Estimating Many Means:
A Mosaic of Recent Methodologies

Rudolf Beran

University of California, Davis

Perspectives on High-dimensional Data Analysis

Fields Institute, University of Toronto

June 11, 2011

THE AGE OF THE ALGORITHM

In the 20th century, mathematical logicians (Gödel, Church, Post, Turing, . . .) made rigorous the concept of algorithm:

An **algorithm** is a finite procedure, written in a fixed symbolic vocabulary, governed by precise instructions, moving in discrete steps, that eventually comes to an end (cf. Berlinski 2000).

This definition of **effective** computation led to digital computers. whose speed in implementing algorithms continues to have consequences that transform the human world.

Marshall McLuhan (1964). "[T]he medium is the message . . . The history of the arts and sciences could be written in terms of the continuing process by which new technologies create new environments for old technologies."

MULTI-WAY LAYOUTS

A fundamental data structure is the k-way layout of observations, complete or incomplete, balanced or unbalanced. The cells of the layout are indexed by all k-fold combinations of the levels of the k covariates (or factors). Replication of observations within cells may be rare or nonexistent. Observations with error may be available for only a subset of the cells. The problem is to estimate the mean observation, or mean potential observable, for each cell in the k-way layout. Equivalently, the problem is to estimate an unknown regression function that depends on k covariates.

Remark: The number of means (cells) in the layout may be large. Substantially many of the unknown means of interest may lack noisy observations.

SOME IDEAS IN STATISTICAL THEORY

Use of probability models in Statistics spans a **spectrum:**

- none in exploratory data analysis $\Leftrightarrow$ fixed effects models $\Leftrightarrow$ random effects models $\Leftrightarrow$ Bayesian randomness.
- Statisticians **distinguish** among data, probability models, pseudorandom numbers, and algorithms.
- Data is not certifiably random (e.g. Kolmogorov, Knuth on definitions of randomness).
- Statistical estimation theory increasingly supports **simpler** fits to data (e.g. through bias-variance tradeoff, sparsity).
- Ongoing is the evolution of Statistics **from** mathematical philosophy **to** empirically supported information science. (cf. the development of scientific Medicine).

SOME HISTORICAL STEPS

Peter Peregrinus of Maricourt (1269). "[An investigator] diligent in the use of his own hands . . . will then in a short time be able to correct an error which he would never do in eternity by his knowledge of natural philosophy and mathematics alone. [However] there are many things subject to the rule of reason that we cannot completely investigate by the hand."

William of Ockham (early 14th century). "It is pointless to do with more what can be done with fewer" (Frustra fit per plura quod potest fieri per pauciora).

Andrey Kolmogorov(1963). "There arises a problem of finding the reasons for applicability of the mathematical theory of probability to the phenomena of the real world."

APPROACHES TO MULTI-WAY LAYOUTS

Statistical methodologies for estimating means in multi-way layouts include:

- Unconstrained least squares fits;
- Submodel fits and model selection;
- Shrinkage estimators for balanced complete layouts with nominal factors (e.g. Stein 1966);
- Functional data analysis (e.g. Wahba, Wang, Gu, Klein, Klein 1995, Ramsay and Silverman 1997, Lin 2000);
- Multiply penalized least squares estimators for discrete incomplete layouts, including regression fits (Beran 2007).

Implicit or explicit in these fits is **regularization**.

FIXED EFFECTS TWO-WAY LAYOUT

Observations $\{y_{ijk}: 1 \leq k \leq n_{ij}\}$ are available at pairs $(i,j) \in B$, a subset of $\{(i,j): 1 \leq i \leq p_1, 1 \leq j \leq p_2\}$.

Assumed is the **Saturated Model**:

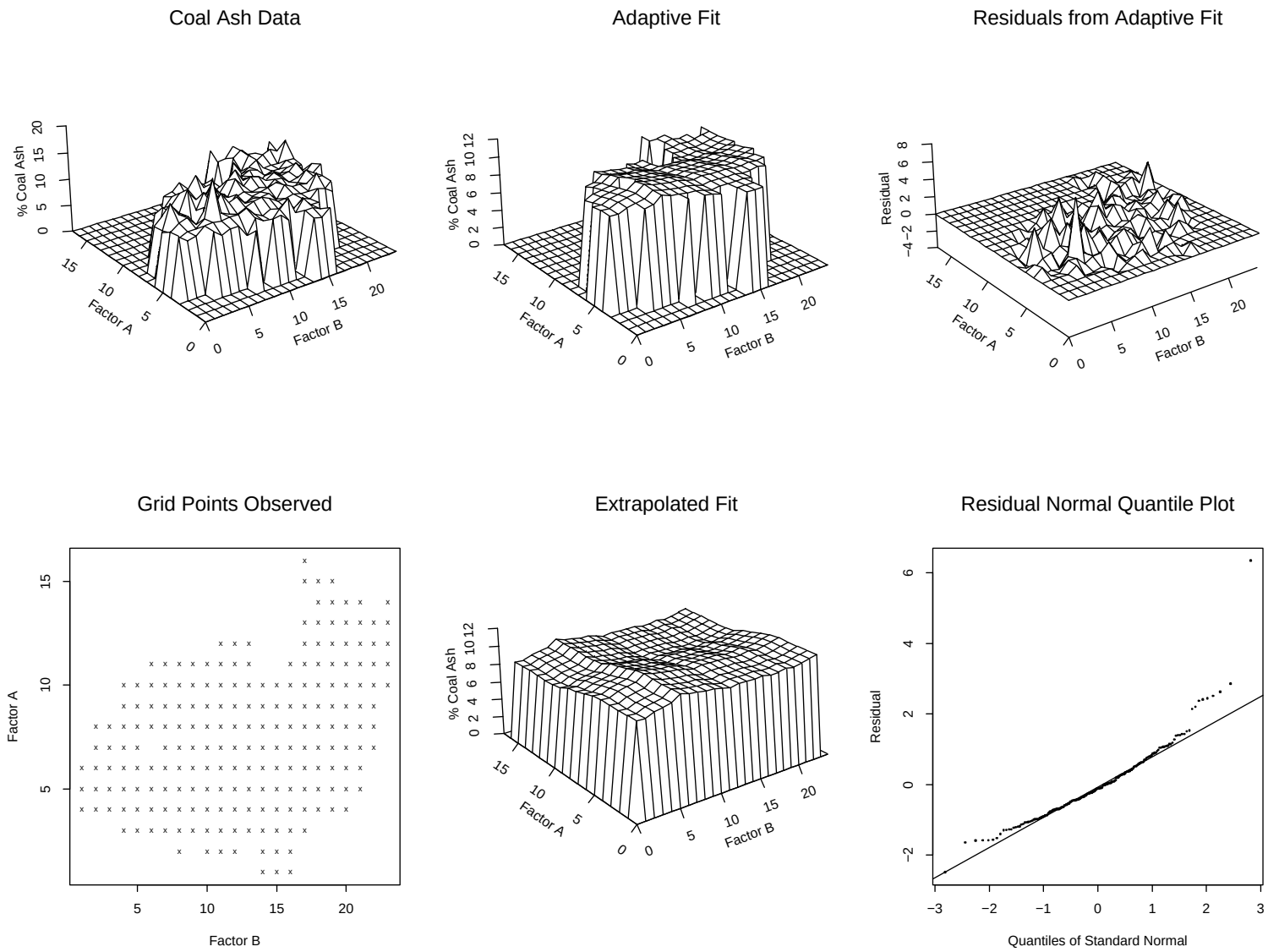
$$y_{ijk} = m_{ij} + e_{ijk}, \qquad m_{ij} = \mu(x_{1i}, x_{2j}), \qquad (i,j) \in B.$$

- The $\{x_{ri}\}$ are the **levels** of **factor** r. They can be **ordinal** (their values matter) or **nominal** (numerical labels).
- The function μ is **unknown and unrestricted**.
- Subscript k labels repeated observations.
- The $\{e_{ijk}\}$ are identically distributed random variables, each having mean 0, unknown variance σ^2, and $E(e_{ijk}^4) < \infty$ (a strong Gauss-Markov model).

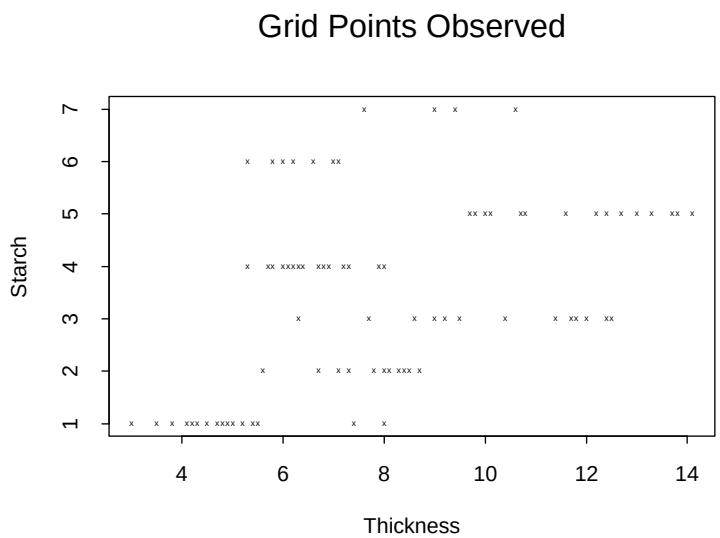
TWO DATA-SETS

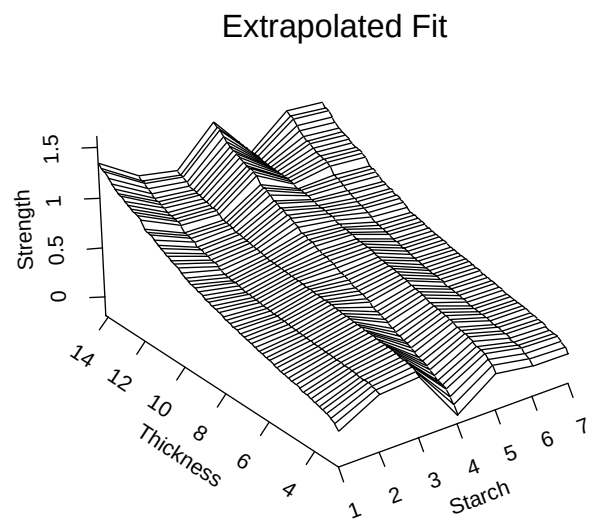
- **Coal Ash Data** (Cressie 1993, p. 34). The geographical factors are **both ordinal** with $p_1 = 23$ and $p_2 = 16$ levels. One coal ash measurement is made at each of $q = 208$ level-pairs (locations) out of $p = p_1 p_2 = 368$ in the complete layout.

- **Starch Data** (Freeman 1942, pp. 120–121). Starch type is **nominal** with $p_1 = 7$. Thickness is **ordinal** with $p_2 = 69$. One or more strength measurements are made at $q = 81$ factor-level pairs out of $p = p_1 p_2 = 483$ in the complete layout.



The **Coal Ash data**, its factor-level grid, the adaptive PLS estimate $\hat{m}_D(\hat{t})$, its extrapolation $\hat{m}(\hat{t})$, and a residual plot.





The observed factor-level grid of the **Starch data** and the extrapolated adaptive PLS estimate $\hat{m}(\hat{t})$ of mean breaking strengths.

VECTORIZED SATURATED MODEL FOR THE MEANS

Let $m = \{\{m_{ij}: 1 \leq i \leq p_1\}, 1 \leq j \leq p_2\}$ be the vectorized means of associated complete layout, a column vector of length $p = p_1 p_2$. Let $q \leq p$ be the cardinality of B.

Let the $q \times p$ **means-incidence** matrix D of zeroes and ones be such that $m_D = Dm$ lists the means on which data is observed.

Let $y = \{\{y_{ijk}: 1 \leq k \leq n_{ij}\}, (i,j) \in B\}$ denote the vectorized observations, a column vector y of length $n = \sum_{(i,j)\in B} n_{ij}$.

Vectorize similarly the errors $\{e_{ijk}\}$ into e.

The $n \times q$ **data-incidence** matrix C of zeroes and ones is such that $\eta = E(y) = Cm_D$. The saturated model asserts:

$$y = \eta + e, \text{ where } \eta = Cm_D = CDm$$

and e satisfies the strong Gauss-Markov model.

In the saturated model, the **least squares estimators** of $\eta = E(y)$ and of m_D are

$$\hat{\eta}_{LS} = CC^{+}y = (CD)(CD)^{+}y$$
$$\hat{m}_{D,LS} = C^{+}\hat{\eta}_{LS} = C^{+}y.$$

(a) When we have one observation at each $(i, j) \in B$, then $\hat{m}_{D,LS} = y$ (the raw data)—a provably inadequate estimator.

(b) We have not used the factor-levels much.

(c) How to extrapolate $\hat{m}_{D,LS}$ to estimate m?

To quantify the performance of an estimator $\hat{\eta}$ for η, we consider its quadratic **risk** $q^{-1}E|\hat{\eta} - \eta|^2$.

The risk of $\hat{\eta}_{LS}$ is σ^2. For Gaussian e, Stein (1956) proved existence of estimators with uniformly smaller risk if $q \geq 3$.

MULTIPLY PENALIZED LEAST SQUARES

For $1 \leq s \leq d$, let $t_s \in [0,1]$ and let Q_s be a $p \times p$ symmetric psd matrix with $\rho(Q_s) = \sup_{x \neq 0} \frac{|Q_s x|}{|x|} = 1$.

For $c > 0$ large, $\epsilon > 0$ very small, $t = \{t_s\}$, define

$$Q(t) = \epsilon I_p + c \sum_{s=1}^{d} t_s Q_s.$$

For every $t \in [0,1]^d$, the **penalized least squares** (PLS) estimator of m is

$$\begin{aligned} \hat{m}(t) &= \mathrm{argmin}_{m \in R^p} [|y - CDm|^2 + m' Q(t) m] \\ &= [D'C'CD + Q(t)]^{-1} D'C'y. \end{aligned}$$

This yields the additional estimators

$$\hat{\eta}(t) = CD\hat{m}(t) = CD[D'C'CD + Q(t)]^{-1} D'C'y$$

$$\hat{m}_D(t) = C^+ \hat{\eta}(t) = D\hat{m}(t).$$

Choice of penalty weights t and matrices $\{Q_s\}$?

BAYES FOMULATION

Suppose that

$$y|m \sim N(CDm, \sigma^2 I_n), \qquad m \sim N(0, \sigma^2 Q^{-1}(t)).$$

The implied **candidate** Bayes estimators (quadratic loss) are

$$\hat{m}(t) = [D'C'CD + Q(t)]^{-1} D'C'y,$$

which does imputation, and

$$\begin{aligned}\hat{m}_D(t) &= D[D'C'CD + Q(t)]^{-1} D'C'y \\ &= [C'C + (DQ^{-1}(t)D')^{-1}]^{-1} C'y\end{aligned}$$

$$\begin{aligned}\hat{\eta}(t) &= CD[D'C'CD + Q^{-1}(t)] D'C'y \\ &= C[C'C + (DQ^{-1}D')^{-1}]^{-1} C'y.\end{aligned}$$

These candidate Bayes estimators for a Gaussian data model **coincide** with the candidate PLS estimators.

RISK. Let $R = C'C$, $U = CR^{-1/2}$,

$S(t) = [I_q + V(t)]^{-1}$, and $V(t) = R^{-1/2}(DQ^{-1}(t)D')^{-1}R^{-1/2}$.

Then

$$\hat{\eta}(t) = US(t)U'y$$

Let $T(t) = S^2(t)$, $\bar{T}(t) = [I_q - S(t)]^2$ and $\xi = U'm$.

The **risk** of $\hat{\eta}(t)$, calculated under the saturated model, is

$$r(t) = q^{-1}E|\hat{\eta}(t) - \eta|^2 = q^{-1}[\sigma^2 \operatorname{tr}\{T(t)\} + \operatorname{tr}\{\bar{T}(t)\xi\xi'\}].$$

ESTIMATED RISK. Let $z = U'y$ and let $\hat{\sigma}^2$ estimate σ^2. The **estimated risk** of $\hat{\eta}(t)$ is

$$\hat{r}(t) = q^{-1}[\hat{\sigma}^2 \operatorname{tr}\{T(t)\} + \operatorname{tr}\{\bar{T}(t)(zz' - \hat{\sigma}^2 I_q)\}].$$

(cf. Mallows 1973, Stein 1981). We will use estimated risk as surrogate for risk, seeking theoretical justification.

ADAPTIVE CHOICE OF PENALTY WEIGHTS

Fix penalty matrices $\{Q_s\}$. Picking penalty weights t to minimize estimated risk yields the **adaptive PLS estimator**

$$\hat{\eta}(\hat{t}), \qquad \hat{t} = \mathrm{argmin}_{t \in [0,1]^d} \, \hat{r}(t).$$

Correspondingly, we may estimate m_D by $\hat{m}_D(\hat{t})$ and m by $\hat{m}(\hat{t})$.

Note. Without additional model assumptions, we cannot estimate the risk of $\hat{m}(\hat{t})$ at factor level combinations that lack data. The extrapolation of $\hat{m}_D(\hat{t})$ to $\hat{m}(\hat{t})$ is a what-if experiment that reveals the regression function implicit in the adaptive PLS fit.

ASYMPTOTICS (cf. Beran 2007)

Limits as $q \to \infty$ support asymptotic trustworthiness of this strategy under the saturated model. These results assess the estimator at covariate-level combinations where observations are available.

Fact 1. *Assume that, for every finite* $a > 0$ *and* $\sigma^2 > 0$,

$$\lim_{q\to\infty} \sup_{q^{-1}|\eta|^2 \le a} E|\hat{\sigma}^2 - \sigma^2| = 0.$$

Let $W(t)$ *denote either the loss* $q^{-1}|\hat{\eta}(t) - \eta|^2$ *or the estimated risk* $\hat{r}(t)$ *of candidate estimator* $\hat{\eta}(t)$. *Then, for every finite* $c > 0$, $a > 0$, *and* $\sigma^2 > 0$,

$$\lim_{q\to\infty} \sup_{q^{-1}|\eta|^2 \le a} E[\sup_{t\in[0,1]^d} |W(t) - r(t)|] = 0.$$

In other words, the loss and estimated risk of $\hat{\eta}(t)$ both converge to the risk function $r(t)$ **uniformly** in t.

Implied by Fact 1 is

Fact 2. *Assume that, for every finite* $a > 0$ *and* $\sigma^2 > 0$,

$$\lim_{q\to\infty} \sup_{q^{-1}|\eta|^2 \le a} E|\hat{\sigma}^2 - \sigma^2| = 0.$$

Then, for every finite $c > 0$, $a > 0$, *and* $\sigma^2 > 0$,

$$\lim_{q\to\infty} \sup_{q^{-1}|\eta|^2 \le a} |q^{-1} E|\hat{\eta}(\hat{t}) - \eta|^2 - r(\tilde{t})| = 0.$$

where $\tilde{t} = \operatorname{argmin}_{t\in[0,1]^d} r(t)$.

Moreover,

$$\lim_{q\to\infty} \sup_{q^{-1}|\eta|^2 \le a} E|\hat{r}(\hat{t}) - r(\tilde{t})| = 0.$$

In other words, the risk and estimated risk of the empirically best adaptive estimator $\hat{\eta}(\hat{t})$ both converge to the oracle risk $r(\tilde{t}) = \min_{t\in[0,1]^d} r(t) \le \sigma^2$.

TENSOR-PRODUCT PENALTY MATRICES

For $r = 1, 2$, let

$$u_r = p_r^{-1/2}(1, 1, \ldots, 1)', \qquad J_r = u_r u_r', \qquad H_r = I_{p_r} - J_r.$$

J_r and H_r are mutually orthogonal projections such that $I_{p_r} = J_r + H_r$. Hence the **ANOVA decomposition** of m:

$$m = (I_{p_2} \otimes I_{p_1})m = P_0 m + P_1 m + P_2 m + P_{12} m,$$

where $P_0 = J_2 \otimes J_1$, $P_1 = J_2 \otimes H_1$, $P_2 = H_2 \otimes J_1$, and $P_3 = H_2 \otimes H_1$.

We construct penalty matrix Q_s to act only on $P_s m$:

$$m' Q(t) m = \epsilon |m|^2 + c \sum_s t_s (P_s m)' Q_s (P_s m).$$

Then $\hat{\eta}(t)$ penalizes departures in the ANOVA components from attributes determined by the $\{Q_s\}$. The nominal or ordinal character of each factor is taken into account.

Construction of Q_s.

Let A_r be an **annihilator** for factor r: $A_r u_r = 0$.

Let

$$Q_1 = u_2 u_2' \otimes A_1' A_1, \quad Q_2 = A_2' A_2 \otimes u_1 u_1',$$
$$Q_3 = A_2' A_2 \otimes A_1' A_1.$$

Motivating this definition are:

$$m' Q_1 m = |(u_2' \otimes A_1) m|^2, \quad m' Q_2 m = |(A_2 \otimes u_1') m|^2,$$
$$m' Q_3 m = |(A_2 \otimes A_1) m|^2.$$

Note that$P_{s_2} Q_{s_1} = Q_{s_1} P_{s_2} = 0$ if $s_1 \neq s_2$.

Hence $m' Q_s m = (P_s m)' Q_s (P_s m)$ as sought.

Both factors nominal. Because the factor levels are labels, $\hat{\eta}(t)$ should be invariant under such permutations of the levels. Let $A_r = H_r$ for $r = 1, 2$.

Both factors ordinal. Suppose that both sets of factor levels are arranged in decreasing order. Construct annihilator A to penalize departures from conjectured smoothness.
For instance, choose A_r to penalize departures from local polynomial behavior of degree $d - 1$. The normalized d-th difference operator does this for equally spaced factor levels and generalizes to handle unequally spaced factor levels.

First factor nominal, second factor ordinal. Mix suitably the preceding choices of A_r.

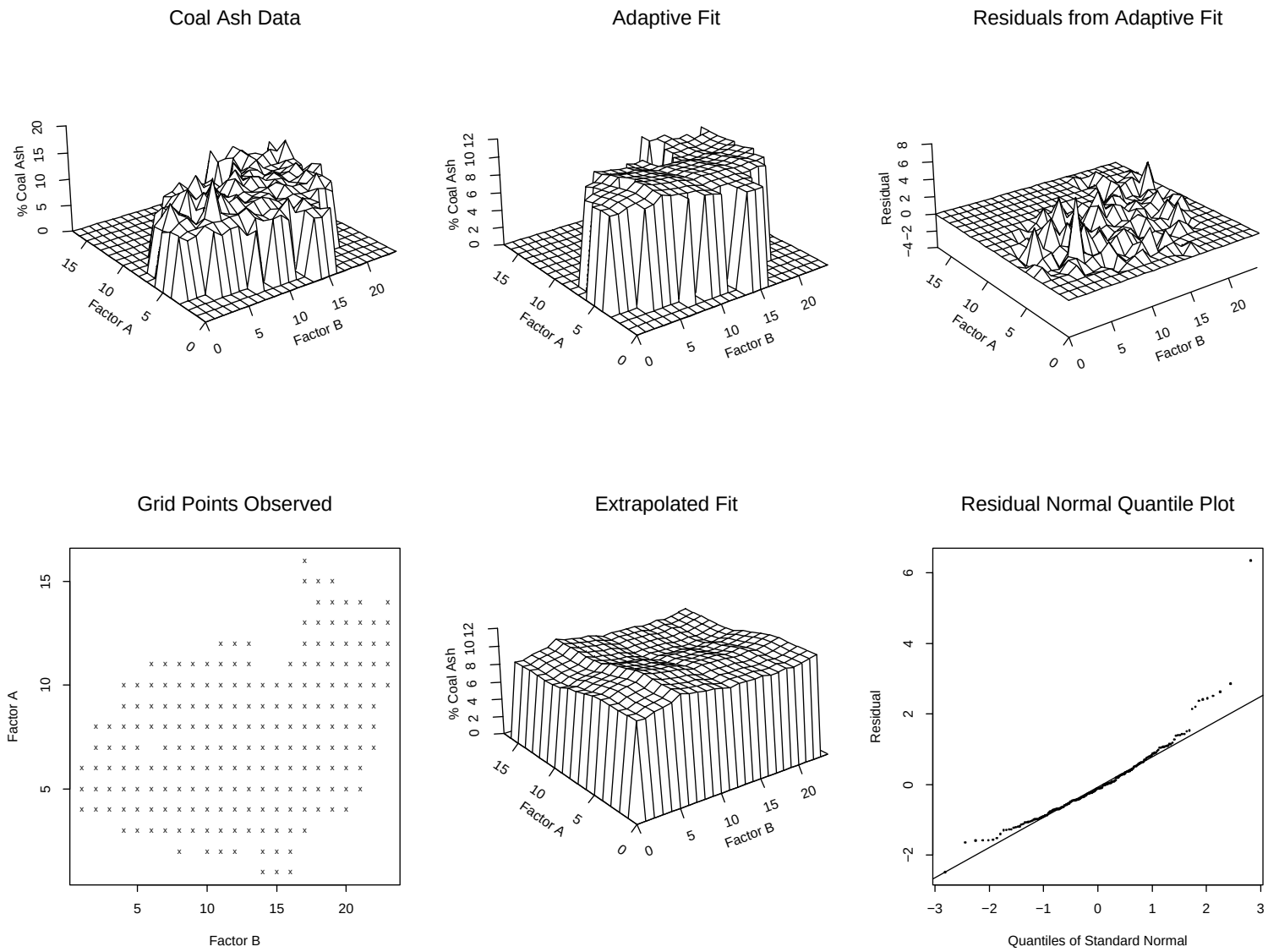
Coal Ash Data. Both geographical factors are ordinal with $p_1 = 23$ and $p_2 = 16$. One coal ash measurement is made at each of $q = 208$ locations.

Set $\epsilon = 10^{-7}$, $c = 10^4$, and $A_1 = A_2 =$ the first difference operator.

The variance estimate $\hat{\sigma}^2 = 1.038$ uses first differences in rows and columns. It assumes slow variation in the means.

The minimal estimated risk is $\hat{r}(\hat{t}) = .117$, achieved at $\hat{t} = (.000349, .000217, 1.)$

The estimated risk of the adaptive estimator is **one ninth** that of the least squares fit. The adaptive fit is nearly additive.



The **Coal Ash data**, its factor-level grid, the adaptive PLS estimate $\hat{m}_D(\hat{t})$, its extrapolation $\hat{m}(\hat{t})$, and a residual plot.

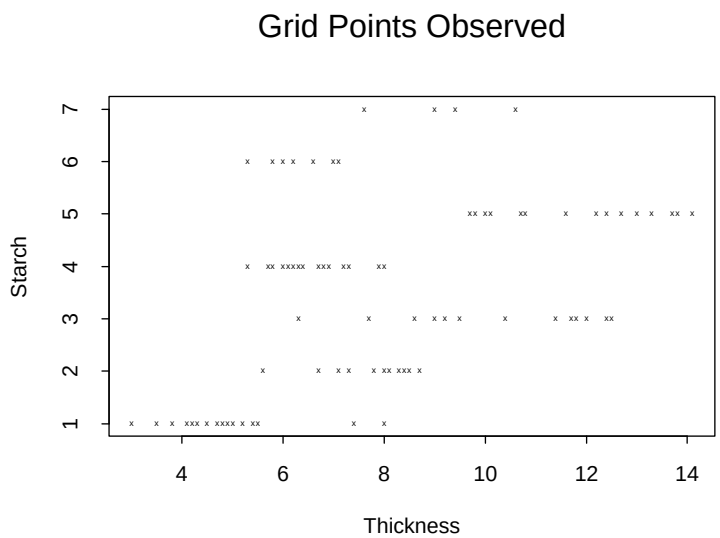
Starch Data. Starch type is nominal with $p_1 = 7$. Thickness is ordinal with $p_2 = 69$. One or more strength measurements are made at $q = 81$ factor-level pairs.

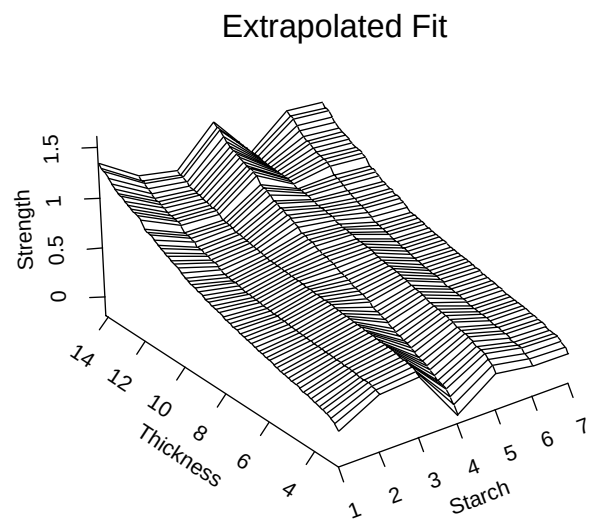
Set $\epsilon = 10^{-7}$, $c = 10^5$, $A_1 = H_1$, and $A_2 =$ the generalized second difference operator.

The variance estimate $\hat{\sigma}^2 = 13976$ is based on straight line fits to each starch.

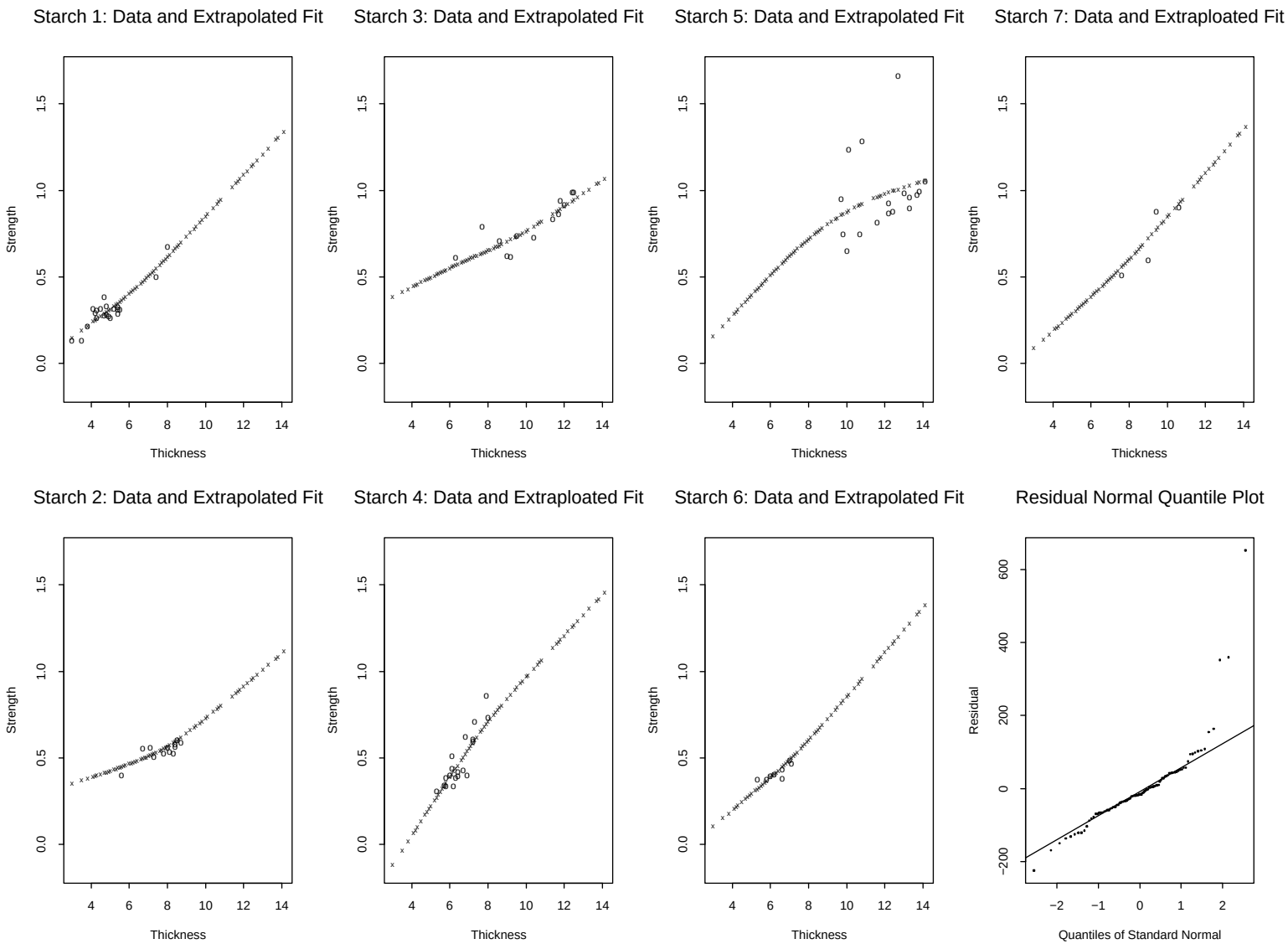
The minimal estimated risk is $\hat{r}(\hat{t}) = 4116$, achieved at $\hat{t} = (1., .7707 \times 10^{-6}, 1.)$

The estimated risk of the adaptive estimator is **one third** that of the least squares fit. Note the curious data for starch five!





The observed factor-level grid of the **Starch data** and the extrapolated adaptive PLS estimate $\hat{m}(\hat{t})$ of mean breaking strengths.



The **Starch data** (small o) by starch type, the extrapolated adaptive PLS estimate $\hat{m}(\hat{t})$ (small x), and a residual plot.

HISTORY OF A SPECIAL CASE

Consider PLS candidate estimators of m in $y = Cm + e$ when

- the design is balanced and complete: $C'C = n_0 I_p$;
- penalty matrix $Q(t) = \sum_{s=1}^{d} t_s P_s$, where the $\{P_s\}$ are mutually orthogonal, symmetric, idempotent, $p \times p$ matrices with $\sum_{s=1}^{d} P_s = I_p$ and $t_s \geq 0$. E.g. projections underlying ANOVA.

The **LS estimator** of m is $\hat{m}_{ls} = n_0^{-1} C'y$.

For each t, the **PLS estimator** of m is

$$\hat{m}(t) = \mathrm{argmin}_{m \in R^p}[|y - Cm|^2 + m'Q(t)m]$$
$$= [n_0 I_p + Q(t)]^{-1} C'y = \sum_{s=1}^{d} (n_0 + t_s)^{-1} n_0 P_s \hat{m}_{ls},$$

Reparametrize and enlarge the candidate estimator class to

$$\hat{m}(a) = \sum_{s=1}^{d} a_s P_s \hat{m}_{ls}$$

where $a = (a_1, \ldots, a_d) \in [0, 1]^s$.

Estimated Risk. Let

$$\hat{\tau}_s = (n_0 p)^{-1} \hat{\sigma}^2 \operatorname{tr}(P_s), \qquad \hat{w}_s = [p^{-1} |P_s \hat{m}_{ls}|^2 - \hat{\tau}_s]_+.$$

By direct argument, the estimated risk of $\hat{m}(a)$ is

$$\hat{r}(a) = \sum_{s=1}^{d} [\hat{\tau}_s a_s^2 + \hat{w}_s (1 - a_s)^2].$$

Adaptive Estimator. The $\hat{a} \in [0,1]^d$ that minimizes estimated risk has components

$$\hat{a}_s = \hat{w}_s (\hat{\tau}_s + \hat{w}_s)^{-1} = [1 - p\hat{\tau}_s / |P_s \hat{m}_{ls}|^2]_+.$$

The adaptive estimator of m is thus

$$\hat{m}(\hat{a}) = \sum_{s=1}^{d} [1 - p\hat{\tau}_s / |P_s \hat{m}_{ls}|^2]_+ P_s \hat{m}_{ls}.$$

Remark. Stein (1966) already obtained a small p refinement of this multiple shrinkage estimator through an exact study of the quadratic risk when the errors e are Gaussian iid

Though highly effective in balanced, complete ANOVA models (for instance), his estimator remains widely unused.

EXTENSIONS AND LINKS

Larger Families of Penalty Matrices

For the one-way layout, we have considered adaptive PLS over the penalty matrix family $Q(t) = \epsilon I_p + ctAA', \qquad t \in [0, 1]$, where A is an annihilator.

To enlarge this family, replace A by $A(\theta) = \sum_{i=1}^{g} \theta_i A_i$, where the $\{A_i\}$ are annihilators, $\theta = (\theta_1, \ldots, \theta_g)$, and $|\theta| = 1$.

- Adaptation by minimizing estimated risk over t and θ works asymptotically (symmetric linear estimators, Beran 2006).
- On a continuous factor domain, the annihilators may be derivatives as in Heckman and Ramsay 2000.
- The enlargement may be extended to multi-way layouts.

Multi-way Layouts with Multivariate Responses

The model is $Y = CDM + V\Sigma^{1/2}$, where

- the rows of $n \times h$ matrix Y are h-variate responses;
- the $p \times h$ mean matrix M is unknown;
- the $h \times h$ covariance matrix Σ may be unknown;
- data-incidence matrix C is $n \times p$;
- the elements of $n \times h$ are iid with means 0, variance 1, and finite 4-th moment.

When $\Sigma = I_h$, the candidate PLS estimator $\hat{M}(N)$ minimizes

$$|Y - CDM|^2 + \sum_{s=1}^{d} |Q_s^{1/2} M N_s^{1/2}|^2,$$

where the $\{N_s\}$ are $h \times h$ psd **affine penalty weights** while the $\{Q_s\}$ are $p \times p$ psd **penalty matrices** as in the univariate case (cf. Beran 2008). Previous theory extends to this and to general Σ. Efron-Morris (1972) estimator is a special case!

SUMMARY

The Core Ideas

- Interactions are to be expected in multi-way layouts or multi-covariate regression.
- Replication matters for trustworthy estimation of variability (cf. R. A. Fisher's development of ANOVA).
- Folk-practice advocates fitting simplified models. Done mindfully, regularized fits reduce risk (cf. C. Stein on shrinkage).

Now

- Adaptive PLS methodologies for estimating many means have a variety of sources. Study of these methodologies in terms of the core ideas sharpens their effectiveness and deepens our understanding.