

LOW RISK FITS TO DISCRETE INCOMPLETE MULTI-WAY LAYOUTS

Rudolf Beran

Key words: submodel selection, multiparametric asymptotics.
COMPSTAT 2004 section: Model Selection, Functional Data Analysis.

Abstract: The discrete multi-way layout is a widespread data-type associated with regression, experimental designs, gene or protein chips, digital images or videos, and more. A discrete multi-way layout has a finite number of factor level combinations. The layout may be unbalanced or incomplete or both. We consider candidate fits to an incomplete layout that are least squares fits to certain submodels induced by tensor product space ANOVA models for a complete layout. The candidate estimator with smallest estimated risk is selected. Multiparametric asymptotics under a general (saturated) Gaussian model show that the selected estimator achieves smallest asymptotic risk over the candidate class through bias-variance trade-off.

1 Introduction

Each factor that influences the responses in a *discrete* multi-way layout has a finite number of levels, as in classical experimental design, digital images or videos, gene or protein chips, and regression. The factors can be either ordinal or nominal or some of each. Pioneering results on low-risk fits to multi-way layouts include Stein's (1966) shrinkage estimators for complete balanced discrete two-way layouts with both factors nominal and Mallow's (1973) study of C_p as a criterion for selecting a submodel fit. Tukey's (1977) computational experiments in discrete one- or two- or three-way layouts with ordinal factors indicated that smoothing can bring out pattern. Beran (2000, 2002) studied low risk adaptive penalized least squares fits to a discrete one-way layout whose factor is either ordinal or nominal.

Related in spirit are spline estimators of a mean function on a one-way layout with a *continuous* ordinal factor that is observed at discrete points (Wahba (1990), Heckman and Ramsay (2000)) and smoothing spline tensor product space ANOVA techniques for functional data analysis (Wahba et al. (1995), Lin (2000)). The methods in this paper are designed for large, incomplete, *discrete* multi-way layouts with little or no replication. Regression models are a leading case.

As candidate estimators, we consider least squares fits to submodels for the observed incomplete layout that are induced by tensor product space ANOVA submodels for an associated complete multi-way layout of means. We estimate the risk of each candidate estimator under a general model that does *not* assume correctness of any submodel. Finally, we select the candidate fit that minimizes estimated risk. We show that the selected submodel fit has relatively low asymptotic risk under the general model as the number of observed factor-level combinations tends to infinity. This low asymptotic risk is achieved through variance-bias trade-off.

2 Candidate fits to the saturated model

Consider k_0 factors, either nominal or ordinal, in which factor k has p_k distinct levels. Let $\mathcal{I}$ denote the set of all k_0-tuples $i = (i_1, i_2, \ldots, i_{k_0})$ such that $1 \le i_k \le p_k$ for $1 \le k \le k_0$. The component i_k indexes the levels of factor k. A *complete* k_0-way layout of means consists of real values $\{m_i : i \in \mathcal{I}\}$. We order the $p = \prod_{k=1}^{k_0} p_k$ elements of the index set $\mathcal{I}$ in mirrored dictionary order: i_{k_0} serves as the first "letter" of the word, i_{k_0-1} as the second "letter", and so forth. Hereafter we assume that $\mathcal{I}$ is so ordered. Taken in this order, the indexed means for the complete multi-way layout form the $p \times 1$ vector

$$m = \{\ldots\{\{m_{i_1,i_2,\ldots,i_{k_0}} : 1 \le i_1 \le p_1\}, 1 \le i_2 \le p_2\}, \ldots, 1 \le i_{k_0} \le p_{k_0}\}. \quad (1)$$

Observations are available on the means $\{m_i : i \in \mathcal{I}_0\}$, where $\mathcal{I}_0$ is a subset of $\mathcal{I}$. In general, these observations $y = \{\{y_{ij} : 1 \le j \le n_i\}, i \in \mathcal{I}_0\}$ form an *incomplete unbalanced* k_0-way layout. The vector y is $n \times 1$ with $n = \sum_{i \in \mathcal{I}_0} n_i$. Let $q \le p$ denote the cardinality of $\mathcal{I}_0$. Define the means-incidence matrix D to be the $q \times p$ matrix of zeroes and ones such that $m_D = Dm$ lists, in vector form, the means $\{m_i : i \in \mathcal{I}_0\}$ for the observed incomplete k_0-way layout. Let C be the $n \times q$ data-incidence matrix that suitably replicates components of the vector $m_D = Dm$ into the vector $\eta = \mathrm{E}(y) = Cm_D$. For a complete layout of data, q equals p and D is just the identity matrix. The general Gaussian *saturated model* for the incomplete layout of observations y puts no restrictions on the mean vector m:

$$y \sim N(\eta, \sigma^2 I_n), \text{ where } \eta = Cm_D = CDm, \quad m \in R^p. \quad (2)$$

2.1 Generic candidate submodel fits

Consider a submodel of the saturated model in which m is restricted to a given subspace of R^p. This condition has several mathematical expressions.

Theorem 1 *Let V be a $p \times r$ matrix whose columns form an orthonormal basis for the r-dimensional subspace $\mathcal{V}$. Let $Q = VV'$. Then the following assertions are equivalent: (a) $m \in \mathcal{V}$; (b) $m = V\gamma$ for some $\gamma \in R^r$; (c) $m = Q\beta$ for some $\beta \in R^p$; (d) $m = Qm$.*

The symmetric idempotent matrix Q is the unique orthogonal projection of R^p into $\mathcal{V}$. Its eigenvalues are either zero or one. The expression $Q = VV'$ is a spectral decomposition of Q. Because Q has eigenvalue one r times and eigenvalue zero $p - r$ times, there exist many eigenvector matrices V such that $Q = VV'$.

For Q a projection as in the Theorem, define *submodel*(Q) by imposing the constraint $m = Qm$ on the saturated model (2). Taking Q to be the identity matrix I_p recovers the saturated model itself. Let $\eta(Q) = \mathrm{E}(y)$, evaluated under submodel(Q). By the foregoing Theorem, $\eta(Q) = CDQ\beta = CDV\gamma$. Consequently, the least squares estimator of $\eta(Q)$ under submodel(Q) has several equivalent expressions in which the superscript $^+$ denotes the Moore-Penrose inverse.

Theorem 2 *Let $M(Q) = CDQ(CDQ)^+ = CD(CDQ)^+ = CDV(CDV)^+$ and let $M = M(I_p) = CD(CD)^+ = CC^+$. The least squares estimator of $\eta(Q)$ under submodel(Q) is $\hat{\eta}(Q) = M(Q)y$. In particular, the least squares estimator of η under the saturated model is $\hat{\eta} = My$.*

This result can be derived thorough properties of the Moore-Penrose inverse. The matrices $M(Q)$ and M are both symmetric idempotent and satisfy $MM(Q) = M(Q) = M(Q)M$.

For each Q in a class of projection matrices that express tentative prior conjectures about m, we will consider $\hat{\eta}(Q)$ as a (usually biased) candidate estimator for η in the saturated model for the unbalanced incomplete k_0-way layout. The associated candidate estimator for the cell means $m_D = Dm$ is then $\hat{m}_D(Q) = C^+\hat{\eta}(Q)$. Although submodel fits generate the candidate estimators, it is not assumed in this paper that any submodel of the saturated model (2) holds.

2.2 Tensor product candidate submodel fits

To generate useful candidate projections Q, we first express the ANOVA decomposition for a complete k_0-way layout of means in projection form. Let $\mathcal{S}$ denote the set of all subsets of $\{1, 2, \ldots, k_0\}$, including the empty set $\emptyset$. The cardinality of $\mathcal{S}$ is 2^{k_0}.

For every k, define the $p_k \times 1$ vector u_k and the $p_k \times p_k$ matrices J_k, H_k by $u_k = p_k^{-1/2}(1, 1, \ldots, 1)'$, $J_k = u_k u_k'$, and $H_k = I_{p_k} - J_k$. For each k, the symmetric idempotent matrices J_k and H_k have rank (or trace) 1 and $p_k - 1$ respectively. Let U_k be any $p_k \times (p_k - 1)$ matrix such that $(u_k|U_k)$ is an orthogonal matrix. Then the foregoing entails that $H_k = U_k U_k'$.

For every set $S \in \mathcal{S}$, define $P_{S,k} = J_k$ if $k \notin S$ and $P_{S,k} = H_k$ if $k \in S$. Define the $p \times p$ Kronecker product matrix $P_S = \bigotimes_{k=1}^{k_0} P_{S,k_0-k+1}$. Evidently P_S is symmetric idempotent for every $S \in \mathcal{S}$; if $S \neq \emptyset$, the rank (or trace) of P_S is $\prod_{k \in S}(p_k - 1)$; the rank (or trace) of $P_\emptyset$ is 1; if S_1 and S_2 are two different sets in $\mathcal{S}$, then $P_{S_1}P_{S_2} = 0 = P_{S_2}P_{S_1}$; and $\sum_{S \in \mathcal{S}} P_S = I_p$.

Consequently, the $\{P_S \colon S \in \mathcal{S}\}$ are orthogonal projections that decompose R^p into 2^{k_0} mutually orthogonal subspaces. The ANOVA decomposition of a complete k_0-way layout of means is the identity, for every $m \in R^p$,

$$m = \sum_{S \in \mathcal{S}} P_S m. \tag{3}$$

Here $P_\emptyset m$ is the overall mean term in the decomposition. If $S \neq \emptyset$, $P_S m$ is the main effect or interaction term defined by the factors $k \in S$. This ANOVA decomposition suggests a rich variety of choices for the projection matrix Q that determines submodel(Q) of the saturated model (2) for the incomplete k_0-way layout. The central ideas are as follows:

All factors nominal. Let $\{\mathcal{A}_j\colon 1 \leq j \leq j_0\}$ denote a collection of subsets of $\mathcal{S}$ that is partially ordered under the inclusion operation. Let

$$Q_{\mathcal{A}_j} = \sum_{S \in \mathcal{A}_j} P_S \tag{4}$$

and let $\mathcal{Q} = \{Q_{\mathcal{A}_j}\colon 1 \leq j \leq j_0\}$. The candidate estimators $\{\hat{\eta}(Q)\colon Q \in \mathcal{Q}\}$ for η are least squares fits to the designated ANOVA submodels.

All factors ordinal. Without loss of generality, assume that the indexing of the levels of an ordinal factor follows their numerical order. Prior conjecture may then be that adjacent means in the k_0-way layout vary smoothly as a function of the ordinal factor levels. Tensor product space submodels that express this are described through the following general scheme.

Let $W_k(c_k)$ be any $p_k \times c_k$ matrix of rank $c_k \leq p_k$ whose first column is proportional to u_k and whose successive columns are increasingly wiggly. For instance, the columns of $W_k(c_k)$ could be successive powers, from zero to $c_k - 1$, of the vector of levels of factor k. Define the orthogonal projections $G_k(c_k) = W_k(c_k)W_k(c_k)^+$ and $K_k(c_k) = G_k(c_k) - J_k$. Define $P_{S,k} = J_k$ if $k \notin S$ and $P_{S,k} = K_k(c_k)$ if $k \in S$. For $c = (c_1, c_2, \ldots, c_{k_0})$, define $P_S(c) = \bigotimes_{k=1}^{k_0} P_{S,k_0-k+1}$.

In analogy to the preceding paragraph, let

$$Q_{\mathcal{A}_j}(c) = \sum_{S \in \mathcal{A}_j} P_S(c) \tag{5}$$

and let $\mathcal{Q} = \{Q_{\mathcal{A}_j}(c)\colon 1 \leq j \leq j_0, 1 \leq c_1 \leq d_1, 1 \leq c_2 \leq d_2, \ldots 1 \leq c_{k_0} \leq d_{k_0}\}$. The estimators $\{\hat{\eta}(Q)\colon Q \in \mathcal{Q}\}$ are a class of candidate submodel estimators for η.

Some factors nominal, some factors ordinal. We proceed in similar fashion, choosing the multiplicands $P_{S,k}$ in the cross-factor Kronecker product according to the nominal or ordinal nature of each factor, as above.

3 Adaptive low risk submodel fits

We will assess the performance of the candidate estimator $\hat{\eta}(Q)$ of $\eta = CDm$ through its risk under normalized quadratic loss, $L(\hat{\eta}(Q), \eta) = q^{-1}|\hat{\eta}(Q) - \eta|^2$. Let $r(Q) = \text{rank}(M(Q)) = \text{rank}(CDQ)$. The risk of candidate estimator $\hat{\eta}(Q) = M(Q)y$ under the saturated model is then

$$R(\hat{\eta}(Q), \eta, \sigma^2) = \mathrm{E}L(\hat{\eta}(Q), \eta) = q^{-1}[\sigma^2 r(Q) + |M(Q)\eta - \eta|^2]. \tag{6}$$

Recall that $\hat{\eta} = My$ is the least squares estimator of η under the saturated model. Let denote $\hat{\sigma}^2$ denote a consistent estimator of σ^2. Apart from possible bias in $\hat{\sigma}^2$, Mallow's (1973) C_p criterion and Stein's (1981) unbiased risk estimator both yield the estimator

$$\hat{R}(\hat{\eta}(Q)) = q^{-1}[|\hat{\eta} - \hat{\eta}(Q)|^2 + (2r(Q) - q)\hat{\sigma}^2] \tag{7}$$

for the risk (6) of $\hat{\eta}(Q)$ under the saturated model.

The *pooling* variance estimator $\hat{\sigma}^2_P$, potentially useful when n equals or is not much larger than q, is $\hat{\sigma}^2_P = [n - r(Q_L)]^{-1}|y - \hat{\eta}(Q_L)|^2$, where Q_L is a projection in $\mathcal{Q}$ with rank $r(Q_L)$. This biased estimator is consistent for σ^2 when $n - r(Q_L)$ tends to infinity and $[n - r(Q_L)]^{-1}|\eta - \eta(Q_L)|^2$ tends to zero.

The *first-difference* variance estimator $\hat{\sigma}^2_{FD}$ is potentially useful when $n = q$, in which case the data forms an incomplete multi-way layout with one observation per cell. Form all possible first differences of adjacent $\{y_{i1}\}$ along each coordinate direction. Then $\hat{\sigma}^2_{FD}$ is defined to be one-half of the average of the squared first differences. This biased estimator is consistent for σ^2 when q tends to infinity and the quantity obtained by replacing each y_{i1} in $\hat{\sigma}^2_{FD}$with m_i tends to zero.

Let $\mathcal{Q}$ denote a finite class of candidate projections Q, constructed as in the previous section, whose cardinality may depend on q. The *adaptive estimator* of η is defined to be the candidate estimator with smallest estimated risk:

$$\hat{\eta}_{\mathcal{Q}} = \hat{\eta}(\hat{Q}), \text{ where } \hat{Q} = \operatorname*{argmin}_{Q \in \mathcal{Q}} \hat{R}(\hat{\eta}(Q)). \tag{8}$$

Asymptotic analysis supports the claim that the risk of $\hat{\eta}_{\mathcal{Q}}$ approximately minimizes risk among all candidate estimators $\{\hat{\eta}(Q): Q \in \mathcal{Q}\}$.

Theorem 3 *Let $\#\mathcal{Q}$ be the number of projections in the class $\mathcal{Q}$. Suppose that $\lim_{q\to\infty} q^{-1/2}\#\mathcal{Q} = 0$ and $\lim_{q\to\infty} \#\mathcal{Q} \sup_{|\eta|^2 \le qa\sigma^2} \mathrm{E}|\hat{\sigma}^2 - \sigma^2| = 0$ for every $a > 0$ and $\sigma^2 > 0$. Then, for W equal to either $R(\hat{\eta}_{\mathcal{Q}}, \eta, \sigma^2)$ or $L(\hat{\eta}_{\mathcal{Q}}, \eta)$,*

$$\lim_{q\to\infty} \sup_{|\eta|^2 \le q\sigma^2 a} \mathrm{E}|W - \min_{Q \in \mathcal{Q}} R(\hat{\eta}(Q), \eta, \sigma^2)| = 0 \tag{9}$$

and

$$\lim_{q\to\infty} \sup_{|\eta|^2 \le q\sigma^2 a} \mathrm{E}|W - \hat{R}(\hat{\eta}_{\mathcal{Q}})| = 0. \tag{10}$$

Under the multiparametric asymptotics of this theorem, in which the number q of unknown means tends to infinity, the loss and corresponding risk of a candidate estimator converge to a common value. According to (9), the risk and loss of the adaptive estimator $\hat{\eta}_{\mathcal{Q}}$ converge to the minimum risk achievable over the class of candidate estimators $\{\hat{\eta}(Q): Q \in \mathcal{Q}\}$. Equation (10) asserts that the estimated risk of the adaptive estimator $\hat{\eta}_{\mathcal{Q}}$ is a trustworthy asymptotic approximation to both its actual risk and loss.

Proof idea. Matrices $M(Q)$ and M are each symmetric and idempotent and satisfy $MM(Q) = M(Q) = M(Q)M$. Evidently $\text{rank}(M) = \text{tr}[C(C'C)^{-1}C'] = q$. The matrix $M - M(Q)$ is therefore symmetric and idempotent; has rank $q - r(Q)$; and satisfies $M(Q)[M - M(Q)] = 0$. We have the spectral decompositions $M(Q) = U_Q U_Q'$ and $M - M(Q) = \bar{U}_Q \bar{U}_Q'$, where U_Q is $n \times r(Q)$, $\bar{U}_Q$ is $n \times (q - r(Q))$, $U_Q' U_Q = I_{r(Q)}$, $\bar{U}_Q' \bar{U}_Q = I_{q-r(Q)}$, and $U_Q' \bar{U}_Q = 0$. It follows that the $n \times q$ matrix $U = (U_Q | \bar{U}_Q)$ is orthogonal and that $M = U_Q U_Q' + \bar{U}_Q \bar{U}_Q' = UU'$.

Let $z = U'y$ and $\xi = U'\eta$. Under the saturated model, z has a $N(\xi, \sigma^2)$ distribution and $\eta = M\eta = U\xi$. Let f_Q denote the q-dimensional vector whose first $r(Q)$ components equal 1 and whose other components equal 0. Let $F_Q = \text{diag}\{f_Q\}$. By Theorem 1 and the preceding paragraph, $\hat{\eta}(Q) = U_Q U_Q' y = UF_Q z$ and $\hat{\eta} = UU'y = Uz$. For any vector h, let $\text{ave}(h)$ denote the average of the components of h. For any two vectors g and h of the same dimension, let gh denote the vector formed by componentwise multiplication. Equations (6), (7), and the preceding notations yield the canonical forms $L(\hat{\eta}(Q), \eta) = q^{-1}|f_Q z - \xi|^2$, $R(\hat{\eta}(Q), \eta, \sigma^2) = \text{ave}[f_Q^2 \sigma^2 + (1 - f_Q)^2 \xi^2]$, and $\hat{R}(\hat{\eta}(Q)) = \text{ave}[f_Q^2 \hat{\sigma}^2 + (1 - f_Q)^2 (z^2 - \hat{\sigma}^2)]$ for loss, risk, and estimated risk.

Let V denote either loss $L(\hat{\eta}(Q))$ or the estimated risk $\hat{R}(\hat{\eta}(Q))$. Applying Theorem 2.1 in Beran and Dümbgen (1998) to the preceding canonical forms establishes existence of a finite constant C such that

$$\text{E}|V - R(\hat{\eta}(Q), \eta, \sigma^2)| \le C[q^{-1/2}(\sigma^2 + \sigma\{\text{ave}(\xi)^2\}^{1/2}) + \text{E}|\hat{\sigma}^2 - \sigma^2|]. \quad (11)$$

for every projection $Q \in \mathcal{Q}$. An argument based on this inequality yields the Theorem conclusions.

4 Example: coal ash data

The coal ash data from Cressie (1993, p. 34) records percentage of coal ash in 208 assay samples. The data forms an incomplete two-way layout with one observation for each coordinate pair at which an assay sample is obtained. The factors row coordinate and column coordinate are both ordinal and range over $p_1 = 23$ and $p_2 = 16$ equally spaced levels respectively. It seems likely, a priori, that the coal ash means vary smoothly with geographical location. We construct candidate projections by the tensor product space method previously described, using the discrete cosine basis: For $1 \le k \le 2$, the first column of $W_k(c_k)$ is the vector u_k already defined and the succeeding columns are

$$w_{kc} = \{(2/p_k)^{1/2} \cos[(2r-1)(c-1)\pi/(2p_k)] \colon 1 \le r \le p_k\}, \quad 2 \le c \le c_k. \quad (12)$$

Let $\mathcal{A}_1 = \{\emptyset, \{1\}, \{2\}\}$ and let $\mathcal{A}_2 = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}$. Let $\mathcal{C} = \bigcup_{j=2}^{8} \{(j,j)\}$. For every $c \in \mathcal{C}$, define $Q_{\mathcal{A}_j}(c)$ by equation (5) and its preceding paragraph and let $\mathcal{Q} = \{Q_{\mathcal{A}_j}(c) \colon 1 \le j \le 2, c \in \mathcal{C}\}$. For the coal-ash data, the first difference variance estimate is $\hat{\sigma}^2_{FD} = 1.038$.

The following table lists estimated risks for the candidate estimators $\{\hat{\eta}(Q): Q \in \mathcal{Q}\}$:

c	(2,2)	(3,3)	(4,4)	(5,5)	(6,6)	(7,7)	(8,8)
$Q_{\mathcal{A}_1}(c)$	.213	.232	.150	.148	.134	.151	.155
$Q_{\mathcal{A}_2}(c)$	.222	.243	.192	.238	.261	.333	.385

The additive submodel fit $\hat{\eta}(Q_{\mathcal{A}_1}(6,6))$ has smallest estimated risk among these competitors. Moreover, its estimated risk .134 is about one-eighth of the estimated risk 1.038 of the least squares fit to the saturated model. The latter fit coincides with the raw data in this example and is clearly not useful. Through variance-bias trade-off, our data-driven algorithm selects a submodel fit that is visually appealing, has the structural simplicity of additivity, and has much lower risk under the saturated model (see the preceding theorem). This adaptive submodel fit compresses the data so as to discard more noise than signal.

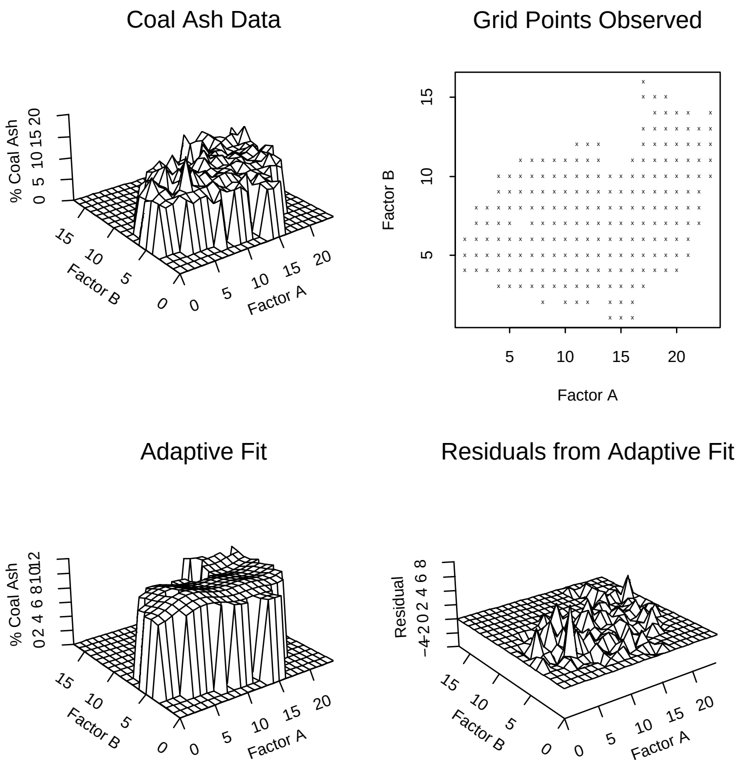

Figure 1. The Coal Ash data, its factor level grid, the low risk adaptive submodel fit $\hat{\eta}(Q_{\mathcal{A}_1}(6,6))$ using the discrete cosine basis, and residuals.

References

[1] Beran, R. (2000). *REACT scatterplot smoothers: superefficiency through basis economy*. Journal of the American Statistical Association **63**,155–171.

[2] Beran, R. (2002). *Improving penalized least squares through adaptive selection of penalty and shrinkage*. Annals of the Institute of Statistical Mathematics **54**, 900–917.

[3] Beran, R. and Dümbgen, L. (1998). *Modulation of estimators and confidence sets*. Annals of Statistics **26**, 1826–1856.

[4] Cressie, N. A. (1993). Statistics for Spatial Data (revised edition), Wiley, New York.

[5] Heckman, N. E. and Ramsay, J. O. (2000). *Penalized regression with model-based penalties*. Canadian Journal of Statistics **28**, 241–258.

[6] Lin, Yi (2000). *Tensor product space ANOVA fits*. Annals of Statistics **28**, 734–755.

[7] Mallows, C. L. (1973). *Some comments on C_p*. Technometrics **15**, 661–676.

[8] Stein, C. (1966). *An approach to the recovery of inter-block information in balanced incomplete block designs*. Festschrift for Jerzy Neyman (ed. F. N. David), Wiley, New York, 351–364.

[9] Stein, C. (1981). *Estimation of the mean of a multivariate normal distribution*. Annals of Statistics **9**, 1135–1151.

[10] Tukey, J. W. (1977). Exploratory Data Analysis, Addison-Wesley, Reading MA.

[11] Wahba, G. (1990). Spline Models for Observational Data, Society for Industrial and Applied Mathematics, Philadelphia.

[12] Wahba, G., Wang Y., Gu, C., Klein, R. and Klein, B. (1995). *Smoothing spline ANOVA for exponential families with application to the Wisconsin epidemiological study of diabetic retinopathy*. Annals of Statistics **23**, 1868–1895.

Acknowledgement: This research was supported in part by National Science Foundation Grant DMS 0300806

Address: Department of Statistics, University of California, Davis, CA 95616, USA

E-mail: `beran@wald.ucdavis.edu`