Prediction of Ordered Random Effects in a Simple Small Area Model

Yaakov Malinovsky* and Yosef Rinott*,**

*The Hebrew University of Jerusalem and **LUISS, Rome

Abstract: Prediction of a vector of ordered parameters or part of it arises naturally in the context of Small Area Estimation (SAE). For example, one may want to estimate the parameters associated with the top ten areas, the best or worst area, or a certain percentile. We use a simple SAE model to show that estimation of ordered parameters by the corresponding ordered estimates of each area separately does not yield good results with respect to MSE. Shrinkage-type predictors, with an appropriate amount of shrinkage for the particular problem of ordered parameters, are considerably better, and their performance is close to that of the optimal predictors, which cannot in general be computed explicitly.

Key words and phrases: Empirical Bayes predictor, Shrinkage, Order statistics, Linear predictor.

1 Introduction

We study the prediction of ordered random effects in a simple model, motivated by Small Area Estimation (SAE), under a quadratic loss function. The model is

$$y_i = \mu + u_i + e_i, \quad i = 1, ..., m, \tag{1.1}$$

where y_i is observed, μ is an unknown constant, $u_i \overset{iid}{\sim} F(0, \sigma_u^2)$ and $e_i \overset{iid}{\sim} G(0, \sigma_e^2)$, and F and G are general distributions with zero means and variances σ_u^2 and σ_e^2. Set $\mathbf{y} = (y_1, ..., y_m), \mathbf{u} = (u_1, ..., u_m)$, and $\mathbf{e} = (e_1, ..., e_m)$, and assume that $\mathbf{u}$ and $\mathbf{e}$ are independent. Set $\theta_i = \mu + u_i$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. The purpose is to predict the ordered random variables $\theta_{(i)}$, $\left(\theta_{(1)} \leq \theta_{(2)} \leq ... \leq \theta_{(m)}\right)$ from the observed y's. In SAE the random effect θ_i represent the ith area parameter.

The above model is a special case of the SAE model of Fay and Herriot (1979) that was presented in the context of estimating per capita income for small places (i.e., population less than 1,000) from the 1970 Census of Population and Housing. The original Fay-Herriot model allows different μ_i of the form $\mu_i = x_i'\beta$,

where x_i is a vector of covariates for area i, β is a vector of coefficients that are common to all areas, u_i is a random effect of area i, and $\theta_i = \mu_i + u_i$, the value of interest in area i, is measured with a sampling error e_i. The SAE literature is concerned with the estimation of θ_i; see, e.g., Rao (2003). However, it is also natural to consider the ordered parameters $\theta_{(i)}$ if one is interested in estimating jointly the best, second best, median, or worst area's parameter, for example, or in studying the best or worst k areas. In these cases, one is interested in many or all ranked parameters, and not just a single $\theta_{(i)}$.

When we have more than one observation per area, the model is known as the Battese-Harter-Fuller model (1988), see also Pfeffermann (2002), which we again simplify as in (1.1) :

$$y_{ij} = \mu + u_i + e_{ij}, \quad j = 1, \ldots, n, \; i = 1, \ldots, m. \tag{1.2}$$

Typically in SAE m is large, while n is small; however, we consider both small and large m. Taking area means, as justified by sufficiency, the latter model reduces to that of (1.1) with σ_e^2 replaced by σ_e^2/n. When the σ_e^2 is unknown, it should be estimated. A main idea in SAE is to borrow strength across the different areas in order to predict effects. This can be applied also to variance estimation when some of the areas have only one observation; however, this is beyond the scope of the present paper, and for simplicity we assume the same number of observations n in each area.

To see the difference between predicting the unordered vector $\boldsymbol{\theta}$, and the ordered vector $\boldsymbol{\theta}_{()} = (\theta_{(1)}, \ldots, \theta_{(m)})$, consider estimating the maximum $\theta_{(m)}$, and two natural unbiased predictors of θ_i, $\widehat{\theta}_i = y_i$ or $\widehat{\theta}_i = E(\theta_i|\mathbf{y})$. By Jensen's inequality $\widehat{\theta}_{(m)} := \max_i \widehat{\theta}_i$ is an overestimate in expectation of $\theta_{(m)}$ in the case of $\widehat{\theta}_i = y_i$, and an underestimate if we use $\widehat{\theta}_i = E(\theta_i|\mathbf{y})$. Such biases increase in m, which in SAE and in many parts of this paper is taken to be large. Similar considerations hold for other ordered parameters.

With different loss functions, prediction of the ordered parameters appears in Wright, Stern and Cressie (2003), and prediction and ranking of small area parameters appear in Shen and Louis (1998). Their Bayesian methods require heavy numerical calculations, and are sensitive to the choice of priors; see Shen and Louis (2000).

If μ and σ_u^2 and/or σ_e^2 are known, we have a Bayesian model in (1.1) or

(1.2), and under quadratic loss, the optimal predictors of the ordered parameters would be of the form $\widehat{\theta}_{(i)} = E(\theta_{(i)}|\mathbf{y})$, where the expectation depends on μ, σ_u^2 and σ_e^2, (and the distribution F and G). If μ and σ_u^2 and/or σ_e^2 are unknown, we adopt an Empirical Bayes approach, and estimate them from the data. However, even in the normal case, analytical computation of $E(\theta_{(i)}|\mathbf{y})$ seems intractable for $m > 2$, and even more so under other distributions. Numerical computations could be done, and in fact, this is the Bayesian approach taken in principle by Wright et al (2003) and Shen and Louis (1998, 2000); the precise quantities they compute are different due to the fact that they use different loss functions.

In this paper we avoid such Bayesian calculations and present simple predictors whose performance is close to optimal; furthermore, due to their simplicity, they are more robust against model misspecification.

Our starting point is the following. Consider the predictor $\widehat{\theta}_i = E(\theta_i|\mathbf{y})$ of θ_i; under the assumption that F and G are normal, we have $\widehat{\theta}_i = \widehat{\theta}_i(\mu, \gamma^*) = \gamma^* y_i + (1 - \gamma^*)\mu$, where $\gamma^* = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$. For unknown μ, we plug in the estimator $\widehat{\mu} = \overline{y}$, and obtain the shrinkage-type predictor $\widehat{\theta}_i(\widehat{\mu}) = \gamma^* y_i + (1-\gamma^*)\overline{y}$, which is the best linear unbiased predictor of θ_i for any F and G; see, e.g., Robinson (1991), Rao (2003). Here γ^* determines the amount of shrinkage toward the mean. We discuss the required amount of shrinkage when the goal is to predict $\theta_{(i)}$ rather than θ_i.

For the problem of predicting the unordered parameters, Bayesian considerations as above, and Stein (1956) and the ensuing huge body of literature suggest shrinkage predictors. In view of the discussion on under and overestimation, it is not surprising that for the present problem of predicting the ordered parameters, shrinkage is also desirable, but to a lesser extent. In fact, it can be shown geometrically that if the coordinates of the predicting vector $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_m)$ happen to have the right order, that is, the same order as the coordinates of $\boldsymbol{\theta}$, then the desirable shrinkage is the same for the two problems, but otherwise it is smaller. The latter case happens with high probability for large m when the parameters are not very different. In this paper we show that rather satisfactory results can be obtained by simple predictors of the type $\gamma y_{(i)} + (1 - \gamma)\overline{y}$, and study the optimal value of γ. In general we have $\gamma^* \leq \gamma$. Specifically, for large m ($m > 25$, say), we propose the predictor $\sqrt{\gamma^*} y_{(i)} + (1 - \sqrt{\gamma^*})\overline{y}$, to be denoted

later by $\widehat{\theta}^{[2]}_{(i)}(\sqrt{\gamma^*})$. This predictor is easy to compute when the variances are either known or estimated, and performs well in comparison to Bayes predictors, and other numerically demanding predictors that appear in the literature.

In most of this paper we consider some predictor $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_m)$, take $\widehat{\theta}_{(i)}$ as a predictor of $\theta_{(i)}$ with a loss function given by

$$L(\widehat{\boldsymbol{\theta}}_{(\,)}, \boldsymbol{\theta}_{(\,)}) = \sum_{i=1}^{m} \left(\widehat{\theta}_{(i)} - \theta_{(i)}\right)^2, \tag{1.3}$$

and compare different predictors in terms of the (Bayesian) risk

$$r(H, \widehat{\theta}\,) = E\{L(\widehat{\boldsymbol{\theta}}_{(\,)}, \boldsymbol{\theta}_{(\,)})\},$$

where $H = (F, G)$ and the expectation is over all random variables involved. Note that by a simple rearrangement inequality we always have $\sum_{i=1}^{m} \left(\widehat{\theta}_{(i)} - \theta_{(i)}\right)^2 \leq \sum_{i=1}^{m} \left(\widehat{\theta}_i - \theta_i\right)^2$.

We also briefly consider the individual mean square error (MSE) of a predictor $\widehat{\theta}_{(i)}$, defined to be $MSE(\widehat{\theta}_{(i)}) = E(\widehat{\theta}_{(i)} - \theta_{(i)})^2$.

Even in the case of $m = 2$ this prediction problem is not trivial. Blumenthal and Cohen (1968) consider the following model: given independent observation X_1, X_2 with $X_i \sim N(\theta_i, \tau^2)$, estimate $\theta_{(2)} = \max(\theta_1, \theta_2)$. They present five different estimators for $\theta_{(2)}$ and evaluate their biases and mean square errors. Generalizing their method to more than two parameters appears to be hard.

Finally we mention that Senn (2008), with reference to Dawid (1994) and others, deals with a different but related problem of estimation of the parameter θ_{i^*} corresponding to $i^* = \arg\max y_i$. In SAE, this is the parameter belonging to the population having the largest sample outcome, while we consider estimation of $\theta_{(m)}$, the parameter of the "largest population" (and likewise for other ordered parameters). The difference is important when m is large, and the parameters vary significantly; our interest is in the ordered parameters and not parameters chosen by the data, as in the above references.

In Section 2 we discuss the model and present several predictors. We also give minimax results that provide some justification to normality assumptions and linear prediction. Section 3 contains the main results on properties and comparisons of the various predictors. The proposed class of predictors contains

a parameter γ. Some of the results apply to the whole class, while others suggest a range where the best value of γ should be, and apply to γ in this range. We describe a conjecture about the optimal value of γ when m is large and provide an approximation for the optimal value of γ in the normal case. The last part of Section 3 deals with a special case when F and G are normal and $m = 2$. In this part we get tighter conclusions than in the general case.

In Section 3 we assumed that the variances in (1.1) are known. Section 4 deals with the case of unknown variances and studies plug-in Empirical Bayes predictors by simulation. In Sections 5 and 6 we study robustness of the proposed predictors against certain misspecifications of the assumptions on the distributions, and compare to other predictors from the literature.

The proofs of results concerning general m are given in Section 7. The rest of the proofs are given in an on-line Supplement at
http://www.stat.sinica.edu.tw/statistica. In the Supplement we provide simulations for Conjectures 1 and 2, we compare various predictors under the assumption of known variances, and when one of the variances is unknown. Theorems 5 and 6 are also proved there.

2 Predictors

2.1 Unordered parameters

In Sections 2 and 3 we assume that σ_u^2 and σ_e^2 are known. Later they are assumed unknown, and plug-in estimators are used. First we review some known results for the unordered case of Model (1.1) and the standard problem of predicting θ_i, $i = 1, ..., m$. The best linear predictor is of the form $\mathbf{a}^t\mathbf{y} + b$, with $\mathbf{a}, b$ that minimize the mean square error. It is easy to see that when μ is known, and recalling that σ_u^2 and σ_e^2 are now also assumed to be known, the best linear predictor of θ_i, that is, the predictor that minimized $E(\widehat{\theta}_i - \theta_i)^2$ and therefore $r^*(H, \widehat{\theta}) := E\left(\sum_{i=1}^m (\widehat{\theta}_i - \theta_i)^2\right)$ among linear predictors, is

$$\widehat{\theta}_i(\mu) = \gamma^* y_i + (1 - \gamma^*)\mu, \tag{2.1}$$

where $\gamma^* = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$. Note that the model (1.1) does not assume normality, and that the best linear predictor is unbiased, that is, $E(\widehat{\theta}_i - \theta_i) = 0$. When

both distributions F and G are normal, the best linear predictor above is the best predictor (or Bayes predictor).

For unknown μ, the best linear unbiased predictor (BLUP) of θ_i (see, for example, Robinson (1991), Rao (2003) and references therein) is

$$\widehat{\theta}_i(\widehat{\mu}) = \gamma^* y_i + (1-\gamma^*)\overline{y}, \tag{2.2}$$

where $\widehat{\mu} := \overline{y} = \frac{1}{m}\sum_{i=1}^m y_i$. The BLUP property means that the predictor (2.2) minimizes $E(\widehat{\theta}_i - \theta_i)^2$ among linear unbiased predictors for all F and G with the prescribed variances. These are shrinkage predictors (with shrinkage towards the mean). Such predictors appear also in Fay and Herriot (1979). We see in Section 3 that for the ordered parameters shrinkage is also required, but in a smaller amount (see also Louis (1984) and Ghosh (1992) for such shrinkage, under different loss functions), showing again that the related problems of predicting the ordered and unordered parameters, are not the same.

A justification of normality and linearity

Using the fact that an equalizer Bayes rule is minimax, Schwarz (1987) proved the following result, which in some sense justifies both linear estimators and the assumption of normality of F and G.

Theorem 1. *Consider (1.1) with μ, σ_u^2, and σ_e^2 all fixed and known, and the risk function $r^*(H,\widehat{\theta}) = E\left(\sum_{i=1}^m(\widehat{\theta}_i-\theta_i)^2\right)$. The predictor $\delta_0 = (\delta_{01},\ldots,\delta_{0m})$ of $\boldsymbol{\theta}$ given by $\delta_{0i} = \gamma^* y_i + (1-\gamma^*)\mu$, $i = 1,...,m$, is minimax and the normal strategy for $H = (F,G)$ is least favorable.*

The next result is closely related to the previous one, and justifies linearity when μ is unknown, which is the case we consider. It can easily be extended to the original Fay-Herriot model with $\mu_i = x_i'\beta$.

Theorem 2. *Under the assumptions of Theorem 1, but with unknown μ, the predictor defined by $\delta_{0i} = \gamma^* y_i + (1-\gamma^*)\overline{y}$, $i = 1,...,m$, is minimax among all linear unbiased predictors of $\boldsymbol{\theta}$.*

Proof. Let $\mathcal{H}$ denote the class of pairs of distributions (F,G) having the given variances. Note that $r^*(H,\delta_0)$ depends only on the fixed variances, and therefore

for $H \in \mathcal{H}$, $r^*(H, \delta_0)$ is constant, say v. Let $\mathcal{L}$ denote the class of linear unbiased predictors.

We know that $\delta_0 = (\delta_{01}, \ldots, \delta_{0m})$ is BLUP. We have

$$\overline{V} = \inf_{\delta \in \mathcal{L}} \sup_{H \in \mathcal{H}} r^*(H, \delta) \leq \sup_{H \in \mathcal{H}} r^*(H, \delta_0) = r^*(H_0, \delta_0) = v,$$
$$\underline{V} = \sup_{H \in \mathcal{H}} \inf_{\delta \in \mathcal{L}} r^*(H, \delta) \geq \inf_{\delta \in \mathcal{L}} r^*(H_0, \delta) = r^*(H_0, \delta_0) = v,$$

for any $H_0 \in \mathcal{H}$, where the penultimate equality holds by the BLUPness of δ_0. Since clearly $\overline{V} \geq \underline{V}$, it follows that $\inf_{\delta \in \mathcal{L}} \sup_{H \in \mathcal{H}} r(H, \delta) = \sup_{H \in \mathcal{H}} r(H, \delta_0)$, so that δ_0 is minimax among predictors in $\mathcal{L}$ as required. □

2.2 Ordered parameters

Let $\vartheta_{(i)}(\mu) = E_\mu(\theta_{(i)}|\mathbf{y})$, the best predictor of $\theta_{(i)}$ when μ is known, and consider its empirical or plug-in version when μ is unknown: $E_{\widehat{\mu}}(\theta_{(i)}|\mathbf{y}) = \vartheta_{(i)}(\widehat{\mu})$, where $\widehat{\mu} = \overline{y}$.

We consider three predictors:

$$\widehat{\theta}^{[1]}_{(i)} = y_{(i)}, \qquad \widehat{\theta}^{[2]}_{(i)}(\gamma) = \gamma y_{(i)} + (1-\gamma)\overline{y}, \qquad \widehat{\theta}^{[3]}_{(i)} = E_{\widehat{\mu}}(\theta_{(i)}|\mathbf{y}), \tag{2.3}$$

where $y_{(1)} \leq ... \leq y_{(m)}$ denote the order statistics of $y_1, ..., y_m$.

Set $\widehat{\boldsymbol{\theta}}^{[k]}_{()} = \left(\widehat{\theta}^{[k]}_{(1)}, \ldots, \widehat{\theta}^{[k]}_{(m)}\right)$ for $k = 1, 2, 3$. The predictors in the class $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma)$ are analogous to the best linear predictor for the unordered case, but as we shall see, the value of γ has to be reconsidered, and $\widehat{\boldsymbol{\theta}}^{[3]}_{()}$ is the empirical best predictor in the ordered case (the best predictor with μ replaced by $\overline{y}$). The latter predictor cannot in general be computed explicitly for $m > 2$, and some of our results are aimed at showing that it can be efficiently replaced by $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma)$ with an appropriate choice of γ for the ordered case at hand. Thus $\widehat{\boldsymbol{\theta}}^{[3]}_{()}$, the empirical best predictor, will be used as a yardstick to which other predictors are compared.

3 Main results, known variances

3.1 General distributions F and G and general m

The proofs of the results of this subsection are given in the Appendix.

The first few results show that shrinkage-type predictors in the class $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma)$ perform better than the predictor $\widehat{\boldsymbol{\theta}}^{[1]}_{()}$. Refined calculations of the range of the optimal γ allow us to understand the amount of shrinkage required for the ordered parameters case.

Theorem 3. *Consider (1.1) with the loss function (1.3) and* $\gamma^* = \sigma_u^2/(\sigma_u^2+\sigma_e^2)$. *If*

$$\frac{m}{m-1}(2\sqrt{\gamma^*}-1) - \frac{1}{m-1}(2\gamma^*-1) \leq \ \gamma \ \leq 1, \tag{3.1}$$

then

$$E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma), \boldsymbol{\theta}_{()})\} \leq E\{L(\widehat{\boldsymbol{\theta}}^{[1]}_{()}, \boldsymbol{\theta}_{()})\}. \tag{3.2}$$

Note that if $\sigma_u^2 \to 0$ then $\gamma^* \to 0$ and the left-hand side of (3.1) tends to -1. If $m \to \infty$, which is of interest in SAE, then the left-hand side of (3.1) tends to $2\sqrt{\gamma^*}-1$. The left-hand side of (3.1) is 1 when $\gamma^* = 1$ and increases in γ^*, hence is bounded by 1.

By verifying the condition for the left-hand side of (3.1) to be nonpositive, we obtain the following.

Corollary 1. *If*

$$\gamma^* \leq \left(\frac{m-\sqrt{(m-1)^2+1}}{2}\right)^2, \tag{3.3}$$

then

$$E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma), \boldsymbol{\theta}_{()})\} \leq E\{L(\widehat{\boldsymbol{\theta}}^{[1]}_{()}, \boldsymbol{\theta}_{()})\} \tag{3.4}$$

for all $\gamma, 0 \leq \gamma \leq 1$.

Note that asymptotically (3.3) becomes $\gamma^* \leq \frac{1}{4}$, since $\lim_{m\to\infty}\left(\frac{m-\sqrt{(m-1)^2+1}}{2}\right)^2 = \frac{1}{4}$. Condition (3.3) is sufficient and may not be necessary. But, (3.4) does not hold without a suitable condition on γ^*; for example, if $m = 100$, the upper bound of (3.3) is 0.2475. For $\gamma^* = 1/3$ and $\gamma = 0.1$, a straightforward simulation using normal variables shows that (3.4) does not hold.

For $\gamma = \gamma^*$, (3.1) holds if and only if $\gamma^* \leq (m-1)^2/(m+1)^2$ (see Appendix). From this result we obtain the following.

Corollary 2. *If*

$$\gamma^* \leq (m-1)^2/(m+1)^2, \tag{3.5}$$

then $E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma), \boldsymbol{\theta}_{()})\} \leq E\{L(\widehat{\boldsymbol{\theta}}^{[1]}_{()}, \boldsymbol{\theta}_{()})\}$ *for all* γ, $\gamma^* \leq \gamma \leq 1$.

Asymptotically, Corollary 2 holds for all γ^* without (3.5), because $\lim_{m\to\infty}\{(m-1)^2/(m+1)^2\} = 1$ and $0 \leq \gamma^* \leq 1$ by definition. A small simulation study indicates that Corollary 2 may hold without the condition $\gamma^* \leq (m-1)^2/(m+1)^2$ for a large variety of F and G. We can prove it only for the extreme case $m = 2$ and normal F and G; see Theorem 5 below. The range of γ's for which shrinkage improves the predictors, $\gamma^* \leq \gamma$, indicates that, for the ordered problem, less shrinkage is required.

The following lemma is used in the proof of Theorem 3, but may be of independent interest.

Lemma 1. *Under (1.1),*

$$m(\sigma_u^2 + \mu^2) \leqslant E\sum_{i=1}^{m} \theta_{(i)} y_{(i)} \leqslant m[(\sigma_u^2 + \mu^2)(\sigma_u^2 + \sigma_e^2 + \mu^2)]^{1/2}.$$

For the predictors $\widehat{\theta}^{[2]}_{(i)}(\gamma)$, it is natural to look for optimal or good values of γ.

Theorem 4. *Under (1.1), let* γ^o *be the optimal choice of* γ *for the predictor* $\widehat{\theta}^{[2]}_{(i)}(\gamma)$ *in the sense of minimizing* $E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma), \boldsymbol{\theta}_{()})\}$. *Then*

$$\gamma^o \in \left[\gamma^*, \frac{m}{m-1}\sqrt{\gamma^*} - \frac{1}{m-1}\gamma^*\right]. \tag{3.6}$$

As $m \to \infty$, the above range for the optimal γ becomes $[\gamma^*, \sqrt{\gamma^*}]$.

Conjecture 1. *The optimal* γ *in the sense of Theorem 4 satisfies* $\lim_{m\to\infty} \gamma^o = \sqrt{\gamma^*}$.

Simulations that justify Conjecture 1 are given in the Supplement.

For $m > 25$ or so, which is common in SAE, we recommend using the predictor $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma)$ with $\gamma = \sqrt{\gamma^*}$. Numerous simulations suggest that the latter choice, or the choice of $\gamma = \gamma^o$, yield essentially the same results. We emphasize that the predictor $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\sqrt{\gamma^*})$ is very easy to compute. The case of $m \leq 25$ is discussed next.

3.2 An approximation for γ^o in the normal case

For practical computation of γ^o for $m \leq 25$ or so, we propose the following approach that we have implemented in the normal case. (For $m > 25$, $\sqrt{\gamma^*}$ provides an excellent approximation to γ^o, see simulation results and Conjecture 1). In view of Theorem 4 we consider the approximation formula

$$\gamma^o \approx \alpha\gamma^* + (1-\alpha)u(m,\gamma^*), \tag{3.7}$$

with $u(m,\gamma^*) = \frac{m}{m-1}\sqrt{\gamma^*} - \frac{1}{m-1}\gamma^*$, and α depending on m and γ^*. For fixed γ^*, and for each m satisfying $2 \leq m \leq 30$ we compute $E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma), \boldsymbol{\theta}_{()})\}$ by simulations and find the minimizer γ^o by an exhaustive search. We then define α_{m,γ^*} to be the solution of (3.7). For fixed γ^* we carry out polynomial regression of the computed values of α_{m,γ^*} on the explanatory variable m in the range $2 \leq m \leq 30$; this is repeated for an array of values of γ^*. It turns out that an excellent approximation is obtained when $\alpha_{m,\gamma^*} = \alpha_m$ is taken to be only a function of m for a large range of values of γ^*. We therefore combine the different regressions for the different values of γ^*, and obtain a polynomial approximation for α_m. The numerical calculations lead to the quadratic polynomial $\alpha_m = 0.8236 - 0.0573m + 0.0012m^2$. Plugging it into (3.7) we obtain the approximation $\widetilde{\gamma^o} = \alpha_m\gamma^* + (1-\alpha_m)u(m,\gamma^*)$ for γ^o.

Numerical simulations show that in the range $2 \leq m \leq 25$ and for all values of γ^*, the resulting $\widetilde{\gamma^o}$ is indeed very close to γ^o. In fact, for $m = 2$ they may differ by about 10%, but for $m \geq 4$ they differ by about $1\% - 2\%$. Using one or the other yields almost identical expected losses.

3.3 Normal distribution of F and G and m=2

When both F and G are normal and $m = 2$, we obtain tighter conclusions for the previous results.

Theorem 5. *For (1.1) with F and G normal and $m = 2$:*

1. *if* $0 \leq \gamma^* \leq c \approx 0.4119$, *then* $E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma), \boldsymbol{\theta}_{()})\} \leq E\{L(\widehat{\boldsymbol{\theta}}^{[1]}_{()}, \boldsymbol{\theta}_{()})\}$ *for all* $0 \leq \gamma \leq 1$;
2. *for all* γ^* *and* γ *satisfying* $\gamma^* \leq \gamma \leq 1$, $E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma), \boldsymbol{\theta}_{()})\} \leq E\{L(\widehat{\boldsymbol{\theta}}^{[1]}_{()}, \boldsymbol{\theta}_{()})\}$;

3. the optimal γ for the predictor $\widehat{\theta}^{[2]}_{(i)}(\gamma)$ (i=1,2) in the sense of minimizing $E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma), \boldsymbol{\theta}_{()})\}$ is

$$\gamma^o = \gamma^* \left(4\psi(a) - 1\right) + (1-\gamma^*)\frac{2}{\pi}\sqrt{\gamma^*(1-\gamma^*)}, \tag{3.8}$$

where $\psi(a) = \int_0^\infty t^2 \Phi\left(at\right)\varphi(t)dt$, and $a = \sqrt{\frac{\gamma^*}{1-\gamma^*}}$.

Thus Part *1* of Theorem 5 shows that we can replace the condition $\gamma^* \leq \left(\frac{2-\sqrt{2}}{2}\right)^2 \approx 0.086$ of Corollary 1 by $\gamma^* \leq c,\ c \approx 0.4119$; Part *2* shows that (3.5) ($\gamma^* \leq 1/9$ for $m = 2$) of Corollary 2 may be omitted; Part *3* of Theorem 5 gives an exact result rather than the range given by (3.6).

Remark. The accurate definition of c and its approximation are given in the proof of Theorem 5 in the Supplement. The function $\psi(a)$ can be computed by Matlab: double($int(x^2 * \mathrm{normpdf}\,(x) * (\mathrm{erf}\,(a * x/sqrt(2)) + 1)/2, x, 0, \mathrm{inf}\,)$).

The results given so far compare $\widehat{\theta}^{[2]}_{(i)}(\gamma)$ with $\widehat{\theta}^{[1]}_{(i)}$ in the sense of minimizing expected loss. In the absence of an explicit expression for $\widehat{\theta}^{[3]}_{(i)}$, it is not easy to compare it with other predictors analyticallly, but it is possible to do this if F and G are normal and $m = 2$, and the result is Theorem 6. For $m > 2$ we provide simulations.

It is obvious that the estimator $\vartheta_{(i)}(\mu) = E_\mu(\theta_{(i)}|y)$ minimizes the MSE. The point of Theorem 6 is that the unknown μ is replaced in $\vartheta_{(i)}(\mu)$ by its estimate $\overline{y}$ to obtain $\widehat{\theta}^{[3]}_{(i)}$.

Theorem 6. *Consider (1.1) with F and G normal. Then for $m = 2$, $E\{L(\widehat{\boldsymbol{\theta}}^{[3]}_{()}, \boldsymbol{\theta}_{()})\} \leq E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma^*), \boldsymbol{\theta}_{()})\}$.*

Conjecture 2. *If F and G are normal and $m \geq 2$, then $E\{L(\widehat{\boldsymbol{\theta}}^{[3]}_{()}, \boldsymbol{\theta}_{()})\} \leq E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma^o), \boldsymbol{\theta}_{()})\}$.*

Various simulations support this conjecture. Some of them are presented in the Supplement. The simulations show that the predictor $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma^o)$ is worse than $\widehat{\boldsymbol{\theta}}^{[3]}_{()}$ in the sense of $E\{L(\widehat{\boldsymbol{\theta}}_{()}, \boldsymbol{\theta}_{()})\}$, as suggested by Conjecture 2. However, they are rather close, while the predictor $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma^*)$ is far worse. This suggests that the linear predictor $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma^o)$ can be used without much loss. As mentioned above, for $m \geq 25$ or so, the predictor $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\sqrt{\gamma^*})$, which is easy to calculate, is as good as $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma^o)$, and the calculation of γ^o can be avoided. See the Supplement.

4 Unknown variances

Until now it was assumed that the variances are known. We now turn to the case of unknown variances. This case will be studied by simulations, whose detailed description is given in the Supplement.

We first make the common assumption in SAE that only σ_u^2 is unknown, and later that both variances, σ_u^2 and σ_e^2 are unknown. We replace each unknown variance by plugging-in its natural estimator. For the case that only σ_u^2 is unknown, it is estimated by

$$\widehat{\sigma}_u^2 = \max\left(\frac{1}{m-1}\sum_{i=1}^{m}(y_i-\overline{y})^2 - \sigma_e^2, 0\right). \tag{4.1}$$

This approach cannot be expected to work for small values of m. We emphasize again the interest in SAE is in large m's.

The notation for the resulting estimates remains as it was for the case of known variances. In this case, and in the case that both variances are unknown (Figure 1 below), we use simulations to compare the risk $E\{L(\widehat{\boldsymbol{\theta}}_{()}, \boldsymbol{\theta}_{()})\}$ for the predictors $\widehat{\boldsymbol{\theta}}^{[3]}_{()}$, $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma^*)$, $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\sqrt{\gamma^*})$, and $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma^o)$, and since in all simulations the risks of the latter two predictors are almost identical, we present only one of them. We also compare the performance of these predictors when only the maximum is predicted.

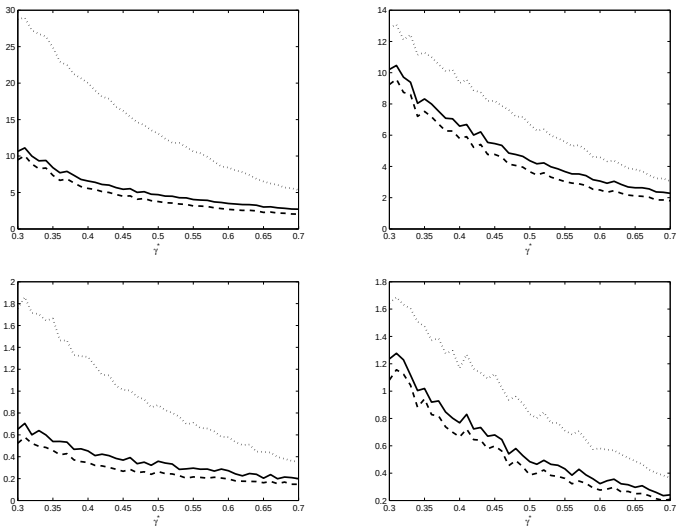

Figure 1:

- Comparison of $E\{L(\widehat{\boldsymbol{\theta}}_{()}, \boldsymbol{\theta}_{()})\}$ as a function of γ^*, for the predictors $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma^*)$, $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\sqrt{\gamma^*})$, $\widehat{\boldsymbol{\theta}}^{[3]}_{()}$ (dotted, solid, dashed lines), where F and G are normal and $m = 100, n = 15$ (upper left), $m = 30, n = 15$ (upper right)
- Comparison of the MSE of $\widehat{\theta}^{[2]}_{(m)}(\gamma^*)$, $\widehat{\theta}^{[2]}_{(m)}(\sqrt{\gamma^*})$, $\widehat{\theta}^{[3]}_{(m)}$ (dotted, solid, dashed lines) for predicting $\theta_{(m)}$, as a function of γ^*, where F and G are normal and $m = 100, n = 15$ (bottom left), $m = 30, n = 15$ (bottom right)

In Figure 3S (given in the Supplement) only σ_u^2 is estimated, and in Figure 1 both σ_u^2 and σ_e^2 are estimated. The figures are rather similar. The results should be compared to those of Figure 2S (Supplement), where the variances are known. Clearly the less one knows, the higher the loss. However, the simple shrinkage predictor $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\sqrt{\gamma^*})$ performed almost as well as the best plug-in predictor $\widehat{\boldsymbol{\theta}}^{[3]}_{()}$, and much better than $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma^*)$. Thus again we conclude that for the problem at hand, shrinkage estimators work, provided one uses the right amount of shrinkage for the ordered parameters problem.

For the case of unknowns σ_u^2 and σ_e^2, consider the model

$$y_{ij} = \mu + u_i + e_{ij}; i = 1, ..., m, j = 1, ..., n,$$

which is a special case of the Nested Error Unit Level Regression model of Battese, Harter and Fuller (1988). We apply our previous estimators, replacing the variances by

$$\widehat{\sigma}_e^2 = \frac{1}{m(n-1)} \sum_{i=1}^{m} \sum_{j=1}^{n} (y_{ij} - \overline{y}_{i.})^2 , \quad \widehat{\sigma}_u^2 = \max\Big\{\frac{1}{mn-1} \sum_{i=1}^{m} \sum_{j=1}^{n} (y_{ij} - \overline{y}_{..})^2 - \widehat{\sigma}_e^2, 0\Big\},$$

and set $\gamma^* = \dfrac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n}$. Simulation results are given Figure 1.

5 Shrinkage type predictor $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\sqrt{\gamma^*})$ in the non normal case

We briefly consider non-normal F, whereas the error distribution G remains normal. We first take the double exponential distribution (Laplace distribution)

for the random effects u_i, with density $\frac{1}{2b}\exp\left(-\frac{|u_i|}{b}\right)$, where $b = \frac{\sigma_u}{\sqrt{2}}$. Direct calculations show that the density function of θ_i given y_i is

$$f_{\theta_i|y_i}(t|y) = \begin{cases} \dfrac{p_1(t)}{\int_{-\infty}^{\mu} p_1(t)dt + \int_{\mu}^{\infty} p_2(t)dt}, & \text{if } t \le \mu \\ \dfrac{p_2(t)}{\int_{-\infty}^{\mu} p_1(t)dt + \int_{\mu}^{\infty} p_2(t)dt}, & \text{if } t > \mu \end{cases} \tag{5.1}$$

where $p_i(t) = exp\left(-\left(t - \left(y + (-1)^{i+1}\sigma_e^2 b^{-1}\right)\right)^2/2\sigma_e^2\right), \ \ i = 1, 2.$

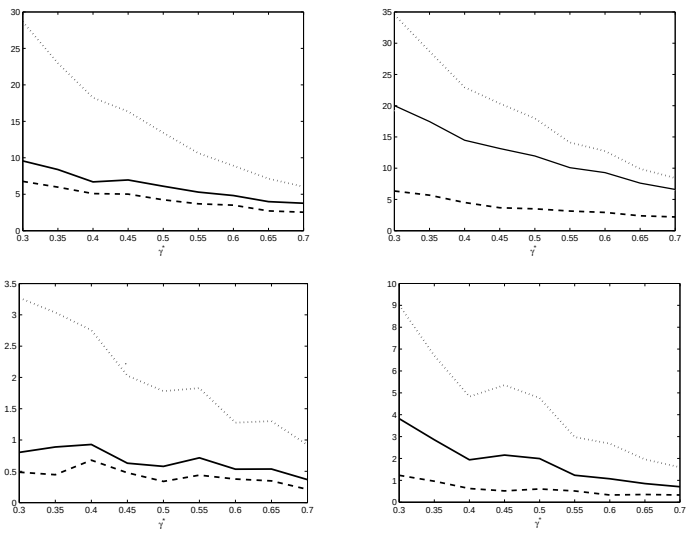

Figure 2:

- Comparison of $E\{L(\widehat{\boldsymbol{\theta}}_{()}, \boldsymbol{\theta}_{()})\}$ as a function of γ^*, for the predictors $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma^*)$, $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma^o)$, $\widehat{\boldsymbol{\theta}}^{[3]}_{()}$ (dotted, solid, dashed lines), where F is the Laplace distribution and G is normal (upper left), and where F is the Location exponential distribution and G is normal (upper right), for $m = 100$.
- Comparison of the MSE of $\widehat{\theta}^{[2]}_{(m)}(\gamma^*)$, $\widehat{\theta}^{[2]}_{(m)}(\gamma^o)$, $\widehat{\theta}^{[3]}_{(m)}$ (dotted, solid, dashed lines) for predicting $\theta_{(m)}$, as a function of γ^*, where F is the Laplace distribution and G is normal (bottom left), and where F is the Location exponential distribution and G is normal (bottom right), for $m = 100$.

We also take a location exponential distribution for the random effects , with density $\frac{1}{b}\exp\left(-\frac{u_i - a}{b}\right)1_{(u_i \ge a)}$, where $b = \sigma_u, \ a = -b$. By direct calculation the

density function of θ_i given y_i is

$$f_{\theta_i|y_i}(t|y) = \frac{exp\left(-\left(t-\left(y-\sigma_e^2\sigma_u^{-1}\right)\right)^2/2\sigma_e^2\right)1_{(t\geq-\sigma_u+\mu)}}{\int_{-\sigma_u+\mu}^{\infty} exp\left(-\left(t-\left(y-\sigma_e^2\sigma_u^{-1}\right)\right)^2/2\sigma_e^2\right)dt} \tag{5.2}$$

The simulations (Figure 2) were done as in Figure 2S (m=100), except that for each value of γ^* we ran 100 simulations and generated 100 random variables from $f_{\theta_i|y_i}(\cdot|\cdot)$, sorted them, and approximated $\widehat{\theta}_{(i)}^{[3]}$.

We can see in Figure 2 that for the symmetric but heavy-tailed Laplace distribution, our shrinkage type predictor $\widehat{\boldsymbol{\theta}}_{()}^{[2]}(\gamma^o)$ (and the same is true for $\widehat{\boldsymbol{\theta}}_{()}^{[2]}(\sqrt{\gamma^*})$) is close to the empirical best predictor $\widehat{\boldsymbol{\theta}}_{()}^{[3]}$, but in the asymmetric case of the Location Exponential distribution, this does not happen.

6 Robustness and comparison with Shen and Louis (1998)

Shen and Louis (1998; henceforth SL) proposed predictors called "Triple-goal estimates" for random effects in two-stage hierarchical models. Their method is in general not analytically tractable, and requires numerical calculations. Moreover, being sensitive to Bayesian assumptions, it is not robust (Shen and Louis (2000)).

The first stage of SL is minimizing $E\int\{A(t;\mathbf{y})-G_m(t)\}^2\,dt$ with the constraint that A is a discrete distribution with at most m mass points, where $G_m(t)$ is the 'empirical' distribution function $G_m(t) = \frac{1}{m}\sum_{i=1}^m I_{(\theta_i\leq t)}$. They show that the solution A is the empirical distribution of $\widehat{\boldsymbol{U}} = (\widehat{U}_1,...,\widehat{U}_m)$, $\widehat{U}_j = \overline{G}_m^{-1}\left(\frac{2j-1}{2m}\right)$, where $\overline{G}_m(t) = E\left(G_m(t)|\mathbf{y}\right) = \frac{1}{m}\sum_{i=1}^m P\left(\theta_i\leq t|y_k\right)$. Therefore $\widehat{U}_j$ is a predictor of $\theta_{(j)},\ j=1,...,m$. The solution $\widehat{U}_j = \overline{G}_m^{-1}\left(\frac{2j-1}{2m}\right)$ depends on the posterior distributions of $\theta_1,....,\theta_m$ and requires estimation of unknown parameters and a solution of nonlinear equations. In order to compute $\overline{G}_m(t)$ in our simulations, we compute $P\left(\theta_i\leq t|y_k\right)$ using the plug-in (or moment) estimator $\overline{y}$ of μ, and (1.1) with the assumption that F and G are normal, and apply Matlab function '**fzero**$'$ for the solution $t=\widehat{U}_j$ of the equations $\overline{G}_m(t) = \frac{2j-1}{2m}$.

For the purpose of checking robustness we generated data taking F to be the Laplace distribution or the asymmetric location exponential distribution, and a normal G. The simulations were done as in Figure 2S (m=100), except

that in the stage of prediction we ignored the true distribution of the random effects and used the normal distribution. Here we compared $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\sqrt{\gamma^*})$, $\widehat{\boldsymbol{\theta}}^{[3]}_{()}$, and the predictor $\widehat{\boldsymbol{U}}$ based on SL. Note that, unlike the estimators in SL, it is not necessary to know the distributions for the predictor $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\sqrt{\gamma^*})$.

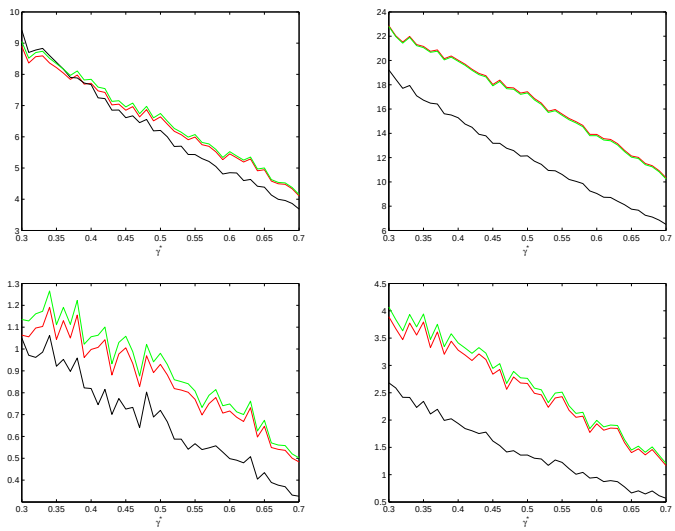

Figure 3:

- Comparison of $E\{L(\widehat{\boldsymbol{\theta}}_{()}, \boldsymbol{\theta}_{()})\}$ as a function of γ^*, for the predictors $\widehat{\boldsymbol{U}}$, $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\sqrt{\gamma^*})$, $\widehat{\boldsymbol{\theta}}^{[3]}_{()}$ (red, black, green lines), where F is the Laplace distribution and G is normal (upper left) and where F is the Location exponential distribution and G is normal (upper right) for, $m = 100$.
- Comparison of the MSE of $\widehat{U}_m$, $\widehat{\theta}^{[2]}_{(m)}(\sqrt{\gamma^*})$, $\widehat{\theta}^{[3]}_{(m)}$ (red, black, green lines) for predicting $\theta_{(m)}$, as a function of γ^*, where F is the Laplace distribution and G is normal (bottom left) and where F is the Location exponential distribution and G is normal (bottom right), for $m = 100$.

In general, the SL estimators and $\widehat{\boldsymbol{\theta}}^{[3]}_{()}$ exhibited very similar performance, see Figure 3. Under the correct assumptions they were somewhat better than our predictor $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\sqrt{\gamma^*})$ (Figure 2); however, they are usually computationally intensive and non-robust against model misspecification. Under misspecification of the distributions in the model, it turned out that $\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\sqrt{\gamma^*})$, which does not

depend on the assumed model was better, as can be sees from the simulations of Figure 3.

7 Appendix: Proofs

Proof of Theorem 3. Without loss of generality take $\mu = 0$. We have

$$E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma), \boldsymbol{\theta}_{()})\} = E\sum_{i=1}^{m}\left(\gamma y_{(i)} + (1-\gamma)\overline{y} - \theta_{(i)}\right)^2$$

$$= E\sum_{i=1}^{m}\left(\gamma y_{(i)} + (1-\gamma)\overline{y} - y_{(i)} + y_{(i)} - \theta_{(i)}\right)^2 = E\sum_{i=1}^{m}\left(y_{(i)} - \theta_{(i)}\right)^2 \tag{7.1}$$

$$+ (1-\gamma)^2 E\sum_{i=1}^{m}\left(y_{(i)} - \overline{y}\right)^2 - 2(1-\gamma)E\sum_{i=1}^{m}\left(y_{(i)} - \theta_{(i)}\right)(y_{(i)} - \overline{y}).$$

Therefore,

$$D(\gamma) := E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma), \boldsymbol{\theta}_{()})\} - E\{L(\widehat{\boldsymbol{\theta}}^{[1]}_{()}, \boldsymbol{\theta}_{()})\} = (1-\gamma)^2 E\sum_{i=1}^{m}\left(y_{(i)} - \overline{y}\right)^2$$

$$- 2(1-\gamma)E\sum_{i=1}^{m}(y_{(i)} - \theta_{(i)})(y_{(i)} - \overline{y}).$$

We calculate each part separately. First note that $E\sum_{i=1}^{m}\left(y_{(i)} - \overline{y}\right)^2 = E\sum_{i=1}^{m}(y_i - \overline{y})^2$ $= (\sigma_u^2 + \sigma_e^2)(m-1)$. Now

$$E\sum_{i=1}^{m}\left(y_{(i)} - \theta_{(i)}\right)(y_{(i)} - \overline{y}) = E\sum_{i=1}^{m} y_{(i)}^2 - E\sum_{i=1}^{m} y_{(i)}\overline{y} - E\sum_{i=1}^{m}\theta_{(i)}y_{(i)} + E\sum_{i=1}^{m}\theta_{(i)}\overline{y}$$

$$= m(\sigma_u^2 + \sigma_e^2) - m\left(\frac{\sigma_u^2 + \sigma_e^2}{m}\right) - E\sum_{i=1}^{m}\theta_{(i)}y_{(i)} + m\left(\frac{\sigma_u^2}{m}\right) = m(\sigma_u^2 + \sigma_e^2) - \sigma_e^2 - E\sum_{i=1}^{m}\theta_{(i)}y_{(i)}. \tag{7.2}$$

Summarizing the above we have

$$D(\gamma) = (1-\gamma)^2(\sigma_u^2+\sigma_e^2)(m-1) - 2(1-\gamma)\Big(m(\sigma_u^2+\sigma_e^2) - \sigma_e^2 - E\sum_{i=1}^{m}\theta_{(i)}y_{(i)}\Big). \tag{7.3}$$

From Lemma 1 (to be proved later)

$$m(\sigma_u^2 + \sigma_e^2) - \sigma_e^2 - m\sqrt{\sigma_u^2(\sigma_u^2 + \sigma_e^2)} \le E\sum_{i=1}^{m}\left(y_{(i)} - \theta_{(i)}\right)(y_{(i)} - \overline{y}) \le (m-1)\sigma_e^2. \tag{7.4}$$

We use the first inequality to deduce that for $\gamma \leq 1$,

$$D(\gamma) \leq (1-\gamma)^2(\sigma_u^2+\sigma_e^2)(m-1) - 2(1-\gamma)\left(m(\sigma_u^2+\sigma_e^2) - \sigma_e^2 - m\sqrt{\sigma_u^2(\sigma_u^2+\sigma_e^2)}\right).$$

Equating the right-hand side to zero and solving the quadratic equation in $1-\gamma$, it is easy to see that $D(\gamma) < 0$ in the interval $\left(\frac{m}{m-1}(2\sqrt{\gamma^*}-1) - \frac{1}{m-1}(2\gamma^*-1), 1\right)$, and the result follows. □

Proof of Corollary 1. Clearly, $E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma), \boldsymbol{\theta}_{()})\} \leq E\{L(\widehat{\boldsymbol{\theta}}^{[1]}_{()}, \boldsymbol{\theta}_{()})\}$ for all $0 \leq \gamma \leq 1$, if $\frac{m}{m-1}(2\sqrt{\gamma^*}-1) - \frac{1}{m-1}(2\gamma^*-1) \leq 0$. Solving the quadratic equation in $\sqrt{\gamma^*}$, we see that the latter inequality holds if either (i) $\gamma^* \leq \left(\frac{m-\sqrt{(m-1)^2+1}}{2}\right)^2$ or (ii) $\gamma^* \geq \left(\frac{m+\sqrt{(m-1)^2+1}}{2}\right)^2$ (≥ 1). Since $\gamma^* \leq 1$ the only possibility is (i), and the proof is complete. □

Proof of Corollary 2. From Theorem 1 it is clear that $E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma), \boldsymbol{\theta}_{()})\} \leq E\{L(\widehat{\boldsymbol{\theta}}^{[1]}_{()}, \boldsymbol{\theta}_{()})\}$ if $\frac{m}{m-1}(2\sqrt{\gamma^*}-1) - \frac{1}{m-1}(2\gamma^*-1) \leq \gamma^* \leq \gamma \leq 1$. The first inequality is equivalent to $\gamma^* \leq \frac{(m-1)^2}{(m+1)^2}$ or $\gamma^* \geq 1$. The case $\gamma^* = 1$ is trivial because in this case $\widehat{\theta}^{[1]}_{(i)} = \widehat{\theta}^{[2]}_{(i)}(\gamma^*)$. □

Proof of Lemma 1. The lower bound is a result of the rearrangement inequality

$$E\sum_{i=1}^m \theta_{(i)} y_{(i)} \geq E\sum_{i=1}^m \theta_i y_i = m(\sigma_u^2+\mu^2).$$

The upper bound follows from

$$E\sum_{i=1}^m \theta_{(i)} y_{(i)} \leq \left(\sum_{i=1}^m E(\theta_i^2)\sum_{i=1}^m E(y_i^2)\right)^{1/2} = m\sqrt{(\sigma_u^2+\mu^2)(\sigma_u^2+\sigma_e^2+\mu^2)},$$

where the inequality follows from the Cauchy-Schwarz inequality. □

Proof of Theorem 4. By the calculations of Theorem 3,

$$E\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma), \boldsymbol{\theta}_{()})\} = E\sum_{i=1}^m \left(y_{(i)} - \theta_{(i)}\right)^2$$
$$+ (1-\gamma)^2(\sigma_u^2+\sigma_e^2)(m-1) - 2(1-\gamma)E\sum_{i=1}^m \left(y_{(i)} - \theta_{(i)})(y_{(i)} - \overline{y}\right).$$

Hence, $dE\{L(\widehat{\boldsymbol{\theta}}^{[2]}_{()}(\gamma), \boldsymbol{\theta}_{()})\}/d\gamma = 0$ if and only if $\gamma = 1 - \frac{E\sum_{i=1}^m \left(y_{(i)}-\theta_{(i)})(y_{(i)}-\overline{y}\right)}{(m-1)(\sigma_u^2+\sigma_e^2)}$, which is a minimum by convexity. We cannot calculate the latter expression exactly, yet the bounds of (7.4) imply the result readily. □

Acknowledgment We thank Danny Pfeffermann for discussions that led to the formulation of the problems studied in this paper. An Associate Editor and the referees made comments that resulted in significant improvements in the paper. This paper is dedicated to the memory of Gideon Schwarz, a teacher and a friend.

This research was supported in part by grant number 473/04 from the Israel Science Foundation.

References

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error component model for predection of county crop areas using survey and satellite data. *Journal of the American Statistical Association* **83**, 28-36.

Blumenthal, S. and Cohen, A. (1968). Estimation of the larger of two normal means. *Journal of the American Statistical Association* **63**, 861-876.

David, H. A. and Nagaraja, N. H. (2003). *Order Statistics* (third edition). Wiley, New York.

Dawid, A. P. (1994). Selection paradoxes of Bayesian inference. *In Multivariate Analysis and its Application* **24**, (eds. T.W. Anderson, K. A-T. A Fang and I. Olkin) Philadelphia, PA:IMS.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269-277 .

Ghosh, M. (1992). Constrained Bayes estimates with application. *Journal of the American Statistical Association* **87**, 533-540.

Kella, O. (1986). On the distribution of the maximum of bivariate normal random variables with general means and variances. *Commun. Statist-Theory Meth.* **15**, 3265-76.

Louis, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association* **79**, 393-398.

Pfeffermann, D. (2002). Small area estimation- new developments and directions. *International Statistical Review* **70**, 125-143.

Rao, J. N. K. (2003). *Small Area Estimation.* Wiley, New York.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**, 15-32.

Rinott, Y. and Samuel-Cahn, E. (1994). Covariance between variables and their order statistics for multivariate normal variables. *Statist. Probab. Lett.* **21**, 153-155.

Senn, S. (2008). A Note concerning a selection Paradox of Dawid's. *The American Statistician* **62**, 206-210.

Shen, W. and Louis, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society B* **60**, 455-471.

Shen, W. and Louis, T. A. (2000). Triple-Goal estimates for Disease Mapping. *Statistics in Medicine* **19**, 2295-2308.

Schwarz, G. (1987). A minimax property of linear regression. *Journal of the American Statistical Association* **82**, 220.

Siegel, A. F. (1993). A surprising covariance involving the minimum of multivariate normal variables. *Journal of the American Statistical Association* **88**, 77-80.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Probability* **1** , 197-206. University of California Press, Berkeley, CA.

Wright, D. L., Stern, H. S., and Cressie, N. (2003). Loss function for estimation of extreme with an application to disease mapping. *The Canadian Journal of Statistics* **31**, 251-266.

first author affiliation

E-mail: (msyakov@mscc.huji.ac.il)

second author affiliation

E-mail: (rinott@mscc.huji.ac.il)