IMS Collections
Contemporary Developments in Bayesian Analysis and Statistical Decision Theory: A Festschrift for William E. Strawderman
Vol. 8 (2012) 64–77

DOI: 10.1214/11-IMSCOLL805

Reducing data nonconformity in linear models

Andrew L. Rukhin

National Institute of Standards and Technology, NIST

Abstract: Procedures to reduce nonconformity in interlaboratory studies by shrinking multivariate data toward a consensus matrix-weighted mean are discussed. Some of them are shown to have a smaller quadratic risk than the ordinary least squares rule. Bayes procedures and shrinkage estimators in random effects models are also considered. The results are illustrated by an example of collaborative studies.

Contents

1 Introduction and summary . . . 64
2 Minimaxity results . . . 66
3 Random effects models . . . 67
4 Bayesian procedure . . . 68
5 Silver vapor pressure study . . . 70
6 Conclusions . . . 71
7 Appendix . . . 72
7.1 Proof of Theorem 2.1 . . . 72
7.2 Proof of Theorem 2.2 . . . 75
References . . . 76

1. Introduction and summary

Consider the situation where a consensus (or reference) vector value is to be established by combining information from several, say, p, studies or laboratories. This is a classical problem of meta-analysis which appears in many diverse fields. See Hedges and Olkin (1985) or Hartung, Knapp and Sinha (2008),

Assume that the n_i-dimensional data from the i-th laboratory satisfies the linear model,

$$Y_i = B_i\theta_i + e_i. \tag{1.1}$$

Here $i = 1, \ldots, p$ indexes the laboratories, B_i is the i-th study design matrix of size $n_i \times q, n_i \geq q$, and of rank q, which is the dimension of the parameter θ_i. It will be assumed that the errors e_i are independent and normally distributed,

Statistical Engineering Division, ITL, National Institute of Standards and Technology, NIST, 100 Bureau Drive, Gaithersburg, MD 20899-0001 USA, e-mail: andrew.rukhin@nist.gov

AMS 2000 subject classifications: Primary 62C20; secondary 62F10, 62P30

Keywords and phrases: Birge ratio, consensus value, jackknife estimator, matrix weighted means, meta-analysis, normal mean, shrinkage estimator

$e_i \sim N_{n_i}(0, \sigma_i^2 I)$. The parameters θ_i are expected to be "close" to the same q-dimensional vector θ.

We define the consensus value of individual parameters as a matrix weighted average (discussed in more detail later),

$$\hat{\theta} = \left(\sum_j \frac{B_j^T B_j}{\sigma_j^2} \right)^{-1} \sum_i \frac{B_i^T B_i}{\sigma_i^2} \theta_i. \tag{1.2}$$

Our problem is to determine a procedure to evaluate (1.2) by using data Y_i, and to estimate the associated uncertainties, i.e., the covariance matrix of the estimator.

Let $X_i = (B_i^T B_i)^{-1} B_i^T Y_i$ be the classical least squares estimator of θ_i. The covariance matrix of X_i is $\sigma_i^2 (B_i^T B_i)^{-1}$. The traditional unbiased estimate of the laboratory variance σ_i^2 is

$$v_i^2 = (n_i - q)^{-1} (Y_i - B_i X_i)^T (Y_i - B_i X_i),$$

$i = 1, \ldots, p$. However, the best equivariant estimator of σ_i^2 under the quadratic loss is $s_i^2 = (n_i - q) v_i^2 / (n_i - q + 2)$, so that $s_i^2 / \sigma_i^2 \sim \chi^2_{n_i - q} / (n_i - q + 2)$. The relative improvement upon v_i^2 is noticeable for small to medium values of $n_i - q$ (Rukhin, 1987).

With $S = (s_1^2, \ldots, s_p^2)$, let

$$\tilde{\theta} = \tilde{\theta}(X, S) = \left(\sum_{i=1}^p s_i^{-2} B_i^T B_i \right)^{-1} \sum_{i=1}^p s_i^{-2} B_i^T B_i X_i, \tag{1.3}$$

be a matrix weighted mean which can be viewed as a plug-in estimator of (1.2). When $q = 1$, this statistic is closely related to the often employed in collaborative studies Graybill-Deal estimator.

Commonly, the full data set $Y = (Y_1^T, \ldots, Y_p^T)^T$ does not comply with the model in which $\theta_i \equiv \hat{\theta}$, as the datum Y_i is influenced by systematic, laboratory specific errors. In this case one needs a procedure to reduce or to remove this non-conformity.

To this end, we first look at Stein-type minimax estimators. When estimating the pq-dimensional vector $(\theta_1^T, \ldots, \theta_p^T)^T$, they shrink towards q-dimensional subspace $\mathcal{V}$ formed by vectors $V\hat{\theta} = (\hat{\theta}^T, \ldots, \hat{\theta}^T)^T$. With our goal in mind, for $n = \sum_i n_i$ it is more natural to think about estimation of the n-dimensional vector formed by stacked vectors, $B_i \theta_i, i = 1, \ldots, p$. In this case shrinkage of the data/estimator Y is performed toward $B\tilde{\theta}$, where $B = (B_1^T, \ldots, B_p^T)^T$, and $\tilde{\theta}$ is the estimator of $\hat{\theta}$ defined by (1.3). These procedures are known to have a smaller mean squared error than $X = (X_1^T, \ldots, X_p^T)^T$.

A nonconformity removal procedure shrinks the data Y toward $B\tilde{\theta}$,

$$\delta = B\tilde{\theta} + \left(1 - \frac{a}{\sum_j ||Y_j - B_j \tilde{\theta}||^2 / s_j^2} \right) (Y - B\tilde{\theta})$$

$$= \left(1 - \frac{a}{\sum_j ||Y_j - B_j \tilde{\theta}||^2 / s_j^2} \right) Y + \frac{a}{\sum_j ||Y_j - B_j \tilde{\theta}||^2 / s_j^2} B\tilde{\theta}, \tag{1.4}$$

where a is a positive constant. Thus, δ is a linear (convex if $a \leq \sum_j ||Y_j - B_j \tilde{\theta}||^2 / s_j^2$) combination of Y and $B\tilde{\theta}$ with $\tilde{\theta}$ defined as in (1.3), i.e., of the original data and the predictor $B\tilde{\theta}$ of this data from the linear model in which $\theta_i \equiv \hat{\theta}$.

Theorem 2.1 shows that the procedure (1.4) dominates Y (and therefore is minimax) under invariant quadratic loss when $a \leq 2(n-q-2.)$ Thus, our approach assumes that $n \geq q+3$. According to a dual result, shifting the least squares estimator $X = (X_1^T, \dots, X_p^T)^T$ toward $V\tilde{\theta} = (\tilde{\theta}^T, \dots, \tilde{\theta}^T)^T$,

$$\psi = V\tilde{\theta} + \left(1 - \frac{b}{\sum_j (X_j - \tilde{\theta})^T (B_j^T B_j)(X_j - \tilde{\theta})/s_j^2}\right)(X - V\tilde{\theta}), \tag{1.5}$$

also leads to a minimax rule, but the loss function is different and the condition on the constant b is now $b \leq 2(pq - q - 2)$. In this case the least squares estimator of the vector $(\theta_1, \dots, \theta_p)$ is replaced by its linear combination with the estimate when in the linear model $\theta_i \equiv \hat{\theta}$. A body of literature deals with a given covariance matrix of Y, or shrinkage applied toward a fixed subspace. In our problem the true metric is unknown (as it depends on σ's), so the projection onto a subspace is performed after an estimated distance. Shrinking toward a subspace chosen from variable selection was investigated by Lee and Birkes (1994).

Strawderman and Rukhin (2010) derived a minimaxity result when $q = 1$, with unknown covariance matrix so that the metric is estimated. The estimators suggested there reduce nonconformity in interlaboratory studies by shrinking data toward a scalar consensus mean. Here we deal with the multivariate version of the problem. As is shown in section 2, the positive part version of Stein estimator can be used for nonconformity reduction. However, in many applications (1.4) may not sufficiently pull all Y's together. As an alternative, in section 3 we suggest a Bayes procedure as well as estimators which take into account a random effects component in a more general version of (1.1) discussed in section 7.1. All proofs are collected in the Appendix.

2. Minimaxity results

We start with estimators of the form (1.4) for model (1.1).

Theorem 2.1. *Suppose $Y_i \sim N_{n_i}(B_i\theta_i, \sigma_i^2 I)$ and an independent s_i^2, $s_i^2/\sigma_i^2 \sim \chi^2_{\nu_i}/(\nu_i + 2), i = 1, \dots, p, \min n_i \geq q$. Then the estimator (1.4) of the vector $\theta = (\theta_1^T, \dots, \theta_p^T)^T$ dominates $\delta^0 = Y$ under the weighted quadratic loss $L(\theta, \sigma, \delta) = \sum_i ||\delta_i - B_i\theta_i||^2/\sigma_i^2$ provided that $0 \leq a \leq 2(n-q-2)$. If the function g of positive argument takes values in the unit interval, $0 \leq g \leq 1$ and is monotonically non-decreasing, the same result holds for estimators of the form*

$$\delta^g = B\tilde{\theta} + \left(1 - \frac{ag(\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2)}{\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2}\right)(Y - B\tilde{\theta})$$

$$= Y - \frac{ag(\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2)}{\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2}(Y - B\tilde{\theta}). \tag{2.1}$$

In particular, the choice of $g(z) = \min(z/a, 1), z \geq 0$, shows that the positive-part estimator

$$\delta^+ = B\tilde{\theta} + \left(1 - \frac{a}{\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2}\right)_+ (Y - B\tilde{\theta}), \tag{2.2}$$

$x_+ = \max(x, 0)$, is minimax. The domination result in Theorem 2.1 implies that $\delta = (\delta_1^T, \ldots, \delta_p^T)^T$ and similarly defined δ^+ are minimax estimators, i.e., for all θ, σ, $EL(\theta, \sigma, \delta) \leq \sup_{\theta,\sigma} EL(\theta, \sigma, X)$. In the context of discussion in section 1, $\nu_i = n_i - q$.

Theorem 2.2. *Suppose X_i has a $N_q(\theta_i, \sigma_i^2(B_i^T B_i)^{-1})$ distribution and an independent s_i^2, $i = 1, \ldots, p$, has the same meaning as in Theorem 2.1. Then the estimator (1.5) of the vector $\theta = (\theta_1^T, \ldots, \theta_p^T)^T$ dominates $\psi_0 = X$ under the quadratic loss $L(\theta, \sigma, \psi) = \sum_i (\psi_i - \theta_i)^T (B_i^T B_i)(\psi_i - \theta_i)/\sigma_i^2$ provided that $0 \leq b \leq 2(pq - q - 2)$.*

The same result holds for estimators of the form

$$\psi^+ = V\tilde{\theta} + \left(1 - \frac{b}{\sum_j (X_j - \tilde{\theta})^T (B_j^T B_j)(X_j - \tilde{\theta})/s_j^2}\right)_+ (X - V\tilde{\theta}), \tag{2.3}$$

and for the counterpart of (2.1) where the function g has the same properties as in Theorem 2.1.

Since

$$\sum_j \frac{||\delta_j^+ - B_j\tilde{\theta}||^2}{s_j^2} \leq \sum_j \frac{||\delta_j - B_j\tilde{\theta}||^2}{s_j^2} \leq \sum_j \frac{||Y_j - B_j\tilde{\theta}||^2}{s_j^2},$$

δ_j^+ are always closer to $B_j\tilde{\theta}$ than the original data. $(n - pq)^{-1} \sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2$ extends to a multivariate setting the concept of *Birge ratio*, which is commonly used in metrology for testing goodness-of-fit. The above shows that the Birge ratio evaluated for the original data Y is always larger than the Birge ratio for δ^+.

For a q-dimensional vector t, we put $||t||_j^2 = t^T B_j^T B_j t$, $j = 1, \ldots, p$. Under this notation,

$$\sum_j \frac{||\psi_j^+ - \tilde{\theta}||_j^2}{s_j^2} \leq \sum_j \frac{||X_j - \tilde{\theta}||_j^2}{s_j^2} \leq \sum_j \frac{||Y_j - B_j\tilde{\theta}||^2}{s_j^2}.$$

The last inequality here is due to the fact that $I - B_i(B_i^T B_i)^{-1} B_i^T$, being a projection matrix, is non-negative definite.

3. Random effects models

Meta-analysis has in its arsenal more versatile statistics than the Graybill-Deal estimator (1.3). One of them introduced by DerSimonian and Laird (1986) to estimate common scalar mean, turned out to be very popular in the meta-analysis of multicenter clinical trials.

A linear model extension of this procedure can be based on the so-called random effects version of (1.1),

$$Y_i = B_i(\theta + \ell_i) + e_i, \tag{3.1}$$

with B_i having the same meaning as in (1.1) (see Rukhin, 2011). Thus $\theta_i = \theta + \ell_i$, where θ is the unknown consensus value, and ℓ_i represents a random study effect which is independent of the errors e_i and is normally distributed with zero mean and some covariance matrix Ξ. Then $Cov(Y_i) = \sigma_i^2 I + B_i \Xi B_i^T$, and as in section 1, $s_i^2 = (n_i - q + 2)^{-1} ||Y_i - B_i X_i||^2 \sim \tau_i^2 \chi_{\nu_i}^2/(\nu_i + 2)$, are independent of X_i, minimum variance equivariant estimators of τ_i^2.

When all variances are known, the best (in terms of the mean squared error) unbiased estimator of $\hat{\theta}$ in the model (3.1) is a matrix weighted mean,

$$\tilde{\theta}_W = \left(\sum_{i=1}^{p} W_i\right)^{-1} \sum_{i=1}^{p} W_i X_i, \tag{3.2}$$

with $W_i = (B_i \Xi B_i^T + \sigma_i^2 I)^{-1}, i = 1, \dots, p$. When $\sigma_i^2 \equiv \sigma^2, B_i \equiv B$, $\tilde{\theta}_W$ reduces to the sample mean $\bar{X}$, and with $W_i = s_i^{-2} B_i B_i^T, i = 1, \dots, p$, (1.3) takes the form (3.2) as well.

It makes sense to estimate the within-lab variances σ_i^2 by the available estimates s_i^2. Put

$$C = \sum_i s_i^{-2} (B_i^T B_i)^{1/2} (X_i - \tilde{\theta}_0)(X_i - \tilde{\theta}_0)^T (B_i^T B_i)^{1/2} - pI$$

$$+ \sum_i s_i^{-2} (B_i^T B_i)^{1/2} \left(\sum_k s_k^{-2} B_k^T B_k\right)^{-1} (B_i^T B_i)^{1/2},$$

With ω_i defined by (7.3), the moment-type equation

$$\sum_i s_i^{-2} (B_i^T B_i)^{1/2} \Big(I - \omega_i\Big) V \Big(I - \omega_i\Big)^T (B_i^T B_i)^{1/2}$$

$$+ \sum_i s_i^{-2} (B_i^T B_i)^{1/2} \left(\sum_{j: j \neq i} \omega_j V \omega_j^T\right) (B_i^T B_i)^{1/2} = C, \tag{3.3}$$

allows to determine a symmetric matrix solution V (an estimator of Ξ). See Rukhin (2007) for details. We take $V_{DL} = V_+$ to be the positive part of V, i.e., let V_{DL} have the same spectral decomposition as V, with its eigenvalues being positive parts of V eigenvalues. The matrix weights of the estimator $\tilde{\theta}_{DL}$ then have the form

$$W_i = (B_i V_{DL} B_i^T + s_i^2 I)^{-1}. \tag{3.4}$$

Many practical examples and simulation results suggest that the DerSimonian-Laird estimator is a better shrinkage center than $\tilde{\theta}$. While our focus so far was on reducing nonconformity among the estimated laboratory means via shrinkage methods, one of the main goals of interlaboratory studies is to establish the consensus value. For this purpose the DerSimonian-Laird estimator $\tilde{\theta}_{DL}$ determined by the matrix weights in (3.4) may be preferable to (1.3), although it is unknown if the positive-part estimator, which shrinks toward $\tilde{\theta}_{DL}$, is minimax.

4. Bayesian procedure

We look now at Bayes procedures to reduce nonconformity of data by using a (generalized) prior distribution with density $\pi(\theta_1, \dots, \theta_p)$ such that

$$\log \pi(\theta_1, \dots, \theta_p | \sigma_1^2, \dots, \sigma_p^2) = -\frac{\beta}{2} \sum_i \frac{||\theta_i - \hat{\theta}||_i^2}{\sigma_i^2}. \tag{4.1}$$

Here $\hat{\theta}$ is the parametric consensus value (1.2), and the prior distribution is concentrated around the subspace of vectors $\mathcal{V}$. The degree of this concentration is measured by a positive parameter β.

For fixed $\sigma_1^2, \ldots, \sigma_p^2$, we evaluate the posterior mode (which in this case coincides with the posterior mean),

$$\arg\min_{\theta_1,\ldots,\theta_p} \Big[\sum_i \frac{||Y_i - B_i\theta_i||^2}{\sigma_i^2} + \beta \sum_i \frac{||\theta_i - \hat{\theta}||_i^2}{\sigma_i^2} \Big]. \tag{4.2}$$

Let the block diagonal matrix C be formed by matrices $B_i^T B_i/\sigma_i^2$, $i = 1, \ldots, p$, and put $A^T = (\sum_j B_j^T B_j/\sigma_j^2)^{-1/2}(B_1^T B_1/\sigma_1^2, \ldots, B_p^T B_p/\sigma_p^2)^T$, $D = (B_1^T Y_1/\sigma_1^2, \ldots, B_p^T Y_p/\sigma_p^2)^T$. Then (4.2) can be written as

$$\theta^T[(1+\beta)C - \beta AA^T]\theta - 2D^T\theta + \sum_j \frac{Y_j^T Y_j}{\sigma_j^2}.$$

It follows that the Bayes estimator of $(\theta_1, \ldots, \theta_p)$ is found as the solution to

$$[(1+\beta)C - \beta AA^T]\theta = D.$$

Since $A^T C^{-1} A = I$,

$$[(1+\beta)C - \beta AA^T]^{-1} = \frac{1}{1+\beta}C^{-1} + \frac{\beta}{(1+\beta)(1+2\beta)}C^{-1}AA^TC^{-1},$$

and the Bayes estimator of θ is

$$\tilde{\theta}_B = \frac{1}{1+\beta}X + \frac{\beta}{(1+\beta)(1+2\beta)}V\left(\sum_j \frac{B_j^T B_j}{\sigma_j^2}\right)^{-1}\left(\sum_j \frac{B_j^T Y_j}{\sigma_j^2}\right).$$

The minimum M in (4.2) is

$$M = \sum_j \frac{Y_j^T Y_j}{\sigma_j^2} - \frac{1}{1+\beta}\sum_j \frac{Y_j^T B_j (B_j^T B_j)^{-1} B_j^T Y_j}{\sigma_j^2}$$

$$-\frac{\beta}{(1+\beta)(1+2\beta)}\left(\sum_j \frac{B_j^T Y_j}{\sigma_j^2}\right)^T\left(\sum_j \frac{B_j^T B_j}{\sigma_j^2}\right)^{-1}\left(\sum_j \frac{B_j^T Y_j}{\sigma_j^2}\right),$$

and this function of $\sigma_1^2, \ldots, \sigma_p^2$ combined with the prior density is to be minimized.

For example, if σ's are assigned a noninformative density $\prod_i \sigma_i^{-\alpha}$, differentiation of M shows that the Bayes estimators of these parameters satisfy simultaneous equations,

$$(n_i + \alpha)\sigma_i^2 = Y_i^T Y_i - \frac{1}{1+\beta}Y_i^T B_i (B_i^T B_i)^{-1} B_i^T Y_i$$

$$-\frac{2\beta}{(1+\beta)(1+2\beta)}Y_i^T B_i\left(\sum_j \frac{B_j^T B_j}{\sigma_j^2}\right)^{-1}\left(\sum_j \frac{B_j^T Y_j}{\sigma_j^2}\right) + \frac{\beta}{(1+\beta)(1+2\beta)}$$

$$\times\left(\sum_j \frac{B_j^T Y_j}{\sigma_j^2}\right)^T\left(\sum_j \frac{B_j^T B_j}{\sigma_j^2}\right)^{-1} B_i^T B_i \left(\sum_j \frac{B_j^T B_j}{\sigma_j^2}\right)^{-1}\left(\sum_j \frac{B_j^T Y_j}{\sigma_j^2}\right),$$

$i = 1, \ldots, p$. These equations can be solved iteratively by using the initial values $\sigma_i^{(0)} = s_i$, and this was done for $\alpha = 1$ in the example discussed in the next section.

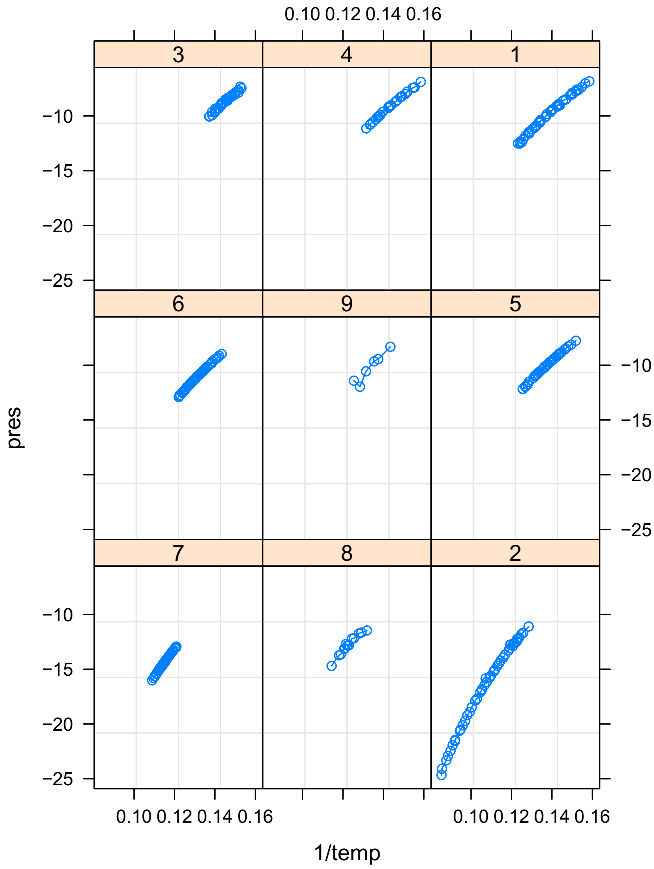

FIG 1. *Silver vapor pressure data.*

5. Silver vapor pressure study

In the silver vapor pressure study (Paule and Mandel, 1971) nine laboratories performed via different techniques measurements of silver vapor pressure P as a function of the absolute temperature T in the (individual for each laboratory) range from 800 to 1600K.

After the heat law, the logarithm of pressure must be a linear function of $1/T$, so we take the design matrix B_i to be formed by pairs $(1, 1/T_{ij}), j = 1, \ldots, n_i, i = 1, \ldots, 9$. A natural assumption is that the error variance depends only on the individual laboratory (and not on the temperature value). This study then fits the model (1.1) with $p = 9$ and $q = 2$. Figure 1 displays the data set.

There are a total of 304 different temperature points $T_{ij}, i = 1, \ldots, 9, j = 1, \ldots, n_i, n_1 + \cdots + n_9 = 304$ given in Paule and Mandel (1971), Table 4 which employs $1/T10^4$ in K^{-1} units. The results of one laboratory (# 9) portrayed in the center of Figure 1 seem to be dubious, so a procedure to remove this nonconformity is of interest.

Here are the estimates of the intercept θ_0 and the slope θ_1

	$\tilde{\theta}$	(1.3)	$\tilde{\theta}_B$	$\tilde{\theta}_{RL}$
θ_0	13.30	12.80	14.16	14.07 .
θ_1	−3.19	−3.13	−3.30	−3.28

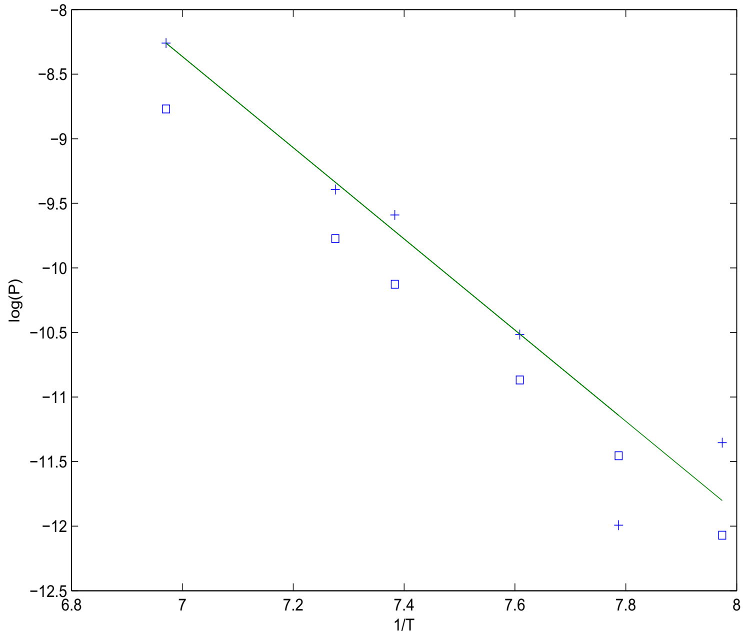

Fig 2. *Silver vapor pressure data for lab 9 (marked by +) and this data with nonconformity removed by (2.2) (marked by 'o') The dash-dotted line corresponds to (3.4).*

The restricted maximum likelihood estimator $\tilde{\theta}_{RL}$ was found from the R-language function *lme*, and its numerical evaluation was possible only after removal of the results of lab 9, as with the full data, an error message in *lme* function indicated false convergence.

The estimated within lab variance V_{DL} is

$$\begin{pmatrix} 1.62 & -0.72 \\ -0.72 & 0.09 \end{pmatrix}.$$

The procedure (2.2) with $a = 2(n-4) = 600$, $a(\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2)^{-1} = 0.004$, virtually leaves the original data intact. The Bayes rule with $\beta = 0.12$ pulls the data towards the line determined by the results of this lab. Larger values of β lead to Bayes estimators which pull all θ_i towards the origin, which shows some shortcomings of the prior (4.1). The DerSimonian-Laird estimator with weights (3.4) defines the consensus line which brings the data of lab 9 in better agreement with other labs data. See Figure 2.

6. Conclusions

Removal of data nonconformity is formulated here as a statistical estimation problem in a rather general context of linear models. Minimaxity of Stein-type procedures is established although the practical importance of this property is not clear. In fact, the matrix weighted means which do not satisfy conditions of Theorems 2.1 or 2.2 seem to do a better job in terms of reconciling the data. Clearly, further study of this challenging practical problem is desirable. In particular, it is of interest to determine if there is a prior distribution such that the Bayes estimator is a convex combination of X and its projection onto $\mathcal{V}$.

7. Appendix

7.1. Proof of Theorem 2.1

We give the proof only for the estimator (1.4). The proof for the general estimator (2.1) is similar. It uses the facts that $0 \le g^2(t) \le g(t)$ and $g'(t) \ge 0$, as discussed in Strawderman and Rukhin (2010), to show that the contribution to the risk difference of terms involving derivatives is negative.

The difference Δ between the risk functions of δ_0 and δ can be written as

$$\Delta = a^2 \sum_i E \frac{||Y_i - B_i\tilde{\theta}||^2}{\sigma_i^2[\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2]^2} - 2a \sum_i E \frac{(Y_i - B_i\theta_i)^T(Y_i - \tilde{\theta})}{\sigma_i^2 \sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2}$$

$$(7.1) \quad = a^2 \sum_i E \frac{s_i^4 ||Y_i - B_i\tilde{\theta}||^2/\sigma_i^2}{s_i^4[\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2]^2} - 2a \sum_i E \text{ div} \left(\frac{Y_i - \tilde{\theta}}{\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2} \right).$$

Here we have used the well known in statistical decision theory, Stein identity (e.g., Lehmann and Casella, 1998),

$$E \frac{(Y_i - B_i\theta_i)^T g_i(Y_i)}{\sigma_i^2} = E \sum_k \frac{\partial}{\partial Y_i(k)} g_i(Y) = E \text{ div}(g_i),$$

$Y_i = (Y_i(1), \ldots, Y_i(n_i))^T$. This formula holds provided that the n_i-dimensional vector function g_i is weakly differentiable and the integral in the left hand side exists.

The motivation for rewriting the first term in (7.1) is another useful identity, according to which under the same conditions on function h,

$$E \frac{s_i^4 h(S)}{\sigma_i^2} = E s_i^2 h(S) + 2(\nu_i + 2) E s_i^4 \frac{\partial}{\partial s_i^2} h(S).$$

Indeed by using this formula, with $h = s_i^{-4} ||Y_i - B_i\tilde{\theta}||^2 [\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2]^{-2}$, $i = 1, \ldots, p$, we can evaluate the first term in (7.1) as follows,

$$\Delta = a^2 \sum_i E \frac{||Y_i - B_i\tilde{\theta}||^2}{s_i^2[\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2]^2}$$

$$+2a^2 \sum_i (\nu_i + 2) E s_i^4 \frac{\partial}{\partial s_i^2} \frac{||Y_i - B_i\tilde{\theta}||^2}{s_i^4[\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2]^2}$$

$$-2a \sum_i E \text{ div} \left(\frac{Y_i - B_j\tilde{\theta}}{\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2} \right)$$

$$(7.2) \quad \le a^2 \sum_i E \frac{||Y_i - B_i\tilde{\theta}||^2/s_i^2}{[\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2]^2} - 2a \sum_i E \text{ div} \left(\frac{Y_i - B_i\tilde{\theta}}{\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2} \right),$$

since, as will be shown below, the term which involves $\partial/\partial s_i^2$ is non-positive.

Note that the first term in (7.2) is $a^2 E\left(\sum_i ||Y_i - B_i\tilde{\theta}||^2/s_i^2\right)^{-1}$. Define the normalized matrix weights

$$\omega_i = \Big(\sum_j B_j^T B_j/s_j^2\Big)^{-1} B_i^T B_i/s_i^2, \tag{7.3}$$

$\sum_i \omega_i = I$, and $n_i \times n_i$ matrices

$$J_i = I - B_i\Big(\sum_j s_j^{-2} B_j^T B_j\Big)^{-1} B_i^T/s_i^2, \tag{7.4}$$

$i = 1, \ldots, p$. Then $\tilde{\theta} = \sum_i \omega_i X_i = \left(\sum_j B_j^T B_j/s_j^2\right)^{-1} \sum B_i^T Y_i/s_i^2$, and $\sum_i B_i^T(Y_i - B_i\tilde{\theta})/s_i^2 = 0$. The second term in (7.2) is the expected value of

$$\sum_i \operatorname{div}\left(\frac{J_i Y_i - B_i \sum_{j\neq i}\omega_j (B_j^T B_j)^{-1} B_j^T Y_j}{\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2}\right)$$

$$= \sum_i \frac{n_i - \operatorname{tr}(\omega_i)}{\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2} - 2\sum_i \frac{(Y_i - B_i\tilde{\theta})^T(Y_i - B_i\tilde{\theta})/s_i^2}{(\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2)^2}$$

$$+2\sum_i \frac{\sum_k (Y_k - B_k\tilde{\theta})^T B_k\Big(\sum_j B_j^T B_j/s_j^2\Big)^{-1} B_i^T(Y_i - B_i\tilde{\theta})/s_i^2}{(\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2)^2}$$

$$= \frac{n - q - 2}{\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2}. \tag{7.5}$$

Hence, if $0 < a \le 2(n - q - 2)$,

$$\Delta \le E\frac{a^2 - 2(n-q-2)a}{\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2} < 0.$$

It remains to be shown that

$$\frac{\partial}{\partial s_i^2}\frac{||Y_i - B_i\tilde{\theta}||^2}{[s_i^2\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2]^2} \le 0.$$

For fixed i and j,

$$\frac{\partial}{\partial s_i^2}\omega_j = \frac{\omega_i}{s_i^2}(\omega_j - \delta_{ij}I),$$

so that

$$\frac{\partial}{\partial s_i^2}\tilde{\theta} = \frac{\omega_i}{s_i^2}(\tilde{\theta} - X_i),$$

$$\frac{\partial}{\partial s_i^2}||Y_j - B_j\tilde{\theta}||^2 = 2(Y_j - B_j\tilde{\theta})^T B_j\omega_i(X_i - \tilde{\theta})/s_i^2,$$

and

$$\frac{\partial}{\partial s_i^2}s_i^2\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2 = \sum_{j\neq i}||Y_j - B_j\tilde{\theta}||^2/s_j^2$$

$$+2\sum_j (Y_j - B_j\tilde{\theta})^T B_j \omega_i (X_i - \tilde{\theta})/s_j^2 = \sum_{j\neq i} ||Y_j - B_j\tilde{\theta}||^2/s_j^2.$$

Thus,

$$\frac{\partial}{\partial s_i^2} \frac{||Y_i - B_i\tilde{\theta}||^2}{[s_i^2 \sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2]^2} = \frac{2(Y_i - B_i\tilde{\theta})^T B_i \omega_i (X_i - \tilde{\theta})}{s_i^6 [\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2]^2}$$

$$-\frac{2||Y_i - B_i\tilde{\theta}||^2 \sum_{j\neq i} ||Y_j - B_j\tilde{\theta}||^2/s_j^2}{s_i^6 [\sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2]^3}.$$

Our goal is to prove that for a fixed i,

$$(Y_i - B_i\tilde{\theta})^T B_i \omega_i (X_i - \tilde{\theta}) \sum_j ||Y_j - B_j\tilde{\theta}||^2/s_j^2$$

$$-||Y_i - B_i\tilde{\theta}||^2 \sum_{j\neq i} ||Y_j - B_j\tilde{\theta}||^2/s_j^2 \leq 0,$$

or that

$$\frac{||Y_i - B_i\tilde{\theta}||^4}{s_i^2} \leq (Y_i - B_i\tilde{\theta})^T (Y_i - B_i\tilde{\theta} - B_i\omega_i(X_i - \tilde{\theta})) \sum_j \frac{||Y_j - B_j\tilde{\theta}||^2}{s_j^2}. \tag{7.6}$$

If J_i is defined by (7.4), then $J_i B_i = B_i(I - \omega_i)$, and

$$B_i \omega_i (X_i - \tilde{\theta}) = (I - J_i)(Y_i - B_i\tilde{\theta}).$$

It is easy to check that

$$J_i^{-1} = I + B_i \Big(\sum_{j\neq i} s_j^{-2} B_j^T B_j\Big)^{-1} B_i^T/s_i^2,$$

and

$$Y_i - B_i\tilde{\theta} = J_i \Big[Y_i - B_i \Big(\sum_j s_j^{-2} B_j^T B_j\Big)^{-1} \sum_{k\neq i} B_k^T Y_k/s_k^2$$

$$-s_i^{-2} B_i \Big(\sum_{k\neq i} s_k^{-2} B_k^T B_k\Big]^{-1} B_i^T B_i \Big(\sum_j s_j^{-2} B_j^T B_j\Big)^{-1} \sum_{k\neq i} B_k^T Y_k/s_k^2\Big]$$

$$= J_i \Big(Y_i - B_i \tilde{\theta}^{(i)}\Big),$$

where

$$\tilde{\theta}^{(i)} = \Big(\sum_{k\neq i} s_k^{-2} B_k^T B_k\Big)^{-1} \sum_{k\neq i} B_k^T Y_k/s_k^2$$

is the least squares estimator of θ based on all data excluding Y_i. Wu (1986) discusses the relationship between $\tilde{\theta}$ and delete-one estimators $\tilde{\theta}^{(i)}, i = 1, \ldots, p$, which are used to form a jackknife estimator of θ.

It follows that for any $j = 1, \ldots, p$,

$$B_j\tilde{\theta} = B_j (B_i^T B_i)^{-1} B_i^T \Big[(I - J_i)Y_i + J_i B_i \tilde{\theta}^{(i)}\Big],$$

$$Y_j - B_j\tilde{\theta} = Y_j - B_j\tilde{\theta}^{(i)} + B_j \left(\sum_k s_k^{-2} B_k^T B_k\right)^{-1} B_i^T (Y_i - B_i\tilde{\theta}^{(i)})/s_i^2.$$

Since $\sum_{j\neq i} B_j^T(Y_j - B_j\tilde\theta^{(i)})/s_j^2 = 0$, we get

$$s_i^2 \sum_{j\neq i} \frac{||Y_j - B_j\tilde\theta||^2}{s_j^2} = s_i^2 \sum_{j\neq i} \frac{||Y_j - B_j\tilde\theta^{(i)}||^2}{s_j^2} + (Y_i - B_i\tilde\theta^{(i)})^T B_i \left(\sum_k s_k^{-2} B_k^T B_k\right)^{-1}$$

$$(\sum_{j\neq i} s_j^{-2} B_j^T B_j) \left(\sum_k s_k^{-2} B_k^T B_k\right)^{-1} B_i^T (Y_i - B_i\tilde\theta^{(i)})/s_i^2.$$

Therefore,

$$s_i^2 \sum_{j\neq i} \frac{||Y_j - B_j\tilde\theta||^2}{s_j^2} \geq (Y_i - \tilde\theta)^T J_i^{-1} B_i \left(\sum_k s_k^{-2} B_k^T B_k\right)^{-1}$$

$$\times \left(\sum_{j\neq i} s_j^{-2} B_j^T B_j\right) \left(\sum_k s_k^{-2} B_k^T B_k\right)^{-1} B_i^T J_i^{-1} (Y_i - B_i\tilde\theta)/s_i^2,$$

as

$$J_i^{-1} B_i \left(\sum_k s_k^{-2} B_k^T B_k\right)^{-1} \left(\sum_{j\neq i} s_j^{-2} B_j^T B_j\right) \left(\sum_k s_k^{-2} B_k^T B_k\right)^{-1} B_i^T = I.$$

This fact establishes (7.6). Indeed it follows from the inequality,

$$||Y_i - B_i\tilde\theta||^4 \leq (Y_i - B_i\tilde\theta)^T J_i^{-1} (Y_i - B_i\tilde\theta)(Y_i - B_i\tilde\theta)^T J_i (Y_i - B_i\tilde\theta),$$

(Beckenbach and Bellman, 1961, Ch 2, Theorem 20.)

7.2. Proof of Theorem 2.2

The modifications needed in the proof in Section 7.1 are as follows. By employing the same, integration by parts identities, we see that the difference Ψ between the risk functions of ψ_0 and ψ can be written as

$$\Psi = b^2 \sum_i E \frac{||X_i - \tilde\theta||_i^2}{\sigma_i^2 (\sum_j ||X_j - \tilde\theta||_j^2/s_j^2)^2} - 2b \sum_i E \frac{(X_i - \theta_i)^T B_i^T B_i (X_i - \tilde\theta)}{\sigma_i^2 \sum_j ||X_j - \tilde\theta||_j^2/s_j^2}$$

$$= b^2 \sum_i E \frac{s_i^4 ||X_i - \tilde\theta||_i^2/\sigma_i^2}{s_i^4 (\sum_j ||X_j - \tilde\theta||_j^2/s_j^2)^2} - 2b \sum_i E \text{ div} \left(\frac{X_i - \tilde\theta}{\sum_j ||X_j - \tilde\theta||_j^2/s_j^2}\right). \tag{7.7}$$

The evaluation of the second term in (7.7) is done similarly to (7.5) by using the fact that $\sum_i B_i^T B_i (X_i - \tilde\theta)/s_i^2 = 0$. It gives

$$\sum_i E \text{ div} \left(\frac{X_i - \tilde\theta}{\sum_j ||X_j - \tilde\theta||_j^2/s_j^2}\right) = E \frac{pq - q - 2}{\sum_j ||X_j - \tilde\theta||_j^2/s_j^2}.$$

The first sum in (7.7) is

$$b^2 \sum_i E \frac{||X_i - \tilde\theta||_i^2/s_i^2}{(\sum_j ||X_j - \tilde\theta||^2/s_j^2)^2} + 2b^2 \sum_i (\nu_i + 2) E s_i^4 \frac{\partial}{\partial s_i^2} \frac{||X_i - \tilde\theta||_i^2}{s_i^4 (\sum_j ||X_j - \tilde\theta||_j^2/s_j^2)^2},$$

so that it suffices to prove that the last term here is non-positive.

With ω_i defined by (7.3), one gets for a fixed i,

$$\frac{\partial}{\partial s_i^2}\frac{||X_i-\tilde\theta||_i^2}{(s_i^2\sum_j ||X_j-\tilde\theta||_j^2/s_j^2)^2}=\frac{2(X_i-\tilde\theta_i)^TB_i^TB_i\omega_i(X_i-\tilde\theta)}{s_i^6(\sum_j ||X_j-\tilde\theta||_j^2/s_j^2]^2}$$

$$-\frac{2||X_i-\tilde\theta||_i^2\sum_{j\neq i}||X_j-\tilde\theta||_j^2/s_j^2}{s_i^6(\sum_j ||X_j-\tilde\theta||_j^2/s_j^2)^3}.$$

Theorem 2.2 will be proven when the inequality

$$\frac{||X_i-\tilde\theta||_i^4}{s_i^2}\le\sum_j\frac{||X_j-\tilde\theta||_j^2}{s_j^2}\left[(X_i-\tilde\theta_i)^TB_i^TB_i(I-\omega_i)(X_i-\tilde\theta)\right],$$

is established for $i=1,\dots,p$.

To show that this inequality holds, note that $\tilde\theta=\omega_iX_i+(I-\omega_i)\ \tilde\theta^{(i)}$, and $X_j-\tilde\theta=X_j-\tilde\theta^{(i)}-\omega_i(X_i-\tilde\theta^{(i)})$, so that

$$\sum_j\frac{||X_j-\tilde\theta||_j^2}{s_j^2}=\frac{||(I-\omega_i)(X_i-\tilde\theta)||_i^2}{s_i^2}+\sum_{j\neq i}\frac{||X_j-\tilde\theta^{(i)}||_j^2}{s_j^2}+\sum_{j\neq i}\frac{||\omega_i(X_i-\tilde\theta^{(i)})||_j^2}{s_j^2}$$

$$\ge\frac{||(I-\omega_i)(X_i-\tilde\theta)||_i^2}{s_i^2}+(X_i-\tilde\theta^{(i)})^T\omega_i^T\left(\sum_{j\neq i}\frac{B_j^TB_j}{s_j^2}\right)\omega_i(X_i-\tilde\theta^{(i)})$$

$$=\frac{||(I-\omega_i)(X_i-\tilde\theta)||_i^2}{s_i^2}+(X_i-\tilde\theta)^T(I-\omega_i^T)^{-1}\omega_i^T\left(\sum_{j\neq i}\frac{B_j^TB_j}{s_j^2}\right)\omega_i(I-\omega_i)^{-1}(X_i-\tilde\theta).$$

It remains to observe that

$$\left(\sum_{j\neq i}\frac{B_j^TB_j}{s_j^2}\right)\omega_i(I-\omega_i)^{-1}=\left(\sum_j\frac{B_j^TB_j}{s_j^2}\right)(I-\omega_i)\omega_i(I-\omega_i)^{-1}$$

$$=\left(\sum_j\frac{B_j^TB_j}{s_j^2}\right)\omega_i=\frac{B_i^TB_i}{s_i^2},$$

which concludes the proof by appealing to the same fact as in Theorem 2.1.

References

Beckenbach, E. F. and Bellman, R. (1961). *Inequalities*, Springer, Berlin.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Control. Clin. Trials*, 7, 177–188.

Fourdrinier, D., Strawderman, W. E. and Wells, M. T. (2003). Robust shrinkage estimation for elliptically symmetric distributions with unknown covariance matrix. *J. Mult. Anal.* 85, 24–39.

Gruber, M. H. J. (1998). *Improving Efficiency by Shrinkage: the James-Stein and Ridge Regression Estimators*, M. Dekker, New York.

Hartung, J., Knapp, G. and Sinha, B. K. (2008). *Statistical Meta-Analysis with Applications*, Wiley, New York.

Hedges, L. V. and Olkin I. (1985). *Statistical Methods for Meta-Analysis*, Academic Press, London.

Lee, Y. and Birkes, D. (1994). Shrinking toward submodels in regression. *J. Statist. Plann. Inf.* 41, 95–111.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed., Springer, New York.

Paule, R. C. and Mandel, J. (1971). Analysis of interlaboratory measurements on the vapor pressure of cadmium and silver. National Bureau of Standards Special Publication, 260-21, Washington DC.

Rukhin, A. L. (1987). How much better are better estimators of a normal variance. *J. Amer. Statist. Assoc.* 82, 925–928.

Rukhin, A. L. (2007). Estimating common vector mean in interlaboratory studies. *J. Mult. Anal.* 98, 435–434.

Rukhin, A. L. (2011). Estimating common parameters in heterogeneous mixed effects model, *J. Statist. Plann. Inf.* 141, 3181–3192.

Strawderman, W. E. and Rukhin, A. L. (2010). Simultaneous estimation and reduction of data nonconformity in interlaboratory studies. *J. Royal Statist. Soc.*, Ser B. 72, 219–234.

Wu, C. J. F. (1986). Jackknife, bootstrap and other resampling methods in regression analysis, (with discussion). *Ann. Statist.* 41, 1261–1350.