Adaptively Denoising Discrete Two-way Layouts

Rudolf Beran*

University of California, Davis

Abstract: The two-way layout with ordinal or nominal factors is a fundamental data-type that is widespread in the sciences, engineering, and informatics. The unrestricted least squares estimator for the means of a two-way layout is usually inadmissible under quadratic loss and the model of homoscedastic independent Gaussian errors. In statistical practice, this least squares estimator may be modified by fitting hierarchical submodels and, for ordinal factors, by fitting polynomial submodels. ASP, an acronym for Adapative Shrinkage on Penalty bases, is an estimation (or denoising) strategy that chooses among submodel fits and more general shrinkage or smoothing fits to two-way layouts without assuming that any submodel is true. ASP fits distinguish between ordinal and nominal factors; respect the hierarchical decomposition of means into overall mean, main effects, and interaction terms; and are designed to reduce risk substantially over the unrestricted least squares estimator. For the balanced complete two-way layout, these points are developed through multiparametric asymptotics, in which the number of factor-level pairs tends to infinity, and through numerical case studies.

1. Introduction

A fundamental data type in the sciences, engineering, and informatics is the discrete two-way layout. Instances include the data recorded in agricultural field trials, in DNA microassays, in digital imaging, and in other settings where regression or ANOVA are established tools. The factors in a two-way layout may be ordinal or nominal. The levels of an ordinal factor are real-values that indicate at least order and possibly more. The levels of a nominal factor are pure labels that convey no ordering information.

In devising trustworthy fits to discrete layouts, it is essential not to make strong unsupported assumptions concerning how the mean response at a grid point depends on the factor levels identifying that grid-point. On the other hand, the unrestricted least squares estimator tends to overfit the means to a two-way layout, especially when there is little replication. A formal theoretical statement of this difficulty is Stein’s (1956) inadmissibility result for the least squares fit to a one-way layout with independent identically distributed Gaussian errors. Better estimation (or denoising) techniques for two-way layouts with unrestricted means rely on biased estimators, such as those generated by the statistical regularization methods of this paper.

The acronym ASP stands for Adapative Shrinkage on Penalty bases. An ASP fit to a discrete two-way layout is constructed in three stages:

- Devise a candidate class of constrained penalized least squares (PLS) estimators whose three quadratic penalty terms express tentative notions about the
two main effects and the interactions in the means of the two-way layout.

- Estimate the risk of each candidate estimator under a general saturated model on the means that does not assume any of the prior notions in step one.
- Define the ASP fit to be a candidate fit that minimizes estimated risk under the general model (or a related criterion)—the adaptive aspect of the procedure.

The design of the three penalty terms in the first step above is a key contribution of this paper. Wood (2000) treated penalized least squares with multiple quadratic penalties. The present paper differs from his work in the construction of the multiple penalty terms to address the possible unimportance or smoothness of some main effects or interactions; in the use of estimated risk under a general model (with unrestricted means) rather than cross-validation to select penalty weights and terms; in treating bi-monotone shrinkage strategies more general than penalized least squares; and in articulating asymptotics for ASP estimators in two-way layouts.

An ASP fit is a biased estimator that trades bias against variance so as to achieve, approximately, the lowest quadratic risk attained over the class of candidate estimators for the means of the two-way layout. Multiparametric asymptotics, in which the total number of cells in the two-way layout tends to infinity, indicate that the estimated risk of an ASP estimator is a trustworthy approximation to its risk (Section 4). This asymptotic analysis relies on results, for abstract shrinkage estimators, that were developed by Beran and Dümbgen (1998) and were applied to one-way layouts by Beran (2000, 2002).

The smoothing spline literature seeks to estimate a mean function that is deemed to be a function of continuous ordinal factors, using observations typically made at a scattered set of factor levels (cf. Wahba (1990), Wahba, Wang, Gu, Klein and Klein (1995), Heckman and Ramsay (2000), Lin (2000)). This scattered set may be viewed as an incomplete subset of the smallest discrete grid that contains it. The discrete fitting techniques of this paper may be compared and contrasted with the spline literature as follows:

a) Our aim is to estimate well a possibly large discrete array of means rather than a smooth mean function. ASP estimators do not assume any smoothness or ANOVA submodel structure in the means. Their risk is evaluated under the saturated model that puts no restrictions on the unknown means. The ASP approach exploits the possibility that a fit which selectively downplays some interactions and some main effects may reduce risk through variance-bias tradeoff.

b) The factors affecting the discrete means in the two-way layout can be either ordinal or nominal or one of each. For both factors ordinal, ASP fits rely on discrete splines that are akin to continuous smoothing splines. For both factors nominal, ASP yields multiple-shrinkage estimators very close to those of Stein (1966).

Tukey (1977) proposed and experimented with certain smoothing algorithms for fitting one- and higher-way layouts with ordinal factors. In ordinal one-way layouts where wavelet bases provide a sparse representation of the means, Donoho and Johnstone (1995) used adaptive shrinkage through soft-thresholding. Beran and Dümbgen (1998) proposed and studied adaptive symmetric linear estimators that perform monotone shrinkage relative to a fixed orthonormal basis. ASP estimators for two-way layouts with either ordinal or nominal factors can be represented canonically as a closed set of bi-monotone shrinkage estimators acting on a tensor product basis determined by the three penalty terms (Section 2).
In discussing ASP fits, we distinguish strictly among data, statistical procedure, and probability model. The simple probability model used to motivate ASP methodology is a theoretical construct that seeks to approximate selected relative frequencies in the data. It is not believed that the probability model generated the data. It is expected that ASP fits reflect properties of the motivating probability model and are effective on data whose salient relative frequencies can be approximated in terms of the model. The quality of the approximation being an empirical matter, mathematical properties of ASP fits under the probability model are supplemented by numerical case studies and certain diagnostic plots.

For the purpose of developing ASP estimators and studying their theoretical properties, we consider the saturated Gaussian model for a complete balanced two-way layout. All risk calculations in this paper are made under the saturated model. The first factor has $p_1$ distinct levels; the second factor has $p_2$ distinct levels; and $q$ observations are taken at each combination of factor levels. Without loss of generality in the theory, we may take $q = 1$. This corresponds to using the averages over replications in place of the raw observations. Subscripting is arranged so that, for an ordinal factor, the factor levels are a strictly increasing function of subscript. This reduction yields the model

$$y_{ij} = \mu(s_{1i}, s_{2j}) + \epsilon_{ij} \quad 1 \leq i \leq p_1, 1 \leq j \leq p_2,$$

where the $\{y_{ij}\}$ are the (averaged) observations, the $\{s_{ki}\}$ are the levels of factor $k$, and the errors $\{\epsilon_{ij}\}$ are independent, identically distributed $N(0, \sigma^2)$ random variables. In the saturated model, both the function $\mu$ and the variance $\sigma^2$ are unknown. If factor $k$ is ordinal, then $s_{k1} < s_{k2} < \ldots < s_{kpk}$. For notational simplicity, we will usually write $m_{ij}$ instead of $\mu(s_{1i}, s_{2j})$.

Let $M$ denote the $p_1 \times p_2$ matrix with elements $\{m_{ij}\}$. The Frobenius matrix norm $| \cdot |$ is defined by $|C|^2 = \text{tr}(C'C) = \text{tr}(CC')$. We will assess any estimator $\hat{M}$ of $M$ through the normalized quadratic loss and corresponding risk

$$L(\hat{M}, M) = (p_1p_2)^{-1} |\hat{M} - M|^2, \quad R(\hat{M}, M, \sigma^2) = EL(\hat{M}, M).$$

The unrestricted least squares estimator of $M$ is the matrix $Y$ with elements $\{y_{ij}\}$ and has risk $\sigma^2$. This least squares estimator underlies classical analysis of variance for the two-way layout but is less useful in fitting response surfaces or analyzing an image. Indeed, Stein (1956) proved that it is inadmissible whenever $p_1p_2 \geq 3$.

### 1.1. Penalized least squares and submodels

Estimators of $M$ that may dominate least squares are suggested by the following class of penalized least squares estimators. For $k = 1$ or 2, define the $p_k \times 1$ unit vector $u_k = p_k^{-1/2}(1, 1, \ldots, 1)'$. Let $A_k$ be any matrix with $p_k$ columns such that $A_k u_k = 0$. Examples of such annihilator matrices that have additional useful properties are presented in Section 1.2 and are treated more fully in Section 3. Let $A = (A_1, A_2)$ and let $\nu = (\nu_1, \nu_2, \nu_{12})$ be any vector in $[0, \infty]^3$. The candidate penalized least squares (PLS) estimator of $M$ is defined to be

$$\hat{M}_{PLS}(\nu, A) = \arg\min_M S(M, \nu, A),$$

where

$$S(M, \nu, A) = |Y - M|^2 + \nu_1 |A_1 M u_1|^2 + \nu_2 |u_1' M A_2|^2 + \nu_{12} |A_1 M A_2|^2.$$


The three penalty terms in 1.4 are designed to measure departures in $M$ from certain submodels. The unrestricted model for the mean matrix of the two-way layout is

- **Full Model:** $M = \gamma_0 u_1' u_2' + \gamma_1 u_1' + \gamma_2 u_2' + \Gamma_{12}$, where $\gamma_0$ is a scalar, $\gamma_k$ is a $p_k \times 1$ vector such that $u_k' \gamma_k = 0$, and $\Gamma_{12}$ is a $p_1 \times p_2$ matrix such that $u_1' \Gamma_{12} = 0$, $\Gamma_{12} u_2 = 0$ (cf. Scheffé (1959)).

The vanishing of one or more penalty terms indicates when $M$ satisfies a designated submodel of the full model.

- **Additive Model:** This is the submodel of the full model for which $\Gamma_{12} = 0$. For this submodel, the penalty term $|A_1 M A_2'|^2$ vanishes.
- **Row-effects Model:** This is the submodel of the full model for which $\gamma_2 = 0$ and $\Gamma_{12} = 0$. For this submodel, the penalty terms $|A_1 M A_2'|^2$ and $|u_1' M A_2'|^2$ both vanish.
- **Column-effects Model:** This is the submodel of the full model for which $\gamma_1 = 0$ and $\Gamma_{12} = 0$. For this submodel, the penalty terms $|A_1 M A_2'|^2$ and $|A_1 M u_2|^2$ both vanish.
- **Constant Model:** This is the submodel of the full model for which each $\gamma_k = 0$ and $\Gamma_{12} = 0$. For this submodel, the penalty terms $|A_1 M A_2'|^2$, $|u_1' M A_2'|^2$, and $|A_1 M u_2|^2$ all vanish.

Thus, if $\nu_{12}$ is very large, the candidate PLS estimator will fit a nearly additive model. If $\nu_1$ and $\nu_{12}$ are both large, the PLS estimator will fit a nearly row-effects model. If $\nu_2$ and $\nu_{12}$ are both large, the candidate PLS estimator will fit a nearly column-effects model. Finally, if every component of $\nu$ is large, the candidate PLS estimator will fit a nearly constant model. Further properties of the PLS fit will depend on the precise choice of $A_k$ and will be discussed in Sections 2 and 3. The value of $\nu$ and hence the extent of submodel fitting will be chosen to minimize estimated risk of the candidate PLS estimator.

### 1.2. Examples of annihilators and fits

The following examples introduce suitable choices of the annihilator matrices $A_k$ and corresponding ASP fits to data.

**Example 1: Two ordinal factors.** Estimating a response surface or denoising an image deals with responses indexed by two ordinal factors. Vague prior information may suggest that the mean function $\mu(s_{1i}, s_{2j})$ behaves locally like a polynomial function of the factor levels. Suppose that each set of factor levels is equally spaced. To have the PLS estimator favor a fit that is locally polynomial of degree $r-1$ in the levels of the first factor and of degree $c-1$ in the levels of the second factor, we take $A_1$ and $A_2$ to be, respectively, the $r$-th and $c$-th difference operators of column dimensions $p_1$ and $p_2$ respectively. More explicitly, consider the $(p - 1) \times p$ matrix $\Delta(p) = \{\delta_{i,j}\}$ in which $\delta_{1,i} = 1$, $\delta_{i,i+1} = -1$ for every $i$ and all other entries are zero. Define

$$D_1(p) = \Delta(p), \quad D_d(p) = \Delta(p - d + 1) D_{d-1} \quad \text{for} \quad 2 \leq d \leq p - 1.$$

The annihilators just mentioned are $A_1 = D_c(p_1)$ and $A_2 = D_c(p_2)$ respectively.

Subplot (1,1) of Figure 1 displays a linearly interpolated $70 \times 50$ two-way layout with one observation per cell. The artificial data was obtained by adding pseudo-random Gaussian white noise to the values at the grid-points on the response surface.
Fig 1. ASP fit and diagnostics for the artificial data. Both factors are ordinal. Both annihilators are second difference.
Fig 2. Replotting of Figure 1 as greyscale images.
in subplot (2,1). Section 3.2 provides mathematical details for this example, in which the means are highly non-additive and have a sharp central dip. Both factors are ordinal. Subplot (3,1) gives a linearly interpolated adaptive PLS estimator that uses the second difference annihilator for each factor. This ASP estimator recovers major features of the true means far more clearly than the unrestricted least squares estimator, which coincides with the raw data in subplot (1,1). The fitting errors displayed in subplot (3,2) are the difference between the ASP estimator and the true mean matrix. The fitting errors appear homogeneously random except at the central dip. Sections 2, 3.1, and 3.2 develop this example by discussing adaptive choice of \( r \), \( c \) and the penalty weights and by presenting more general constructions of annihilator matrices.

Figure 2 replots Figure 1 as greyscale images. The subplots in the left column suggest that ASP fits have potential for denoising images. Indeed, a JPEG technique for denoising an image would apply hard-thresholding to the coefficients obtained by bivariate discrete cosine transform of the pixel values. Unless \( p_1 \) and \( p_2 \) are small, it may be verified that the basis provided by the bivariate discrete cosine transform essentially coincides with the penalty basis generated when both \( A_1 \) and \( A_2 \) are first difference operators.

**Example 2: Two nominal factors.** Classical analysis of variance deals with such data. When both factors are purely nominal, permutation of the subscripts (labels) should not affect the estimator of \( M \). The matrix

\[
A_k = I_{p_k} - u_k u_k' \tag{1.6}
\]

is an annihilator that is invariant under permutations of row and column labels. We call it the flat annihilator for reasons that will become clear in Section 3.3. When both \( A_k \) are flat annihilators, the corresponding candidate PLS estimators are equivariant under permutations of row and column labels.

Subplot (1,1) of Figure 3 displays a linearly interpolated \( 6 \times 8 \) two-way layout with one observation by cell. The data comes from p. 238 of Anderson and Bancroft (1952) and is reprinted on p. 138 of Scheffé (1959). Cell \((i, j)\) in the layout reports the amount of cooking fat number \( j \) that is absorbed in baking a batch of donuts on day \( i \) of the experiment. Both factors in this example are treated as nominal. Subplot (1,2) gives an interpolated adaptive PLS estimator that uses the flat annihilator (1.6) for each factor. The cross-sections in the second row of Figure 3 show how this ASP fit, unlike the unrestricted least squares fit, recovers near-additivity in the dependence of fat-absorption on the day and the oil used. Sections 2 and 3.3 develop this example.

**Example 3: One nominal and one ordinal factor.** Classical analysis of covariance deals with such data. If the first factor is nominal with equally spaced levels while the second factor is ordinal, we may take \( A_1 = I_{p_1} - u_1 u_1' \) and \( A_2 = D_c(p_2) \). The resulting candidate PLS estimators are equivariant under permutations of the levels of the first factor, shrink the least squares estimator for the main effects of the first factor, and favor a fit that is locally of degree \( c - 1 \) in the levels of the second factor.

Subplot (1,1) of Figure 4 displays a linearly interpolated \( 52 \times 3 \) two-way layout with one observation by cell. The data comes from Chatterjee et al. (1995). Cell \((i, j)\) in the layout reports the grape yield harvested in year \( j \) from row \( i \) of a vineyard with 52 rows. Vineyard row in this example is an ordinal factor. Harvest year is treated as a nominal factor because weather and viticulture can vary considerably from year to year. Subplot (1,2) gives an interpolated adaptive PLS estimator that
Fig 3. ASP fit and diagnostics for the data on fat-absorption by donuts. The factors day and fat number are both nominal. Both annihilators are flat.
Fig 4. ASP fit and diagnostics for the data on vineyard grape yields. The factor vineyard row is ordinal while the factor year is nominal. The annihilators are respectively third difference and flat.
uses the third difference annihilator for the ordinal factor and the flat annihilator for the nominal factor. The cross-sections in the second row of Figure 4 show how this ASP fit, more clearly than the unrestricted least squares fit, brings out leading features of the grape yield over the three harvest years. Among these is a dip in yield at and near vineyard row 33. Sections 2 and 3.4 develop this example.

1.3. Outline of the paper

Section 2 defines ASP estimators in several stages. Candidate PLS estimators are expressed in canonical form with respect to an orthogonal basis determined by the three penalty terms in (1.4). This representation suggests larger classes of candidate estimators that contain the candidate PLS estimators and have the mathematical advantage of forming closed convex sets. ASP estimators are then defined as estimators that minimize estimated risk over the class of candidate estimators being considered. A theorem develops conditions under which estimated risk is a trustworthy surrogate for the unknown risk as the number of cells $p_1p_2$ tends to infinity. It is shown that ASP estimators will greatly dominate unrestricted least squares estimators when the selected penalty basis is economical. Section 3 discusses important algorithmic aspects, including how to devise appropriate annihilator matrices that express vague prior information about $M$ and how to minimize the estimated risk when the factors are both ordinal or both nominal or mixed. In the case of two nominal factors, a simple closed form solution exists that essentially coincides with an estimator proposed by Stein (1966). Section 4 provides proofs of theorems stated in Sections 2 and 3.

The mathematical results of this paper may be extended to $k$-way layouts. The vectorized PLS criterion for the $k$-way layout, which generalizes (2.1), contains a separate penalty term for each main effect and for each interaction term that appears in the full model. The vectorized candidate PLS estimators then parallel those for two-way layouts: a $k$-fold Kronecker product replaces $U_2 \otimes U_1$ in (2.8) and a $k$-way array replaces the shrinkage matrix (2.9). Asymptotic analysis of adaptation for PLS and larger classes of candidate shrinkage estimators proceeds in analogy to the treatment for $k = 2$. The exposition in this paper is limited to the two-way layout for notational and computational simplicity. New issues that arise in computing and displaying ASP fits to three- or higher-way layouts will be treated elsewhere.

2. Defining ASP Estimators

We motivate the definition of ASP estimators by studying the form and risk of candidate PLS estimators in terms of a canonical penalty basis for the regression space.

2.1. Candidate PLS estimators and penalty bases

The PLS criterion may be vectorized as follows. Let $y = \text{vec}(Y) = \{(y_{ij}: 1 \leq i \leq p_1, 1 \leq j \leq p_2)\}$, the column vector obtained by sequentially stacking the columns of $Y$ with first column on top and last column at the bottom. Similarly, let $m = \text{vec}(M)$. Then, from (1.4),

$$S(M, \nu, A) = |y - m|^2 + \nu_1 m'(u_2u_2' \otimes A_1'A_1)m + \nu_2 m'(A_2'A_2 \otimes u_1u_1')m + \nu_{12} m'(A_2'A_2 \otimes A_1'A_1)m.$$  

(2.1)
Let \( \hat{m}_{PLS}(\nu, A) = \text{vec}(\hat{M}_{PLS}(\nu, A)) \), the right side being defined through (1.3). By calculus,
\[
\hat{m}_{PLS}(\nu, A) = [I_{p_1p_2} + \nu_1(u_2u'_2 \otimes A'_1A_1) + \nu_2(A'_2A_2 \otimes u_1u'_1) + \nu_{12}(A'_2A'_2 \otimes A'_2A'_1)]^{-1}y.
\]

(2.2)

This expression for the candidate PLS estimator can be simplified to reveal its essential structure. Suppose that the \( p_k \times p_k \) symmetric matrix \( A_k'A_k \) has the spectral decomposition \( A_k'A_k = U_k\Lambda_kU_k' \), where the eigenvector matrix satisfies \( U_kU'_k = U'_kU_k = I_{p_k} \) and the diagonal matrix \( \Lambda_k = \text{diag}\{\lambda_k\} \) gives the ordered eigenvalues \( 0 = \lambda_{k1} \leq \lambda_{k2} \leq \ldots \leq \lambda_{kp_k} \). This eigenvalue ordering, the reverse of the customary, is adopted here because the eigenvectors associated with the smallest eigenvalues play the greatest role in determining the numerical value and risk of the candidate PLS estimator. Because the annihilator \( A_k \) satisfies \( A_ku_k = 0 \), the eigenvalue \( \lambda_{k1} \) is necessarily zero and has \( u_k \) as corresponding eigenvector. Thus, the first column of \( U_k \) is \( u_k \). It follows from this discussion that
\[
(A'_2A_2 \otimes A'_1A_1)(U_2 \otimes U_1) = U_2A_2 \otimes U_1A_1 = (U_2 \otimes U_1)(A_2 \otimes A_1).
\]

(2.3)

Consequently,
\[
A'_2A_2 \otimes A'_1A_1 = (U_2 \otimes U_1)(A_2 \otimes A_1)(U_2 \otimes U_1)' = (U_2 \otimes U_1)(A_2 \otimes A_1)(U_2 \otimes U_1)'.
\]

(2.4)

gives a spectral decomposition of the symmetric matrix on the right side.

The \( p_k \times p_k \) matrix \( u_ku'_k \) is symmetric, idempotent, has eigenvalue 1 associated with the eigenvector \( u_k \), and has eigenvalue 0 repeated \( p_k - 1 \) times. Let \( E_k = \text{diag}\{e_{ki}\} \) denote the \( p_k \times p_k \) diagonal matrix that has 1 in the \((1, 1)\) cell and zeroes elsewhere. Because \( u_k \) is the first column of \( U_k \), we may write \( u_ku'_k = U_kE_kU'_k \), a spectral decomposition of the left-hand side. As in the preceding paragraph,
\[
u_2u'_2 \otimes A'_1A_1 = (U_2 \otimes U_1)(E_2 \otimes \Lambda_1)(U_2 \otimes U_1)',
\]
\[
A'_2A_2 \otimes u_1u'_1 = (U_2 \otimes U_1)(A_2 \otimes E_1)(U_2 \otimes U_1)'.
\]

(2.5)

Combining (2.2), (2.4) and (2.5) yields
\[
\hat{m}_{PLS}(\nu, A) = U[I_{p_1p_2} + \nu_1(E_2 \otimes \Lambda_1) + \nu_2(A_2 \otimes E_1) + \nu_{12}(A_2 \otimes \Lambda_1)]^{-1}U'y.
\]

(2.6)

for \( U = U_2 \otimes U_1 \). Let
\[
f_{ij}(\nu) = [1 + \nu_1\lambda_{1i}\epsilon_{2j} + \nu_2\epsilon_{1i}\lambda_{2j} + \nu_{12}\lambda_{1i}\lambda_{2j}]^{-1}.
\]

(2.7)

The matrix inverse in (2.6) is a diagonal matrix whose main diagonal is the vector \( f(\nu) = \{f_{ij}(\nu): 1 \leq i \leq p_1; 1 \leq j \leq p_2\} \). Let \( z = (U_2 \otimes U_1)'y \). Then
\[
\hat{m}_{PLS}(\nu, A) = (U_2 \otimes U_1) \text{diag}\{f(\nu)\} z.
\]

(2.8)

The columns of \( U_2 \otimes U_1 \) constitute the penalty basis generated by the annihilator matrices \( A_1 \) and \( A_2 \). Equation (2.8) shows that the PLS candidate estimator maps the data-vector \( y \) into its coefficient vector \( z \) with respect to the penalty basis, then shrinks \( z \) through componentwise multiplication by \( f(\nu) \), then maps the result back to the original basis.

To obtain a compact matrix expression for the candidate PLS estimator of \( M \), define the matrix \( F(\nu) = \{f_{ij}(\nu)\} \). Because \( \lambda_{k1} = 0 \) and \( \epsilon_{ki} = 0 \) if \( i \geq 2 \), expression
Equation (2.8) is equivalent to
\[ F(\nu) = \begin{pmatrix} 1 & (1 + \nu_1 \lambda_{12})^{-1} & \ldots & (1 + \nu_2 \lambda_{2p_2})^{-1} \\ (1 + \nu_1 \lambda_{11})^{-1} & (1 + \nu_1 \lambda_{12} \lambda_{22})^{-1} & \ldots & (1 + \nu_2 \lambda_{12} \lambda_{2p_2})^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ (1 + \nu_1 \lambda_{1p_1})^{-1} & (1 + \nu_1 \lambda_{1p_1} \lambda_{2p_2})^{-1} & \ldots & (1 + \nu_2 \lambda_{1p_1} \lambda_{2p_2})^{-1} \end{pmatrix}. \] (2.9)

Let \( Z = U'_1 Y U_2 \) and let \( F(\nu), Z \) denote the componentwise product of the two matrices. Equation (2.8) is equivalent to
\[ \tilde{M}_{PLS}(\nu, A) = U_1 [F(\nu), Z] U'_2. \] (2.10)

Note that the least squares estimator \( Y \) of \( M \) is the special case of (2.10) when every component of \( F(\nu) \) equals 1, or equivalently, when \( \nu_1 = \nu_2 = \nu_{12} = 0 \).

2.2. Candidate shrinkage estimators and their risks

A shrinkage class \( \mathcal{F} \) consists of \( p_1 \times p_2 \) matrices \( F = \{ f_{ij} \} \) such that \( 0 \leq f_{ij} \leq 1 \) for every \( i \) and \( j \). The associated candidate shrinkage estimator of \( M \) is defined to be
\[ \hat{M}(F, A) = U_1 [F, Z] U'_2. \] (2.11)

In this paper we will consider the following shrinkage classes, each of which is inspired by (2.9) and each of which generates a class of candidate shrinkage estimators for \( M \):

- The Unrestricted shrinkage class \( \mathcal{F}_U \) consists of all \( p_1 \times p_2 \) shrinkage matrices with elements in \([0, 1]\).
- The PLS shrinkage class \( \mathcal{F}_{PLS} \) is the subset of shrinkage matrices defined in (2.9) by \( \{ F(\nu): \nu \in [0, \infty]^3 \} \). The candidate PLS estimator \( \tilde{M}_{PLS}(\nu, A) \) described in (2.10) coincides with \( \hat{M}(F(\nu), A) \) in the notation (2.11).
- The Bi-Flat shrinkage class \( \mathcal{F}_{BF} \) is the subset of \( \mathcal{F}_U \) defined by
  \[ f_{ij} = \begin{cases} 1 & \text{if } i = j = 1 \\ c_1 & \text{if } j = 1, i \geq 2 \\ c_2 & \text{if } i = 1, j \geq 2 \\ c_{12} & \text{if } i \geq 2, j \geq 2, \end{cases} \] (2.12)
  where \( c_1, c_2 \) and \( c_{12} \) are any constants in \([0, 1]\). This is the specialization of PLS shrinkage obtained when \( \lambda_{ki} = 1 \) for \( i \geq 2 \) and every \( k \).
- The Submodel shrinkage class \( \mathcal{F}_{SM} \) is the subset of \( \mathcal{F}_{BF} \) in which the possible values of \( c_1, c_2, \) and \( c_{12} \) are restricted to either 0 or 1. This shrinkage class is suggested by classical techniques for choosing a hierarchical submodel.
- The Monotone Score shrinkage class \( \mathcal{F}_{MS} \) is the subset of \( \mathcal{F}_U \) defined by
  \[ f_{ij}(\lambda_{1i}, \lambda_{2j}) = \begin{cases} 1 & \text{if } i = j = 1 \\ g_1(\lambda_{1i}) & \text{if } j = 1, i \geq 2 \\ g_2(\lambda_{2j}) & \text{if } i = 1, j \geq 2 \\ g_{12}(\lambda_{11}, \lambda_{22}) & \text{if } i \geq 2, j \geq 2, \end{cases} \] (2.13)
  where \( g_1, g_2, g_{12} \) are any functions nonincreasing in their arguments.
• The Bi-Monotone shrinkage class \( F_{BM} \) is the subset of \( F_U \) defined by \( f_{11} = 1 \); \( \{f_{ij} : i \geq 2\} \) is nonincreasing in \( i \); \( \{f_{ij} : j \geq 2\} \) is nonincreasing in \( j \); and \( \{f_{ij} : i, j \geq 2\} \) is nonincreasing in \( i \) for each fixed \( j \) and nonincreasing in \( j \) for each fixed \( i \).

• The Bi-Nested shrinkage class \( F_{BN} \) is the subset of \( F_{BM} \) such that: \( f_{11} = 1 \); each \( f_{ii} \) is either \( c_1 \) or 0 for \( i \geq 2 \); each \( f_{ij} \) is either \( c_2 \) or 0 for \( j \geq 2 \); and each \( f_{ij} \) is either \( c_{12} \) or 0 for \( i \geq 2 \) and \( j \geq 2 \). Here \( c_1, c_2 \) and \( c_{12} \) are any constants in \([0, 1]\).

• The Flat \( \times \) Monotone shrinkage class \( F_{FM} \) is the subset of \( F_U \) defined by

\[
\begin{align*}
f_{ij} &= \begin{cases} 
1 & \text{if } i = j = 1 \\
c & \text{if } j = 1, i \geq 2 \\
g_{ij} & \text{if } i = 1, j \geq 2 \\
h_{ij} & \text{if } i \geq 2, j \geq 2,
\end{cases}
\end{align*}
\]  

(2.14)

where \( c \) is any constant in \([0, 1]\) and \( \{g_{ij}\}, \{h_{ij}\} \) are each any nonincreasing sequence.

Evidently, \( F_{SM} \subset F_{BF} \subset F_{PLS} \subset F_{MS} \subset F_{BM} \). Also \( F_{BF} \subset F_{BN} \subset F_{BM} \) and \( F_{FM} \subset F_{BM} \). With the exception of \( F_{PLS} \) (in general), \( F_{SM} \) and \( F_{BN} \), these shrinkage classes are closed convex subsets of \( F_U \). PLS, monotone score, bi-nested and bi-monotone shrinkage are useful when both factors are ordinal. Bi-flat shrinkage, a specialization of PLS, is useful when both factors are nominal. PLS and Flat \( \times \) monotone shrinkage are useful when the row factor is nominal while the column factor is ordinal. However, the unrestricted shrinkage class \( F_U \) does not generate low risk ASP estimators. These matters will be developed in the remainder of the paper.

Let \( f = \text{vec}(F) \). The risk of the candidate estimator \( \hat{M}(F, A) \) defined in (2.1) may be expressed simply through the penalty basis representation of \( \hat{m}(f, A) = \text{vec}(\hat{M}(F, A)) \). Let \( \xi = E(z) = (U_2 \otimes U_1)'m \) and \( \hat{\xi}(f) = \text{diag}\{f\}z \). Then

\[
\hat{m}(f, A) = (U_2 \otimes U_1)\hat{\xi}(f), \quad m = (U_2 \otimes U_1)\xi.
\]  

(2.15)

The normalized quadratic loss (1.2) thus reduces to

\[
L(\hat{M}(F, A), M) = (p_1p_2)^{-1}|\hat{m}(f, A) - m|^2 = (p_1p_2)^{-1}|\hat{\xi}(f) - \xi|^2.
\]  

(2.16)

For any vector \( x \), let \( \text{ave}(x) \) denote the average of its components. From (2.16), the risk of candidate shrinkage estimator \( \hat{M}(F, A) \) is

\[
R(\hat{M}(F, A), M, \sigma^2) = r(f, A, \xi^2, \sigma^2),
\]  

(2.17)

where

\[
r(f, A, \xi^2, \sigma^2) = \text{ave}[f^2\sigma^2 + (1 - f)^2\xi^2].
\]  

(2.18)

Multiplication of vectors on the right side of (2.18) is done componentwise, as in the S language.

### 2.3. Estimated risks and ASP estimators

If the risk function (2.18) were known, we would seek an oracle estimator of \( M \)—the candidate estimator that minimizes risk over the class of shrinkage vectors and the
class of annihilator matrices under consideration. Of course, the oracular strategy is usually unavailable. Instead, we will estimate the risk function from the data, then choose the candidate estimator that minimizes estimated risk. The result is called an ASP estimator of $M$. A basic theoretical question is to provide conditions on the shrinkage class and on the class of annihilator matrices such that the risk of an ASP estimator converges asymptotically to the risk of the pertinent oracle estimator. Though one may not believe that the probability model fully explains the data, such theoretical study throws light on the intellectual coherence, potential benefits, and potential limitations of ASP estimators.

The risk function (2.18) contains two quantities, $\sigma^2$ and $\xi^2$, that are usually unknown. The sampling scheme and the ordinal or nominal character of the factors both influence methods for estimating $\sigma^2$. Basic possibilities include:

- **Replicated layout.** Fundamental in this setting is the least squares estimator of $\sigma^2$, the normalized residual sum of squares in the ANOVA table for the two-way layout.
- **One observation per combination of factor levels.** If the penalty basis is economical in the sense that the coefficients $\{\xi_{ij}: q_1 < i \leq p_1, q_2 < j \leq p_2\}$ are close to zero, then the high-component estimator is

$$
\hat{\sigma}^2 = \frac{[(p_1 - q_1)(p_2 - q_2)]^{-1}}{\sum_{i>q_1} \sum_{j>q_2} \xi_{ij}^2}.
$$

The classical pooled interaction estimator of ANOVA, suitable when the means approximately follow the additive model described in the Introduction, is equivalent to (2.19) with $q_1 = q_2 = 1$.

For the variance estimator (2.19), $E(\hat{\sigma}^2 - \sigma^2)^2$ converges to zero if and only if $(p_1 - q_1)(p_2 - q_2)$ tends to infinity and the sum of squared biases $[(p_1 - q_1)(p_2 - q_2)]^{-1} \sum_{i>q_1} \sum_{j>q_2} \xi_{ij}^2$ tends to zero as $p_1p_2$ tends to infinity. When the number of replications is greater than one but not large enough to make the least squares estimator of variance accurate, it may be useful to combine it with a pooled interaction estimator. Robust analogs of these variance estimators are the medians of the respective sets of $\{z_{ij}\}$ divided by $\Phi^{-1}(0.75)$. Here $\Phi^{-1}$ denotes the quantile function of the standard normal distribution.

Having devised a variance estimator $\hat{\sigma}^2$, we may estimate $\xi^2$ by $z^2 - \hat{\sigma}^2$ and hence the risk function $r(f, A, \xi^2, \sigma^2)$ by

$$
r(f, A) = \text{ave}[\hat{\sigma}^2 f^2 + (1 - f)^2(z^2 - \hat{\sigma}^2)] = \text{ave}[(f - \hat{g})^2 z^2] + \hat{\sigma}^2 \text{ave}(\hat{g}),
$$

where $\hat{g} = (z^2 - \hat{\sigma}^2)/z^2$. Apart from considerations entering into the estimation of $\sigma^2$, this equation is an application of the Stein (1981) unbiased estimator of risk or of the risk estimator that underlies Mallow’s (1973) discussion of $C_P$.

For fixed annihilator pair $A$ and shrinkage class $F$, the shrinkage-adaptive estimator is defined to be $M(\hat{F}, A)$, where

$$
\hat{f} = \text{vec}(\hat{F}) = \arg \min_{f \in F} r(f, A) = \arg \min_{f \in F} \text{ave}[(f - \hat{g})^2 z^2].
$$

Computation of $\hat{F}$ is a weighted least squares problem that will be discussed further in Section 3 because the details depend upon the shrinkage class. When clarity requires, we will add a subscript to $\hat{F}$ to indicate the shrinkage class being used.

In general, ASP estimators involve adaptation over annihilators as well as over the shrinkage vector. Let $A$ be a class of of annihilator pairs. The ASP estimator
of \( M \) determined by annihilator class \( \mathcal{A} \) and shrinkage class \( \mathcal{F} \) is defined to be \( \hat{M}(\hat{f}, \hat{A}) \), where

\[
(\hat{f}, \hat{A}) = \arg\min_{A \in \mathcal{A}, \hat{f} \in \mathcal{F}} \hat{r}(f, A).
\] (2.22)

The following theorem gives conditions under which shrinkage adaptation to minimize estimated risk approximately minimizes true risk as \( p_1 p_2 \) tends to infinity. Section 4 gives the proof, which draws on abstract results for shrinkage estimators established by Beran and Dümbgen (1998).

**Theorem 2.1.** Fix the annihilator pair \( A \) and let \( \mathcal{F} \) be a subset of \( \mathcal{F}_{BM} \) that is closed in \([0,1]^{p_1 p_2}\). In particular, \( \mathcal{F} \) can be any shrinkage class listed in Section 2.2 other than \( \mathcal{F}_U \). Suppose that \( \hat{\sigma}^2 \) is consistent in that, for every \( a > 0 \) and \( \sigma^2 > 0 \),

\[
\lim_{p_1 p_2 \to \infty} \sup_{\text{ave}(\xi^2) \leq \sigma^2} \mathbb{E}|\hat{\sigma}^2 - \sigma^2| = 0.
\] (2.23)

a) Let \( V(f) \) denote either the loss \( L(\hat{M}(F,A), M) \) or the estimated risk \( \hat{r}(f, A) \). Then for every annihilator pair \( A \), every \( t > 0 \), and every \( \sigma^2 > 0 \),

\[
\lim_{p_1 p_2 \to \infty} \sup_{\text{ave}(\xi^2) \leq \sigma^2} \mathbb{E} \sup_{f \in \mathcal{F}} |V(f) - R(\hat{M}(F,A), M, \sigma^2)| = 0.
\] (2.24)

b) If \( \hat{f} = \text{vec}(\hat{F}) = \arg\min_{f \in \mathcal{F}} \hat{r}(f, A) \), then

\[
\lim_{p_1 p_2 \to \infty} \sup_{\text{ave}(\xi^2) \leq \sigma^2} |R(\hat{M}(\hat{F},A), M, \sigma^2) - \min_{f \in \mathcal{F}} R(\hat{M}(F,A), M, \sigma^2)| = 0.
\] (2.25)

c) For \( W \) equal to either \( L(\hat{M}(\hat{F},A), M) \) or \( R(\hat{M}(\hat{F},A), M, \sigma^2) \),

\[
\lim_{p_1 p_2 \to \infty} \sup_{\text{ave}(\xi^2) \leq \sigma^2} \mathbb{E}|\hat{r}(\hat{f}, A) - W| = 0.
\] (2.26)

d) Let \( \#(\mathcal{A}) \) denote the cardinality of the class \( \mathcal{A} \). If \( \#(\mathcal{A}) \cdot \min\{p_1^{-1/2}, p_2^{-1/2}\} \) and \( \#(\mathcal{A}) \cdot \mathbb{E}|\hat{\sigma}^2 - \sigma^2| \) both converge to zero as \( p_1 p_2 \to \infty \), then convergences (2.24) to (2.26) hold for the ASP estimator \( \hat{M}(\hat{F}, \hat{A}) \), defined in (2.22).

Because \( \max\{p_1, p_2\} \leq p_1 p_2 \leq [\max\{p_1, p_2\}]^2 \), the condition \( p_1 p_2 \to \infty \) is equivalent to \( \max\{p_1, p_2\} \to \infty \). By part a, the loss, risk, and estimated risk of a candidate estimator converge together asymptotically. Uniformity of this convergence over the shrinkage class \( \mathcal{F} \) makes the estimated risk of a candidate estimator a trustworthy surrogate for its true risk or loss. By part b, the risk of the shrinkage adaptive estimator \( \hat{M}(\hat{F}, A) \) converges to that of the best candidate estimator. Thus, when \( \mathcal{F} \) is a closed subset of \( \mathcal{F}_{BM} \), shrinkage adaptation works as intended. This covers every shrinkage class defined in Section 2.2 except \( \mathcal{F}_U \). Moreover, because the unrestricted least squares estimator is one of the candidate estimators indexed by these shrinkage classes, its asymptotic risk is at least as large as that of the best-shrinkage adaptive estimator. In practice, the risk of the best shrinkage-adaptive estimator is often much smaller than that of the unrestricted least squares estimator and this is the point. Part c shows that the loss, risk, and plug-in estimated risk of an adaptive estimator converge together asymptotically. Part d preserves these conclusions for ASP estimators in which the cardinality of the annihilator class is finite or slowly growing in the sense described.

The pleasant properties stated in Theorem 2.1 break down when the shrinkage class is \( \mathcal{F}_U \). Then the estimator \( \hat{M}(\hat{F}, A) \) is dominated by the least squares estimator \( Y \) (see Beran and Dümbgen (1998), p. 1829). Adaptation works when the shrinkage class is not too large, in a sense made precise in Section 4.
3. Annihilators and Algorithms

This section treats methods for constructing annihilator matrices and algorithms for minimizing estimated risk so as to construct ASP estimators. Case studies illustrate what ASP estimators can accomplish on data.

3.1. Role of basis economy

The following discussion motivates techniques for selecting annihilator matrices. Let \( \Xi = \{ \xi_{ij} \} = U_1' M U_2 \), so that \( \xi = \text{vec}(\Xi) \). Heuristically, a penalty basis is economical if all components outside the upper left corner of \( \Xi \) are close to zero. In that case, we need only to identify and estimate from the data the relatively few non-zero components of \( \xi \), estimating the remaining components by zero. The quadratic risk then accumulates small squared biases from ignoring the nearly zero components of \( \xi \) but does not accumulate the many variance terms that would arise in attempting to estimate these unbiasedly.

An idealized formulation of basis economy facilitates mathematical analysis of how economy affects estimation risk in the two-way layout. Let \( S \) denote the set of all subsets of \( \{(i, j) : 1 \leq i \leq p_1, 1 \leq j \leq p_2\} \). For given subset \( S \in S \), let \( F(S) = \{ f_{ij}(S) \} \) where \( f_{ij}(S) = 1 \) or \( 0 \) according to whether or not \( (i, j) \in S \).

Define \( S_0 = \{ S \in S : F(S) \in F_{BM} \} \).

(3.1)

For every \( S \in S_0 \), every \( a > 0 \), and every \( \sigma^2 > 0 \), consider the projected ball \( B(a, S, \sigma^2) = \{ \xi \in R^{p_1 p_2} : \text{ave}(\xi^2) \leq \sigma^2 a \text{ and } \xi_{ij} = 0 \text{ for } (i, j) / \in S \} \).

(3.2)

Formally, we will say that the penalty basis associated with the annihilator pair \( A \) is economical if \( \xi \in B(a, S, \sigma^2) \) for some finite \( a > 0 \) and \( \#(S) \), the cardinality of \( S \), is small relative to \( p_1 p_2 \). Though this formulation is too simple to serve as a complete definition of basis economy, it yields the following quantitative result that shows how basis economy affects the risk of estimators of \( M \).

**Theorem 3.1.** Suppose that \( S \in S_0 \) and

\[
\lim_{p_1 p_2 \to \infty} (p_1 p_2)^{-1} \#(S) = b.
\]

(3.3)

Then, for every \( a > 0 \) and every \( \sigma^2 > 0 \), the asymptotic minimax quadratic risk over all estimators of \( M \) is

\[
\lim \inf_{p_1 p_2 \to \infty} \inf_{\xi \in B(a, S, \sigma^2)} \sup_{\hat{M}} R(\hat{M}, M, \sigma^2) = \sigma^2 \frac{ab}{a+b}.
\]

(3.4)

The bi-monotone shrinkage-adaptive estimator \( \hat{M}(\hat{F}_{BM}, A) \) satisfies

\[
\lim_{p_1 p_2 \to \infty} \sup_{\xi \in B(a, S, \sigma^2)} R(\hat{M}(\hat{F}_{BM}, A), M, \sigma^2) = \sigma^2 \frac{ab}{a+b}.
\]

(3.5)

The same holds for the bi-nested shrinkage-adaptive estimator \( \hat{M}(\hat{F}_{BN}, A) \).

This theorem reveals substantially more than formal asymptotic minimaxity of the bi-monotone shrinkage-adaptive estimator \( \hat{M}(\hat{F}_{BM}, A) \). When \( b \in [0, 1] \) is close to zero—in which case the penalty basis is highly economical—the right side of (3.5) is much smaller than the risk \( \sigma^2 \) of the unrestricted least squares estimator of \( M \). To the extent that the PLS and other shrinkage-adaptive estimators defined in Section 2 approximate \( \hat{M}(\hat{F}_{BM}, A) \), their performance also benefits strongly from economy of the basis.
3.2. Both factors ordinal

The ideal choice of penalty basis $U_2 \otimes U_1$ would have its first basis vector proportional to the unknown mean vector $m$ so that only the first component of $\xi$ would be nonzero. Though unrealizable, this ideal selection suggests that prior information or conjecture about $m$ should be exploited in devising the annihilator matrices $A_k$ that generate the penalty basis. The following discussion relates prior notions about the local behavior of the mean function $\mu$ in (1.1) to constructions of $A_1$ and $A_2$ for two ordinal factors.

Let $t = (t_1, t_2, \ldots, t_p)$ denote the levels of an ordinal factor, where $p$ may be either $p_1$ or $p_2$. Let $g_0, g_1, \ldots, g_{d-1}$ be a given set of real-valued functions defined on the real line such that $g_0 \equiv 1$. We will construct a sparse matrix $B_d = B_d(t, p)$ to annihilate functions that behave locally like a linear combination of the $\{g_h : 0 \leq h \leq d-1\}$. For each $i$ such that $1 \leq i \leq p - d$, let $\mathcal{G}_i$ denote the subspace of $R^{d+1}$ that is spanned by the $d$ vectors $\{(g_h(t_i), \ldots, g_h(t_{i+d})) : 0 \leq h \leq d - 1\}$. Assume that the dimension of $\mathcal{G}_i$ is $d$. This condition is satisfied, for instance, when $g_h(t_i) = t_i^h$. Define the $(p - d) \times p$ local annihilator matrix $B_d = \{b_{ij}\}$ as follows: In the $i$-th row of $B_d$, the subvector $\{b_{ij} : i \leq j \leq i + d\}$ is the unit vector in $R^{d+1}$, unique up to sign, that is orthogonal to $\mathcal{G}_i$. The remaining elements of $B_d$ are zero.

**Theorem 3.2.** Let $\bar{g}_h = (g_h(t_1), g_h(t_2), \ldots, g_h(t_p))'$. Each row vector of the local annihilator matrix $B_d$ has unit length and

$$B_d \bar{g}_h = 0 \quad \text{for } 0 \leq h \leq d - 1. \quad (3.6)$$

**Proof.** The definition of $B_d$ ensures that its rows have unit length and that

$$\sum_{j=1}^{p} b_{ij} g_h(t_j) = \sum_{j=i}^{i+d} b_{ij} g_h(t_j) = 0 \quad \text{for } 0 \leq h \leq d - 1. \quad (3.7)$$

Of frequent utility is the local polynomial annihilator, which is obtained by setting $g_h(t_i) = t_i^h$ in the foregoing definition of $B_d$. If we conjecture that the unknown mean function $\mu(s_{11}, s_{21})$ behaves locally like a polynomial of degree $r - 1$ in the first ordinal factor and like a polynomial of degree $c - 1$ in the second ordinal factor, we would take the annihilators that generate the penalty basis to be

$$A_1 = B_r(s_{11}, p_1), \quad A_2 = B_c(s_{21}, p_2), \quad (3.8)$$

where $s_1 = (s_{11}, \ldots, s_{1p_1})'$ and $s_2 = (s_{12}, \ldots, s_{1p_2})'$. When the factor levels are equally spaced, the local polynomial annihilator $B_d$ becomes a scalar multiple of the $d$-th difference matrix defined in display (1.5) of Example 1.

We turn next to the computation of shrinkage-adaptive estimators for the case of two ordinal factors once the annihilators $A$ have been fixed.

**Penalized least squares.** From Section 2.2 and definition (2.21), the shrinkage-adaptive PLS estimator is

$$\hat{M}(F(\hat{\nu}), A) = U_1[F(\hat{\nu})Z]'U_2', \quad (3.9)$$

where

$$\hat{\nu} = \text{argmin}_{\nu \in [0, \infty]^3} \text{ave}[(f(\nu) - \hat{g})^2 z^2]. \quad (3.10)$$
Let \( \hat{g}_{ij} = (z_{ij}^2 - \hat{\sigma}^2)/z_{ij}^2 \). Because of (2.9), equation (3.10) is equivalent to

\[
\hat{\nu}_1 = \arg\min_{\nu_1 \in [0, \infty]} \sum_{i=2}^{p_1} [(1 + \nu_1 \lambda_{1i})^{-1} - \hat{g}_{1i}]^2 z_{1i}^2,
\]

\[
\hat{\nu}_2 = \arg\min_{\nu_2 \in [0, \infty]} \sum_{j=2}^{p_2} [(1 + \nu_2 \lambda_{2j})^{-1} - \hat{g}_{1j}]^2 z_{1j}^2,
\]

\[
\hat{\nu}_{12} = \arg\min_{\nu_{12} \in [0, \infty]} \sum_{i=2}^{p_1} \sum_{j=2}^{p_2} [(1 + \nu_{12} \lambda_{1i} \lambda_{2j})^{-1} - \hat{g}_{1j}]^2 z_{1j}^2.
\]  

(3.11)

Calculation of \( \hat{\nu} = (\hat{\nu}_1, \hat{\nu}_2, \hat{\nu}_{12}) \) thus amounts to solving three nonlinear, weighted least squares problems, each of which can be treated with minimization algorithms for a function of a single variable.

Example 1 (continued). The data matrix for this example, displayed in Figures 1 and 2, is constructed as \( Y = M + E \) with \( p_1 = 70 \) and \( p_2 = 50 \). The components of the error matrix \( E \) are pseudo-random independent Gaussian with mean 0 and standard deviation \( \sigma = .15 \). The mean matrix \( M \) has components \( m_{ij} = \mu [i - (p_1 + 1)/2, j - (p_2 + 1)/2] \) for \( 1 \leq i \leq p_1, 1 \leq j \leq p_2 \), where \( \mu(u, v) = 2 t^{-3/4} \sin(t) \) with \( t = \sqrt{u^2 + v^2} \). The penalty basis is generated by using a second difference annihilator for each factor.

Subplot (1,2) in both figures explores the economy of this basis empirically by plotting the transformed components \( \{|z_{ij}|^{1/2}\} \) of Z as surrogates for the likewise transformed components of \( \Xi \). The transformation reduces the vertical range and makes it easier to see what is happening when \( z_{ij} \) is close to zero. The components outside the upper right corner of the matrix Z are relatively small, supporting the conclusion that the chosen penalty basis is economical here. The nature of this economy motivates using variance estimator (2.19) with \( q_1 = 29 \) and \( q_2 = 24 \). The estimated risk of the least squares estimator is then .0220, in good agreement with the actual risk .15. The estimated risk .0084 of the shrinkage-adaptive PLS estimator—the ASP estimator—is much smaller than that of the unrestricted least squares estimator. In this example, reduction of estimated risk accompanies visually better recovery of the response surface or image. The actual loss incurred by the ASP estimator is .0082, in approximate agreement with the estimated risk. This is to be expected from part c of Theorem 2.1.

The shrinkage matrix that defines the adaptive PLS estimator is shown in subplot (2,2). Shrinkage of the interaction coefficients \( \{z_{ij} : i \geq 2, j \geq 2\} \) increases pronouncedly with \( i \) and \( j \), though is less dramatic than the shrinkage of the main-effect coefficients \( \{z_{1j} : j \geq 2\} \) and \( \{z_{ij} : j \geq 2\} \). The shrinkage matrix reflects the non-additivity of the true means in this example. Experimenting with \( d \)-th difference annihilators of orders 1 through 4 did not reduce estimated risk below that achieved with second differences.

Bi-monotone shrinkage. The shrinkage-adaptive BM estimator is

\[
\hat{M}(\hat{F}_{BM}, A) = U_1 [\hat{F}_{BM}.Z] U'_2,
\]  

(3.12)

where vec(\( \hat{F}_{BM} \)) = \( \hat{f}_{BM} \) and

\[
\hat{f}_{BM} = \arg\min_{f \in F_{BM}} \text{ave}[(f - \hat{g})^2 z^2].
\]  

(3.13)
Consider the generic decomposition

\[ \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} a_{i,j}^2 = a_{11}^2 + \sum_{i=2}^{p_1} a_{i,1}^2 + \sum_{j=2}^{p_2} a_{1,j}^2 + \sum_{i=2}^{p_1} \sum_{j=2}^{p_2} a_{i,j}^2. \]  

(3.14)

To evaluate (3.13), we may proceed as follows:

a) Decompose the left side of (3.13) into minimizations of three separate sums formed as in (3.14). Minimize each of these sums without the constraint that \( f \in [0,1]^p \). Weighted isotonic regression with the pool adjacent violators (PAV) algorithm accomplishes this for the first two sums. Iterative use of the PAV algorithm handles the third sum. Roberston, Wright and Dykstra (1988) describe both algorithms. Bril et al. (1984) provide a Fortran implementation of the latter. Burdakov et al. (2004) review more efficient algorithms for isotonic regression in several variables.

b) Then each component of \( \hat{f}_{BM} \) is the positive part of the unconstrained minimizer found in part a. An extension of the argument in Section 5 of Beran and D"umbgen (1998) establishes this point.

Bi-nested shrinkage. The components of \( \hat{f}_{BN} = \arg\min_{f \in \mathcal{F}_{BN}} \text{ave}[(f - \hat{g})^2 z^2] \) are given by \( \hat{f}_{BN,ij} = 1(\hat{f}_{BM,ij} \geq 1/2) \) or may be found directly by finite search.

Monotone score shrinkage. The shrinkage-adaptive MS estimator is

\[ \hat{M}(\hat{F}_{MS}, A) = U_1 [\hat{F}_{MS}, Z] U'_2, \]  

(3.15)

where \( \text{vec}(\hat{F}_{MS}) = \hat{f}_{MS} \) and

\[ \hat{f}_{MS} = \arg\min_{f \in \mathcal{F}_{MS}} \text{ave}[(f - \hat{g})^2 z^2]. \]  

(3.16)

The components \( \{\hat{f}_{ij}\} \) of the matrix \( \hat{F}_{MS} \) may be found as follows:

First step. Set \( \hat{f}_{11} = 1 \).

Second step. Let \( w = \{z_{1i} : 2 \leq i \leq p_1\} \) and let \( \hat{h} = (w^2 - \hat{\sigma}^2)/w^2 \). Let \( \mathcal{K} = \{k \in R^q : k_1 \geq k_2 \geq \ldots \geq k_q\} \), where \( q = p_1 - 1 \). Find \( \hat{k} = \arg\min_{k \in \mathcal{K}} \text{ave}[(k - \hat{h})^2 w^2] \), using an algorithm for weighted isotonic least squares such as the PAV. Set \( \hat{f}_{1i} = \max\{\hat{k}_{i-1}, 0\} \) for \( 2 \leq i \leq p_1 \).

Third step. Repeat the second step, letting \( w = \{z_{1j} : 2 \leq j \leq p_2\} \) and \( q = p_2 - 1 \). Having found \( \hat{k} \), set \( \hat{f}_{ij} = \max\{\hat{k}_{j-1}, 0\} \) for \( 2 \leq j \leq p_2 \).

Fourth step. Let \( y = \text{vec}(\{z_{ij} : 2 \leq i \leq p_1, 2 \leq j \leq p_2\}) \), let \( q = (p_1 - 1)(p_2 - 1) \), and let \( v = \text{vec}(\{\lambda_1, \lambda_2 : 2 \leq i \leq p_1, 2 \leq j \leq p_2\}) \) be the vector of corresponding scores. Suppose first that these scores contain no ties. Let \( \rho \) denote the rank vector of \( v \) and define the \( q \) dimensional vector \( w \) through \( w_{p_1} = y_1 \). Repeat the second step using these definitions of \( w \) and \( q \). Having found \( \hat{k} \), define the vector \( \hat{n} \) to have \( i \)-th component \( \max\{k_{p_i}, 0\} \). Let \( \hat{N} = \{\hat{n}_{ij}\} \) be the \( (p_1 - 1) \times (p_2 - 1) \) matrix such that \( \hat{n} = \text{vec}(\hat{N}) \). Set \( \hat{f}_{ij} = \hat{n}_{i-1,j-1} \) for \( 2 \leq i \leq p_1, 2 \leq j \leq p_2 \). In the presence of ties among the components of \( v \), we pool the corresponding components of \( y^2 \) in constructing \( w^2 \) and reduce \( q \) accordingly.

Monotone score shrinkage, a special case of bi-monotone shrinkage, has the computational advantage that the PAV algorithm converges in a finite number of steps.
3.3. Both factors nominal

As was noted after (1.6), the flat annihilator \( A_k = I_{p_k} - u_k u_k' \) is invariant under permutations of row and column labels. This makes \( A_1 \) and \( A_2 \) suitable for defining candidate PLS estimators when both factors are nominal. Let \( U_k \) denote any orthogonal matrix whose first column is the vector \( u_k \). We may write

\[
U_k = (u_k, C_k),
\]

(3.17)

where \( u_k' C_k = 0 \) and \( C_k' C_k = I_{p_k-1} \). The columns of \( C_k \) are any set of orthonormal contrasts in \( R^{p_k} \). The matrix \( A_k \) is symmetric and idempotent. The eigenvalues of \( A_k' A_k \) are \( \lambda_{k1} = 0, \lambda_{k2} = \ldots \lambda_{kp_k} = 1 \) and the columns of the matrix \( U_k \) defined above give corresponding eigenvectors.

**PLS and bi-flat shrinkage.** It follows from (2.9) and (2.12) that the class of candidate PLS estimators generated by the flat annihilators coincides with the class of BF candidate estimators

\[
\hat{M}(c, A) = U_1[F,Z]U_2' \text{ for } F \in F_{BF}
\]

(3.18)

through the correspondence \( c_1 = (1+c_1)^{-1}, c_2 = (1+c_2)^{-1} \) and \( c_{12} = (1+c_{12})^{-1} \).

By calculus, the shrinkage-adaptive bi-flat estimator is \( \hat{M}(\hat{c}, A) \), where \( \hat{c} = (\hat{c}_1, \hat{c}_2, \hat{c}_{12}) \)

\[
\begin{align*}
\hat{c}_1 &= \arg \min_{c_1 \in [0, \infty]} \sum_{i=2}^{p_1} [c_1 - \hat{g}_{i1}]^2 z_{i1}^2 = [1 - (p_1 - 1)\hat{\sigma}^2 / \sum_{i=2}^{p_1} z_{i1}^2],
\hat{c}_2 &= \arg \min_{c_2 \in [0, \infty]} \sum_{j=2}^{p_2} [c_2 - \hat{g}_{j2}]^2 z_{j2}^2 = [1 - (p_2 - 1)\hat{\sigma}^2 / \sum_{j=2}^{p_2} z_{j2}^2],
\hat{c}_{12} &= \arg \min_{c_{12} \in [0, \infty]} \sum_{i=2}^{p_1} \sum_{j=2}^{p_2} [c_{12} - \hat{g}_{ij}]^2 z_{ij}^2 = [1 - (p_1 - 1)(p_2 - 1)\hat{\sigma}^2 / \sum_{i=2}^{p_1} \sum_{j=2}^{p_2} z_{ij}^2].
\end{align*}
\]

(3.19)

To express the shrinkage-adaptive BF estimator more simply in terms of \( y \), let

\[
P_1 = u_2 u_2' \otimes A_1, \quad P_2 = A_2 \otimes u_1 u_1', \quad P_{12} = A_2 \otimes A_1.
\]

(3.20)

For the flat annihilators (1.6), these three matrices are symmetric, idempotent and \( P_1 P_2 = P_1 P_{12} = P_2 P_{12} = 0 \). Because

\[
\begin{align*}
P_1 m &= (u_2 u_2' \otimes A_1)m = \text{vec}(A_1 M u_2 u_2'), \\
P_2 m &= (A_2 \otimes u_1 u_1')m = \text{vec}(u_1 u_1' M A_2), \\
P_{12} m &= (A_2 \otimes A_1)m = \text{vec}(A_1 M A_2),
\end{align*}
\]

(3.21)

it follows that \( |P_1 m|^2 = |A_1 m u_2|^2 \), \( |P_2 m|^2 = |u_1' m A_2|^2 \), and \( |P_{12} m|^2 = |A_1 M A_2|^2 \). Thus, the penalized least squares criterion (1.4) reduces here to

\[
S(M, \nu, A) = |y - m|^2 + \nu_1 |P_1 m|^2 + \nu_2 |P_2 m|^2 + \nu_{12} |P_{12} m|^2.
\]

(3.22)

By calculus, the PLS candidate estimator that minimizes (3.22) is

\[
\hat{m}_{PLS}(\nu, A) = (I_{P_1 P_2} + \nu_1 P_1 + \nu_2 P_2 + \nu_{12} P_{12})^{-1} y
\]

(3.23)

\[
= [P_0 + (1 + \nu_1)^{-1} P_1 + (1 + \nu_2)^{-1} P_2 + (1 + \nu_{12})^{-1} P_{12}] y,
\]

where \( P_0 = I_{p_1 p_2} \).
where \( P_0 = u_2u_2' \otimes u_1u_1' \). The equivalent candidate BF estimator \( \hat{M}(c, A) \) is

\[
\text{vec}(\hat{M}(c, A)) = [P_0 + c_1P_1 + c_2P_2 + c_{12}P_{12}]y
\]  

(3.24)

Computing and minimizing the estimated risk of (3.24) shows that (3.19), which determines the shrinkage adaptive BF estimator, is equivalent to

\[
\hat{c}_1 = \left[1 - (p_1 - 1)\hat{\sigma}^2/|P_1y|_2^2\right]_+,
\]

\[
\hat{c}_2 = \left[1 - (p_2 - 1)\hat{\sigma}^2/|P_2y|_2^2\right]_+,
\]

\[
\hat{c}_{12} = \left[1 - (p_1 - 1)(p_2 - 1)\hat{\sigma}^2/|P_{12}y|_2^2\right]_+.
\]

(3.25)

On the other hand,

\[
P_{1m} = \text{vec}(\gamma_1u_1'), \quad P_{2m} = \text{vec}(u_1\gamma_2'), \quad P_{12m} = \text{vec}(\Gamma_{12}).
\]

(3.26)

From this and remarks following (3.20), \( P_0, P_1, P_{12} \) are the respective the orthogonal projections into the subspaces determined by the constant, row effects, column effects, and interaction terms of the full model for the two-way layout. Consequently, in (3.25),

\[
|P_1y|^2 = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} (y_{i.} - y_{.-})^2,
\]

\[
|P_2y|^2 = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} (y_{.-} - y_{..})^2,
\]

\[
|P_{12}y|^2 = \sum_{i=1}^{p_1} \sum_{j=1}^{p_2} (y_{ij} - y_{i.} - y_{.-} + y_{..})^2.
\]

(3.27)

Equations (3.24) through (3.27) entail that the shrinkage-adaptive BF estimator \( \hat{M}(\hat{c}, A) \) has components

\[
\hat{m}_{ij}(\hat{c}, A) = y_{i.} + \left[1 - \frac{(p_1 - 1)\hat{\sigma}^2}{\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} (y_{i.} - y_{.-})^2}\right]_+ (y_{i.} - y_{.-})
\]

\[
+ \left[1 - \frac{(p_2 - 1)\hat{\sigma}^2}{\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} (y_{.-} - y_{..})^2}\right]_+ (y_{.-} - y_{..})
\]

\[
+ \left[1 - \frac{(p_1 - 1)(p_2 - 1)\hat{\sigma}^2}{\sum_{i=1}^{p_1} \sum_{j=1}^{p_2} (y_{ij} - y_{i.} - y_{.-} + y_{..})^2}\right]_+ (y_{ij} - y_{i.} - y_{.-} + y_{..}).
\]

(3.28)

By different reasoning, Stein (1966, p. 358) obtained an estimator akin to this for the case when \( \hat{\sigma}^2 \) is an independent least squares estimator of variance. In that setting, Stein refined the right side of (3.28) slightly—subtracting 2 from the three factors \( (p_1 - 1), (p_2 - 1) \) and \( (p_1 - 1)(p_2 - 1) \)—so as to reduce risk in estimating \( M \). The effects of his modification decrease to vanishing as \( p_1 \) and \( p_2 \) increase. Devising such improvements to the other adaptive estimators treated in this paper is an open question.

**Example 2 (continued).** A flat annihilator for each nominal factor generates the penalty basis, whose dramatic empirical economy is revealed by subplot (3.1) in Figure 3. The nature of this economy motivates using variance estimator (2.19) with \( q_1 = q_2 = 3 \). The shrinkage matrix that defines the adaptive PLS or BF estimator
is shown in subplot (3,2). Shrinkage of the main effects is slight while shrinkage of the interactions is great, making this ASP fit nearly additive as noted earlier. The estimated risk of the least squares estimator is .4616 while the considerably smaller estimated risk of the ASP estimator is .1522. In this example, reduction of estimated risk accompanies clarification of how fat absorption depends on fat number and day.

Submodel shrinkage. The components of $\hat{f}_{SM} = \arg\min_{f \in \mathcal{F}_{SM}} \text{ave}[(f - \hat{g})^2 z^2]$ are given by $\hat{f}_{SM,ij} = 1 (\hat{f}_{BF,ij} \geq 1/2)$ or may be found directly by finite search.

### 3.4. One nominal and one ordinal factor

Suppose that the first factor is nominal while the second factor is ordinal. A suitable pair of annihilators is then $A_1 = I_{p_1} - u_1u_1'$ as in Section 3.3 and $A_2 = B_c(s_2, p_2)$ as in Section 3.2.

Penalized least squares. Here the shrinkage-adaptive PLS estimator is a specialization of (3.9) and (3.11) that is obtained by setting $\lambda_{11} = 0, \lambda_{12} = \ldots = \lambda_{1p_1} = 1$. The value of $\hat{\nu}_1$ is given by

$$
(1 + \hat{\nu}_1)^{-1} = \left[1 - (p_1 - 1)\hat{\sigma}^2 / \sum_{i=2}^{p_1} z_{i1}^2 \right]_+. 
$$

Calculating $\hat{\nu}_2$ and $\hat{\nu}_{12}$ amounts to minimizing the respective nonlinear weighted least squares criteria in (3.11).

Example 3 (continued). A third difference annihilator for the ordinal factor and a flat annihilator for the nominal factor generate the penalty basis, whose empirical economy is revealed by subplot (3,1) in Figure 4. The nature of this economy motivates using variance estimator (2.19) with $q_1 = 24$ and $q_2 = 0$. The shrinkage matrix that defines this ASP estimator is shown in subplot (3,2). Shrinkage is negligible for the main effects of the nominal factor but is pronounced for the higher order coefficients, whether main effect or interaction, of the ordinal factor. Strong shrinkage of the highest order interaction coefficients makes the ASP fit roughly additive, as seen in subplot (2,2). The estimated risk of the least squares estimator is 1.8751 while the much smaller estimated risk of the ASP estimator is .3918. In this example, reduction of estimated risk accompanies greater understanding of how grape yield depends on row number and year.

Experimenting with $d$-th difference annihilators of orders one through four on the row factor does not reduce estimated risk below that achieved with the third difference annihilator. If the factor year is treated as ordinal rather than nominal, the first difference annihilator on that factor best controls estimated risk of PLS candidate estimators. However, the corresponding ASP fit virtually coincides with that obtained in the preceding paragraph.

Flat $\times$ monotone shrinkage. Larger than the PLS class is the shrinkage class $\mathcal{F}_{F \times M}$ defined in (2.14). The shrinkage-adaptive $F \times M$ estimator is

$$
\hat{M}(\hat{F}_{F \times M}, A) = U_1[\hat{F}_{F \times M} Z]U_2',
$$

where $\text{vec}(\hat{F}_{F \times M}) = \hat{f}_{F \times M}$ and

$$
\hat{f}_{F \times M} = \arg\min_{f \in \mathcal{F}_{F \times M}} \text{ave}[(f - \hat{g})^2 z^2].
$$
The components \( \{ \hat{f}_{ij} \} \) of the matrix \( \hat{F}_{F \times M} \) may be found as follows:

First step. Set \( \hat{f}_{11} = 1 \).

Second step. For \( i \geq 2 \), set \( \hat{f}_{ii} = \left[ 1 - (p_1 - 1)\sigma^2 / \sum_{i=2}^{p_1} z_{ij}^2 \right]_+ \).

Third step. Let \( w = \{ z_{ij} : 2 \leq j \leq p_2 \} \) and let \( \hat{h} = (w^2 - \sigma^2)/w^2 \). Let \( K = \{ h \in R^q : k_1 \geq k_2 \geq \ldots \geq k_q \} \), where \( q = p_2 - 1 \). Find \( \hat{k} = \text{argmin}_{k \in K} \text{ave}[(k - \hat{h})^2w^2] \), using an algorithm for weighted isotonic least squares. Set \( \hat{f}_{ij} = \max\{k_{j-1}, 0\} \) for \( 2 \leq j \leq p_2 \).

Fourth step. Letting \( w^2 = \{ \sum_{i=2}^{p_1} z_{ij}^2 : 2 \leq j \leq p_2 \} \) and \( q = p_2 - 1 \), find \( \hat{k} \) as in the third step. Set \( \hat{f}_{ij} = \max\{k_{j-1}, 0\} \) for \( 2 \leq i \leq p_1, 2 \leq j \leq p_2 \).

4. Multiparametric Asymptotics

Adaptation works when estimated risk converges to actual risk uniformly over the class of candidate estimators. Empirical process theory provides sufficient conditions for such uniform convergence. For our purpose, the richness of a shrinkage class \( \mathcal{F} \subset \mathcal{F}_U \) is characterized through the covering number \( J(\mathcal{F}) \) that is defined as follows. For any probability measure \( Q \) on the set \( T = \{(i, j) : 1 \leq i \leq p_1, 1 \leq j \leq p_2 \} \), consider the pseudo-distance

\[
N(u, \mathcal{F}, d_Q) = \min\{\#\mathcal{F}_0 : \mathcal{F}_0 \subset \mathcal{F}, \inf_{f_0 \in \mathcal{F}_0} d_Q(f_0, f) \leq u \quad \forall f \in \mathcal{F} \}. \tag{4.1}
\]

Let

\[
N(u, \mathcal{F}) = \sup_Q N(u, \mathcal{F}, d_Q), \tag{4.2}
\]

where the supremum is taken over all probabilities on \( T \). Define

\[
J(\mathcal{F}) = \int_0^1 [\log N(u, \mathcal{F})]^{1/2} du. \tag{4.3}
\]

Important in proving Theorem 2.1 is the fact \( J(\mathcal{F}_{BM}) = O(\min\{p_1^{1/2}, p_2^{1/2}\}) \), which follows from Example 5 on p. 1832 of Beran and Dümbgen (1998) and implies

\[
(p_1 p_2)^{-1/2} J(\mathcal{F}_{BM}) = O(\min\{p_1^{-1/2}, p_2^{-1/2}\}). \tag{4.4}
\]

In particular, because \( \max\{p_1, p_2\} \leq p_1 p_2 \leq \max\{p_1, p_2\}^2 \), the right side of (4.4) tends to zero as \( p_1 p_2 \to \infty \).

**Proof of Theorem 2.1.** Part a. By Theorem 2.1 in Beran and Dümbgen (1998), there exists a finite constant \( C \) such that

\[
E \sup_{f \in \mathcal{F}} |V(f) - R(\hat{M}(F, A), M, \sigma^2)| \leq C \left[ J(\mathcal{F}) \sigma^2 + \sigma \sqrt{\text{ave}(\xi^2)} + E|\hat{\sigma}^2 - \sigma^2| \right]. \tag{4.5}
\]

Limit (2.24) follows from this, the inclusion of \( \mathcal{F} \) in \( \mathcal{F}_{BM} \), (4.4), and (2.23).

Parts b and c. In analogy to \( \hat{f} = \text{argmin}_{f \in \mathcal{F}} \hat{r}(f, A) \), let

\[
\hat{f} = \text{argmin}_{f \in \mathcal{F}} r(f, A, \xi^2, \sigma^2). \tag{4.6}
\]
Then \( \min_{\mathcal{F}} R(\hat{M}(F, A), M, \sigma^2) = r(\hat{f}, A, \xi^2, \sigma^2) \). Let \( \hat{F} \) be the shrinkage matrix vectorized by \( \hat{f} \). We first show that (2.24) implies

\[
\lim_{p_1, p_2 \to \infty} \sup_{\text{ave}(\xi^2) \leq \sigma^2} \mathbb{E}[W - r(\hat{f}, A, \xi^2, \sigma^2)] = 0, \tag{4.7}
\]

where \( W \) can be \( L(\hat{M}(\hat{F}, A), M) \) or \( L(\hat{M}(\hat{F}, A), M) \) or \( \hat{r}(\hat{f}, A) \).

Indeed, (2.24) with \( V(f) = \hat{r}(f, A) \) entails

\[
\lim_{p_1, p_2 \to \infty} \sup_{\text{ave}(\xi^2) \leq \sigma^2} \mathbb{E}[\hat{r}(\hat{f}, A) - r(\hat{f}, A, \xi^2, \sigma^2)] = 0, \]

\[
\lim_{p_1, p_2 \to \infty} \sup_{\text{ave}(\xi^2) \leq \sigma^2} \mathbb{E}[\hat{r}(\hat{f}, A) - r(\hat{f}, A, \xi^2, \sigma^2)] = 0. \tag{4.8}
\]

Hence, (4.7) holds for \( W = \hat{r}(\hat{f}, A) \) and

\[
\lim_{p_1, p_2 \to \infty} \sup_{\text{ave}(\xi^2) \leq \sigma^2} \mathbb{E}[r(\hat{f}, A, \xi^2, \sigma^2) - r(\hat{f}, A, \xi^2, \sigma^2)] = 0. \tag{4.9}
\]

On the other hand, (2.24) with \( V(f) = L(\hat{M}(F, A), M) \) gives

\[
\lim_{p_1, p_2 \to \infty} \sup_{\text{ave}(\xi^2) \leq \sigma^2} \mathbb{E}[L(\hat{M}(\hat{F}, A), M) - r(\hat{f}, A, \xi^2, \sigma^2)] = 0, \]

\[
\lim_{p_1, p_2 \to \infty} \sup_{\text{ave}(\xi^2) \leq \sigma^2} \mathbb{E}[L(\hat{M}(\hat{F}, A), M) - r(\hat{f}, A, \xi^2, \sigma^2)] = 0. \tag{4.10}
\]

These limits together with (4.9) establish the remaining two cases of (4.7).

The limits (2.25) and (2.26) are immediate consequences of (4.7).

**Part d.** This conclusion follows by combining the separate results for \( \mathcal{F} = \mathcal{F}_{MS} \) and \( \mathcal{F} = \mathcal{F}_{ST} \).

Proving Theorem 3.1 requires a preliminary result. Let \( \mathcal{E} = \{c \in \mathbb{R}^{p_1 p_2} : c_i \in [1, \infty], 1 \leq i \leq p_1 p_2 \} \). For every \( c \in \mathcal{E} \), define the ellipsoid

\[
E(a, c, \sigma^2) = \{\xi \in \mathbb{R}^{p_1 p_2} : \text{ave}(c \xi^2) \leq \sigma^2 a\}. \tag{4.11}
\]

When \( \xi \in E(a, c, \sigma^2) \) and \( c_i = \infty \), it is to be understood that \( \xi_i = 0 \) and \( c_i^{-1} = 0 \). Let

\[
\xi_0^2 = \sigma^2[(\alpha/c)^{1/2} - 1]_+, \quad g_0 = \xi_0^2/(\sigma^2 + \xi_0^2) = [1 - (c/\alpha)^{1/2}]_+, \tag{4.12}
\]

where \( \alpha \) is the unique positive number such that \( \text{ave}(c \xi_0^2) = \sigma^2 a \). Define

\[
\tau(a, c, \sigma^2) = r(f, A, \xi^2, \sigma^2) = \sigma^2 \text{ave}[g_0^2/(\sigma^2 + \xi_0^2)]. \tag{4.13}
\]

Evidently, \( \tau(a, c, \sigma^2) \in [0, \sigma^2] \) for every \( a > 0 \) and every \( c \in \mathcal{E} \).

The following theorem, specialized from the argument of Pinsker (1980), establishes that the linear estimator \( g_0 z \) is typically asymptotically minimax among all estimators of \( \xi \).

**Theorem 4.1.** Suppose that \( \lim \inf_{p_1, p_2 \to \infty} \tau(a, c, \sigma^2) > 0 \). Then,

\[
\lim_{p_1, p_2 \to \infty} \left[ \inf_{\xi} \sup_{\xi \in E(a, c, \sigma^2)} (p_1 p_2)^{-1} \mathbb{E}[\xi^2 - \tau(a, c, \sigma^2)] \right] = 0 \tag{4.14}
\]

and

\[
\lim_{p_1, p_2 \to \infty} \left[ \sup_{\xi \in E(a, c, \sigma^2)} (p_1 p_2)^{-1} \mathbb{E}[g_0 z - \xi^2 - \tau(a, c, \sigma^2)] \right] = 0. \tag{4.15}
\]
Proof of Theorem 3.1. Limit (3.4) is the specialization of (4.14) when $c_{ij} = 1$ for $(i, j) \in S$ and is infinite otherwise. In that case, $\lim_{p_1p_2 \to \infty} \tau(a, c, \sigma^2) = \sigma^2 [ab/(a + b)]$ by specialization of (4.12) and (4.13).

The coefficients of $g_0$ are $g_{0,ij} = [1 - \alpha^{-1/2}]_+$ for $(i, j) \in S$ and are zero otherwise. By the definition of $S$, $g_0 \in F_{BN} \subset F_{BM}$. Consequently the oracle estimator $\tilde{f}_z$ that is defined by (4.7) when $\hat{F}$ is either $F_{BN}$ or $F_{BM}$ satisfies

$$\sup_{\xi \in E(a, c, \sigma^2)} (p_1p_2)^{-1}E|\tilde{f}_z - \xi|^2 \leq \sup_{\xi \in E(a, c, \sigma^2)} (p_1p_2)^{-1}E|g_0z - \xi|^2. \quad (4.16)$$

From this, (4.15), and the preceding evaluation of $\tau(a, c, \sigma^2)$,

$$\lim_{p_1p_2 \to \infty} \sup_{\xi \in E(a, c, \sigma^2)} E|\tilde{f}_z - \xi|^2 = \sigma^2 [ab/(a + b)]. \quad (4.17)$$

Limit (3.5) follows from (4.17) and limit (2.25) in Theorem 2.1.

References


Department of Statistics
University of California, Davis
Davis, CA 95616
USA
e-mail: beran@wald.ucdavis.edu