Statistica Sinica Preprint No: SS-2014-0034R1

Title	An Optimal Shrinkage Factor in Prediction of Ordered Random Effects
Manuscript ID	SS-2014-0034R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202014.0034
Complete List of Authors	Nilabja Guha Anindya Roy Yaakov Malinovsky and Gauri Datta
Corresponding Author	Nilabja Guha
E-mail	nguha@math.tamu.edu

Notice: Accepted version subject to English editing.

An Optimal Shrinkage Factor in Prediction of Ordered Random Effects

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

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

Abstract: The problem of predicting a vector of ordered parameters or its part arises in the context of a variety of applications in 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 generalized simple shrinkage estimator for predicting ordered random effects. The proposed approach is distribution free and has significant advantage when there is model misspecification. One of the main contributions is an expression 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 also evaluate the relative efficiency gain by comparing 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, Shrinkage, Order statistics, Linear predictor.

1 Introduction

A common model of interest is

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

where $\sigma_i^{-1} e_i$ are assumed to be distributed independently and identically as $H(0,1)$, a mean zero unit variance distribution where the constants σ_i are assumed to be known. Independent of e_i the quantities θ_i are assumed to be distributed as $G(\cdot - \mu_i)$, i.e., $\theta_i = \mu_i + u_i$ where u_i are independently and identically distributed (iid) as mean zero and 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. In this paper, we develop methodology for predicting the ordered random effects, $\theta_{(i)}$, in the context of model (1).

Model (1) is a special case of the Fay-Herriot model (Fay and Herriot; 1979) in small area estimation (SAE), given by 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 and empirical distribution of the area means (Shen and Louis; 1998) , 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 $e_i \overset{iid}{\sim} N(0, \sigma^2)$ and $\theta_i \overset{iid}{\sim} N(\mu, \sigma_\theta^2)$ 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), 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 may result in substantial bias in prediction. Wright, Stern and Cressie (2003) considered a Bayesian scheme for predicting ordered means. However, the 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. However, for the ordered parameters no suitable predictors are available. When error variances are assumed to be equal, Malinovsky and Rinott (2010) proposed a class of shrinkage estimators:

$$\theta_{(i)}(\lambda) = \lambda y_{(i)} + (1-\lambda)\mu. \tag{2}$$

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 simple shrinkage estimators under model (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, the main contribution of the paper is to 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 (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. We provide a framework for estimation of the optimal predictor and 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) in which the design variances are all assumed to be equal. Under the assumed model, constant error variance would imply that the errors are iid. 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) with constant design variances, i.e., $\sigma_1^2 = \cdots = \sigma_m^2 = \sigma^2$. Let the marginal distribution of y_i be denoted by F which under the assumed model will have 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 (2). Under the squared error loss the PMSE for a sample of size m is

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

Based on the theory of ordinary least squares estimators, it is immediate that the optimal risk minimizing value of λ can be derived as

$$\lambda_m^* = \frac{m^{-1} E\left(\sum_{i=1}^{m} (y_{(i)} - \mu)(\theta_{(i)} - \mu)\right)}{m^{-1} E\left(\sum_{i=1}^{m} (y_{(i)} - \mu)^2\right)}. \qquad (3)$$

To investigate the limiting form of the optimal shrinkage coefficient (3) as $m \to \infty$ and to draw comparison and evaluate the relative efficiency of the optimal shrinkage coefficient with respect to other predictors we first define a few quantities. Let $W(F, G)$ be defined as

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

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

$$\begin{aligned}\theta_{(i)}(\gamma) &= \gamma y_{(i)} + (1-\gamma)\mu,\\ \theta_{(i)}(\sqrt{\gamma}) &= \sqrt{\gamma} y_{(i)} + (1-\sqrt{\gamma})\mu.\end{aligned}$$

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 $\boldsymbol{\theta}^{(1)}_{()} = (\theta_{(1)}(\sqrt{\gamma}) \leq \cdots \leq \theta_{(m)}(\sqrt{\gamma}))'$ would be the form conjectured in Malinovsky and Rinott (2010) to have the asymptotically optimum performance. The PMSE associated with the predictors $\boldsymbol{\theta}^{(1)}_{()}$ and $\boldsymbol{\theta}^{(2)}_{()}$ are $R^{(1)}_m = R_m(\sqrt{\gamma})$ and $R^{(2)}_m = R_m(\gamma)$, respectively. For any estimator of the form $\theta_{(i)}(\lambda)$ define the relative efficiency with respect to $\theta_{(i)}(\sqrt{\gamma})$ as $RE^{(1)}_m(\lambda) = \frac{R^{(1)}_m}{R_m(\lambda)}$. Similarly, define the relative efficiency of $\theta_{(i)}(\lambda)$ with respect to $\theta_{(i)}(\gamma)$ as $RE^{(2)}_m(\lambda) = \frac{R^{(2)}_m}{R_m(\lambda)}$.

Let the distribution of the standardized observations, $\frac{y_i-\mu}{\sigma_y}$, be F^* and that of the standardized parameters, $\frac{\theta_i-\mu}{\sigma_\theta}$, be G^* where $\sigma^2_y = \sigma^2_\theta + \sigma^2$. Suppose the following conditions are satisfied.

- (A1): The distributions F^* and G^* have finite fourth moments.
- (A2): 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.

Then the following result holds.

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

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

The gain in PMSE at the optimal shrinkage value over that at $\sqrt{\gamma}$, is $[1 - \quad W^2(F^*, G^*)/4]^{-1}$. This improvement can be quite significant if the Wasserstein distance between F^* and G^* is large. Since $0 \leq W^2(F^*, G^*) \leq 2$, potentially there can be a two fold reduction in the PMSE of the optimal predictor over that of $\theta_{(i)}(\sqrt{\gamma})$. However, given that F^* and G^* are related via the convolution equation, the maximum value of $W^2(F^*, G^*)$ is possibly smaller than two, putting an upper bound on the gain in efficiency of the optimal shrinkage predictor over that with shrinkage coefficient $\sqrt{\gamma}$. We show in the simulation section 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 hence $\lambda^* = \sqrt{\gamma}$ and the PMSE of the optimal predictor will go to zero as m goes to infinity. Intuitively, given that the distribution of the centered y_i is a scaled version of that of the centered θ_i, 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 (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 the necessary and sufficient condition for $W(F^*, G^*) = 0$, without loss of generality, we assume $\mu = 0$.

Theorem 2. *Let the model* (1) *hold and let the errors e_i be 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 $c = \sqrt{\gamma} = \frac{\sigma_\theta}{\sigma_y}$.*

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, 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 will 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

A more general model is one where the variances, σ_i^2, are different. With a slight modification, a shrinkage predictor similar to (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 iid case considered in the section 2.1. Let $v_i^2 = var(y_i) = \sigma_\theta^2 + \sigma_i^2$ and $z_i = \frac{y_i - \mu}{v_i}$. Then we propose a class of shrinkage predictors for the ordered area means as

$$\theta_{(i)}(\lambda) = \lambda z_{(i)} + \mu. \tag{4}$$

The PMSE at λ is defined as

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

If σ_i^2's are the same, the class of predictors in (4) reduces to the class (2). Suppose $\gamma_i = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_i^2}$. 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 (4) 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. Also let $RE_m^{(1)}(\lambda) = \frac{R_m^{(1)}}{R_m(\lambda)}$ and $RE_m^{(2)}(\lambda) = \frac{R_m^{(2)}}{R_m(\lambda)}$

denote the relative efficiencies of the predictors $\theta_{(i)}(\lambda)$ with respect to $\boldsymbol{\theta}_{()}^{(1)}$ and $\boldsymbol{\theta}_{()}^{(2)}$, respectively. Let $w_i = \frac{\theta_i - \mu}{\sigma_\theta}$ denote the standardized area means. Then the shrinkage coefficient with minimum PMSE is given by

$$\lambda_m^* = \sigma_\theta \frac{m^{-1} E\left(\sum_{i=1}^m w_{(i)} z_{(i)}\right)}{m^{-1} E\left(\sum_{i=1}^m z_i^2\right)}. \qquad (5)$$

As before, we can establish a simpler limiting form for the optimal shrinkage coefficient, thereby letting us propose a 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. In order to establish a limiting value of the shrinkage coefficient we assume:

- (A3) The sequence of distributions F_m^* and K_m^* converge in distribution to mean zero distributions F^* and K^* with finite fourth moments, respectively. Moreover, F_m^* and K_m^* are assumed to be uniformly integrable. Let G^* be the distribution of w_i with finite fourth moment.

Then the limit of λ_m^* is given by

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

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

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

$$\theta_{(i)}^* = \sigma_\theta \left[1 - \frac{W^2(F^*, G^*)}{2}\right] z_{(i)} + \mu. \qquad (6)$$

Remark 3. *The results of Section 2 and also, the approximation results following hold for more general loss function where different ordered effects have different weights for their corresponding risks i.e the risk is of the form* $R_m(\lambda) = m^{-1}E\left(\sum_{i=1}^{m} \xi_i(\theta_{(i)} - \lambda z_{(i)} - \mu)^2\right)$ *and* $\xi_i = \xi^*(\frac{i}{m+1})$*, where* ξ^* *is a positive integrable function on* $(0,1)$*. A more detailed account of this result is given in the supplementary document.*

2.3 An application to small area estimation

The method of estimating ordered random effect can be extended to 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 and hence may be different for different areas. Specifically, 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 covariate information, $\boldsymbol{x}_i$ are available and the area specific fixed effects are modeled as $\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$. Often in SAE u_i are iid $N(0, \sigma_\theta^2)$ and e_i are iid $N(0, \sigma_i^2)$. Let the standardized response be $z_i = \frac{y_i - \boldsymbol{x}_i'\beta}{v_i}$ 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

$$\theta_i^* = \sigma_\theta z_i + \boldsymbol{x}_i'\beta. \tag{7}$$

and let $\boldsymbol{\theta}_{()}^* = (\theta_{(1)}^* \leq \cdots \leq \theta_{(m)}^*)'$ 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 the rank estimation is included in the supplementary materials, in the context of model 1.*

3 Empirical version of the predictors

In practice, the unknown parameters in the expression for the optimal shrinkage predictor have to be replaced with their estimators. Thus, in order to use the optimal predictor (6) one has to plug in the estimated values of μ, $W(F^*, G^*)$ and σ_θ.

3.1 Empirical Predictor

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

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

Based on the estimated σ_θ, we can replace v_i by $\hat{v}_i = \sqrt{\hat{\sigma}_\theta^2 + \sigma_i^2}$. We will also use $\hat{z}_i = \frac{y_i - \bar{y}}{\hat{v}_i}$ 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 in the following manner.

We will 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 will also use a finite normal location scale mixture representation for the distribution of θ, thereby invoking a similar representation for the distribution of y. Suppose,

$$e_i \sim \sum_{l=1}^L p_{e,l,i} N(\mu_{e,l,i}, \sigma_{e,l,i}^2), \qquad (8)$$

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

$$\theta_i \sim G = \sum_{k=1}^K p_{\theta,k} N(\mu_{\theta,k}, \sigma_{\theta,k}^2). \qquad (9)$$

Then

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

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

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. Define, $W^2_m(F,G) = \frac{1}{m}\sum_{i=1}^{m}(F^{-1}(\frac{i}{m+1}) - G^{-1}(\frac{i}{m+1}))^2$ and $\widetilde{W}^2_m(F,G) = E(\frac{1}{m}\sum_{i=1}^{m}(F_m^{-1}(\frac{i}{m+1}) - G_m^{-1}(\frac{i}{m+1}))^2)$. 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

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

where R is the number of replications.

Let $\hat{\lambda}^*$ be the value of the optimal shrinkage coefficient when both σ_θ and $\widetilde{W}(F^*, G^*)$ have been replaced by their estimators $\hat{\sigma}_\theta$ and $\widehat{\widetilde{W}}(F^*, G^*)$, respectively. Then the estimated optimal predictor (6) for the ordered area means will be $\hat{\boldsymbol{\theta}}^*_{()} = (\hat{\theta}^*_{(1)} \leq \cdots \leq \hat{\theta}^*_{(m)})'$ where

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

3.2 Accuracy of the empirical predictor

The empirical estimator is a plug-in version of the optimal predictor. The natural question in terms of performance of the empirical predictor is how well does the plug-in version fare with respect to the oracle predictor. To judge the accuracy we derive asymptotic expression for the PMSE of the empirical predictor. We make the following assumptions that are needed 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^*, respectively. 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): Assume $\int_0^1 \frac{t(1-t)}{f^*(F^{*-1}(t))^2} dt < \infty$ and $\int_0^1 \frac{t(1-t)}{g^*(G^{*-1}(t))^2} dt < \infty$.

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

- (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 $\int \hat{\sigma}_{\theta,m}^{-2} \mathbf{1}_{\hat{\sigma}_{\theta,m}>0} < K$ 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 $W_m^2(,)$ and $\widetilde{W}_m^2(,)$, given by $\widehat{W}_m^2(,)$ and $\widehat{\widetilde{W}}^2_m(,)$, respectively are available.

Assumptions (A4-A8) can be found in Barrio et al.(2005) in the context of convergence of integrated quantile differences. Assumption (A9) is needed for the case when σ_θ is being estimated. We also assume the existence of $\sqrt{m}$ consistent estimators of $W_m^2(,)$ and $\widetilde{W}_m^2(,)$, given by $\widehat{W}_m^2(,)$ and $\widehat{\widetilde{W}}^2_m(,)$, respectively. This assumption is reasonable if the assumed location scale representation is correct, as in that case the MLE of the parameters in the mixture model will be $\sqrt{m}$ consistent for the true value and $W_m(,)$ is a continuous function of the parameters.

Proposition 1. *Under (A1)-(A2),(A10) for equal error variance case* (3), $\frac{\sigma_\theta}{\sigma_y}\left(1-\frac{\widetilde{\widehat{W}}^2_m(F^*,G^*)}{2}\right) = \lambda^*_m + O_P(m^{-\frac{1}{2}})$. *In addition if (A4-A8) and (A10) hold then,* $\frac{\sigma_\theta}{\sigma_y}\left(1-\frac{\widehat{W}^2_m(F^*,G^*)}{2}\right) = \lambda^*_m + O_P(m^{-\frac{1}{2}})$.

Proposition 2. *Under(A2)-(A3),(A10) for the unequal error variance case, from equation*(5), $\sigma_\theta\left(1-\frac{\widetilde{\widehat{W}}^2_m(F^*,G^*)}{2}\right) = \lambda^*_m + O_p(m^{-\frac{1}{2}})$. *Under (A2)-(A8) and (A10),* $\sigma_\theta\left(1-\frac{\widehat{W}^2_m(F^*,G^*)}{2}\right) = \lambda^*_m + O_p(m^{-\frac{1}{2}})$.

Most often we need to estimate σ_θ by the estimate $\hat{\sigma}_\theta$. The following results generalize Proposition 1-2 in the case when the optimal shrinkage predictor is based on plugged-in estimators for W_m, σ_θ and μ.

Theorem 4. *Suppose assumptions (A1)-(A10) hold. For the equal variance case of equation* (3),

$$\frac{\hat{\sigma}_\theta}{\hat{\sigma}_y}\left(1-\frac{\widetilde{\widehat{W}}^2_m(F^*,G^*)}{2}\right) = \lambda^*_m + O_p(m^{-\frac{1}{2}}),$$

$$\frac{\hat{\sigma}_\theta}{\hat{\sigma}_y}\left(1-\frac{\widehat{W}^2_m(F^*,G^*)}{2}\right) = \lambda^*_m + O_p(m^{-\frac{1}{2}}).$$

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

$$[1-\frac{\widetilde{\widehat{W}}^2_m(F^*,G^*)}{4}]^{-1} = RE^{(1)}_m(\lambda^*_m) + O_p(m^{-\frac{1}{2}}),$$

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

Theorem 5. *Suppose assumptions (A2)- (A10) hold. Then for the unequal design variance case* (5),

$$\hat{\sigma}_\theta\left(1-\frac{\widetilde{\widehat{W}}^2_m(F^*,G^*)}{2}\right) = \lambda^*_m + O_P(m^{-\frac{1}{2}}),$$

$$\hat{\sigma}_\theta\left(1-\frac{\widehat{W}^2_m(F^*,G^*)}{2}\right) = \lambda^*_m + O_P(m^{-\frac{1}{2}}).$$

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

$$[1 - \frac{\widehat{\widetilde{W}}^2_m(F^*, G^*)}{4}]^{-1} = RE^{(1)}_m(\lambda^*_m) + O_p(m^{-\frac{1}{2}}),$$

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

Theorem 4-5 provide approximation 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 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 (4). 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

In this section, we investigate the finite sample performance of the optimal shrinkage predictor via a limited simulation study. First we consider the case without covariates under different distributional assumptions on the errors and the area means. Then we consider a typical SAE scenario with covariates in the Fay-Herriot model with normally distributed errors and area means.

4.1 Optimum shrinkage and Wasserstein correction

We consider three examples. The sample sizes considered are $m = 2000$ and $m = 10000$. The larger sample size is chosen to evaluate the accuracy of the estimation of the Wasserstein distance and compare with the theoretical asymptotic value. Each reported Monte Carlo value is based on 500 replications.

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

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

The parameterization gives $E(\theta) = 0$ and $Var(\theta) = 1$ and $W(F^*, G^*)$ as an increasing function of $a \in [2, \infty)$ with $W = 0$ for $a = 2$. The error distribution is fixed as standard normal. For the simulation we look at $a \in \{2, 5, 10, 100, 1000\}$. The scale mixture models for different values of a are denoted by $Nmix(a)$.

Example 2: In addition we look at two possible distributions for the area means, a Double Exponential distribution to reflect possible heavy tail in the distribution and also a two component location mixture of normal to account for possible multimodality in the area mean distribution. The specific distribution considered are

$$\theta_i \sim 0.5N(4, 1) + 0.5N(-4, 1), \text{ and } \theta_i \sim DE(\sqrt{2}),$$

where in each case the errors are generated independently using $e_i \sim N(0, \sigma_i^2)$. For the normal mixture case, we consider two cases, one where the error variances are constant, $\sigma_i^2 = 16$, and another where σ_i^2's are generated from $Uniform(0, 16)$. For the double exponential, we consider a constant error variance scenario with $\sigma_i^2 = 1$ and an unequal error variance situation where σ_i^2's are generated from $Uniform(0, 1)$. 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 is used in each example. We consider the shrinkage predictors discussed earlier, namely $\boldsymbol{\theta}^*_{()}, \boldsymbol{\theta}^{(1)}_{()}$ and $\boldsymbol{\theta}^{(2)}_{()}$.

For prediction we use 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 $K = 6$ has been 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_{2000}	RE_{10000}	RE_∞	RE_{2000}	RE_{10000}	RE_∞
$Nmix(a)$								
$a = 2$	0	.02	1.00	1.00	1.00	40.4	199	∞
$a = 5$	.25	.24	1.01	1.01	1.01	2.19	2.12	2.15
$a = 10$	.41	.40	1.04	1.05	1.05	1.25	1.24	1.26
$a = 20$	.53	.47	1.07	1.08	1.08	1.09	1.08	1.07
$a = 50$	.62	.52	1.09	1.09	1.10	1.02	1.02	1.02
$a = 100$	.67	.61	1.10	1.12	1.13	1.00	1.01	1.01
DE_E	.14	.13	1.00	1.00	1.00	4.84	5.13	5.13
DE_U	.11	.10	1.00	1.00	1.00	3.67	3.94	3.94
$Nmix_E$	.37	.32	1.03	1.03	1.04	1.33	1.35	1.35
$Nmix_U$	.29	.28	1.02	1.02	1.03	1.18	1.18	1.18

Table 1 gives the values of relative efficiency (ratio of PMSE) of the optimal predictors compared with the other predictors at the two different 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 $m = 2000, 10000$ and $m = \infty$, respectively. The value at $m = \infty$ is the theoretical value given in the Theorem 1. Similarly, columns 7-9 give the relative efficiency value of the optimal estimator compared to $\boldsymbol{\theta}^{(2)}_{()}$ at $m = 2000, 10000$ 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, as obvious from the relative efficiency values given in Theorem 1, the BLUP-type estimator, $\boldsymbol{\theta}_{()}^{(2)}$ is much inferior to the other estimators in the small W scenario. For cases 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. In such situation the BLUP-type estimator $\boldsymbol{\theta}_{()}^{(2)}$ is nearly identical to the optimal predictor. For moderate values W, the optimal shrinkage provide substantial gains over both $\boldsymbol{\theta}_{()}^{(1)}$ and $\boldsymbol{\theta}_{()}^{(2)}$.

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

Example 3: In this example, we investigate the effect of skewness in the area mean distribution. We construct $\theta_i \sim Gamma(1.5, 1.5) - 1$, which gives an asymmetric distribution for the random effect part around mean zero and with support $(-1, \infty)$. Error terms follow normal distribution with error structure given as follows. We assume $\sigma_i^2 = b(\alpha + (1-\alpha)c_i)$, with $c_i = |1 - 2(\frac{i}{m})|$, $b = 3$ and $0 \leq \alpha \leq 1$. Changing α from 0 to 1 we can have equal and non equal variance scenarios where $\alpha = 1$ boils down to the equal variance case. We consider three cases $\alpha = 0, .5$ and 1.

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 is 100.

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_{∞}
$\alpha = 0$	.30	.27	1.02	1.02	1.02	2.40	2.70	2.64
$\alpha = 0.5$	.33	.31	1.02	1.02	1.03	2.82	3.31	3.03
$\alpha = 1$	.34	.33	1.02	1.03	1.03	2.97	3.72	3.38

For all the examples we looked at how the well the proposed mixture model approach is able to estimate G. Figure 1 shows the estimated density in two different cases. In all cases the proposed estimator provided reasonable approximation to G.

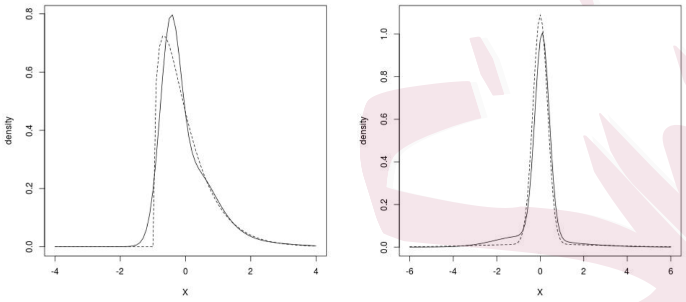

Figure 1: Plot for nonpaparanetric fit for density of G. Left hand panel shows a typical fit for shifted Gamma distrbition for skewed G. Right hand panel is a typical fit for example 1 with $a = 10$. Solid line shows the fitted and dashed line shows the true target density.

4.2 Small area estimation

For the SAE model, we only consider normal distribution because it is the most popular choice in the small area literature. Thus, θ_i and e_i are both assumed to be normally distributed. 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 $\boldsymbol{\theta}^*_{()}$ is identical to $\boldsymbol{\theta}^{(1)}_{()}$. We consider two cases.

Case 1: In this case $m = 100, 300$ and 500 small areas are considered. We consider the case with a single covariate. For the i th area we observe $y_i = \alpha + \beta x_i + u_i + e_i$ where x_i is the observed covariate for the i th area. We assume $u_i \sim N(0, 16)$ and assume $\alpha = 1, \beta = 2$. We generate $x_i \sim N(0, 1)$ and $e_i \sim N(0, \sigma_i^2)$. The error variance values are generated as $\sigma_i^2 \sim U(0, c)$. We choose c from $c = 1, 3, 5$. The simulation results

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

are reported for 500 Monte Carlo replications. The proposed estimator $\boldsymbol{\theta}^*_{()}$ is compared with the BLUP-type predictor $\boldsymbol{\theta}^{(2)}_{()}$. The relative efficiency values for the optimal predictor compared to the other predictor is reported in Table 3.

The proposed predictor outperforms the BLUP-type predictor in predicting the ordered area means. The difference is significantly higher when γ_i are further away from one, that is when c is 3 or 5. Also 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 will be 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 consider a more general case with different value of β. With higher β the fixed effect $\alpha+\beta x_i$ will dominate and the response y_i will be higher and this will effect the performance of 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 vary the number of small areas as $m = 100j^2$ with $j = 1, 2, 3, 4, 5$. The area specific variances are 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 is 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 essentially become like the fixed effect, and 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 bigger than one for cases when the area means are essentially of the same magnitude as that of the fixed part $\alpha + \beta x_i$, i.e, u_i is much smaller than $\alpha + \beta x_i$ in magnitude.

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. Because we are actually minimizing the Bayes risk, the posterior means is the best estimator in terms of PMSE. It will be interesting to compare the Bayesian method to our distribution free approach. However, we also study the sensitivity of the Bayesian estimator to model misspecification, especially in the model for the random effects parameter θ. We consider $Students - t$ distribution with various degrees of freedom for G and assume it to be misspecified as normal distribution. Also a mixture normal distribution is considered for θ with $G \sim .5N(1.5, 1) + .5N(-1.5, 1)$. For error variances, σ_i^2, two cases are considered, equal variances (E) $\sigma_i^2 = 1$ and unequal variances (U) $\sigma_i^2 \sim U(0, 1)$. In the Bayesian methodology we assume the prior on θ, G, to be $N(\mu, \sigma^2)$ and assume 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. The error due to model misspecification may be more than the MSE for the shrinkage and in that case the shrinkage estimator can perform better. Unimodal distributions on θ, such as $Students-t$ distribution, may be misspecified as normal. Normal mixtures, if the means are not far apart, convoluted with error can result in a unimodal distribution and model misspecification can easily occur in such situations.

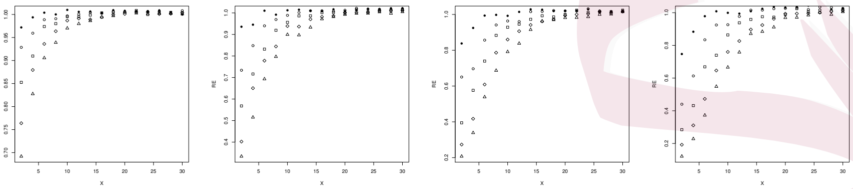

Figure 2: From left to right the figures correspond to $c = 1, 3, 5, 8$, respectively. X axis denotes $i = 1, \ldots, 30$. Here $m = 100j^2$ and $j = 1, 2, 3, 4, 5$ are denoted by $\bullet, \circ, \Box, \diamond$ and $\triangle$, respectively.

5 Conclusion

We propose an optimal estimator of ordered random effect 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 the simulation the distribution free estimator

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

G	m=100		m=400		m=900	
	U	E	U	E	U	E
$T_{2.2}$	.87	.81	.71	.68	.66	.63
T_3	.92	.87	.77	.73	.69	.68
T_4	1.00	.93	.82	.79	.75	.72
T_6	1.02	1.04	.92	.92	.82	.81
$N(0,1)$	1.19	1.22	1.23	1..22	1.23	1.21
Mixture	1.04	.98	.86	.85	.77	.78

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 derive 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 the Wasserstein distance. 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.

Other alternative classes of estimators could involve linear shrinkage estimators with different shrinkage for the different order statistics or alternative classes of shrinkage estimators that directly account for the joint dependence among the order statistics. Inference in such classes may be more difficult and 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. Nevertheless, the proposed estimator gives a good starting point and a preliminary framework for more general classes of distribution free estimators.

Supplementary materials

We have added a supplementary document. It includes four sections. The details about Remark 3 (generalized loss function) are given in section 1 of the supplementary materials. The justification and proof of result related to small area model from Remark 4 are given in section 2. In the third section, we prove some results from the text and in the fourth section, we discuss briefly about the rank estimation issue.

Acknowledgment

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

Appendix

Proof of Theorem 1

From (3)

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

Hence,

$$\lambda^*_m = \frac{m^{-1}\frac{\sigma_\theta}{\sigma_y}E\left(\sum_{i=1}^m \frac{(y_{(i)}-\mu)}{\sigma_y}\frac{(\theta_{(i)}-\mu)}{\sigma_\theta}\right)}{m^{-1}E\left(\sum_{i=1}^m (\frac{y_{(i)}-\mu}{\sigma_y})^2\right)} = \frac{\sigma_\theta}{\sigma_y}E(S^*_{(m)}),$$

where

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

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

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

where $T_1 = \frac{1}{m}\sum_{i=1}^m \left({F^*_m}^{-1}(\frac{i}{m+1}) - {G^*_m}^{-1}(\frac{i}{m+1})\right)^2\}$. Then,

$$T_1 = d_m(F^*_m, F^*) + d_m(G^*_m, G^*) + d_m(F^*, G^*) + C_1 + C_2 + C_3 \qquad (10)$$

where

$$\begin{aligned}
d_m(F^*_m, F^*) &= \frac{1}{m}\sum_{i=1}^m \left({F^*_m}^{-1}(\frac{i}{m+1}) - {F^*}^{-1}(\frac{i}{m+1})\right)^2, \\
d_m(G^*_m, G^*) &= \frac{1}{m}\sum_{i=1}^m \left({G^*_m}^{-1}(\frac{i}{m+1}) - {G^*}^{-1}(\frac{i}{m+1})\right)^2, \\
d_m(F^*, G^*) &= \frac{1}{m}\sum_{i=1}^m \left({F^*}^{-1}(\frac{i}{m+1}) - {G^*}^{-1}(\frac{i}{m+1})\right)^2,
\end{aligned}$$

$$
\begin{aligned}
C_1 &= \frac{2}{m}\sum_{i=1}^{m}(F_m^{*\,-1}(\frac{i}{m+1}) - F^{*-1}(\frac{i}{m+1}))(G_m^{*\,-1}(\frac{i}{m+1}) - G^{*-1}(\frac{i}{m+1})),\\
C_2 &= \frac{2}{m}\sum_{i=1}^{m}(F^{*-1}(\frac{i}{m+1}) - G^{*-1}(\frac{i}{m+1}))(G_m^{*\,-1}(\frac{i}{m+1}) - G^{*-1}(\frac{i}{m+1})),\\
C_3 &= \frac{2}{m}\sum_{i=1}^{m}(F_m^{*\,-1}(\frac{i}{m+1}) - F^{*-1}(\frac{i}{m+1}))(F^{*-1}(\frac{i}{m+1}) - G^{*-1}(\frac{i}{m+1})).
\end{aligned}
$$

From Lemma 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 1 later.

For C_1,

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

$$
\times E(\frac{2}{m}\sum_{i=1}^{m}(G_m^{*\,-1}(\frac{i}{m+1}) - G^{*-1}(\frac{i}{m+1}))^2). \qquad (11)
$$

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

From (10), Lemma 1 and the conclusion of equation (11) we have

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

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 proceed as follows:

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

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

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

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

We now give a lemma needed for the proof of Theorem 1.

Lemma 1. *Under the assumption of A1, 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.*

proof: See supplementary materials.

Proof of Theorem 3

Proof. Let,

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

$$= \frac{1}{2}\{\frac{1}{m}\sum_{i=1}^{m} {z_i}^2 + \frac{1}{m}\sum_{i=1}^{m} {w_i}^2 - \frac{1}{m}\sum_{i=1}^{m}\left({F_m^*}^{-1}(\frac{i}{m+1}) - {G_m^*}^{-1}(\frac{i}{m+1})\right)^2\}. \tag{12}$$

$E(\frac{1}{m}\sum_{i=1}^m z_i^2) = 1$ and $E(\frac{1}{m}\sum_{i=1}^m w_i^2) = 1$. The last part, $\frac{1}{m}\sum_{i=1}^m ({F^*}^{-1}_{m}$ $(\frac{i}{m+1}) - {G^*}^{-1}_{m}(\frac{i}{m+1}))^2$, converges to $W^2(F^*, G^*)$ by Lemma 1 similar to Theorem 1.

Hence, we have, $R_m^{(1)} = \sum_{i=1}^m(\theta_{(i)} - \mu - z_{(i)})^2 = \sigma_\theta^2 \sum_{i=1}^m (w_{(i)} - z_{(i)})^2 \to \sigma_\theta^2 W^2(F^*, G^*)$. Similarly, $R_m^{(2)} \to \sigma_\theta^2 W^2(K^*, G^*)$. Also, $\sigma_\theta - \lambda^* = \sigma_\theta \frac{W^2(F^*,G^*)}{2}$. Hence, $R_m(\lambda) \to \sigma_\theta^2 W^2(F^*, G^*) - \sigma_\theta^2 \frac{W^4(F^*,G^*)}{4}$. □

Refrences

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

Brown, L. D., 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. *Journal of the American Statistical Association,* **74**, 269–277.

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

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.

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

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

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science,* **28,** 40–68.

Shanbhag, D.N., Sreehari, M. (1977). On certain self-decomposable distributions. Z. Wahr. verw. Geb., **38**, 217–222.

Shen, W. and Louis, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *J. Royal Stat. Soc. 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. Prob.***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. *The Canadian Journal of Statistics,* **31,** 251–266.