Statistica Sinica **26** (2016), 1709-1728
doi:http://dx.doi.org/10.5705/ss.202014.0034

AN OPTIMAL SHRINKAGE FACTOR IN PREDICTION OF ORDERED RANDOM EFFECTS

Nilabja Guha[1], Anindya Roy[2], Yaakov Malinovsky[2] and Gauri Datta[3]

[1] *Texas A & M University,* [2] *University of Maryland, Baltimore County and* [3] *University of Georgia*

Abstract: The problem of predicting a vector of ordered parameters or its part arises in contexts such as measurement error models, signal processing, data disclosure, and small area estimation. Often estimators of functions of the ordered random effects are obtained under strong distributional assumptions, e.g., normality. We discuss a simple generalized shrinkage estimator for predicting ordered random effects. The proposed approach is distribution free and has significant advantage when there is model misspecification. We give expression to and characterization of the optimal shrinkage parameter; the expression involves the Wasserstein distance between two model-related distributions. We provide a framework for estimating the distance and thereby estimating an empirical version of the oracle optimal estimator. We compare the risk for the optimal predictor to that of other distribution-free estimators. Extensive simulation results are provided to support the theoretical results.

Key words and phrases: Empirical Bayes predictor, linear predictor, order statistics, shrinkage.

1. Introduction

A common model of interest is

$$y_i = \theta_i + e_i, \quad i = 1, 2, \ldots, m, \tag{1.1}$$

where the $\sigma_i^{-1} e_i$ are assumed independent and identically distributed as $H(0, 1)$, a mean zero unit variance distribution, with the constants σ_i assumed to be known. Independent of the e_i, the θ_i are given as $\theta_i = \mu_i + u_i$, where the u_i are independently and identically distributed (i.i.d.) as mean zero, finite variance random variables with distribution G. This model finds wide-spread applications ranging from measurement error models, signal processing, data disclosure, small area estimation to name a few. Here, we develop methodology for predicting the ordered random effects, $\theta_{(i)}$, in the context of Model (1.1).

Model (1.1) is a special case of the Fay-Herriot model (Fay and Herriot (1979)) in small area estimation (SAE), where the area means θ_i are further modeled using area specific covariate information $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})'$, and area

specific random effects u_i as $\theta_i = \boldsymbol{x}_i'\beta + u_i$. Following SAE terminology we refer to $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$ as the vector of area means. While predicting $\eta(\boldsymbol{\theta}) = \boldsymbol{\theta}$ is common, investigators have also studied prediction of other functions; vector of ranks, the empirical distribution of the area means (Shen and Louis (1998)) and the range of area means (Judkins and Liu (2000)). Here we are interested in predicting the vector of order statistics $\eta(\boldsymbol{\theta}) = \boldsymbol{\theta}_{()} = (\theta_{(1)} \leq \theta_{(2)} \leq \cdots \leq \theta_{(m)})$. Prediction of the ordered means is significantly harder than prediction of the linear function of the area means (Pfeffermann (2013)). When G and H are correctly specified, the posterior mean $\hat{\eta} = E(\eta(\boldsymbol{\theta})|\boldsymbol{y})$ minimizes the prediction mean squared error (PMSE), $R(\hat{\eta}) = E[(\hat{\eta} - \eta(\boldsymbol{\theta}))'(\hat{\eta} - \eta(\boldsymbol{\theta}))]$ where $\boldsymbol{y} = (y_1, \ldots, y_m)'$ is the data. When $e_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and $\theta_i \overset{i.i.d.}{\sim} N(\mu, \sigma_\theta^2)$ the Bayes estimator of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}_B = \boldsymbol{y} - (1-\gamma)(\boldsymbol{y} - \mu\mathbf{1})$, where

$$\gamma = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma^2}$$

and $\mathbf{1}$ is a vector of ones. The empirical Bayes (EB) estimator is obtained by replacing μ by $\bar{y}$, $\hat{\boldsymbol{\theta}}_{EB} = \boldsymbol{y} - (1-\gamma)(\boldsymbol{y} - \bar{y}\mathbf{1})$, which is also the Best Linear Unbiased Predictor (BLUP) in the class of all $\{(H, G)\}$ with finite second moments. Brown (1971) and Brown and Greenshtein (2009) have looked at Bayes/empirical Bayes estimation under general prior. The plugged-in version $\eta(\hat{\boldsymbol{\theta}}_B)$, however is not the Bayes predictor and can result in substantial bias in prediction. Wright, Stern, and Cressie (2003) considered a Bayesian scheme for predicting ordered means but their procedure is sensitive to prior choice and requires substantial computation.

When G and H are partially specified up to lower order moments Stein's shrinkage estimators (Stein (1956)) can be used for a variety of parametric functions but for the ordered parameters no suitable predictors are available. When error variances are assumed equal, Malinovsky and Rinott (2010) proposed a class of shrinkage estimators

$$\theta_{(i)}(\lambda) = \lambda y_{(i)} + (1-\lambda)\mu. \tag{1.2}$$

They showed that the risk minimizing value of λ lies in the interval $[\gamma, \sqrt{\gamma}]$ and, based on simulation evidence, conjectured the asymptotic optimal value to be $\sqrt{\gamma}$. The weight $\sqrt{\gamma}$ also appears in Louis (1984), who proposed Bayes and empirical Bayes predictors that minimize an expected distance function between the empirical cdf of predictors of $\boldsymbol{\theta}$ and empirical cdf of its true value.

We use similar shrinkage estimators under model (1.1) and derive expressions for the optimal shrinkage parameter. The optimum estimator is shown to have good finite sample performance with respect to mean squared prediction error, even in comparison to the "best" estimator when G and H are known to be normal. Thus, we provide an estimator that can predict the order parameters with

reasonable accuracy and does not make strong distributional assumptions. For the equal error variance case we show that the optimal choice of λ in (1.2) is not necessarily $\sqrt{\gamma}$, and characterize the cases when $\sqrt{\gamma}$ is indeed the asymptotically optimal choice for λ. Based on the derived expression for the optimal value of λ, we propose a new class of predictors for the ordered parameters and provide a framework for estimation of the optimal predictor. We illustrate its finite sample performance via simulation.

2. Prediction of Ordered Random Effects

It is instructive to begin with a special case of model (1.1) in which the design variances are all assumed to be equal. Under the assumed model, constant error variance would imply that the errors are i.i.d.. Since we later consider the case when the error variances are not equal, we do not separately consider the case where the errors have equal variance but are not necessarily identically distributed. We assume that θ_i arise following some distribution G, with mean μ and variance σ_θ^2, but we do not specify the forms of G and H.

2.1. Prediction in the equal variance model

Assume model (1.1) with constant variances, $\sigma_1^2 = \cdots = \sigma_m^2 = \sigma^2$. The marginal distribution of y_i is denoted by F which under the assumed model has mean μ and variance $\sigma_y^2 = var(y_i) = \sigma_\theta^2 + \sigma^2$. For prediction of the ordered parameters, we consider the class of shrinkage predictors (1.2). Under squared error loss the PMSE for a sample of size m is

$$R_m(\lambda) = m^{-1}E\Big(\sum_{i=1}^{m}(\theta_{(i)} - \lambda y_{(i)} - (1-\lambda)\mu)^2\Big)$$

and the optimal risk minimizing value of λ is

$$\lambda_m^* = \frac{m^{-1}E\left(\sum_{i=1}^{m}(y_{(i)} - \mu)(\theta_{(i)} - \mu)\right)}{m^{-1}E\left(\sum_{i=1}^{m}(y_{(i)} - \mu)^2\right)}. \tag{2.1}$$

We study the limiting form of (2.1) as $m \to \infty$ and evaluate the relative efficiency of the optimal shrinkage coefficient with respect to other predictors. Let

$$\mathrm{W(F,G)} = \Big\{\int_0^1 [\mathrm{F}^{-1}(\mathrm{t}) - \mathrm{G}^{-1}(\mathrm{t})]^2 \mathrm{dt}\Big\}^{1/2}$$

denote the L_2 Wasserstein metric between the distributions F and G, assumed to have finite variance. We consider the predictors

$$\theta_{(i)}(\gamma) = \gamma y_{(i)} + (1-\gamma)\mu \text{ and}$$
$$\theta_{(i)}(\sqrt{\gamma}) = \sqrt{\gamma} y_{(i)} + (1-\sqrt{\gamma})\mu.$$

Following the form of the BLUP for the unordered parameters, a natural choice for the predictor of the ordered quantities would be $\boldsymbol{\theta}_{()}^{(2)} = (\theta_{(1)}(\gamma) \leq \cdots \leq \theta_{(m)}(\gamma))'$, while the predictor $\boldsymbol{\theta}_{()}^{(1)} = (\theta_{(1)}(\sqrt{\gamma}) \leq \cdots \leq \theta_{(m)}(\sqrt{\gamma}))'$ is the form conjectured in Malinovsky and Rinott (2010) to have asymptotically optimum performance. The PMSE associated with the predictors $\boldsymbol{\theta}_{()}^{(1)}$ and $\boldsymbol{\theta}_{()}^{(2)}$ are $R_m^{(1)} = R_m(\sqrt{\gamma})$ and $R_m^{(2)} = R_m(\gamma)$, respectively. For any estimator of the form $\theta_{(i)}(\lambda)$ define the relative efficiency with respect to $\theta_{(i)}(\sqrt{\gamma})$ as $RE_m^{(1)}(\lambda) = R_m^{(1)}/R_m(\lambda)$, and that of $\theta_{(i)}(\lambda)$ with respect to $\theta_{(i)}(\gamma)$ as $RE_m^{(2)}(\lambda) = R_m^{(2)}/R_m(\lambda)$.

Let the distribution of the standardized observations, $(y_i - \mu)/\sigma_y$, be F^* and that of the standardized parameters, $(\theta_i - \mu)/\sigma_\theta$, be G^* where $\sigma_y^2 = \sigma_\theta^2 + \sigma^2$. We require some conditions on F^* and G^*.

(**A**1): The distributions F^* and G^* have finite fourth moments.

(**A**2): For all $0 < t < 1/2$, $F^*(x)$ and $G^*(x)$ have continuous, positive derivatives on $x \in (F^{*-1}(t), F^{*-1}(1-t))$ and $x \in (G^{*-1}(t), G^{*-1}(1-t))$, respectively.

Theorem 1. *Under* (A1−A2)*, as* $m \to \infty$,

$$\begin{aligned}
\lambda_m^* \to \quad \lambda^* &= \sqrt{\gamma}\Big(1 - \frac{W^2(F^*, G^*)}{2}\Big),\\
R_m(\lambda^*) \to \quad R^* &= \sigma_\theta^2\Big(W^2(F^*, G^*) - \frac{W^4(F^*, G^*)}{4}\Big),\\
RE_m^{(1)}(\lambda^*) \to RE^{(1)} &= \Big[1 - \frac{W^2(F^*, G^*)}{4}\Big]^{-1},\\
RE_m^{(2)}(\lambda^*) \to RE^{(2)} &= 1 + \frac{[(1 - W^2(F^*, G^*)/2) - \sqrt{\gamma}]^2}{[1 - (1 - W^2(F^*, G^*)/2)^2]}.
\end{aligned}$$

The gain in PMSE at the optimal shrinkage value over that at $\sqrt{\gamma}$, is $[1 - W^2(F^*, G^*)/4]^{-1}$. This improvement can be quite significant if the Wasserstein distance between F^* and G^* is large and, as $0 \leq W^2(F^*, G^*) \leq 2$, potentially there can be a two fold reduction in the PMSE of the optimal predictor over that of $\theta_{(i)}(\sqrt{\gamma})$. We find in the simulations that the gain in efficiency from the optimal predictor can be substantial.

Remark 1. If F^* and G^* are equal, then $W(F^*, G^*) = 0$ and $\lambda^* = \sqrt{\gamma}$, and the PMSE of the optimal predictor goes to zero as m goes to infinity. Here, the distribution of the centered y_i is a scaled version of that of the centered θ_i, and a simple scaling of the observed values provides the optimal prediction.

In the context of equal error variance Malinovsky and Rinott (2010) conjectured the optimum value of λ in (1.2) to be $\sqrt{\gamma}$. Theorem 1 implies that the

result holds iff $W(F^*, G^*) = 0$. As F^* and G^* are distributions of the standardized quantities, to derive necessary and sufficient condition for $W(F^*, G^*) = 0$, without loss of generality, we take $\mu = 0$.

Theorem 2. *Suppose* (1.1) *holds and the errors* e_i *are independently and identically distributed as* H *with mean zero and variance* σ^2. *Then the Wasserstein distance metric* $W(F^*, G^*)$ *between the distributions of standardized* θ *and standardized* y *is zero if and only if* θ_i *has the same distribution as that of* $\sum_{k=1}^{\infty} c^k e_k$ *where* $c = \sqrt{\gamma} = \sigma_\theta/\sigma_y$.

Proof.
If part : If $\theta = \sum_{k=1}^{\infty} c^k e_k$, then $y = \sum_{k=0}^{\infty} c^k e_k$, where e_k's are i.i.d for all k. Hence, cy has same distribution as θ and after standardization u and y have the same distribution. Hence, $\mathrm{W}(F^*, G^*) = 0$.

Only if part : We can write $cy_i = c\theta_i + ce_i$. From $W(F^*, G^*) = 0$, follows that $y_i^* = cy_i$ has same distribution as θ_i. Iterating the procedure we see that θ has the same distribution as $\sum_{k=1}^{\infty} c^k e_k$. Because $c < 1$, the series representation is valid in mean squared sense.

Remark 2. Normal distributions on θ and e give $W(F^*, G^*) = 0$, and hence Gaussianity is a sufficient condition for Theorem 2 to hold. However, as shown in Theorem 2, the class of distribution pairs (H, G) that will give $W(F^*, G^*) = 0$ is a much wider class containing the normal distribution. In such cases, F^* is a self-decomposable distribution (Lukacs (1970)) and examples of such distribution could be found in Shanbhag and Sreehari (1977).

2.2. Optimal prediction with unequal design variances

With a slight modification, a shrinkage predictor similar to (1.2) can be proposed in the unequal variance case as well. In order to derive the limiting form of the optimal estimator we have to make assumptions about the convergence of the empirical distribution of the standardized responses. Such assumptions automatically hold in the i.i.d. case considered in Section 2.1.

Let $v_i^2 = var(y_i) = \sigma_\theta^2 + \sigma_i^2$ and $z_i = (y_i - \mu)/v_i$. Then we propose a class of shrinkage predictors for the ordered area means as

$$\theta_{(i)}(\lambda) = \lambda z_{(i)} + \mu. \tag{2.2}$$

The PMSE at λ is

$$R_m(\lambda) = m^{-1} E\Big(\sum_{i=1}^{m} (\theta_{(i)} - \lambda z_{(i)} - \mu)^2 \Big).$$

If σ_i^2's are the same, the class of predictors in (2.2) reduces to the class (1.2). Let $\gamma_i = \sigma_\theta^2/(\sigma_\theta^2 + \sigma_i^2)$. Then analogous to the equal variance case, one could look at the predictors $\boldsymbol{\theta}_{()}^{(1)} = (\theta_{(1)}^{(1)} \leq \cdots \leq \theta_{(m)}^{(1)})'$ where $\theta_{(i)}^{(1)} = \sigma_\theta z_{(i)} + \mu$ and $\boldsymbol{\theta}_{()}^{(2)} = (\theta_{(1)}^{(2)} \leq \cdots \leq \theta_{(m)}^{(2)})'$ where $\theta_i^{(2)} = \gamma_i y_i + (1-\gamma_i)\mu$.

Unlike the equal variance case, the predictor $\theta_{(i)}^{(2)}$ does not belong to the class (2.2) but rather it is the ordered version of the area specific BLUP for the unordered area means. Let $R_m^{(1)}$ and $R_m^{(2)}$ denote the PMSE of $\boldsymbol{\theta}_{()}^{(1)}$ and $\boldsymbol{\theta}_{()}^{(2)}$, respectively, with $RE_m^{(1)}(\lambda) = R_m^{(1)}/R_m(\lambda)$ and $RE_m^{(2)}(\lambda) = R_m^{(2)}/R_m(\lambda)$. Let $w_i = (\theta_i - \mu)/\sigma_\theta$ denote the standardized area means. Then the shrinkage coefficient with minimum PMSE is given by

$$\lambda_m^* = \sigma_\theta \frac{m^{-1}E\left(\sum_{i=1}^m w_{(i)} z_{(i)}\right)}{m^{-1}E\left(\sum_{i=1}^m z_i^2\right)}. \tag{2.3}$$

We establish a simpler limiting form for the optimal shrinkage coefficient, leading to suitable predictor that can be used once the unknown parameters have been substituted with data estimates. Let F_m^* and K_m^* denote the empirical distributions of z_i and $\sqrt{\gamma_i} z_i$, respectively.

(**A**3) The sequence of distributions F_m^* and K_m^* are uniformly integrable and converge in distribution to mean zero distributions F^* and K^* with finite fourth moments, respectively. The distribution of w_i, G^* has finite fourth moment.

Theorem 3. *Under* (A3)*, if* (A2) *holds for* F^*, G^* *and* K^**, as* $m \to \infty$

$$\begin{aligned}
\lambda_m^* \to \quad \lambda^* &= \sigma_\theta \left[1 - \frac{W^2(F^*, G^*)}{2}\right], \\
R_m(\lambda^*) \to \quad R^* &= \sigma_\theta^2 \left[1 - \left(1 - \frac{W^2(F^*, G^*)}{2}\right)^2\right], \\
RE_m^{(1)}(\lambda^*) \to RE^{(1)} &= \left[1 - \frac{W^2(F^*, G^*)}{4}\right]^{-1}, \\
RE_m^{(2)}(\lambda^*) \to RE^{(2)} &= \frac{W^2(K^*, G^*)}{[1 - (1 - W^2(F^*, G^*)/2)^2]}.
\end{aligned}$$

Based on the optimal value of the shrinkage coefficient, the proposed predictor for the ordered θ_i is $\boldsymbol{\theta}_{()}^* = (\theta_{(1)}^* \leq \cdots \leq \theta_{(m)}^*)'$ where

$$\theta_{(i)}^* = \sigma_\theta \left[1 - \frac{W^2(F^*, G^*)}{2}\right] z_{(i)} + \mu. \tag{2.4}$$

Remark 3. Our results hold for the more general loss function in which different ordered effects have different weights for their corresponding risks. More details of this can be found in the supplementary document.

2.3. An application to small area estimation

We consider SAE where a fixed area level effect is present, and the mean value of the ith area, $\theta_i = E(y_i \,|\, \theta_i)$, potentially depends on the characteristics of the area. Let $\theta_i = \mu_i + u_i$ and hence $y_i = \mu_i + u_i + e_i$. The μ_i are fixed effects and the u_i are random effects. Typically area specific fixed effects are modeled as $\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$. We take the u_i as i.i.d. $N(0, \sigma_\theta^2)$ and the e_i as i.i.d. $N(0, \sigma_i^2)$. Write the standardized response as $z_i = (y_i - \boldsymbol{x}_i'\beta)/v_i$ where $v_i^2 = \sigma_\theta^2 + \sigma_i^2$ is the variance of y_i. Following the generalized shrinkage estimation development, we can predict $u_{(i)} = \sigma_\theta z_{(i)}$. For predicting $\theta_{(i)}$'s we propose

$$\theta_i^* = \sigma_\theta z_i + \boldsymbol{x}_i'\beta, \tag{2.5}$$

and let $\boldsymbol{\theta}_{()}^* = (\theta_{(1)}^* \le \cdots \le \theta_{(m)}^*)'$ be the ordered values of θ_i^*.

Remark 4. Use of $\boldsymbol{\theta}_{()}^*$ in the equal variance case is justified because maximum a posteriori order for the latent random effects is the same as the order of the observed quantities (under mild distributional assumptions). More details are provided in the supplementary document. We do not address the rank estimation issue directly. A short discussion on rank estimation is included in the supplementary materials, in the context of model 1.1.

3. Empirical Version of the Predictors

In practice, the unknown parameters in the expression for the optimal shrinkage predictor are replaced with their estimators. Thus, at (2.4), one plugs in the estimated values of μ, $W(F^*, G^*)$ and σ_θ.

3.1. Empirical predictor

Unless otherwise mentioned we use the sample mean $\bar{y}$ to estimate μ. Other estimators, such as the sample median can be considered. Estimation of σ_θ is straightforward, but estimation of W is more involved. A consistent method-of-moment estimator of σ_θ^2 is

$$\hat{\sigma}_\theta^2 = \max\Big\{m^{-1}\sum_{i=1}^m y_i^2 - \bar{y}^2 - m^{-1}\sum_{i=1}^m \sigma_i^2, 0\Big\}.$$

Based on the estimated σ_θ, we can replace v_i by $\hat{v}_i = \sqrt{\hat{\sigma}_\theta^2 + \sigma_i^2}$. We also use $\hat{z}_i = (y_i - \bar{y})/\hat{v}_i$ as the observed standardized response in order to compute the Wasserstein distance.

If the family of distributions $G(0, \sigma_\theta^2)$ is known up to σ_θ, then $W(F^*, G^*)$ can be estimated empirically once F^* is estimated based on $\hat{\sigma}_\theta^2$. In cases where G is unknown we can proceed as follows.

We assume that the error distribution is a known finite location-scale mixture of normal distributions (a good approximation to $H(0, \sigma_i^2)$) and that each mixture component is independent of the unobserved θ. We also use a finite normal location scale mixture representation for the distribution of θ. Thus, we take

$$e_i \sim \sum_{l=1}^{L} p_{e,l,i} N(\mu_{e,l,i}, \sigma_{e,l,i}^2), \tag{3.1}$$

where $p_{e,l,i}, \mu_{e,l,i}$, and $\sigma_{e,l,i}^2$ are all known and

$$\theta_i \sim G = \sum_{k=1}^{K} p_{\theta,k} N(\mu_{\theta,k}, \sigma_{\theta,k}^2). \tag{3.2}$$

Then

$$y_i \sim F_i = \sum_{k=1}^{K} \sum_{l=1}^{L} p_{k,l,i} N(\mu_{k,l,i}, \sigma_{k,l,i}^2),$$

where $p_{k,l,i} = p_{\theta,k} p_{e,l,i}$, $\mu_{k,l,i} = \mu_{\theta,k} + \mu_{e,l,i}$, and $\sigma_{k,l,i}^2 = \sigma_{\theta,k}^2 + \sigma_{e,l,i}^2$. One can then use the EM algorithm to estimate the distributions and hence estimate the Wasserstein distance based on the estimated distributions, say $\widehat{W}(F^*, G^*)$.

For computation and implementation, it is more efficient to use the finite sample version of the Wasserstein metric (associated with the finite sample version of the optimal shrinkage) and estimate that to plug-in into the predictor. Set, $W_m^2(F, G) = (1/m) \sum_{i=1}^{m} (F^{-1}(i/(m+1)) - G^{-1}(i/(m+1)))^2$ and $\widetilde{W}_m^2(F, G) = E((1/m) \sum_{i=1}^{m} (F_m^{-1}(i/(m+1)) - G_m^{-1}(i/(m+1)))^2)$. Given the normal location scale mixture representation of G we can generate m independent observations from the distribution of θ and generate a copy of observed y's using the known error distribution. Then, $\widetilde{W}_m(F, G)$ is estimated by its Monte-Carlo estimator. Let, $F_{m,j}^*$ and $G_{m,j}^*$ be the empirical distribution for standardized θ and y in j th replication. We estimate

$$\widehat{\widetilde{W}}_m^2(F^*, G^*) = \frac{1}{R} \sum_{j=1}^{R} \left(\frac{1}{m} \sum_{i=1}^{m} \left({F^*}_{m,j}^{-1}(\frac{i}{m+1}) - {G^*}_{m,j}^{-1}(\frac{i}{m+1}) \right)^2 \right),$$

where R is the number of replications.

Let $\hat{\lambda}^*$ be the value of the optimal shrinkage coefficient when σ_θ and $\widetilde{W}(F^*, G^*)$ have been replaced by their estimators $\hat{\sigma}_\theta$ and $\widehat{\widetilde{W}}(F^*, G^*)$. Then the estimated optimal predictor (2.4) for the ordered area means is $\hat{\boldsymbol{\theta}}_{()}^* = (\hat{\theta}_{(1)}^* \leq \cdots \leq \hat{\theta}_{(m)}^*)'$ where

$$\hat{\theta}^*_{(i)} = \hat{\sigma}_\theta \Big[1 - \frac{\widetilde{\widetilde{W}}^2_m(F^*, G^*)}{2}\Big] \hat{z}_{(i)} + \bar{y}.$$

3.2. Accuracy of the empirical predictor

The empirical estimator is a plug-in version of the optimal predictor. To judge its accuracy we derive asymptotic expression for the PMSE of the empirical predictor. We need some assumptions to establish the asymptotic rates.

(**A4**): For all $0 < t < 1/2$, $F^*(x)$ and $G^*(x)$ have continuous and positive derivative on $x \in (F^{*-1}(t), F^{*-1}(1-t))$ and $x \in (G^{*-1}(t), G^{*-1}(1-t))$, respectively.

(**A5**): Let $S^*_F = \{x : 0 < F^*(x) < 1\}$ and $S^*_G = \{x : 0 < G^*(x) < 1\}$ be the open supports of F^* and G^*. F^* and G^* are twice differentiable on their open supports and their corresponding densities, f^* and g^*, are strictly positive on their respective open supports.

(**A6**): $\displaystyle\int_0^1 \frac{t(1-t)}{f^*(F^{*-1}(t))^2} dt < \infty$ and $\displaystyle\int_0^1 \frac{t(1-t)}{g^*(G^{*-1}(t))^2} dt < \infty$.

(**A7**): $\displaystyle\sup_{0<t<1} \frac{t(1-t)|f^{*\prime}(F^{*-1}(t))|}{f^*(F^{*-1}(t))^2} < \infty$ and $\displaystyle\sup_{0<t<1} \frac{t(1-t)|g^{*\prime}(G^{*-1}(t))|}{g^*(G^{*-1}(t))^2} < \infty$.

(**A8**): The densities $f(x)$ and $g(y)$ are monotone for $x \notin (F^{*-1}(t), F^{*-1}(1-t))$ and $y \notin (G^{*-1}(t), G^{*-1}(1-t))$ for some $0 < t < 1/2$.

(**A9**): There exists $c > 0$, such that $\inf_i \sigma_i^2 > c$, and $\int \hat{\sigma}^{-2}_{\theta,m} \mathbf{1}_{\hat{\sigma}_{\theta,m}>0} < K$ for all $m > m_0$ for some m_0 and K, where $\hat{\sigma}_{\theta,m}$ is the estimate of σ_θ based on m observations.

(**A10**): Assume that $\sqrt{m}$ consistent estimators of $W^2_m(,)$ and $\widetilde{W}^2_m(,)$ – $\widehat{W}^2_m(,)$ and $\widetilde{\widetilde{W}}^2_m(,)$ are available.

Assumptions (A4−A8) can be found in Barrio, Gin, and Utzet (2005) in the context of convergence of integrated quantile differences. Here (A9) is needed for the case when σ_θ is being estimated. Assumption (A10) is plausible since, if the assumed location scale representation is correct, as in that case the MLE of the parameters in the mixture model is $\sqrt{m}$ - consistent for the true value and $W_m(,)$ is a continuous function of the parameters.

Proposition 1. *Under* (A1)−(A2), (A10), *for the equal error variance case at* (2.1), $\sigma_\theta/\sigma_y \left(1 - \widetilde{\widetilde{W}}^2_m(F^*, G^*)/2\right) = \lambda^*_m + O_P(m^{-1/2})$. *If* (A4−A8) *and* (A10) *hold, then* $\sigma_\theta/\sigma_y \left(1 - \widehat{W}^2_m(F^*, G^*)/2\right) = \lambda^*_m + O_P(m^{-1/2})$.

Proposition 2. *Under* (A2)−(A3), (A10), *for the unequal error variance case at* (2.3), $\sigma_\theta\left(1-\widetilde{\widetilde{W}}_m^2(F^*,G^*)/2\right)=\lambda_m^*+O_p(m^{-1/2})$. *Under* (A2)−(A8) *and* (A10), $\sigma_\theta\left(1-\widehat{W}_m^2(F^*,G^*)/2\right)=\lambda_m^*+O_p(m^{-1/2})$.

The following results generalize Proposition 1 and 2 when the optimal shrinkage predictor is based on plugged-in estimators for W_m, σ_θ and μ.

Theorem 4. *Under* (A1)−(A10), *for the equal variance case at* (2.1),

$$\frac{\hat\sigma_\theta}{\hat\sigma_y}\Big(1-\frac{\widetilde{\widetilde{W}}_m^2(F^*,G^*)}{2}\Big)=\lambda_m^*+O_p(m^{-1/2}),$$
$$\frac{\hat\sigma_\theta}{\hat\sigma_y}\Big(1-\frac{\widehat{W}_m^2(F^*,G^*)}{2}\Big)=\lambda_m^*+O_p(m^{-1/2}).$$

If $W(F^*,G^*)>0$,

$$[1-\frac{\widetilde{\widetilde{W}}_m^2(F^*,G^*)}{4}]^{-1}=RE_m^{(1)}(\lambda_m^*)+O_p(m^{-1/2}),$$
$$\frac{\widetilde{\widetilde{W}}_m^2(K^*,G^*)}{[1-(1-\widetilde{\widetilde{W}}_m^2(F^*,G^*)/2)^2]}=RE_m^{(2)}(\lambda_m^*)+O_p(m^{-1/2}).$$

Theorem 5. *Under* (A2)−(A10), *for the unequal design variance case at* (2.3),

$$\hat\sigma_\theta\Big(1-\frac{\widetilde{\widetilde{W}}_m^2(F^*,G^*)}{2}\Big)=\lambda_m^*+O_P(m^{-1/2}),$$
$$\hat\sigma_\theta\Big(1-\frac{\widehat{W}_m^2(F^*,G^*)}{2}\Big)=\lambda_m^*+O_P(m^{-1/2}).$$

If $W(F^*,G^*)>0$,

$$\Big[1-\frac{\widetilde{\widetilde{W}}_m^2(F^*,G^*)}{4}\Big]^{-1}=RE_m^{(1)}(\lambda_m^*)+O_p(m^{-1/2}),$$
$$\frac{\widetilde{\widetilde{W}}_m^2(K^*,G^*)}{[1-(1-\widetilde{\widetilde{W}}_m^2(F^*,G^*)/2)^2]}=RE_m^{(2)}(\lambda_m^*)+O_p(m^{-1/2}).$$

Theorems 4 and 5 provide approximations to the relative efficiency of the estimated version of the optimal shrinkage predictor. In terms of PMSE the proposed optimum shrinkage estimator performs better than the BLUP type estimator in the equal error variance case. But for the unequal variances this may not happen as the BLUP type estimator does not belong to the class of estimators represented by (2.2). In our model-based approach to estimating

$W(F^*, G^*)$, once the normal location-scale mixture model for θ is estimated we can estimate $W(K^*, G^*)$ as well. Thus, the relative efficiency of the BLUP type estimator compared to the proposed optimum estimator can be estimated and the estimator with lower value of estimated asymptotic PMSE can be used.

4. Simulation Study

4.1. Optimum shrinkage and Wasserstein correction

We carried out three examples. The sample sizes considered were $m = 2{,}000$ and $m = 10{,}000$. The larger sample size was chosen to evaluate the accuracy of the estimation of the Wasserstein distance, and compare with the theoretical asymptotic value. Each reported Monte Carlo value was based on 500 replications.

Example 1. The first experimental scenario was designed to evaluate the effect of the Wasserstein distance on the performance of the different predictors. The area mean distributions were chosen to be two component normal scale mixtures parameterized by a single parameter:

$$\theta_i \sim a^{-1}N(0, a-1) + (1-a^{-1})N(0, (a-1)^{-1}).$$

Here $E(\theta) = 0$ and $Var(\theta) = 1$, and $W(F^*, G^*)$ is an increasing function of $a \in [2, \infty)$, with $W = 0$ for $a = 2$. The error distribution was fixed as standard normal. For the simulation we took $a \in \{2, 5, 10, 100, 1{,}000\}$. The scale mixture models for different values of a are denoted by $Nmix(a)$.

Example 2. We look at two distributions for the area means: a Double Exponential distribution to reflect possible heavy tails in the distribution and a two component location mixture of normals to account for possible multimodality in the area mean distribution. In particular we took $\theta_i \sim 0.5N(4,1) + 0.5N(-4,1)$ and $\theta_i \sim DE(\sqrt{2})$, where, in each case, the errors were generated independently using $e_i \sim N(0, \sigma_i^2)$. For the normal mixture case, we considered two cases, constant error variances, $\sigma_i^2 = 16$, and $\sigma_i^2 \sim Uniform(0, 16)$. For the double exponential, we took $\sigma_i^2 = 1$ and $\sigma_i^2 \sim Uniform(0, 1)$. The double exponential models and normal location mixture models with equal and unequal variances are denoted by $DE_E, DE_U, Nmix_E$ and $Nmix_U$, respectively.

The optimum shrinkage coefficient was used in each example. We considered the shrinkage predictors $\boldsymbol{\theta}^*_{()}, \boldsymbol{\theta}^{(1)}_{()}$, and $\boldsymbol{\theta}^{(2)}_{()}$. For prediction we used the estimated version of the predictors where σ^2_θ and W are obtained following the procedure described in Section 3 and plugged into the expression of the predictors. The value $K = 6$ was used.

Table 1. Relative performance of the different shrinkage predictors.

			$\boldsymbol{\theta}^*_{()}$ *vs* $\boldsymbol{\theta}^{(1)}_{()}$			$\boldsymbol{\theta}^*_{()}$ *vs* $\boldsymbol{\theta}^{(2)}_{()}$		
Model	W	$\widehat{W}$	$RE_{2,000}$	$RE_{10,000}$	RE_{∞}	$RE_{2,000}$	$RE_{10,000}$	RE_{∞}
$Nmix(a)$								
$a=$ 2	0	0.02	1.00	1.00	1.00	40.4	199	∞
$a=$ 5	0.25	0.24	1.01	1.01	1.01	2.19	2.12	2.15
$a=$ 10	0.41	0.40	1.04	1.05	1.05	1.25	1.24	1.26
$a=$ 20	0.53	0.47	1.07	1.08	1.08	1.09	1.08	1.07
$a=$ 50	0.62	0.52	1.09	1.09	1.10	1.02	1.02	1.02
$a=100$	0.67	0.61	1.10	1.12	1.13	1.00	1.01	1.01
DE_E	0.14	0.13	1.00	1.00	1.00	4.84	5.13	5.13
DE_U	0.11	0.10	1.00	1.00	1.00	3.67	3.94	3.94
$Nmix_E$	0.37	0.32	1.03	1.03	1.04	1.33	1.35	1.35
$Nmix_U$	0.29	0.28	1.02	1.02	1.03	1.18	1.18	1.18

Table 1 gives the ratio of PMSE of the optimal predictors compared with the other predictors at the two sample sizes. Column 2 gives the value of the true Wasserstein distance and column 3 gives the estimate of W averaged over the Monte Carlo replications. Columns 4-6 give the relative efficiency of the optimal shrinkage estimator $\boldsymbol{\theta}^*_{()}$ compared to the estimator $\boldsymbol{\theta}^{(1)}_{()}$ at sample sizes $m =$ 2,000, 10,000, and $m = \infty$, respectively. The value at $m = \infty$ is the theoretical value given in the Theorem 1. Columns 7−9 give the relative efficiency value of the optimal estimator compared to $\boldsymbol{\theta}^{(2)}_{()}$ at $m =$ 2,000, 10,000, and $m = \infty$, respectively.

In the normal scale mixture models, for smaller values of the Wasserstein distance, the optimal shrinkage predictor $\boldsymbol{\theta}^*_{()}$ and the one ignoring the Wasserstein correction, $\boldsymbol{\theta}^{(1)}_{()}$, are nearly identical. This is expected since for values of W close to zero, the correction factor is close to one and the two predictors essentially coincide. However, the BLUP-type estimator $\boldsymbol{\theta}^{(2)}_{()}$ is much inferior to the other estimators in the small W scenario. When W is large, the optimal estimator is considerably better than $\boldsymbol{\theta}^{(1)}_{()}$, because ignoring the Wasserstein correction has a significant effect on the predictor. Here BLUP-type estimator $\boldsymbol{\theta}^{(2)}_{()}$ is nearly identical to the optimal predictor. For moderate W, the optimal shrinkage provide substantial gains over both $\boldsymbol{\theta}^{(1)}_{()}$ and $\boldsymbol{\theta}^{(2)}_{()}$.

In the normal mean-mixture example for unequal or equal variances $\boldsymbol{\theta}^*_{()}$ and $\boldsymbol{\theta}^{(1)}_{()}$ perform better than the $\boldsymbol{\theta}^{(2)}_{()}$ and, with Wasserstein correction, $\boldsymbol{\theta}^*_{()}$ performs better than $\boldsymbol{\theta}^{(1)}_{()}$ in the equal variance case with relatively higher value of W. For double exponential scenario $\boldsymbol{\theta}^*_{()}$ and $\boldsymbol{\theta}^{(1)}_{()}$ perform much better than the $\boldsymbol{\theta}^{(2)}_{()}$ and, because of the small value of W, the Wasserstein correction is unnecessary for all practical purposes.

Table 2. Relative performance of the different shrinkage predictors.

			$\boldsymbol{\theta}^*_{()}$ *vs* $\boldsymbol{\theta}^{(1)}_{()}$			$\boldsymbol{\theta}^*_{()}$ *vs* $\boldsymbol{\theta}^{(2)}_{()}$		
α	W	$\widehat{W}$	RE_{500}	$\hat{RE}$	RE_∞	RE_{500}	$\hat{RE}$	RE_∞
$\alpha = 0$	0.30	0.27	1.02	1.02	1.02	2.40	2.70	2.64
$\alpha = 0.5$	0.33	0.31	1.02	1.02	1.03	2.82	3.31	3.03
$\alpha = 1$	0.34	0.33	1.02	1.03	1.03	2.97	3.72	3.38

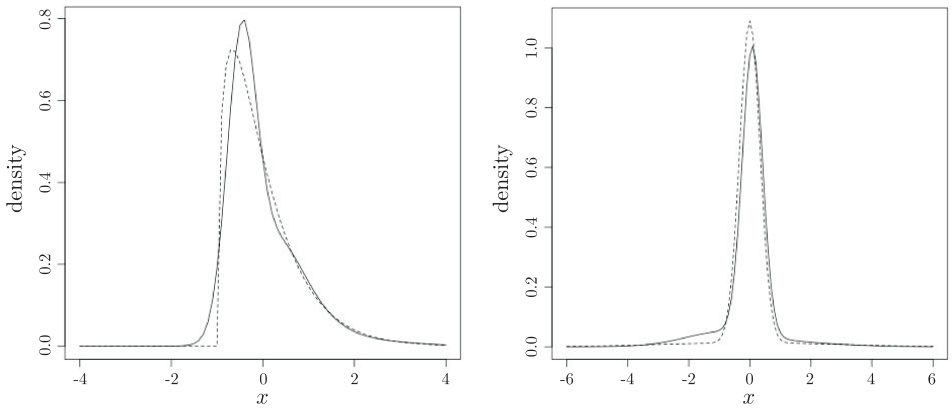

Figure 1. Plot for the nonparametric fit for density of G. The left hand panel shows a typical fit for shifted Gamma distribution for skewed G, the right hand panel is a typical fit in example 1 with $a = 10$. The solid line is the fitted and the dashed line is the true target density.

Example 3. We investigated the effect of skewness in the area mean distribution. We took $\theta_i \sim Gamma(1.5, 1.5) - 1$, an asymmetric distribution for the random effect part around mean zero and with support $(-1, \infty)$. Error terms were normal and we took $\sigma_i^2 = b(\alpha + (1-\alpha)c_i)$, with $c_i = |1 - 2(i/m)|$, $b = 3$ and $0 \le \alpha \le 1$. We considered the cases $\alpha = 0, 0.5$ and 1.

The relative efficiency of $\boldsymbol{\theta}^*_{()}$ with respect to $\boldsymbol{\theta}^1_{()}$ and $\boldsymbol{\theta}^2_{()}$ is given in the Table 2 for $m = 500$ with the data estimate of the relative efficiencies given by $\hat{RE}$, where the number of replication was 100.

In these examples the proposed mixture model approach was able to estimate G. Figure 1 shows the estimated density in two cases; in all cases the proposed estimator provided a reasonable approximation to G.

4.2. Small area estimation

For the SAE model θ_i and e_i are assumed to be normal. By Theorem 2, the Wasserstein distance between the standardized distributions of the area means and the responses is zero. Thus, by Theorem 3, the optimal shrinkage estimator

Table 3. Relative efficiency of the proposed predictor to BLUP-type predictor.

c	$m = 100$	$m = 300$	$m = 500$
1	1.02	1.08	1.09
3	1.08	1.26	1.44
5	1.18	1.62	1.87

$\boldsymbol{\theta}^*_{()}$ is identical to $\boldsymbol{\theta}^{(1)}_{()}$. We considered two cases.

Case 1: In this case $m = 100, 300$, and 500 small areas were considered. We took the case with a single covariate and, for the i th area we had $y_i = \alpha + \beta x_i + u_i + e_i$. We took $u_i \sim N(0, 16)$ with $\alpha = 1, \beta = 2$. We generated $x_i \sim N(0, 1)$ and $e_i \sim N(0, \sigma_i^2)$. The error variance values were generated as $\sigma_i^2 \sim U(0, c)$. We chose c from $c = 1, 3, 5$. The simulation results are reported for 500 Monte Carlo replications. The proposed estimator $\boldsymbol{\theta}^*_{()}$ is compared with the BLUP-type predictor $\boldsymbol{\theta}^{(2)}_{()}$. The relative efficiencies are reported in Table 3.

The proposed predictor outperforms the BLUP-type predictor. The difference is significantly higher when the γ_i are further from one, that is when c is 3 or 5. For large m, the percentage of improvement is generally greater with the optimal predictor providing $10-50\%$ gain in efficiency of prediction.

Case 2: The gain in the performance of the optimal predictor is due to better prediction of the order statistics of the random effects u_i. When the observed values are highly influenced by the fixed effects, then the BLUP-type predictor is expected to perform comparably with the optimal predictor since the main reduction in risk is achieved by accurate prediction of the fixed effect part. To evaluate the effect of the correlation of the responses with the fixed effect on the performance of the optimal predictor we considered a more general case with different values of β. For higher β the fixed effect $\alpha + \beta x_i$ dominates and the response y_i is higher and this effects the performance of the shrinkage estimator. We use the same model as before (Case 1), with $\alpha = 1$, $\beta = 2*i$, where $i = 1, 2, \ldots, 15$. Also, we varied the number of small areas as $m = 100j^2$ with $j = 1, 2, 3, 4, 5$. The area specific variances were generated as $U(0, c)$, with $c = 1, 3, 5, 8$. The relative efficiencies of the BLUP-type predictor $\boldsymbol{\theta}^{(2)}_{()}$ compared to the optimal predictor $\boldsymbol{\theta}^{(1)}_{()}$ are given in Figure 2. The lower the relative efficiency, the better the performance of the proposed predictor in predicting the order statistics.

The proposed predictor has significantly smaller PMSE for a variety of cases where the correlation between the fixed effect and the response is moderate to small. For larger values of β the responses are essentially the fixed effects, and

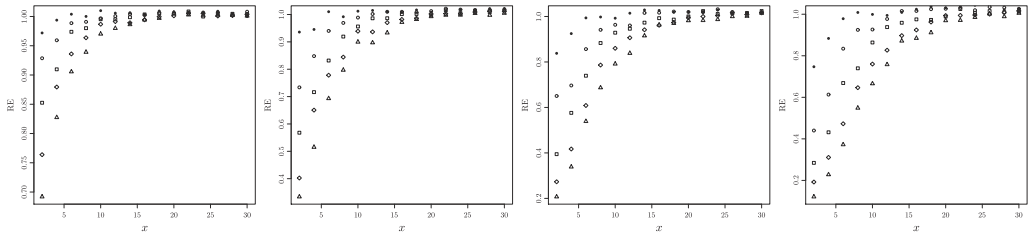

Figure 2. From left to right the figures correspond to $c = 1, 3, 5, 8$, respectively. The X-axis denotes $i = 1, \ldots, 30$. Here $m = 100j^2$ and $j = 1, 2, 3, 4, 5$ are denoted by $\bullet, \circ, \square, \diamond$ and $\triangle$, respectively.

the relative efficiency approaches one. Since the optimal predictor is being replaced by its plug-in version, the relative efficiency of the BLUP-type estimator is actually slightly larger than one, when the area means are essentially of the same magnitude as that of the fixed part $\alpha + \beta x_i$.

4.3. Comparison with the full Bayesian estimator

If the distributions G and H are known, Bayesian computation can be used to generate the posterior samples of θ_i and estimate the posterior expectation of the ordered means by taking the mean of the ordered posterior samples. The posterior mean is the best estimator in terms of PMSE. It is interesting to compare the Bayesian method to our distribution free approach. We also study the sensitivity of the Bayesian estimator to model misspecification, especially in the model for the random effects parameter θ. We considered *Students t* distribution with various degrees of freedom for G and assumed it to be misspecified as normal distribution. Also, a mixture normal distribution was considered for θ with $G \sim .5N(1.5, 1) + .5N(-1.5, 1)$. We took, equal variances (E) $\sigma_i^2 = 1$ and unequal variances (U) $\sigma_i^2 \sim U(0, 1)$. In the Bayesian methodology we assumed the prior on θ, G, to be $N(\mu, \sigma^2)$ and assumed the following non-informative priors $\Pi(\mu, \sigma^2) \propto 1$. The ratio of the square root of the PMSE's for the shrinkage and the Bayesian methods are reported in Table 4. If the model is truly specified then the ratio should be greater than one.

Under model misspecification the shrinkage generally performs better than the Bayesian method, and for the correctly specified case the Bayesian is better, as expected. However, even in the correctly specified model, the model free estimator continues to have reasonable performance.

Table 4. Relative efficiency of the optimal shrinkage and Bayes predictor.

G	m=100		m=400		m=900	
	U	E	U	E	U	E
$T_{2.2}$	0.87	0.81	0.71	0.68	0.66	0.63
T_3	0.92	0.87	0.77	0.73	0.69	0.68
T_4	1.00	0.93	0.82	0.79	0.75	0.72
T_6	1.02	1.04	0.92	0.92	0.82	0.81
$N(0,1)$	1.19	1.22	1.23	1.22	1.23	1.21
Mixture	1.04	0.98	0.86	0.85	0.77	0.78

5. Conclusion

We propose an optimal estimator of ordered random effects in the class of simple shrinkage estimators. The main attraction of the proposed estimator is that it is distribution free. It is very robust to model misspecification and, as evident from simulations the distribution free estimator is reasonably efficient with respect to the best (Bayesian) estimator in a correctly specified normal model. The estimator provides an easy and direct way for predicting ordered random effects under both equal and unequal error variances. From simulations, in many setups we see significant relative gain in efficiency for the optimal shrinkage estimator over other shrinkage estimators in the same class. We also derived limiting form of the risk of the estimator and proposed a method for estimating the risk. The estimator depends on the Wasserstein distance between two standardized distributions and a method based on the observed y_is is also given for estimating it. The optimal estimator based on the estimated Wasserstein distance is shown to have reasonable asymptotic properties. We also address the situation when area specific covariate information is available.

Alternative classes of estimators could involve linear shrinkage estimators with different shrinkage for the different order statistics, or classes of shrinkage estimators that directly account for the joint dependence among the order statistics. Inference in such classes may be more difficult, and this is an interesting area for further exploration. Generalization to classes of estimators that can effectively account for area specific covariates while having the advantage of being distribution free is a topic of future research. The proposed estimator gives a good starting point and a preliminary framework for more general classes of distribution-free estimators.

Supplementary Materials

The supplementary document has four sections. Details about Remark 3 are given in Section 1. The justification and proof of results related to the small area model in Remark 4 are given in Section 2. In the third section we prove some

results from the manuscript and in the fourth, we discuss the rank estimation issue.

Acknowledgement

We thank an associate editor and two reviewers for their thoughtful comments and suggestions. The authors would like to thank Dr. A. M. Kagan for pointing out the connection to self-decomposable distributions.

Appendix

Proof of Theorem 1. From (2.1)

$$\lambda_m^* = \frac{m^{-1}E\left(\sum_{i=1}^m (y_{(i)}-\mu)(\theta_{(i)}-\mu)\right)}{m^{-1}E\left(\sum_{i=1}^m (y_{(i)}-\mu)^2\right)}.$$

Hence,

$$\lambda_m^* = \frac{m^{-1}\frac{\sigma_\theta}{\sigma_y}E\left(\sum_{i=1}^m [(y_{(i)}-\mu)/\sigma_y][(\theta_{(i)}-\mu)/\sigma_\theta]\right)}{m^{-1}E\left(\sum_{i=1}^m ((y_{(i)}-\mu)/\sigma_y)^2\right)} = \frac{\sigma_\theta}{\sigma_y}E(S_{(m)}^*),$$

where

$$S_{(m)}^* = \frac{1}{2}\Big\{\frac{1}{m}\sum_{i=1}^m {z_i}^2 + \frac{1}{m}\sum_{i=1}^m {w_i}^2 - \frac{1}{m}\sum_{i=1}^m \Big(z_{(i)} - w_{(i)}\Big)^2\Big\}.$$

Because $E((1/m)\sum_{i=1}^m z_i^2) = 1$ and $E((1/m)\sum_{i=1}^m w_i^2) = 1$ we have

$$S_{(m)}^* = \frac{1}{2}\Big\{\frac{1}{m}\sum_{i=1}^m {z_i}^2 + \frac{1}{m}\sum_{i=1}^m {w_i}^2 - T_1\Big\},$$

where

$$T_1 = \frac{1}{m}\sum_{i=1}^m \Big({F_m^*}^{-1}(\frac{i}{m+1}) - {G_m^*}^{-1}(\frac{i}{m+1})\Big)^2.$$

Then,

$$T_1 = d_m(F_m^*, F^*) + d_m(G_m^*, G^*) + d_m(F^*, G^*) + C_1 + C_2 + C_3, \qquad \text{(A.1)}$$

where

$$d_m(F_m^*, F^*) = \frac{1}{m}\sum_{i=1}^m ({F_m^*}^{-1}(\frac{i}{m+1}) - {F^*}^{-1}(\frac{i}{m+1}))^2,$$

$$d_m(G_m^*, G^*) = \frac{1}{m}\sum_{i=1}^m ({G_m^*}^{-1}(\frac{i}{m+1}) - {G^*}^{-1}(\frac{i}{m+1}))^2,$$

$$d_m(F^*, G^*) = \frac{1}{m}\sum_{i=1}^m ({F^*}^{-1}(\frac{i}{m+1}) - {G^*}^{-1}(\frac{i}{m+1}))^2,$$

$$C_1 = \frac{2}{m}\sum_{i=1}^{m}({F_m^*}^{-1}(\frac{i}{m+1}) - {F^*}^{-1}(\frac{i}{m+1}))({G_m^*}^{-1}(\frac{i}{m+1}) - {G^*}^{-1}(\frac{i}{m+1})),$$

$$C_2 = \frac{2}{m}\sum_{i=1}^{m}({F^*}^{-1}(\frac{i}{m+1}) - {G^*}^{-1}(\frac{i}{m+1}))({G_m^*}^{-1}(\frac{i}{m+1}) - {G^*}^{-1}(\frac{i}{m+1})),$$

$$C_3 = \frac{2}{m}\sum_{i=1}^{m}({F_m^*}^{-1}(\frac{i}{m+1}) - {F^*}^{-1}(\frac{i}{m+1}))({F^*}^{-1}(\frac{i}{m+1}) - {G^*}^{-1}(\frac{i}{m+1})).$$

From Lemma A.1, we have $E(d_m(F_m^*, F^*)) \to 0$, $E(d_m(G_m^*, G^*)) \to 0$, and $d_m(F^*, G^*) \to \mathrm{W}^2(\mathrm{F}^*, \mathrm{G}^*)$. We state and prove Lemma A.1 later.

For C_1,

$$E(C_1^2) \le E\Big(\frac{2}{m}\sum_{i=1}^{m}({F_m^*}^{-1}(\frac{i}{m+1}) - {F^*}^{-1}(\frac{i}{m+1}))^2\Big) \times E\Big(\frac{2}{m}\sum_{i=1}^{m}({G_m^*}^{-1}(\frac{i}{m+1}) - {G^*}^{-1}(\frac{i}{m+1}))^2\Big). \tag{A.2}$$

From Lemma A.1 $E(C_1) \to 0$. Similarly $E(C_2), E(C_3) \to 0$.

From (A.1), Lemma A.1, and (A.2) we have

$$E\Big(\frac{1}{m}\sum_{i=1}^{m}\Big({F_m^*}^{-1}(\frac{i}{m+1}) - {G_m^*}^{-1}(\frac{i}{m+1})\Big)^2\Big) \to W^2(F^*, G^*).$$

Then, $\lambda_m^* \to \lambda^* = \sqrt{\gamma}(1 - W^2(F^*, G^*)\ /2)$ and thus,

$$R_m^{(1)} = \sum_{i=1}^{m}(\theta_{(i)} - \mu - \sqrt{\gamma}(y_{(i)} - \mu))^2 = \sigma_\theta^2\sum_{i=1}^{m}(w_{(i)} - z_{(i)})^2 \to \sigma_\theta^2 W^2(F^*, G^*).$$

Also,

$$\sqrt{\gamma} - \lambda_m^* = \frac{\sigma_\theta}{\sigma_y}\frac{W^2(F^*, G^*)}{2}.$$

Hence,

$$R_m(\lambda) \to R_m^{(1)} - \sigma_\theta^2\frac{W^4(F^*, G^*)}{4} = R^*.$$

To find a direct expression for $R_m^{(2)}(\gamma)$, we use

$$R_m^{(2)}(\gamma) = \sigma_\theta^2 m^{-1}E\Big(\sum_{i=1}^{m}(w_{(i)} - \sqrt{\gamma}z_{(i)})^2\Big) = \sigma_\theta^2\Big(1+\gamma-2\sqrt{\gamma}m^{-1}E\Big(\sum_{i=1}^{m}z_{(i)}w_{(i)}\Big)\Big).$$

We have shown that $m^{-1}E(\sum_{i=1}^{m} z_{(i)}w_{(i)})) \to (1 - W^2(F^*, G^*)/2)$. Hence,

$$R_m^{(2)}(\gamma) \to \sigma_\theta^2\left(1 + (\sqrt{\gamma} - (1 - \frac{W^2(F^*, G^*)}{2}))^2 - (1 - \frac{W^2(F^*, G^*)}{2})^2\right)$$

$$= R^* + \sigma_\theta^2 \left(\sqrt{\gamma} - (1 - \frac{W^2(F^*, G^*)}{2}) \right)^2 .$$

Lemma A.1. *Under* A1*, and* A2*, as* m *goes to infinity,* $d_m(F^*, G^*) \to \mathrm{W}^2(\mathrm{F}^*, \mathrm{G}^*)$ *and* $E(d_m(F_m^*, F^*))$*,* $E(d_m(G_m^*, G^*))$ *converges to zero.*

The proof is in the supplementary materials.

Proof of Theorem 3. Let

$$\begin{aligned} S_{(m)}^* &= \frac{1}{2}\{\frac{1}{m}\sum_{i=1}^m {z_i}^2 + \frac{1}{m}\sum_{i=1}^m {w_i}^2 - \frac{1}{m}\sum_{i=1}^m (z_{(i)} - w_{(i)})^2\} \\ &= \frac{1}{2}\{\frac{1}{m}\sum_{i=1}^m {z_i}^2 + \frac{1}{m}\sum_{i=1}^m {w_i}^2 - \frac{1}{m}\sum_{i=1}^m ({F_m^*}^{-1}(\frac{i}{m+1}) - {G_m^*}^{-1}(\frac{i}{m+1}))^2\}. \quad \text{(A.3)} \end{aligned}$$

Here, $E((1/m)\sum_{i=1}^m z_i^2) = 1$ and $E((1/m)\sum_{i=1}^m w_i^2) = 1$. The last summation converges to $W^2(F^*, G^*)$ by Lemma A.1, similar to Theorem 1.

Thus, $R_m^{(1)} = \sum_{i=1}^m (\theta_{(i)} - \mu - z_{(i)})^2 = \sigma_\theta^2 \sum_{i=1}^m (w_{(i)} - z_{(i)})^2 \to \sigma_\theta^2 W^2(F^*, G^*)$. Similarly, $R_m^{(2)} \to \sigma_\theta^2 W^2(K^*, G^*)$. As, $\sigma_\theta - \lambda^* = \sigma_\theta W^2(F^*, G^*)/2$, $R_m(\lambda) \to \sigma_\theta^2 W^2(F^*, G^*) - \sigma_\theta^2 W^4(F^*, G^*)/4$.

References

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42**, 855-904.

Brown, L. D. and Greenshtein, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high dimensional vector of normal means. *Ann. Statist.* **37**, 1685-1704.

Barrio, E. D., Gin, E. and Utzet, F. (2005). Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli* **11,** 131-189.

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.* **74**, 269-277.

Judkins, D. R. and Liu, J. (2000). Correcting the bias in the range of a statistic across small areas. *J. Official Statist.* **16**, 1-13.

Louis, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *J. Amer. Statist. Assoc.* **79**, 393-398.

Lukacs, E. (1970). *Characteristic Functions.* 2nd edition. Griffin, London.

Malinovsky, Y. and Rinott, Y. (2010). Prediction of ordered random effects in a simple small area model. *Statist. Sinica* **20,** 697-714.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statist. Sci.* **28,** 40-68.

Shanbhag, D. N. and Sreehari, M. (1977). On certain self-decomposable distributions. Z. Wahrsch. Verw. Gebiete **38**, 217-222.

Shen, W. and Louis, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *J. Roy. Statist. Soc. Ser. B* **60**, 455-471.

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

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

Department of Statistics, Texas A & M University, College Station, TX 77843, USA.

E-mail: nguha@math.tamu.edu

Department of Mathematics and Statistics, University of Maryland, Baltimore County Baltimore, MD 21250, USA.

E-mail: anindya@umbc.edu

Department of Mathematics and Statistics, University of Maryland, Baltimore County Baltimore, MD 21250, USA.

E-mail: yaakovm@umbc.edu

Department of Statistics, University of Georgia, 101 Cedar Street, Athens, GA 30602, USA.

E-mail: gauri@uga.edu

(Received August 2014; accepted November 2015)