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.

March 7, 2008

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. 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 have the oracle property in that asymptotically it performs as well as if the exact structure were known beforehand. A simulation and real data examples show the strong performance of the method.

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

email: bondell@stat.ncsu.edu

1. Introduction

Analysis of Variance (ANOVA) is a commonly used technique to determine a relationship between a continuous response and categorical predictors. 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 differ from one another. This second question is typically answered via a post-hoc analysis involving pairwise comparisons within factors. Some common approaches 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).

Accomplishing the goal of judging importance of a factor 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 resulting in shrinkage and variable selection. Alternative constraints have also been proposed (Frank and Friedman, 1993; Fan and Li, 2001; Tibshirani et al., 2005; Zou and Hastie, 2005; Zou, 2006; Bondell and Reich, 2007).

However, in ANOVA, a factor corresponds to a group of coefficients, typically coded as dummy variables. Naive use of a penalization technique can set some of the dummy variables for a factor to zero, while others are not. Yuan and Lin (2006) propose the Group LASSO to perform variable selection for factors by appropriately treating the dummy variables as a group and use a sum of group norms as opposed to a single overall norm. This approach accomplishes the first goal in ANOVA of selecting the factors. However, the post-hoc analysis of pairwise comparisons must still be performed.

This paper proposes a novel constrained regression approach called CAS-ANOVA (for

Collapsing And Shrinkage in ANOVA) to simultaneously perform the two main goals of ANOVA within a single procedure. The form of the constraint encourages similar effect sizes within a factor to be estimated with exact equality, thus collapsing levels. In addition, via the typical ANOVA parametrization with this constraint, entire factors can be collapsed to a zero effect, hence providing 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 one step as opposed to a two step analysis.

As a motivating example, consider a recent study of six species of deposit feeders from the West Antarctic Peninsula Continental Shelf. To determine feeding selectivity, the activity of Th-234, a naturally-occurring radionuclide, was measured in gut samples collected on five trips at different times of the year. There were a total of 47 observations collected in an unbalanced design, as unequal numbers of each species were collected during the trips. This data was collected as part of a program called FOODBANCS (Food for Benthos on the Antarctic Continental Shelf). For this particular piece of the study, there are two factors of interest: species (6 levels) and time of year (5 levels). The goal of this data collection is to compare feeding habits at various times of the year, as well as, most importantly, between the species. A specific question of interest for the investigators is to classify these species into a smaller number of subgroups based upon similar feeding selectivity. Due to the highly unbalanced nature of the design, the two factors should be handled simultaneously. The typical ANOVA analysis identifies an overall species effect. A post-hoc analysis then finds significant differences among several pairs of species. However, the results do not determine distinct subgroups, as the groups overlap based on the significant/non-significant pairwise differences.

Standard ANOVA approaches based on pairwise difference comparisons do not necessarily yield results that form non-overlapping groups, as illustrated by the species feeding selectivity

example. This is a common complaint among applied scientists. Quoting John Tukey (1949), 'We wish to separate the varieties into distinguishable groups as often as we can without too frequently separating varieties which should stay together'. To accomplish the separation into distinguishable groups for a set of means, as in the one-way ANOVA, Tukey proposed sequential splitting based on the gaps between successive means. Since that time, many other approaches for clustering means in one-way ANOVA have been proposed (Scott and Knott, 1974; Cox and Spotvoll, 1982; Calinski and Corsten, 1985; Bautista, Smith, and Steiner, 1997, for example). However, none of the methods directly generalize to the multi-factor ANOVA problem.

In contrast, the CAS-ANOVA approach automatically forms non-overlapping groups of levels within each factor. Based on the discovered structure, the investigator may look further into the similarities within each group and the difference across groups. For example, from an analysis of the deposit-feeder data, the CAS-ANOVA approach showed a clear distinction between two types of species, surface feeders versus sub-surface feeders.

In addition to its intuitive appeal, it is shown that the CAS-ANOVA approach has attractive theoretical properties. An adaptive version of the CAS-ANOVA is proposed and is shown to have the oracle property, in that it asymptotically performs as well as if the underlying structure were known beforehand and the ANOVA analysis was then performed on the reduced design. This property demonstrates that, asymptotically, the proposed method not only identifies the correct structure, but its estimates of the resulting coefficients are fully efficient. This allows for asymptotic inference to be based on standard ANOVA theory after the reduction in the dimension.

The remainder of the paper proceeds as follows. The CAS-ANOVA procedure is introduced in §2, while computation is discussed in §3. A data adaptive version of the procedure is then introduced in §4. This approach is shown to have the oracle property. A simulation

and two real data examples are given in §5 to illustrate the method. All proofs are given in the Web Appendix.

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 the intercept can be omitted in a balanced design. Let the $n \times p$ design matrix X be the typical over-parameterized ANOVA design matrix of zeros and ones denoting the combination of levels for each observation. This parametrization is useful as equality of coefficients corresponds exactly to collapsing levels.

The least squares procedure estimates the p coefficients, which can be interpreted as the effect sizes for each level, by the solution to

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

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

An additional constraint can be added to this optimization problem to shrink the co-efficients 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, 2007) uses a mixture of L_1 and pairwise L_∞ penalty terms to simultaneously perform variable selection and find clusters among the important predictors.

5

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

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

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 account for the fact that the number of levels for each factor may not be the same and the design may be unbalanced. In the balanced design case, the weights for all pairs of levels within a factor would be equal, so that $w_j^{(km)} = w_j$. An appropriate choice of weights is discussed in the next section.

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

2.2   *Choosing the weights*

In penalized regression, having the predictors on a comparable scale is essential so that the penalization is done equally. This is done by standardization, so that each column of the design matrix has unit L_2 norm. 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 weighting the coefficient in the penalty term by the inverse of this scaling factor. In 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 straightforward, due to the pairwise differences. Although the design matrix consists of zeros and ones, the number of terms in the penalty changes dramatically with the number of levels per factor, while the amount of information 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 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 pairwise differences taken within each factor. Hence $\boldsymbol{\theta}$ is 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 differences have parameter vector

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

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, $M_j = [I_{p_j \times p_j} \ D_j^T]^T$, with D_j the $d_j \times p_j$ matrix of ± 1 that creates $\boldsymbol{\theta}_j$ from $\boldsymbol{\beta}_j$ for a given factor j by picking off each pairwise difference for that factor.

The corresponding design matrix Z for this overparameterized design is an $n \times q$ matrix such that $Z\boldsymbol{\gamma} = X\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$, i.e. $ZM = X$. 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. In particular, choose

$$Z = XM^-,$$

7

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 work with this design matrix and the additional set of constraints defined via $\boldsymbol{\gamma} = M\boldsymbol{\beta}$ and then standardize the columns directly. However, the resulting parameters after standardization no longer have the same interpretation as differences between effects, so that collapsing levels is no longer accomplished. Hence the approach to weighting the terms in the penalty is advocated. The appropriate weights are then the Euclidean norm of the columns of Z that correspond to each of the differences. The proposed weights are exactly these norms.

PROPOSITION. The Moore-Penrose generalized inverse of the matrix M is block diagonal with the jth diagonal block corresponding to the jth 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 allows the determination of the weights via the standardization in this new design space.

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. 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 the column 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. In an unbalanced design, the weighting is approximately equivalent to standardization, as the mean of the column is typically negligible due to the factor of n^{-1}.

3. Computation and Tuning

3.1 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 nonzero. Then $|\theta_k| = \theta_k^+ + \theta_k^-$. 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. 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

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

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 given by

$$
L_j = \left[\begin{array}{ccc} D_j & I_{d_j} & -I_{d_j} \\ \mathbf{1}_{p_j}^T & \mathbf{0}_{d_j}^T & \mathbf{0}_{d_j}^T \end{array} \right]
$$

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

The optimization problem in (4) is a quadratic programming problem, as all constraints are linear. This quadratic programming problem has $p + 2d$ parameters and $p + 3d$ linear constraints. 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.

3.2 Cross-validation and degrees of freedom

The choice of tuning parameter t yields a trade-off of fit to the data with model complexity. This choice can be accomplished via minimizing any of the standard criteria such

9

as AIC, BIC, Generalized Cross-Validation (GCV), or via k-fold Cross-Validation. To use AIC, BIC, or GCV, an estimate of the degrees of freedom is needed. For the LASSO, the number of non-zero coefficients is an unbiased estimate of the degrees of freedom (Zou et al., 2007). For the Fused LASSO (Tibshirani et al., 2005), the number of non-zero blocks of coefficients is used to estimate the degrees of freedom. In this ANOVA design, the natural estimate of degrees of freedom for each factor is the number of unique coefficients minus one for the constraint. Specifically, one has

$$\hat{\mathrm{df}} = \sum_{j=1}^{J} \left(p_j^* - 1 \right),\tag{5}$$

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

It is known that under general conditions, BIC is consistent for model selection if the true model belongs to the class of models considered, while although AIC is minimax optimal, it is not consistent in selection (see for example: Shao, 1997; Yang, 2005). In this ANOVA framework, the emphasis is on identifying the model structure, and not necessarily prediction accuracy. Hence, as discussed later in the examples, BIC is the recommended tuning method for the CAS-ANOVA procedure due to its selection consistency.

4. An adaptive CAS-ANOVA and its asymptotic properties

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

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||^2$$
$$\text{subject to} \tag{6}$$
$$\sum_{k=1}^{p} |\hat{\beta}_k|^{-1}|\beta_k| \leq t,$$

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 plus asymptotic efficiency for the non-zero components.

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

$$\tilde{\boldsymbol{\beta}}^A = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||^2$$
$$\text{subject to} \tag{7}$$
$$\sum_{k=1}^{p_j} \beta_{jk} = 0 \text{ for all } j = 1, ..., J \text{ and } \sum_{j=1}^{J} \sum_{1 \leq k < m \leq p_j} w_j^{(km)*}|\beta_{jk} - \beta_{jm}| \leq t,$$

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 in (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. Then the computation in the adaptive CAS-ANOVA procedure proceeds exactly as before, using this choice of weights.

Like the adaptive LASSO, the adaptive CAS-ANOVA procedure enjoys the oracle property in that its performance is asymptotically equivalent to knowing the correct grouping structure beforehand, and then applying the standard least squares ANOVA fit to the collapsed design. This allows for asymptotic inference on the coefficient estimates, along with

11

the knowledge that as the sample size increases, the probability of selecting the correct model structure tends to one. In comparison with post-hoc procedures to compare pairwise differences by controlling an error rate, such as type I error or false discovery rate, use of the adaptive CAS-ANOVA procedure will, by default, have any error rate go to zero with the sample size.

To discuss the asymptotic properties of the procedure, it is more convenient to rewrite the constrained optimization problem in an equivalent Lagrangian formulation,

$$
\tilde{\boldsymbol{\beta}}^A = \arg\min_{\boldsymbol{\beta}} \left\{ ||\mathbf{y} - X\boldsymbol{\beta}||^2 + \lambda_n \sum_{j=1}^J \sum_{1 \le k < m \le p_j} \frac{w_j^{(km)*}}{\sqrt{n}} |\beta_{jk} - \beta_{jm}| \right\}
$$
$$
\text{subject to}
$$
$$
\sum_{k=1}^{p_j} \beta_{jk} = 0 \text{ for all } j = 1, ..., J,
$$

$$(8)$$

where there is a one-to-one correspondence between the bound t in (7) and the penalty parameter λ_n in (8). Note that the factor of $\sqrt{n}$ in the second term accounts for the fact that the weights grow as $\sqrt{n}$ due to their dependence on the sample size.

Let $\mathcal{A} = \{(j, k, m) : \beta_{jk} \ne \beta_{jm}\}$ denote the set of indices for the differences that are truly non-zero and let $\mathcal{A}_n$ denote the set of indices for those differences that are estimated to be non-zero. Let $\boldsymbol{\gamma}_{\mathcal{A}}$ denote the vector of the pairwise differences that are in $\mathcal{A}$. Note that the set $\mathcal{A}$ defines the truly significant level and factor structure. If this were known beforehand, estimation would proceed by collapsing to this structure and performing the standard ANOVA analysis. Let $\hat{\boldsymbol{\gamma}}_{\mathcal{A}}$ denote this resulting 'oracle' estimator of $\boldsymbol{\gamma}_{\mathcal{A}}$. Then under standard conditions, $\sqrt{n}(\hat{\boldsymbol{\gamma}}_{\mathcal{A}} - \boldsymbol{\gamma}_{\mathcal{A}}) \to N(0, \Sigma)$. Note that the covariance matrix Σ is singular due to the fact that the pairwise differences is an overparameterization. Let $\tilde{\boldsymbol{\gamma}}_{\mathcal{A}}$ denote the CAS-ANOVA estimator of $\boldsymbol{\gamma}_{\mathcal{A}}$. The following theorem shows that the adaptive CAS-ANOVA obtains the oracle property.

THEOREM. Let the design be such that $\frac{n_j^{(k)}}{n} \to \rho_{jk}$, with $0 < \rho_{jk} < 1$, for all j, k. Suppose that $\lambda_n = O_p(n^{1/2})$. Then the adaptive CAS-ANOVA estimator $\tilde{\boldsymbol{\beta}}^A$ and its corresponding

estimator of the differences $\tilde{\boldsymbol{\gamma}}$ has the following properties:

    a. $P(\mathcal{A}_n = \mathcal{A}) \to 1$

    b. $\sqrt{n}(\tilde{\boldsymbol{\gamma}}_\mathcal{A} - \boldsymbol{\gamma}_\mathcal{A}) \to N(0, \Sigma)$.

The above theorem guarantees that the procedure will determine the correct structure with probability tending to one. In addition, it allows for inference to be done based on the same asymptotic variance that one would obtain via least squares estimation after constructing the appropriate design matrix that would correspond to collapsing levels and removing unimportant factors. Based on part (a) of the theorem, a secondary inference may not be necessary at least for determining the significant factors and levels. Note that the condition on the design is needed to ensure that each level of each factor contains a non-vanishing proportion of the data. However, it is not necessary to have this non-vanishing proportion for each combination, only marginally.

5. Examples
5.1 *Simulation study*

A simulation study was carried out to examine the performance of the CAS-ANOVA procedure in terms of selecting 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 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 error variance is set to 1. 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 used a balanced design with a single observation per combination, and then repeated the setup for the two additional cases of 2 and 4 replications per combination.

As mentioned previously, based on the results of Shao (1997) and Yang (2005), BIC is the tuning method recommended due to its consistency in model selection. Table 1 compares

13

the standard version of the CAS-ANOVA procedure with the adaptive version, both tuned via BIC 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. This is due to the known phenomenon of the L_1 penalty without the adaptive weights overshrinking the non-zero coefficients (Fan and Li, 2001; Zou, 2006). Due to the overshrinking, in order to eliminate the unimportant differences, the coefficient estimates of the remaining ones are too small, so that the residual squared error is inflated. Any criterion based on the residual squared error, such as BIC, is then less likely to choose the correct model, particularly if the correct model is relatively sparse. The adaptive weights alleviate this overshrinking, by adapting the penalty to penalize the non-zero coefficients less, thus allowing for the unimportant differences to be removed with less shrinkage on the important ones, and hence less inflation of residual squared error. AIC and GCV were also used as tuning approaches. However, the results based on BIC were much better, as expected due to the asymptotic selection consistency. Therefore, only the results tuning via BIC are shown here.

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 Benjamini and Hochberg (BH, 1995) procedure to control False Discovery Rate is used to adjust for all pairwise differences. For both the HSD procedure and the BH correction, dropping a factor is defined as finding no significant pairwise differences within that factor. Note that neither of these two procedures were designed to create non-overlapping groups. The resulting pairwise comparisons can result in impossible grouping structures, i.e. groups that overlap when pairwise comparisons are translated into a grouping structure. This helps to contribute to the lower percentages of selecting the true model ('Oracle'), since the coefficients in any true model necessarily partitions each factor into non-overlapping groups. Overall, the adaptive CAS-ANOVA procedure does a good job in picking out the appropriate model structure. Note that the adaptive CAS-ANOVA procedure has both its TPR going to one and its FSR going to zero as the sample size increases, as anticipated from the results in Theorem 1. Meanwhile, the performance of the non-adaptive procedure is relatively poor. Hence for the remainder of the examples, only the adaptive version is used, and this is the approach that is recommended in practice due to both its asymptotic theory and empirical performance.

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

5.2 *Antarctic shelf deposit-feeder example*

The data for this example comes from a study of deposit feeders from the west Antarctic Peninsula continental shelf. The Th-234 activity was measured in gut samples collected on five trips at different times of the year. The data was collected as part of the FOODBANCS program. The two factors are species (6 levels) and time of year (5 levels). There are a total

15

of 47 observations in this unbalanced design. The six species are: peniagone sp. ($n = 8$), bathyplotes sp. ($n = 9$), protelpidia murrayi ($n = 11$), echiuran worm ($n = 3$), urchin ($n = 8$), and molpadia musculus ($n = 8$). The numbers of observations collected on each of the five trips were: 4, 10, 11, 10, and 12, respectively. An important goal of this study is to use the gut Th-234 activity to help determine similarities and differences between feeding selectivity of the six species.

Using the adaptive CAS-ANOVA procedure tuned via BIC on this data resulted in the grouping structure shown in Figure 1. The six species are plotted in Figure 1a. The urchin and molpadia formed a distinct group from the others. This discovery is scientifically justified; these two species are sub-surface feeders, while the other four species are surface feeders. Among the four surface feeders, the peniagone was separated out from the others, so further investigation into that difference is warranted. Figure 1b shows that the five trips were all combined so that the time of year factor is deemed to be not significant.

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

An analysis using Tukey's HSD procedure at both the .95 and the .65 levels was performed and also plotted in Figure 1. At the .95 level, the surface and subsurface feeders are no longer separated. The groups are overlapping with no clear distinction between the types of species. At the .65 level, the subsurface feeders are now separated out. However, at this level, the time of year factor is now found to be significant, with the trips forming two overlapping groups. Using the Benjamini and Hochberg correction also resulted in no clear separation between the species, as in the HSD procedure at the .95 level. In addition, the time of year is split into two overlapping groups, as in the HSD procedure at the .65 level. For all three of these procedures, the echiuran worm appears in multiple groups. This is partly due to the fact that the sample size for the echiuran worm is very small, and hence the classical approaches based on significance testing lack power to detect significant differences.

16

5.3 *Minnesota barley example*

This second 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).

Again, Tukey's HSD procedure at both the .95 and the .65 levels along with the Benjamini and Hochberg correction were performed and plotted in Figure 2. All three procedures produce overlapping groups of levels for the location factor (Figure 2b). For example, each finds that the Morris and Duluth locations are significantly different from one another, but neither location significantly differs from the University Farm. Two of the three approaches also find overlapping groups for the varieties as well.

The newly proposed adaptive CAS-ANOVA procedure tuned via BIC was used. The grouping structure is shown in Figure 2. By nature of the method, the groups are non-overlapping. For example, the overlapping-group problem mentioned above is resolved by creating a separate group for the University Farm.

Finally, note that all of the approaches find the two years significantly different from one another.

***** FIGURE 2 GOES HERE ******

Based on the simulation and the real data examples, 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.

6. 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. An adaptive version of this procedure with tuning parameter chosen via BIC has shown strong empirical performance. In addition, it asymptotically obtains the oracle property in that inference may be conducted by collapsing the design based on the discovered structure and using the asymptotic distribution of the least squares estimator for that design.

Although the CAS-ANOVA procedure obtains the correct structure asymptotically, one drawback is that it does not guarantee control of an error rate in finite samples, as in the methods based on typical pairwise comparisons. A modification of the approach of Wu, Boos, and Stefanski (2007) is possible to tune the approach to control a false discovery rate for the pairwise differences. Instead of adding noise predictors, cells in the ANOVA design can be randomly split. This creates differences that are known to be zero. These differences in split cells can be monitored to estimate a false discovery rate. However, by fixing the error rate at a prespecified threshold, the resulting procedure will no longer obtain the oracle property in that it will not select the correct structure with probability tending to one. To do so, the threshold on the error rate must go to zero with the sample size. The existing procedures based on pairwise comparisons also cannot obtain the oracle property by using a fixed error rate.

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 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.

Supplementary Materials

The Web Appendix containing the proofs is available under the paper information link at the Biometrics website http://www.tibs.org/biometrics. An explicit solution for a simple 3-level one-way ANOVA model is also in the Web Appendix.

Appendix A

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

19

$d \times p_j$ matrix of ± 1 that picks off each of the pairwise differences from factor j. 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 $M_j^- M_j = (p_j + 1)^{-1}[(I_{p_j \times p_j} + \mathbf{1}_{p_j}\mathbf{1}_{p_j}^T) + D_j^T D_j]$. 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 $M_j^- M_j = I_{p_j \times p_j}$.

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),

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

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, and M^- is the Moore-Penrose inverse of M. The Euclidean norm of the columns of $Z = XM^-$ corresponding to the differences follows, 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.

Proof of Theorem 1a.

Due to the form of the penalty in that only differences within a factor are considered, for ease of presentation, it suffices to consider the one-way ANOVA case, and the result directly generalizes to the multi-factor case.

For the one-way ANOVA, (8) becomes

$$\tilde{\boldsymbol{\beta}}^A = \arg\min_{\boldsymbol{\beta}} \left\{ ||\mathbf{y} - X\boldsymbol{\beta}||^2 + \lambda_n \sum_{1 \leq k < l \leq p} \frac{w^{(kl)*}}{\sqrt{n}} |\beta_l - \beta_k| \right\}$$
$$\text{subject to}$$
$$\sum_{k=1}^p \beta_k = 0. \tag{9}$$

Let $\mathcal{B}_n = \mathcal{A}_n \cap \mathcal{A}^c$, i.e. the differences that should be zero, but were incorrectly set to non-zero. It will now be shown that $P(\mathcal{B}_n \neq \emptyset) \to 0$ i.e., that the true zeros will be set to zero with probability tending to one. It then remains to show that $P(\mathcal{A}_n^c \cap \mathcal{A} \neq \emptyset) \to 0$, i.e. none of the nonzero differences are mistakenly set to zero. This second assertion, will follow directly from the $\sqrt{n}$-consistency of the estimators for the differences in $\mathcal{A}$, which will be proved in part b of the theorem.

Without loss of generality, suppose that the levels were ordered such that at the solution we have $\tilde{\beta}_1 \leq \tilde{\beta}_2 \leq ... \leq \tilde{\beta}_p$. Now, suppose that $\mathcal{B}_n$ were not empty, then let $\tilde{\beta}_m$ denote the largest coefficient among those involved in the differences that are indexed by $\mathcal{B}_n$ i.e., $m = \max\{l : (k,l) \in \mathcal{B}_n \text{ for some } k\}$. Let $\tilde{\beta}_q$ denote the smallest coefficient such that the pair $(q,m) \in \mathcal{B}_n$, so that $q < m$. Consider the full-rank reparameterization of the ANOVA design to treat level q as the baseline, i.e. for $k \neq q$ let $\eta_k = \beta_k - \beta_q$, and let $\eta_q = \mu + \beta_q$. Note that by assumption, $\eta_m = 0$ and $\tilde{\eta}_m \neq 0$ and that $\tilde{\eta}_m - \tilde{\eta}_k \geq 0$ for all $(k,m) \in \mathcal{B}_n$. From this reparameterization it follows that (for $k < l$) at the solution

$$|\beta_l - \beta_k| = \begin{cases} \eta_l - \eta_k & \text{for } k \neq q, l \neq q \\ \eta_l & \text{for } k = q, l > q \\ -\eta_k & \text{for } k < q, l = q \end{cases} \tag{10}$$

21

Hence,

$$\tilde{\boldsymbol{\eta}}^A = \underset{\boldsymbol{\eta}}{\arg\min} \left[||\mathbf{y} - Z\boldsymbol{\eta}||^2 + \lambda_n \left\{ \begin{array}{l} \sum_{1 \le k < l \le p, k \ne q, l \ne q} \frac{w^{(kl)*}}{\sqrt{n}} (\eta_l - \eta_k) \\ + \sum_{q < k \le p} \frac{w^{(qk)*}}{\sqrt{n}} \eta_k \\ - \sum_{1 \le k < q} \frac{w^{(kq)*}}{\sqrt{n}} \eta_k \end{array} \right\} \right], \qquad (11)$$

where Z is the typical design matrix for the parameterization that treats level q as the baseline.

Since $\tilde{\eta}_m \ne 0$, the optimization criteron is differentiable with respect to η_m at the solution. Consider this derivative in a neighborhood of the solution on which the coefficients that are set equal remain equal i.e. the differences remain zero. On this neighborhood, the terms involving $(k, m) \in \mathcal{A}_n^c \cap \mathcal{A}^c$ can be omitted, as these terms will vanish in the objective function. A contradiction will be obtained on this neighborhood of the solution.

Due to the differentiability, the solution $\tilde{\boldsymbol{\eta}}$ must satisfy

$$\frac{2}{\sqrt{n}} z_m'(\mathbf{y} - Z\tilde{\boldsymbol{\eta}}) = \frac{\lambda_n}{\sqrt{n}} \sum_{k \ne m, (k,m) \in \mathcal{A}} (-1)^{I[m<k]} \frac{w^{(km)*}}{\sqrt{n}} + \frac{\lambda_n}{\sqrt{n}} \sum_{k \ne m, (k,m) \in \mathcal{B}_n} \frac{w^{(km)*}}{\sqrt{n}}, \qquad (12)$$

where z_m denotes the mth column of Z.

Note that each term in the second sum on the right hand side will be nonnegative. Now, $\frac{\lambda_n}{\sqrt{n}} \frac{w^{(km)*}}{\sqrt{n}} = \frac{\lambda_n}{\sqrt{n}} |\hat{\beta}_m - \hat{\beta}_k|^{-1} (p+1)^{-1} \sqrt{\frac{n^{(k)}}{n} + \frac{n^{(m)}}{n}}$. By assumption on the design, $\sqrt{\frac{n^{(k)}}{n} + \frac{n^{(m)}}{n}} = O_p(1)$. Now, for all k such that $(k, m) \in \mathcal{A}$, it follows that $|\hat{\beta}_m - \hat{\beta}_k|^{-1} = O_p(1)$, while for $(k, m) \in \mathcal{A}^c$, $|\sqrt{n}(\hat{\beta}_m - \hat{\beta}_k)|^{-1} = O_p(1)$, due to the $\sqrt{n}$-consistency of the initial OLS estimator. Hence each term in the second sum on the right hand side is $O_p(\lambda_n)$, while the terms in the first sum are $O_p(\lambda_n n^{-1/2})$. By assumption, the second sum is nonempty, and hence the right hand side is $O_p(\lambda_n)$, since at least one term must be.

Meanwhile, the left hand side is $O_p(1)$. Since, by assumption $\lambda_n = O_p(n^{1/2})$, we have a contradiction, so it must be that $P(\mathcal{B}_n \ne \emptyset) \to 0$.

Proof of Theorem 1b.

The proof of asymptotic normality follows that of Zou (2006). The sketch of how it is adapted to this setting follows below, for complete details see Zou (2006). Let $\tilde{u} = \sqrt{n}(\tilde{\eta} - \eta_0)$, where η_0 denotes the true parameter vector. Then $\tilde{u} = \arg\min V_n(u)$, where

$$V_n(u) = u'\left(\tfrac{1}{n}Z'Z\right)u - 2\tfrac{\epsilon' Z}{\sqrt{n}}u + \tfrac{\lambda}{\sqrt{n}}P(u),$$

with

$$P(u) = \sum_{1\le k<l\le p, k\neq q, l\neq q} \tfrac{w^{(kl)*}}{\sqrt{n}}\sqrt{n}\left(\left|\eta_{l0} - \eta_{k0} + \tfrac{u_l - u_k}{\sqrt{n}}\right| - |\eta_{l0} - \eta_{k0}|\right) +$$
$$\sum_{k\neq q}(-1)^{I[k<q]}\tfrac{w^{(qk)*}}{\sqrt{n}}\sqrt{n}\left(\left|\eta_{k0} + \tfrac{u_k}{\sqrt{n}}\right| - |\eta_{k0}|\right).$$

By the same argument as in Zou (2006), $\tfrac{\lambda}{\sqrt{n}}P(u)$ will go to zero for the correct structure and diverge under the incorrect structure. Furthermore, if $\tilde{\gamma}_{\mathcal{A}}^c = 0$ then $V_n(u) = V_n^O(u_O)$ where $V_n^O(u_O)$ is the value of the objective function obtained after collapsing the design matrix Z to the correct structure (the 'Oracle') determined by $\mathcal{A}$ i.e., $Z = Z_O$ and the corresponding η_O. This is accomplished by collapsing the columns of the original design matrix into the sum of the two indicator variables for those differences that are truly zero i.e., combining the two levels for each pair in $\mathcal{A}^c$.

Now, $\tfrac{1}{n}Z_O'Z_O \to C$, where C is positive definite and $\tfrac{\epsilon' Z_O}{\sqrt{n}} \to W = N(0, \sigma^2 C)$. This follows directly from the assumption that $\tfrac{n^{(k)}}{n} \to \rho_k$, with $0 < \rho_k < 1$, for all k and that $\operatorname{Var}(\epsilon) = \sigma^2$. As in Zou (2006), $V_n(u) \to V(u)$, where

$$V(u) = \begin{cases} u_O'Cu_O - 2u_O'W & \text{if } \tilde{\gamma}_{\mathcal{A}}^c = 0 \\ \infty & \text{otherwise} \end{cases}. \tag{13}$$

Hence the asymptotic normality follows by noting that $V_n(u)$ is convex and the unique minimizer of $V(u)$ is $(C^{-1}W, 0)'$. Hence $\tilde{u}_{\mathcal{A}} \to N(0, \sigma^2 C^{-1})$. The result for all pairwise differences, $\gamma_{\mathcal{A}}$, then follows after the (singular) transformation.

Appendix B

An explicit solution for a special case.

Consider the balanced one-way ANOVA with 3 levels and r observations per level. Without loss of generality, suppose that the means of the groups are ordered so that $\bar{y}_1 > \bar{y}_2 > \bar{y}_3$, and $\sum_{j=1}^{3} \bar{y}_j = 0$. Note that the OLS estimates of $\beta_1, \beta_2, \beta_3$ are these means. In the balanced design, the ordering of the estimates under the CAS-ANOVA penalty will remain constant.

After simplification, the constraint on the pairwise differences in the standard non-adaptive CAS-ANOVA procedure (2) is $\frac{\sqrt{2r}}{2}(\beta_1 - \beta_3) \leq t$. Rewrite $t = \frac{\sqrt{2r}}{2}\rho t^*$, where $0 \leq \rho \leq 1$, and $t^* = \hat{\beta}_1 - \hat{\beta}_3 = \bar{y}_1 - \bar{y}_3$, the OLS estimate of the range. Then the constraint becomes $\beta_1 - \beta_3 \leq \rho(\bar{y}_1 - \bar{y}_3)$, so that ρ is a shrinkage factor on the range. Let $\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3$ be the CAS-ANOVA solution. Then it must be that $\sum_{j=1}^{3} \tilde{\beta}_j = 0$ and $\tilde{\beta}_1 - \tilde{\beta}_3 = \rho(\bar{y}_1 - \bar{y}_3)$, so there is only one free parameter. Substitution of the constraints into the quadratic loss function in (2) and differentiation yields, after simplification, the possible solution at the critical point $\tilde{\beta}_1 = \bar{y}_1 - \frac{1-\rho}{2}(\bar{y}_1 - \bar{y}_3)$, a shrunken version of the original. Then $\tilde{\beta}_2 = \rho t^* - 2\tilde{\beta}_1$ and $\tilde{\beta}_3 = \tilde{\beta}_1 - \rho t^*$. Now the boundary must be checked to ensure $\tilde{\beta}_1 \geq \tilde{\beta}_2 \geq \tilde{\beta}_3$, this yields $\rho(\bar{y}_1 - \bar{y}_3)/3 \leq \tilde{\beta}_1 \leq 2\rho(\bar{y}_1 - \bar{y}_3)/3$, so that $\tilde{\beta}_1$ must take up at least $1/3$, but no more than $2/3$ of the range. This gives the solution:

$$\tilde{\beta}_1 = \min\left\{\max\left\{\rho(\bar{y}_1 - \bar{y}_3)/3, \bar{y}_1 - \tfrac{1-\rho}{2}(\bar{y}_1 - \bar{y}_3)\right\}, 2\rho(\bar{y}_1 - \bar{y}_3)/3\right\}.$$

and $\tilde{\beta}_2, \tilde{\beta}_3$ are as above.

The same calculations may be done for the adaptive version given by (7), however, the form of the solution is not as concise due to the weights being reciprocals of the differences, so that not much simplification is possible. Once beyond the three level one-way ANOVA, each boundary condition must be checked, corresponding to each of the possible groupings, so that the solution is thus not available in a simple form.

References

Bautista, M. G., Smith, D. W., and Steiner, R. L. (1997), A cluster-based approach to means separation, *Journal of Agricultural, Biological and Environmental Statistics*, **2**, 179-197.

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. (2007), Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR, *Biometrics*, doi: 10.1111/j.1541-0420.2007.00843.x

Calinski, T. and Corsten, L. C. A. (1985), Clustering means in ANOVA by simultaneously testing, *Biometrics*, **41**, 39-48.

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

Cox, D. R. and Spotvoll, E. (1982), On partitioning means into groups, *Scandinavian Journal of Statistics: Theory and Applications*, **9**, 147-152.

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**, 109-148.

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.

Scott, A. J. and Knott, M. (1974), A cluster analysis method for grouping means in the analysis of variance, *Biometrics*, **30**, 507-512.

Shao, J. (1997), An asymptotic theory for linear model selection, *Statistica Sinica*, **7**, 221-264.

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

Tukey, J. W. (1949), Comparing individual means in the analysis of variance, *Biometrics*, **5**, 99-114

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. (2007), Controlling variable selection by the addition of pseudo-variables, *Journal of the American Statistical Association*, **102**, 235-243.

Yang, Y. (2005), Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation, *Biometrika*, **92**, 937-950.

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. (2007), On the degrees of freedom of the lasso, *Annals of Statistics*, **35**, 2173-2192.

Figure 1: Group structure for the Antarctic shelf data using Tukey's HSD (with 95% and 65% confidence levels), Benjamini and Hochberg correction, and adaptive CAS-ANOVA. Each distinct letter represents a set of levels that are grouped together.

(a) Species

	CAS–ANOVA	Tukey's HSD (0.95)	Tukey's HSD (0.65)	Benjamini & Hochberg
peniagone sp.	A	A	A	A
bathyplotes sp.	B	A B	B	B
protelpidia	B	B	B	B
echiuran worm	B	A B C	A B	A B C
urchin	C	B C	C	C
molpadia	C	C	C	C

(b) Trips

	CAS–ANOVA	Tukey's HSD (0.95)	Tukey's HSD (0.65)	Benjamini & Hochberg
Trip I	A	A	A	A B
Trip II	A	A	A	A
Trip III	A	A	A B	A B
Trip IV	A	A	A B	A B
Trip V	A	A	B	B

Figure 2: Group structure for the Minnesota barley data using Tukey's HSD (with 95% and 65% confidence levels), Benjamini and Hochberg correction, and adaptive CAS-ANOVA. Each distinct letter represents a set of levels that are grouped together.

(a) Varieties

	CAS–ANOVA	Tukey's HSD (0.95)	Tukey's HSD (0.65)	Benjamini & Hochberg
Manchuria	A	A	A	A
Svansota	A	A	A	A
Velvet	A	A B	A	A
Peatland	A	A B	A	A B
Trebi	B	B	B	B

(b) Locations

	CAS–ANOVA	Tukey's HSD (0.95)	Tukey's HSD (0.65)	Benjamini & Hochberg
Crookston	A	A	A	A
Morris	A	A B	A B	A
University Farm	B	A B C	B C	A B
Duluth	C	B C	C	B
Grand Rapids	C	C	C	B
Waseca	D	D	D	C

Table 1: *Performance of the CAS-ANOVA procedure and the adaptive CAS-ANOVA procedure. Tukey's Honestly Significant Difference (HSD) and the Benjamini and Hochberg (BH) correction are 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$_{Adaptive}$	21.0	83.0	95.8	11.4
	CAS-ANOVA$_{Standard}$	3.0	48.0	97.8	22.8
	HSD$_{.95}$	1.0	92.0	70.0	0.8
	HSD$_{.65}$	4.0	39.0	83.0	7.2
	BH	12.0	77.5	86.6	4.8
2	CAS-ANOVA$_{Adaptive}$	41.5	93.5	99.1	6.4
	CAS-ANOVA$_{Standard}$	6.0	62.0	99.6	19.2
	HSD$_{.95}$	12.5	93.0	85.1	0.7
	HSD$_{.65}$	22.0	46.5	95.9	5.7
	BH	36.5	82.5	95.8	4.6
4	CAS-ANOVA$_{Adaptive}$	58.5	94.5	100	4.0
	CAS-ANOVA$_{Standard}$	4.0	59.5	100	19.4
	HSD$_{.95}$	62.5	88.5	98.9	0.8
	HSD$_{.65}$	29.0	38.5	99.8	5.3
	BH	45.7	81.8	99.8	5.2