $\bar{U}_P^*$ used to estimate the FSR is based solely on the number of nonzero estimates of the components of $\boldsymbol{\nu}$ for this split. If $r > 2$ then the cells can be split in multiple ways so that $\bar{U}_P^*$ is based on averaging the number of nonzero estimates over the splits, either via all possible splits, or by random splits.

Remark: This approach to tuning is not advocated for the adaptive CAS-ANOVA as, in that case, one must also adaptively weight each $|\nu_j|$. These weights would be different depending on the split, thus changing the meaning of the tuning parameter t for each split.

This method proposed to estimate the FSR by Wu et al. (2006) is based on the assumption that on average, the real unimportant variables and the added unimportant variables have the same probability of being selected. Wu et al. (2006) recognize that this assumption is almost surely to be violated and only claim the method to be an approximation. In this ANOVA setting, since the 'variables' under consideration are actually the pairwise differences and hence tied together in a complex manner, this assumption is even less likely to hold. However, the simulation results in the next section show that, although it is only an approximation, tuning the CAS-ANOVA procedure in this manner works well in practice for control of the FSR.

6. Examples

6.1 *Simulation study*

A simulation study was carried out to examine the performance of the CAS-ANOVA procedure in terms of selecting both complete factors and appropriately collapsing levels.

The first example is a three-factor experiment having 8, 4, and 3 levels respectively. The response is generated to have overall mean zero and true effect vector for the first factor given by $\beta = (2, 2, -1, -1, -1, -1, 0, 0)^T$ and the remaining two factors have zero effect. The first factor truly has 3 distinct subgroups, hence an 'oracle' procedure would eliminate the two extraneous factors and collapse the 8 levels into 3 groups. The simulation was done for a balanced design with a single observation per combination, and then repeated for the two additional cases of 2 and 4 replications per combination.

Table 1 compares the standard version tuned by AIC, BIC, and GCV, with the adaptive version tuned in the analagous manner for this example. 'Oracle' represents the percent of time the procedure chose the fully correct model in terms of which factors should be eliminated and which levels collapsed. 'Correct Factors' represents the percent of time that the complete nonsignificant factors are eliminated and the true factors are kept. True Positive Rate (TPR) represents the average percent of true differences found (out of the 20 total true differences), whereas FSR represents the average percent of declared differences that are mistaken (the observed False Selection Rate).

***** TABLE 1 GOES HERE ******

Clearly, the adaptive weights yield a large improvement in terms of the model selection as expected. In addition, the adaptive version using BIC is much preferred as the other criteria regularly choose too large of a model. This was seen in each of the examples that were tried, so that for the remainder of the simulations and real example, the adaptive version tuned via BIC is the method used, and is the method recommended in practice.

Table 2 shows the results for the proposed procedure (CAS-ANOVA) tuned via controlling the estimated False Selection Rate, so that the tuning parameter is chosen to have an

estimated FSR of 0.05 in the standard CAS-ANOVA procedure. Additionally is shown the results for the proposed adaptive version of the procedure tuned via BIC. Note that the CAS-ANOVA procedure tuned via the FSR is not applicable for the unreplicated design so the corresponding entries in the table are omitted. It is also again noted that the adaptive version tuned via FSR is not appropriate to choose the tuning parameter, as this parameter does not have the same meaning in each newly created data set.

Tukey's HSD procedure based on the 95% confidence level is shown for comparison. As seen in the table, the HSD procedure is overly conservative in terms of its FSR and is thus selecting a smaller model in general. This is seen from the much lower TPR and Oracle. The HSD procedure using the 65% confidence level was empirically found to yield an approximate FSR of 0.05, and thus is given as an additional comparison. This procedure is too aggressive in that too many terms are included in the model. This can be seen by the poor performance in not dropping the irrelevant factors often enough. Additionally, the fact that impossible grouping structures can be found by the HSD procedures help to contribute to the lower percentages of selecting the true model ('Oracle'). Overall, the CAS-ANOVA procedures tuned by both methods do a good job in picking out the appropriate model structure. For larger samples, the adaptive version tuned via BIC exhibits better performance than that of tuning by FSR, but does not allow for control of the error rate. The tuning by FSR allows the user to approximately control the error rate, but cannot be used in all situations.

***** TABLE 2 GOES HERE ******

A second simulation setup with 4 factors having smaller effect sizes was also examined. Again the response has overall mean zero with the first two factors having effects given by $\beta_1 = (.75, .75, -.5, -.5, -.5, 0)^T$ and $\beta_2 = (.25, .25, -.5)^T$, respectively. The other two factors had 4 levels and 2 levels each with no effect.

Table 3 shows the results of this simulation for the CAS-ANOVA procedure tuned by the two previous methods, as well as the two versions of Tukey's HSD. Again one can see that the newly proposed method does well at accurately detecting the underlying model structure.

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

6.2 *Real data example*

The data for this example comes from a three-way analysis of variance for yields of barley. The response is the barley yield for each of five varieties at six experimental farms in Minnesota for each of the years 1931 and 1932 giving a total of 60 observations. Note that there is a single observation for each variety/location/year combination representing the total barley yield for that combination. This data was first introduced by Immer et al. (1934), and variations of this dataset have been used by Fisher (1971), Cleveland (1993) and Venables and Ripley (2002).

A practitioner would proceed by conducting an overall analysis of variance to decide on the significant factors. The standard ANOVA table for testing the overall factors is given in Table 4. Clearly all three factors are significant, so a typical analysis would next proceed by considering the pairwise differences to determine which factor levels differ from one another. To accomplish this goal, a post-hoc analysis using Tukey's HSD procedure at both the .95 and the .65 levels was performed and is plotted in Figure 1. Tukey's procedure (at both confidence levels) produces overlapping groups of levels for the location factor (Figure 1b). For example, Tukey's procedure finds that the Morris and Duluth locations are significantly different, but neither location significantly differs from the University Farm.

Instead of performing a first analysis to determine significance and then the post-hoc analysis on the pairwise differences, the newly proposed CAS-ANOVA procedure allows for a

Table 4: *ANOVA table for the barley data.*

Source	Df	SS	MS	F	p-value
Variety	4	5310.0	1327.5	4.516	.0035
Location	5	21220.9	4244.2	14.439	<.0001
Year	1	3798.5	3798.5	12.923	.0008
Error	49	14403.0	293.9		

single combined analysis. Since there is no replication, the tuning via FSR is not applicable, so the adaptive version of the procedure tuned via BIC was used. As with the standard ANOVA, the CAS-ANOVA procedure includes all three factors. The grouping structure is shown in Figure 1. By construction, the groups are non-overlapping. The overlapping-group problem of Tukey's procedure is resolved by creating a separate group for the University Farm.

***** FIGURE 1 GOES HERE ******

Based on the simulations and the real data example, the new proposal of this paper appears to be a competitive alternative to the standard post-hoc analysis in ANOVA. An additional benefit is that it allows the investigator to directly determine a grouping structure among the levels of a factor that may not be clearly demonstrated by using a standard pairwise comparison procedure.

7. Discussion

This paper has proposed a new procedure to simultaneously include pairwise comparisons into the estimation procedure when performing an analysis of variance. By combining the overall testing with the collapsing of levels, it creates distinct groups of levels within a factor, which unlike standard post-hoc procedures will always correspond to feasible grouping

structure. The procedure avoids the use of multiple testing corrections by performing the comparisons directly. In the case of a design with replication, a tuning method based on the idea of controlling False Selection Rate has been introduced. In cases with or without replication, the reweighted version of this procedure with tuning parameter chosen via BIC has shown strong performance.

In a similar manner, if desired, the collapsing of levels can be accomplished on interaction terms as well. Currently an approach to collapse levels in a hierarchical framework is under investigation. In this case it may be desired to collapse two levels of a main effect only if it is also reasonable to completely collapse all interactions of those two levels with each of the other terms. This enforcing of the hierarchical structure would require a more complex penalty form.

An additional important idea from this approach is that, in general, it may be possible to combine individual components of an analysis into a single step by an appropriately constructed constraint, or penalty function. In this manner, constrained regression can be tailored to accomplish multiple statistical goals simultaneously.

The ANOVA model has also been revisited recently from the Bayesian perspective. For example, Nobile and Green (2000) model the factor levels as draws from a mixture distribution, thus ensuring non-overlapping levels at each MCMC iteration. The CAS-ANOVA procedure also has a Bayesian interpretation. The CAS-ANOVA solution is the posterior mode assuming each factor's levels have a Markov random field prior (popularized for spatial modelling by Besag et al., 1991) with L1-norm, i.e.,

$$p(\boldsymbol{\beta}_j) \propto \exp\left(-\lambda \sum_{1 \le k < m \le p_j} w_j^{km} |\beta_{jk} - \beta_{jm}|\right),$$

where λ is the Lagrangian multiplier corresponding to the tuning parameter t.

Appendix

Proof of Proposition 1.

The matrix M is block diagonal with j^{th} block given by $M_j = [I_{p_j \times p_j} \; D_j^T]^T$, with D_j the $d \times p_j$ matrix of ± 1 that picks off each of the pairwise differences from factor j. Hence it suffices to consider each block individually. Consider the matrix $M_j^- = (p_j + 1)^{-1}[(I_{p_j \times p_j} + \mathbf{1}_{p_j}\mathbf{1}_{p_j}^T) \; D_j^T]$. Then $M_j^- M_j = (p_j+1)^{-1}[(I_{p_j \times p_j} + \mathbf{1}_{p_j}\mathbf{1}_{p_j}^T) + D_j^T D_j]$. Now via direct calculation, one obtains $D_j^T D_j = p_j I_{p_j \times p_j} - \mathbf{1}_{p_j}\mathbf{1}_{p_j}^T$, so that $M_j^- M_j = I_{p_j \times p_j}$.

Now to show that M_j^- is the Moore-Penrose inverse of M_j, it suffices to show

1. $M_j^- M_j M_j^- = M_j^-$.
2. $M_j M_j^- M_j = M_j$.
3. $M_j^- M_j$ is symmetric.
4. $M_j M_j^-$ is symmetric.

Clearly (1), (2), and (3) follow directly from the fact that $M_j^- M_j = I_{p_j \times p_j}$. For (4),

$$M_j M_j^- = (p_j + 1)^{-1} \begin{bmatrix} (I_{p_j \times p_j} + \mathbf{1}_{p_j}\mathbf{1}_{p_j}^T) & D_j^T \\ D_j(I_{p_j \times p_j} + \mathbf{1}_{p_j}\mathbf{1}_{p_j}^T) & D_j D_j^T \end{bmatrix}.$$

Now from the form of D_j, it follows that $D_j \mathbf{1}_{p_j}\mathbf{1}_{p_j}^T = 0_{d \times p_j}$. Thus the matrix is symmetric. Hence M^- is the Moore-Penrose inverse of M. The Euclidean norm of the columns of $Z = XM^-$ corresponding to the differences follows directly, as these columns of Z contain the elements $\{0, \pm 1\}$, with the number of nonzero elements for a column being exactly the total number of observations at the two corresponding levels. This completes the proof.

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.

Besag, J., York, J. C., and Mollié, A (1991), Bayesian image restoration, with two applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics*, **43**, 1-59.

Bondell, H. D. and Reich, B. J. (2006), Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR, *Institute of Statistics Mimeo Series # 2583*, North Carolina State University.

Cleveland, W. S. (1993), *Visualizing Data*, New Jersey: Hobart Press.

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

Fisher, R. A. (1971), *The Design of Experiments*, New York: Hafner, 9th edition.

Frank, I. E. and Friedman, J. (1993), A statistical view of some chemometrics regression tools, *Technometrics*, **35**, 109148.

Immer, F. R., Hayes, H. D., and Powers, L. (1934), Statistical determination of barley varietal adaptation, *Journal of the American Society for Agronomy*, **26**, 403419

Nobile, A. and Green, P. J. (2000), Bayesian analysis of factorial experiments by mixture modelling, *Biometrika*, **87**, 15-35.

Storey, J. D. (2002), A direct approach to false discovery rates, *Journal of the Royal Statistical Society* B, **64**, 479-498.

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 smooth-

ness via the fused lasso, *Journal of the Royal Statistical Society* B, **67**, 91-108.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, New York: Springer, 4th edition.

Wu, Y., Boos, D. D. and Stefanski, L. A. (2006), Controlling Variable Selection By the Addition of Pseudo-Variables, *Journal of the American Statistical Association*, **102**, 235-243.

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

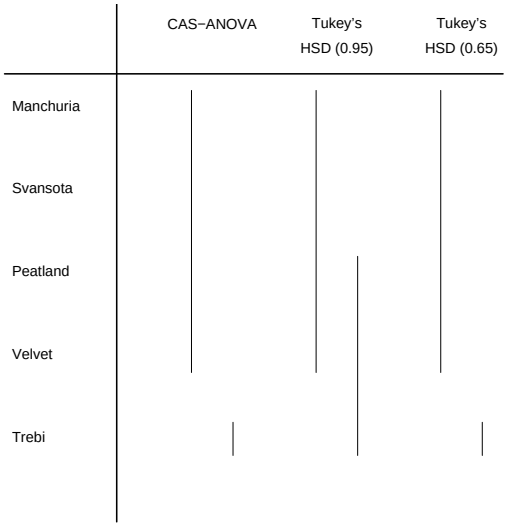
Zou, H. (2006), The adaptive LASSO and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418-1429.

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. (2004), On the degrees of freedom of the lasso, *Technical report, Department of Statistics, Stanford University.*

Figure 1: Group structure for the barley data using Tukey's HSD (with 95% and 65% confidence levels) and CAS-ANOVA. Each vertical line represents a set of levels that are grouped.

(a) Varieties



(b) Locations

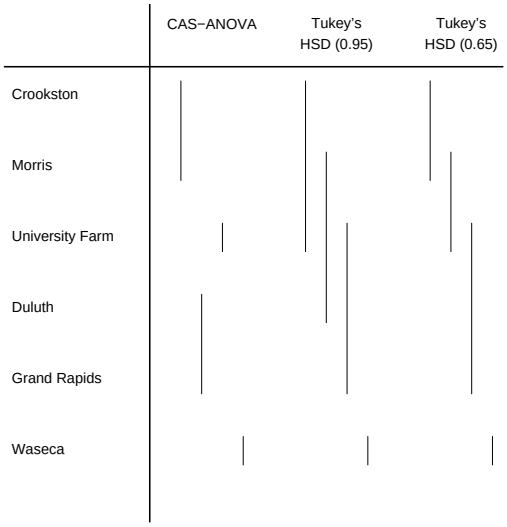


Table 1: *Example 1. Comparison of the adaptive and standard CAS-ANOVA procedures for the selection criteria AIC, BIC, and GCV. Oracle represents the percent of time the procedure chose the entirely correct model. Correct Factors represents the percent of time the nonsignificant factors are completely removed. TPR represents the average percent of true differences found, whereas FSR represents the average percent of declared differences that are mistaken.*

Replications		Oracle (%)	Correct Factors (%)	TPR (%)	FSR (%)
1	Standard$_{AIC}$	0.5	8.5	99.5	36.2
	Standard$_{GCV}$	0.5	10.0	99.4	35.2
	Standard$_{BIC}$	3.0	48.0	97.8	22.8
	Adaptive$_{AIC}$	4.5	43.0	98.4	22.0
	Adaptive$_{GCV}$	5.5	46.0	98.3	21.1
	Adaptive$_{BIC}$	21.0	83.0	95.8	11.4
2	Standard$_{AIC}$	1.0	14.5	99.9	34.4
	Standard$_{GCV}$	1.0	15.0	99.9	34.1
	Standard$_{BIC}$	6.0	62.0	99.6	19.2
	Adaptive$_{AIC}$	14.5	53.0	99.8	18.4
	Adaptive$_{GCV}$	14.5	55.5	99.8	17.8
	Adaptive$_{BIC}$	41.5	93.5	99.1	6.4
4	Standard$_{AIC}$	1.0	11.5	100	35.3
	Standard$_{GCV}$	1.0	12.0	100	35.1
	Standard$_{BIC}$	4.0	59.5	100	19.4
	Adaptive$_{AIC}$	6.0	50.0	100	19.3
	Adaptive$_{GCV}$	7.0	50.5	100	19.0
	Adaptive$_{BIC}$	58.5	94.5	100	4.0

Table 2: *Example 1. Performance of the CAS-ANOVA procedure using the False Selection Rate criterion and the adaptive CAS-ANOVA procedure using BIC. Tukey's Honestly Significant Difference (HSD) is shown for comparison. Oracle represents the percent of time the procedure chose the entirely correct model. Correct Factors represents the percent of time the nonsignificant factors completely removed. TPR represents the average percent of true differences found, whereas FSR represents the average percent of declared differences that are mistaken.*

Replications		Oracle (%)	Correct Factors (%)	TPR (%)	FSR (%)
1	CAS-ANOVA$_{FSR}$	-	-	-	-
	CAS-ANOVA$_{BIC}$	21.0	83.0	95.8	11.4
	HSD$_{.95}$	1.0	92.0	70.0	0.8
	HSD$_{.65}$	4.0	39.0	83.0	7.2
2	CAS-ANOVA$_{FSR}$	27.3	97.0	94.5	5.5
	CAS-ANOVA$_{BIC}$	41.5	93.5	99.1	6.4
	HSD$_{.95}$	12.5	93.0	85.1	0.7
	HSD$_{.65}$	22.0	46.5	95.9	5.7
4	CAS-ANOVA$_{FSR}$	33.0	97.5	99.8	5.4
	CAS-ANOVA$_{BIC}$	58.5	94.5	100	4.0
	HSD$_{.95}$	62.5	88.5	98.9	0.8
	HSD$_{.65}$	29.0	38.5	99.8	5.3

Table 3: *Example 2. Performance of the CAS-ANOVA procedure using the False Selection Rate criterion and the adaptive CAS-ANOVA procedure using BIC. Tukey's Honestly Significant Difference (HSD) is shown for comparison. Oracle represents the percent of time the procedure chose the entirely correct model. Correct Factors represents the percent of time the nonsignificant factors are completely removed. TPR represents the average percent of true differences found, whereas FSR represents the average percent of declared differences that are mistaken.*

Replications		Oracle (%)	Correct Factors (%)	TPR (%)	FSR (%)
1	CAS-ANOVA$_{FSR}$	-	-	-	-
	CAS-ANOVA$_{BIC}$	15.5	86.0	89.3	11.0
	HSD$_{.95}$	0	89.0	67.1	1.8
	HSD$_{.65}$	0.5	43.0	81.4	10.7
2	CAS-ANOVA$_{FSR}$	25.0	94.5	89.7	5.9
	CAS-ANOVA$_{BIC}$	35.5	87.0	95.1	7.4
	HSD$_{.95}$	4.5	90.0	82.1	1.4
	HSD$_{.65}$	11.0	38.0	91.7	10.4
4	CAS-ANOVA$_{FSR}$	39.0	97.5	98.6	5.1
	CAS-ANOVA$_{BIC}$	58.0	93.0	99.4	5.2
	HSD$_{.95}$	35.5	89.5	93.8	1.3
	HSD$_{.65}$	26.0	44.0	98.7	8.7