Penalized Fits to a Multiway Layout with Multivariate Responses

Rudolf Beran

University of California, Davis

Workshop on Model Selection and Related Areas

University of Vienna

24 July 2008

Multivariate Linear Model

$Y = CM + E$, where

- the rows of $n \times d$ matrix Y are d-variate responses;
- the $n \times p$ design matrix C has rank $p \leq n$;
- the $p \times d$ matrix M is unknown;
- the $n \times d$ error matrix $E = V\Sigma^{1/2}$, where Σ is an unknown p.d. covariance matrix and the elements of V are iid with mean 0, variance 1, and finite 4-th moment.

The **least squares estimator** of M is $\hat{M}_{ls} = C^+ Y$.

Let $y = \mathrm{vec}(Y)$, $m = \mathrm{vec}(M)$, $e = \mathrm{vec}(E)$ and $\tilde{C} = I_d \otimes C$.
The vectorized model asserts $y = \tilde{C}m + e$.
The **least squares estimator** of m is $\hat{m}_{ls} = \tilde{C}^+ y = \mathrm{vec}(\hat{M}_{ls})$.

For now, assume $\Sigma = I_d$.

Quadratic Loss and Risk

Let $\hat{\eta}$ be any estimator of $\eta = \tilde{C}m = \mathrm{E}(y)$.

The **loss** of η is $L(\hat{\eta}, \eta) = p^{-1}|\hat{\eta} - \eta|^2$ and the corresponding **risk** is $R(\hat{\eta}, \eta) = \mathrm{E}L(\hat{\eta}, \eta)$. Equivalently, these are loss and risk functions on estimators of m through the 1-to-1 map $\hat{\eta} = \tilde{C}\hat{m}$.

The least squares estimator $\hat{\eta}_{ls} = \tilde{C}\hat{m}_{ls} = \tilde{C}\tilde{C}^{+}y$ has risk $R(\hat{\eta}_{ls}, \eta) = d$.

Biased estimators of η can reduce risk substantially: Stein (1956), James and Stein (1961), Stein (1966); also papers on symmetric linear estimators such as Stein (1981), Li and Hwang (1984), Buja, Hastie and Tibshirani (1989), Kneip (1994), Beran (2007) $\ldots$

Penalized least squares (PLS) generates promising, biased, candidate symmetric linear estimators of η.

General Structure of PLS for the Multivariate Linear Model

Let $\mathcal{S}$ be an index set of fixed cardinality.

Let $\{Q_s : s \in \mathcal{S}\}$ be $p \times p$ p.s.d. **penalty matrices**.

$N = \{N_s : s \in \mathcal{S}\}$ be $d \times d$ p.s.d. **affine penalty weights**.

PLS criterion: $G(m, N) = |y - \tilde{C}m|^2 + m'Q(N)m$,

where $Q(N) = \sum_{s \in \mathcal{S}} (N_s \otimes Q_s)$.

The **PLS estimators** of m and η are then

$\hat{m}_{pls}(N) = \operatorname{argmin}_m G(m, N) = [\tilde{C}'\tilde{C} + Q(N)]^{-1}\tilde{C}'y$,

$\hat{\eta}_{pls}(N) = \tilde{C}\hat{m}_{pls} = \tilde{C}[\tilde{C}'\tilde{C} + Q(N)]^{-1}\tilde{C}'y$, a symmetric linear estimator (generalized ridge).

These estimators can be derived as Bayes estimators in a normal error version of the multivariate linear model. Kimeldorf and Wahba (1970) make the general point.

- When $d = 1$, the penalty weights are non-negative scalars. E.g. Wood (2000), Beran (2005) use multiple penalty terms with scalar weights.
- Functional data-analysis treats penalized estimation of a function m of **continuous** covariates. E.g. Wahba, Wang, Gu, Klein, Klein (1995), Li (2000), Ramsay and Silverman (2002).

To be considered:

- Data-based choice of the affine penalty weights $\{N_s : s \in \mathcal{S}\}$;
- Supporting asymptotic theory for the foregoing, as $p \to \infty$;
- Penalty matrices $\{Q_s : s \in \mathcal{S}\}$ suitable for the multiway layout with d-variate responses;
- Modifications for the case of a general unknown covariance matrix Σ.

Canonical Form and Risk of $\hat{\eta}_{pls}(N)$

Let $\tilde{R} = I_d \otimes C'C$, a $pd \times pd$ matrix of full rank.

Let $\tilde{U} = I_d \otimes C(C'C)^{-1/2}$, a $nd \times pd$ matrix.

Then $\tilde{C} = I_d \otimes C = \tilde{U}\tilde{R}^{1/2}$ and $\tilde{U}'\tilde{U} = I_{pd}$. Hence,

$\hat{\eta}_{pls}(N) = \tilde{C}[\tilde{C}'\tilde{C} + Q(N)]^{-1}\tilde{C}'y = \tilde{U}S(N)\tilde{U}'y$,

where $S(N) = [I_{pd} + \tilde{R}^{-1/2}Q(N)\tilde{R}^{-1/2}]^{-1}$ is symmetric.

Because $\mathcal{R}(\tilde{C}) = \mathcal{R}(\tilde{U})$ and $\tilde{U}'\tilde{U} = I_{pd}$, $\eta = \tilde{C}m = \tilde{U}\xi$, with $\xi = \tilde{U}'\eta$. Let $z = \tilde{U}'y$. Then $\hat{\eta}_{pls}(N) = \tilde{U}S(N)z$.

This is the **canonical form** of $\hat{\eta}_{pls}(N)$.

The **risk** of $\hat{\eta}_{pls}(N)$ is thus

$R(\hat{\eta}(N), \eta) = p^{-1}\mathrm{E}|S(N)z - \xi|^2 = p^{-1}[\mathrm{tr}(T(N)) + \mathrm{tr}(\bar{T}(N)\xi\xi')]$,

where $T(N) = S^2(N)$ and $\bar{T}(N) = [I_{pd} - S(N)]^2$.

Estimated Risk

The **estimated risk** of $\hat{\eta}_{pls}(N)$ is

$\hat{R}(N) = p^{-1}[\mathrm{tr}(T(N)) + \mathrm{tr}(\bar{T}(N)(zz' - I_{pd})')]$,

(cf. Mallows (1973), Stein (1981)). Let $\hat{N} = \mathrm{argmin}_N \hat{R}(N)$.

E.g. Use Cholesky $N_s = L_s L_s'$ with $\{l_{s,i,i} \geq 0\}$.

The **adaptive PLS estimators** of η and of m are

$\hat{\eta}_{apls} = \hat{\eta}_{pls}(\hat{N})$ and $\hat{m}_{apls} = C^{+}\hat{\eta}_{apls}$.

Supporting Asymptotics

Let $|\cdot|_{sp}$ denote spectral matrix norm: $|B|_{sp} = \sup_{x \neq 0}[|Bx|/|x|]$.

- Let $W(N)$ denote either the loss or estimated risk of $\hat{\eta}_{pls}(N)$.

  Let $\mathcal{N} = \{N \colon \max_{s \in \mathcal{S}} |N_s|_{sp} \leq b\}$. Then, for every finite $a > 0$,

$$\lim_{p\to\infty} \sup_{p^{-1}|\eta|^2 \leq a} \mathrm{E}[\sup_{N \in \mathcal{N}} |W(N) - R(\hat{\eta}_{pls}(N), \eta)|] = 0.$$

- For every finite $a > 0$,

$$\lim_{p \to \infty} \sup_{p^{-1}|\eta|^2 \leq a} |R(\hat{\eta}_{apls}, \eta) - \min_{N \in \mathcal{N}} R(\hat{\eta}(N), \eta)| = 0.$$

- Let V denote either the loss or risk of $\hat{\eta}_{apls}$, Then, for every finite $a > 0$,

$$\lim_{p \to \infty} \sup_{p^{-1}|\eta|^2 \leq a} \mathrm{E}|\hat{R}(\hat{N}) - V| = 0.$$

The loss, risk and estimated risk of the candidate estimator $\hat{\eta}_{pls}(N)$ converge together, as $p \to \infty$, **uniformly** over $N \in \mathcal{N}$.

Estimated risk is here a trustworthy surrogate for loss or risk.

The risk of $\hat{\eta}_{apls}$ converges, as $p \to \infty$, to the minimal risk achievable by the PLS candidate estimators

The **plug-in risk estimator** $\hat{R}(\hat{N})$ converges to the loss or risk of $\hat{\eta}_{apls}$ as $p \to \infty$.

Complete k_0-way Layout with Multivariate Responses

Now the d dimensional responses depend on k_0 covariates.

Covariate k has p_k distinct levels $x_{k,1} < x_{k,2} < \ldots x_{k,p_k}$.

Let $\mathcal{I}$ denote all k_0-tuples $i = (i_1, i_2, \ldots, i_{k_0})$, where $1 \le i_k \le p_k$ for $1 \le k \le k_0$. Thus, i_k indexes the levels of covariate k and $\mathcal{I}$ lists all possible covariate-level combinations.

We put the elements of $\mathcal{I}$ in **mirror-dictionary order**.

We observe $Y = CM + E$, the assumptions on E as before.

Here C is the $n \times p$ data-incidence matrix of 0's and 1's that suitably replicates rows of the $p \times d$ matrix M into the rows of $\mathrm{E}(Y) = CM$.

The design is **complete:** $\mathrm{rank}(C) = p$.

Row $i \in \mathcal{I}$ of M equals $f(x_{1,i_1}, x_{2,i_2}, \ldots, x_{k_0,i_{k_0}})$ where f is an **unknown** vector-valued function.

Constructing Penalty Matrices $\{Q_s\colon s \in \mathcal{S}\}$

We devise a scheme that penalizes individually the main effects and interactions in the MANOVA decomposition of M.

For $1 \le k \le k_0$, define the $p_k \times 1$ vector $u_k = p_k^{-1/2}(1, 1, \ldots, 1)'$.

Let A_k be an **annihilator:** a matrix such that $A_k u_k = 0$.

Let $\mathcal{S}$ denote the set of all subsets of $\{1, 2, \ldots, k_0\}$, including $\emptyset$.

Let $Q_{s,k} = u_k u_k'$ if $k \notin s$; and $Q_{s,k} = A_k' A_k$ if $k \in s$. Define

$$Q_s = \bigotimes_{k=1}^{k_0} Q_{s,k-k_0+1}, \qquad s \in \mathcal{S}.$$

Special case: $A_k = I_{p_k} - u_k u_k'$. Denote Q_s in this case by $P_{AN,s}$. The matrices $\{P_{AN,s}\colon s \in \mathcal{S}\}$ are mutually orthogonal, orthogonal projections such that $\sum_{s \in \mathcal{S}} P_{AN,s} = I_p$.

MANOVA decomposition: $M = \sum_{s \in \mathcal{S}} P_{AN,s} M$.

From the foregoing definitions, $P_{AN,s}Q_s = Q_s P_{AN,s} = Q_s$ for every $s \in \mathcal{S}$; and $P_{AN,s_1}Q_{s_2} = Q_{s_2}P_{AN,s_1} = 0$ if $s_1 \neq s_2$. Thus, $m'(N_s \otimes Q_s)m = |Q_s^{1/2} M N_s^{1/2}|^2 = |Q_s^{1/2}(P_{AN,s}M)N_s^{1/2}|^2$.

The **penalty term** in the PLS criterion is seen to operate on the summands in the MANOVA decompostion of M:

$m'Q(N)m = \sum_{s \in \mathcal{S}} m'(N_s \otimes Q_s)m = \sum_{s \in \mathcal{S}} |Q_s^{1/2}(P_{AN,s}M)N_s^{1/2}|^2$.

Spectral Form of the Penalty Matrices $\{Q_s\}$

$A_k' A_k = U_k \Lambda_k U_k'$, where $\Lambda_k = \mathrm{diag}\{l_{k,i_k}\colon 1 \leq i_k \leq p_k\}$ and $0 = \lambda_{k,1} \leq \lambda_{k,2} \leq \ldots \leq \lambda_{k,p_k}$. The first column of U_k is chosen to be u_k. Then $u_k u_k' = U_k E_k U_k'$, where $E_k = \mathrm{diag}\{e_{k,i_k}\colon 1 \leq i_k \leq p_k\}$, with $e_{k,1} = 1$ and $e_{k,i_k} = 0$ if $i_k \geq 2$. Hence, $Q_{s,k} = U_k \Gamma_{s,k} U_k'$, where $\Gamma_{s,k} = \mathrm{diag}\{\gamma_{s,k,i_k}\colon 1 \leq i_k \leq p_k\}$, with $\gamma_{s,k,i_k} = e_{k,i_k}$ if $k \notin s$; $\gamma_{s,k,i_k} = \lambda_{k,i_k}$ if $k \in s$.

Write $U_k = [u_{k,1}, \ldots u_{k,p_k}]$. Then, $Q_{s,k} = \sum_{i_k=1}^{p_k} \gamma_{s,k,i_k} P_{k,i_k}$, where $P_{k,i_k} = u_{k,i_k} u'_{k,i_k}$ is a rank one orthogonal projection. For $i \in \mathcal{I}$, let $P_i = \bigotimes_{k=1}^{k_0} P_{k_0-k+1,i_k}$ and $\gamma_{s,i} = \bigotimes_{k=1}^{k_0} \gamma_{s,k_0-k+1,i_k}$.

Let $\mathcal{I}_s = \{i \in \mathcal{I} \colon i_k = 1$ if $k \notin s$ and $i_k \geq 2$ if $k \in s\}$. This defines a partition of $\mathcal{I}$. Then,

$$Q_s = \bigotimes_{k=1}^{k_0} Q_{s,k-k_0+1} = \sum_{i \in \mathcal{I}_s} \gamma_{s,i} P_i.$$

Here, $\gamma_{\{\emptyset\},i} = 1$ if $i \in \mathcal{I}_\emptyset$ and $\gamma_{s,i} = \prod_{k \in s} \lambda_{s,i_k}$ if $s \neq \emptyset$ and $i \in \mathcal{I}_s$.

Note: The $\{P_i\}$ are mutually orthogonal projections such that $\sum_{i \in \mathcal{I}} P_i = I_{pd}$. The MANOVA projection $P_{AN,s} = \sum_{i \in \mathcal{I}_s} P_i$.

Next steps

- Structure of the PLS estimators in **balanced** layouts.
- Construction of suitable annihilator matrices.
- Extension of PLS estimators to a general covariance matrix Σ.

Balanced k_0-way Layout with Multivariate Responses

In a balanced layout $C'C = n_0 I_p$ for some $n_0 \geq 1$. Then, $\hat{m}_{ls} = (\tilde{C}'\tilde{C})^{-1}\tilde{C}'y = n_0^{-1}\tilde{C}'y$ (averaging responses over replications) and, for $Q(N) = \sum_{s \in \mathcal{S}}(N_s \otimes Q_s)$,

$\hat{m}_{pls} = [\tilde{C}'\tilde{C} + Q(N)]^{-1}\tilde{C}'y = [I_{pd} + n_0^{-1}Q(N)]^{-1}\hat{m}_{ls}$.

Using also $Q_s = \sum_{i \in \mathcal{I}_s} \gamma_{s,i} P_i$ yields

$I_{pd} + n_0^{-1}Q(N) = \sum_{s \in \mathcal{S}} \sum_{i \in \mathcal{I}_s} [(I_d + n_0^{-1}\gamma_{s,i}N_s) \otimes P_i]$.

Hence, for a balanced layout,

$\hat{m}_{pls}(N) = \sum_{s \in \mathcal{S}} \sum_{i \in \mathcal{I}_s} [(I_d + n_0^{-1}\gamma_{s,i}N_s)^{-1} \otimes P_i]\hat{m}_{ls}$.

In matrix form,

$\hat{M}_{pls}(N) = \sum_{s \in \mathcal{S}} \sum_{i \in \mathcal{I}_s} P_i \hat{M}_{ls}(I_d + n_0^{-1}\gamma_{s,i}N_s)^{-1}$.

The annihilators determine the projections $\{P_i\}$ and the $\{\gamma_{s,i}\}$ in the affine shrinkage factors. Estimated risk also simplifies.

Constructing Annihilators

Recall that row $i \in \mathcal{I}$ of M equals $f(x_{1,i_1}, x_{2,i_2}, \ldots, x_{k_0,i_{k_0}})$ where f is unknown; and that $x_{k,1} < \ldots x_{k,p_k}$.

Covariate k is nominal. Permutation of the covariate levels $\{x_{k,j}: 1 \leq j \leq p_k\}$ should not affect the candidate estimator. Set $A_k = I_{p_k} - u_k u_k'$, an orthogonal projection. If all covariates are nominal, this annihilator choice generates candidate estimators that affinely penalize the individual terms in the MANOVA decomposition of M.

Covariate k is ordinal. In choosing A_k, we might hypothesize that f varies locally in ordinal covariate k like a polynomial of degree $r-1$. Right or wrong, $R(\hat{\eta}_{apls}, \eta) \leq R(\hat{\eta}_{ls}, \eta)$ as $p \to \infty$. The estimated risk $\hat{R}(\hat{\eta}_{apls})$ keeps score!

The relevant **local polynomial annihilator** A_k is a $(p_k - r) \times p_k$ matrix charactrized by three conditions:

- All elements in row t of A_k that are not in columns $t, t+1, \ldots, t+r$ are zero.
- Let $x = (x_{k,1}, x_{k,2}, \ldots, x_{k,p_k})'$. Then $A_k x_k^h = 0$ for $0 \leq h \leq r-1$.
- Each row of A_k has unit Euclidean length.

To meet these conditions, set the non-zero elements in row t of A_k equal to the basis vector of degree r in the orthonormal polynomial basis that is defined on the $r+1$ design points $(x_{k,t}, \ldots, x_{k,t+r})$. E.g. use the R function `poly`.

Note: When the ordinal covariate values $\{x_{k,j}\colon 1 \leq j \leq p_k\}$ are equally spaced, this construction makes A_k a multiple of the r-th difference matrix with p_k columns.

The Case of General Σ

Model $Y = CM + V\Sigma^{1/2}$ is equivalent to $y_\Sigma = \eta_\Sigma + v$, where $y_\Sigma = (\Sigma^{-1/2} \otimes I_p)y,\ \eta_\Sigma = (\Sigma^{-1/2} \otimes I_p)\eta$, and $v = \mathrm{vec}(V)$. The compents of v are iid with mean 0, variance 1 and finite 4-th moment—the model already treated.

Suppose Σ is **known**. Because $\eta = (\Sigma^{1/2} \otimes I_p)\eta_\Sigma$,

- Estimate η_Σ by $\hat{\eta}_{\Sigma,apls}$ based on y_Σ.
- Estimate η by $\hat{\eta}_{apls} = (\Sigma^{1/2} \otimes I_p)\hat{\eta}_{\Sigma,apls}$; and m by $\hat{m}_{apls} = \tilde{C}^+\hat{\eta}_{apls}$.
- The previous asymptotic theory carries over to the general Σ model when the loss function is

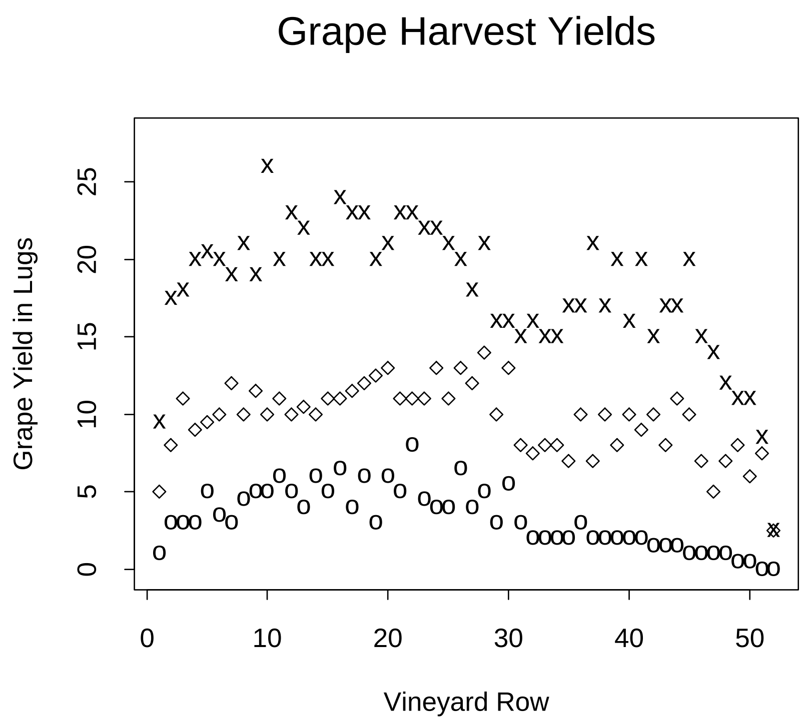
$$p^{-1}|\hat{\eta}_\Sigma - \eta_\Sigma|^2 = p^{-1}(\hat{\eta} - \eta)'(\Sigma^{-1} \otimes I_p)(\hat{\eta} - \eta).$$

If Σ is **unknown**, replace it by a consistent estimator $\hat{\Sigma}$ in constructing $\hat{\eta}_{apls}$ and $\hat{m}_{apls}$.

Remarks on Estimating Σ

- If $\hat{\Sigma}$ is consistent for Σ, the earlier asymptotics for the case $\Sigma = I_d$ can be extended. Loss and estimated risk converge together. Under stronger conditions on $\hat{\Sigma}$, the risk, loss and estimated risk converge together.
- When $n > p$, least squares theory provides the estimator $\hat{\Sigma}_{ls} = (n-p)^{-1}Y'(I_n - CC^+)Y$. This is consistent for Σ when $n - p \to \infty$.
- In the absence of adequate replication, **pooling** may provide a useful estimator of Σ: fit a plausible linear **submodel** for M by least squares and construct the least squares estimator of Σ associated with this fit. This $\hat{\Sigma}$ will be consistent if its bias tends to zero in the asymptotics.
- Obviously, replication is desirable in estimating Σ.

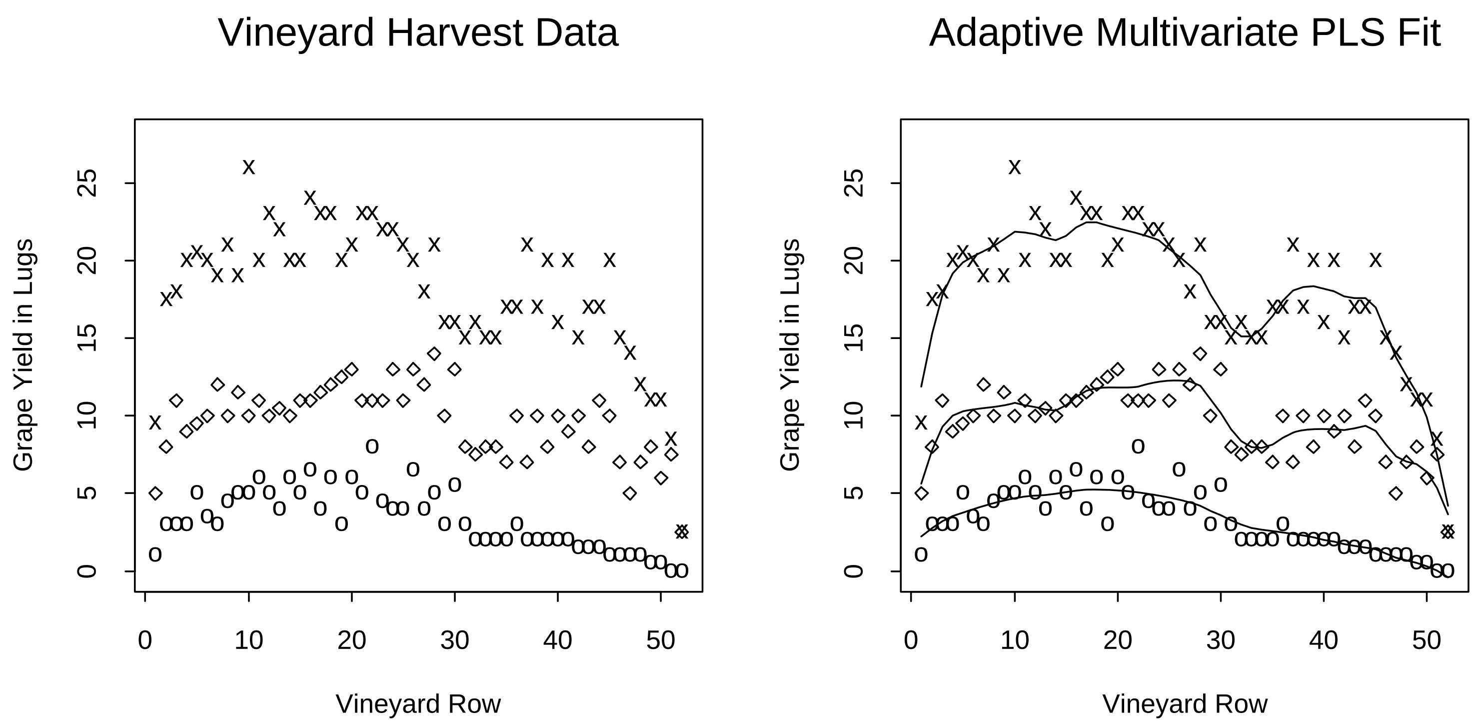
The Vineyard Data



Row i of data matrix Y reports the grape yields harvested in three different years from physical row i of a vineyard. This is a balanced one-way layout with trivariate responses. Both year and row may affect the harvest yields observed. We look for persistent pattern by estimating mean yields.

- In this one-way layout, $p = n = 52$, $d = 3$, and $k_0 = 1$. Hence, $\mathcal{S} = \{\{\emptyset\}, \{1\}\}$, $\mathcal{I} = \{i: 1 \le i \le p\}$, and $\mathcal{I}_{\{\emptyset\}} = \{1\}$, $\mathcal{I}_{\{1\}} = \{i: 2 \le i \le p\}$.
- Set the annihilator A_1 to be the second-difference matrix.
- The eigenvectors of $A_1' A_1$, ordered from smallest to largest eigenvalue, give the basis U that supports spectral representations of the two penalty matrices $\{Q_s: s \in \mathcal{S}\}$.
- Estimate Σ from the residuals after the least squares fit of Y to the first 20 columns of U (pooling strategy).
- Take $N_{\{\emptyset\}} = 0$. Then the candidate PLS estimators do not shrink the mean response vector. Adaptation is over all p.d. affine penalty weights $N_{\{1\}}$.

Some Findings



- $\hat{\Sigma}$ indicates slightly correlated heteroscedastic errors:

$$\hat{\Sigma} = \begin{pmatrix} 0.994 & 0.191 & 0.160 \\ 0.191 & 1.782 & -.268 \\ 0.160 & -.268 & 3.054 \end{pmatrix}$$

- The estimated risks of $\hat{M}_{apls}$ and $\hat{M}_{ls}$ are 0.364 and 3.000. In this example, $\hat{M}_{apls}$ reduces estimated risk more than eightfold!

A Non-statistical Example

Portrait of Kaiser Rudolf II by Hans von Aachen

Superb biased estimator: Kaiser Rudolf II by Giuseppe Arcimboldo