Statistica Sinica **23** (2013), 333-357
doi:http://dx.doi.org/10.5705/ss.2011.071

EMPIRICAL BAYES IN THE PRESENCE OF EXPLANATORY VARIABLES

Noam Cohen[1], Eitan Greenshtein[1] and Ya'acov Ritov[2]

[1]*Israeli CBS and* [2]*The Hebrew University*

Abstract: We study the problem of incorporating covariates in a compound decision setup. It is desired to estimate the means of n response variables that are independent and normally distributed, each accompanied by a vector of covariates. We suggest a method that involves non-parametric empirical Bayes techniques and may be viewed as a generalization of the celebrated Fay-Herriot (1979) method. Some optimality properties of our method are proved. We also compare it numerically with Fay-Herriot and other methods, in a real data situation where the goal is to estimate certain proportions in many small areas. We also demonstrate our approach through the baseball data set originally analyzed by Brown (2008).

Key words and phrases: Compound decision, empirical Bayes.

1. Introduction

The main purpose of this paper is to study and demonstrate how to incorporate compound decision techniques (CD) or almost equivalently, empirical Bayes (EB) methods, in the presence of explanatory variables. The ideas of CD/EB were developed in the 1950's by Robbins (1951, 1955, 1964), see the review papers by Copas (1969) and Zhang (2003). Compound decision and Empirical Bayes procedures were shown to produce very efficient estimators in the simple setup where we have independent observations, $Y_1, \ldots, Y_n$, $Y_i \sim F_{\mu_i}$, and it is desired to estimate μ_i, $i = 1, \ldots, n$. A major case, on which we concentrate, is when $F_{\mu_i} = N(\mu_i, 1)$.

We focus on two types of EB procedures. One is Parametric Empirical Bayes (PEB), where μ_i, $i = 1, \ldots, n$, are assumed to be realizations of independent random variables M_i, $i = 1, \ldots, n$, $M_i \sim G$, $G = N(0, \tau^2)$, where τ^2 is unknown and should be estimated from the data. When n is large, the corresponding estimator, the exact variant of which, depends on the method of estimating τ^2, resembles the James-Stein estimator, cf., Efron and Morris (1973). The other type is Non-Parametric Empirical Bayes (NPEB), where the above distribution G is a member of a large non-parametric family $\mathcal{G}$ of distributions. Two recent NPEB methods and approaches are in Brown and Greenshtein (2009) and Jiang and Zhang (2009).

The advantage of EB procedures, relative to more elementary procedures, occurs as n grows, and may become very significant in high dimensional problems when n is large. A special advantage of NPEB procedures is expected in situations where the vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ is sparse, see e.g., Greenshtein, Park, and Ritov (2008), Brown and Greenshtein (2009).

Since modern statistical problems often involve high dimensional and sparse estimation problems, EB techniques should be embraced for such purposes, cf. Efron (2003). However, apart from literature in small area estimation, e.g., Rao (2003), which follows the seminal paper of Fay and Herriot (1979), EB is hardly used in modern data analysis. A recent approach, which is very much related to ours is Jiang and Zhang (2010). We became aware of it after completing most of this paper, we will elaborate on it in the sequel. One reason that EB is hardly used in practice is that in most applied problems, we have explanatory variables $X_{i1}, \ldots, X_{ip}$ for each observation Y_i and in such cases EB has no appeal, since standard EB procedures are permutation invariant, while $Y_1, \ldots, Y_n$ are not permutation invariant in the presence of the explanatory variables.

In our motivating example observations are binomial, $Y_i \sim B(m_i, p_i)$, and we need to estimate $p_1, \ldots, p_n$ — certain proportions in n (small) areas. The values of $p_1, \ldots, p_n$ are unknown constants to be estimated. In addition to the sample $Y_1, \ldots, Y_n$, we have a set of variables $\mathbf{X_1}, \ldots, \mathbf{X_n}$ (fixed or random, but independent of $Y_1, \ldots, Y_n$) and hope that $\mathbf{X_i}$ can serve as proxy to p_i, $i = 1, \ldots, n$. For example, consider one dimensional covariates $X_i \sim B(k_i, \tilde{p}_i)$ where the $\tilde{p}_i$ are "typically" close to p_i; alternatively $\boldsymbol{X}_i$ may be a vector of known parameters of area i that might be "relevant" to the parameter of interest p_i, for example, the socio-economic level of the region, its size, or mean age. We emphasize two elements. First, because of the proxies, $Y_1, \ldots, Y_n$ cannot be considered as "permutation invariants" or "exchangeable". Second, we do not believe that the observations follow standard regression models. The covariates are considered as proxies to the proportions, but they are statistically independent of the Y's (whose only stochastic aspect comes from the binomial sampling), and may be only a rough approximation to $p_1, \ldots, p_n$.

Simple symmetric and permutation invariant procedures. In cases of total ignorance regarding the parameters of the variables in relation to their identity, e.g., a situation where $Y_i \sim N(\mu_i, 1)$ and there is an exchangeable multivariate prior on $(\mu_1, \ldots, \mu_n)$, procedures that are permutation invariant have a special appeal. Permutation invariant procedures Δ are such that for every permutation π,

$$\Delta(Y_1, \ldots, Y_n) = (a_1, \ldots, a_n) \iff \Delta(Y_{\pi(1)}, \ldots, Y_{\pi(n)}) = (a_{\pi(1)}, \ldots, a_{\pi(n)});$$

here $a_i \in A$, where A is the action space. A simple class of exchangeable priors is where μ_i are realizations of i.i.d $M_i \sim G$, $i = 1, \ldots, n$. The optimal procedures

then belong to the class of 'simple symmetric decision functions', procedures Δ which are of the form

$$\Delta(Y_1, \ldots, Y_n) = (\delta(Y_1), \ldots, \delta(Y_n)),$$

for a given δ. For natural losses, given G, the optimal δ corresponds to the corresponding one dimensional Bayes procedure. On the relation and asymptotic equivalence between the two classes, see Greenshtein and Ritov (2009). Given a loss function, consider an 'oracle' that knows the values of $\mu_1, \ldots, \mu_n$, but is required to use a permutation invariant procedure. EB and CD procedures may be viewed as an attempt to imitate the procedure that an oracle would use. This is a very natural goal under 'total ignorance' or 'exchangeability'.

The appeal in using permutation invariant procedures and consequently EB procedures, is lost when exchangeability is lost, as in cases where there are explanatory variables. Assume $n = n_1 + n_2$ and it is known that the first n_1 observations, were taken from men, while the last n_2 were taken from women. Applying a permutation invariant procedure is equivalent to ignoring this potentially important information/explanatory-variable. However not all is lost, one may still apply EB procedure separately on the first n_1 observations and on the last n_2 observations. The idea is that after accounting for the explanatory variable in this trivial manner, we arrive at (two groups of) exchangeable variables, and applying EB procedures separately on each group becomes appealing. In a similar manner, we account for the information in the explanatory variables and then, after the information from the explanatory variables is accounted for and the "accounted observations" are closer to being exchangeable, we apply an EB procedure.

EB and CD are closely related notions and approaches. Under an EB formulation the parameters μ_i, $i = 1, \ldots, n$ are independent realizations from an unknown distribution G and the aim is to approximate the corresponding Bayes rule; under a CD formulation the aim is to approximate the best decision rule within a class of procedures (e.g., simple-symmetric, permutation invariant), for the given $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$. In this paper we emphasize the CD approach. However, we often use the more familiar EB notion, motivation, and terminology.

Applying a variant of the PEB method after accounting for the covariates, is in the spirit of the paper of Fay and Herriot, as shown in Sub-section 2.2; it is currently the most common practice. Another approach for inference in the presence of explanatory variables is that of Lindley and Smith (1972); this is a parametric empirical Bayes approach, though different than that of Fay and Herriot.

In Section 2, we suggest how EB could naturally be incorporated in problems with explanatory variables. We extend the Fay-Herriot approach and present its PEB and NPEB versions. We show the asymptotic optimality of NPEB.

In Section 3, we demonstrate the application of our suggested methods. We model sampling in the small statistical areas of the city of Tel Aviv-Yafo, Israel as considered in the recent Israeli census, and evaluate the performance of the different estimators. We also introduce results under some perturbations of the model. The application involves estimation of certain population's proportions in small areas. The explanatory variables available when estimating the proportion p_i in statistical area i, are 'Spatial' and 'Temporal', based on historical data, and data from neighboring statistical areas. We elaborate on comparing PEB procedures, versus the more recent NPEB procedure, suggested by Brown and Greenshtein (2009). In Section 4, we demonstrate the performance of our method on the baseball data set, studied by Brown (2008) and by Jiang and Zhang (2010).

Our ideas and techniques are meaningful in a general setup where $Y_i \sim F_{\mu_i}$, but will be presented for the case $F_{\mu_i} \equiv N(\mu_i, 1)$, $i = 1, \ldots, n$. In fact, as mentioned we apply our method for estimating the proportions of $B(m_i, p_i)$ distributions, but we do that after applying an arcsin transformation which will bring us to the normal setup.

2. Collections of Estimators Induced by Affine Transformations

Suppose the observations are vectors $\boldsymbol{V}_i = (Y_i, X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$, where $Y_1, \ldots, Y_n$ are independent, $Y_i \sim N(\mu_i, 1)$, $i = 1, \ldots, n$, and X_{ij} are explanatory variables statistically independent of Y_i, $i = 1, \ldots, n$, $j = 1, \ldots, p$, but related to the μ_i's. Denote by $X_{n\times p}$ the matrix of the explanatory variables and take $\boldsymbol{Y'} = (Y_1, \ldots, Y_n)$. The goal is to find a 'good' estimator $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)$ under the risk

$$E||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||_2^2.$$

The motivation and approach of the paper are as follows. Ideally it could be desired to approximate the Bayes procedure, assuming (at least formally) that $(\boldsymbol{V}_i, \mu_i)$, $i = 1, \ldots, n$, are independent random vectors sampled from an unknown distribution Γ that belongs to a large non-parametric family of distributions $\mathcal{G}$. Then the goal is to approximate the Bayes decision $\delta^* = \text{argmin}_\delta E_\Gamma||\delta(\boldsymbol{V_i}) - \mu_i||^2$ by $\hat{\delta}^*$, and let $\hat{\boldsymbol{\mu}} = (\hat{\delta}^*(\boldsymbol{V_1}), \ldots, \hat{\delta}^*(\boldsymbol{V_n}))$. However, this goal may be too ambitious for $(p+1)$ dimensional observations $\boldsymbol{V}_i$ when n is moderate, due to the "curse of dimensionality". A possible approach, in the spirit of Lindley and Smith (1972), is then to assume that Γ belongs to a convenient parametric family, in order to circumvent such difficulties. The approach of Fay and Herriot (1979) may also be interpreted this way. We, on the other hand, aim for the best permutational invariant estimator with respect to $Z_1, \ldots, Z_n$, where Z_i are one-dimensional random variables obtained by a suitable transformation of $(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n)$. This transformation is estimated from the data.

2.1. Preliminaries and definitions

We start from a general point of view, where initially there are no covariates. We observe independent $Y_i \sim N(\mu_i, 1)$, $i = 1, \ldots, n$. Let $\{T\}$ be a collection of affine transformations $T(\boldsymbol{Y}) = T_{A,\boldsymbol{b}}(\boldsymbol{Y}) = A\boldsymbol{Y} - \boldsymbol{b}$, where A is an orthonormal matrix and $\boldsymbol{b}$ is a vector. Then $\boldsymbol{Z} = T(\boldsymbol{Y})$ is distributed as a multivariate normal with mean vector $\boldsymbol{\nu} = A\boldsymbol{\mu} - \boldsymbol{b}$, and covariance matrix the identity. Let $\Delta = \Delta(\boldsymbol{Y})$ be a fixed estimator of the vector $\boldsymbol{\mu}$, that is not invariant under the group of affine transformations, i.e., $\Delta(T(\boldsymbol{Y})) \neq T(\Delta(\boldsymbol{Y}))$. Then, the pair Δ and $\{T\}$ defines a class of decision functions $\{\Delta_T\}$, $T \in \{T\}$,

$$\Delta_T(\boldsymbol{Y}) = T^{-1}(\Delta(T(\boldsymbol{Y})).$$

Let

$$T^{opt} = \operatorname*{argmin}_{T \in \{T\}} E_{\boldsymbol{\mu}} ||\Delta_T(\boldsymbol{Y}) - \boldsymbol{\mu}||_2^2 \equiv \operatorname*{argmin}_{T \in \{T\}} R(T, \boldsymbol{\mu});$$

here

$$R(T, \boldsymbol{\mu}) = E_{\boldsymbol{\mu}} ||\Delta_T(\boldsymbol{Y}) - \boldsymbol{\mu}||_2^2.$$

Our goal is to approximate T^{opt}, and then estimate $\boldsymbol{\mu}$ by an approximation of $\Delta_{T^{opt}}(\boldsymbol{Y})$.

For every $T \in \{T\}$, suppose we have a good estimator $\hat{R}(T, \boldsymbol{\mu})$ for $R(T, \boldsymbol{\mu})$. Let $\hat{T} = \operatorname{argmin}_{T \in \{T\}} \hat{R}(T, \boldsymbol{\mu})$. The usual approach, which we follow, is to use the estimator $\hat{\boldsymbol{\mu}} = \Delta_{\hat{T}}(\boldsymbol{Y})$. When the class $\{T\}$ is not too large, we expect only a minor affect of overfitting, i.e., $R(\hat{T}, \boldsymbol{\mu}) \approx R(T^{opt}, \boldsymbol{\mu})$.

Example 1 (Wavelet transform)**.** Our formulation describes many standard techniques, for example any harmonic analysis of the data that starts with transforming the data (e.g., Fourier transform). A special case is $T(\boldsymbol{Y}) = A\boldsymbol{Y}$, where A is the matrix that transforms $\boldsymbol{Y}$ to a certain wavelet representation; then, typically, the mean of the transformed vector is estimated and transformed back, see Donoho and Johnstone (1994). Suppose that, $\{T\} = \{A\}$ is a collection of matrices that correspond to a collection of wavelet bases/"dictionaries". The problem of finding the most appropriate basis/transformation, is related to that of basis-pursuit, see e.g., Chen, Donoho, and Saunders (2001). The permutational invariant and non-linear decision functions Δ in those studies is soft/hard-thresholds, Lasso, etc. Procedures of a special interest for us are parametric and non-parametric EB.

Example 2 (Regression)**.** Suppose that in addition to $\boldsymbol{Y}$ there is a fixed (deterministic!) matrix $X \in R^{n \times p}$. Consider the class of transformations $T(\boldsymbol{Y}) = \boldsymbol{Y} - \boldsymbol{b}$, $\boldsymbol{b} \in \{\boldsymbol{b}\}$, where $\{\boldsymbol{b}\}$ is the collection of all vectors of the form $\boldsymbol{b} = X\boldsymbol{\beta}$, $\boldsymbol{\beta} \in R^p$. Note, in particular, that these transformations are non-random.

Remark 1. The formulation for a random set $\{T\}$, that is independent of $\boldsymbol{Y}$ is the same. In the example when $X_{n\times p}$ is random, we condition on the explanatory variables and arrive at a conditional inference version of the development. From a Bayesian perspective, assuming a joint distribution Γ as above, conditional independence of the random set $\{T\}$ and $\boldsymbol{Y}$, conditional on the covariates, follows when we assume that $\boldsymbol{Y}$ and $X_{n\times p}$ are independent conditional on $\boldsymbol{\mu}$. We remark later on the case where the random set of transformations is 'weakly dependent' on $\boldsymbol{Y}$.

The following fact is useful. Let $\boldsymbol{Z} = T(\boldsymbol{Y})$. Then $Z_i \sim N(\nu_i, 1)$ where $\boldsymbol{\nu} = T(\boldsymbol{\mu})$, and

$$R(T, \boldsymbol{\mu}) = E_{\boldsymbol{\mu}}||\Delta_T(\boldsymbol{Y}) - \boldsymbol{\mu}||_2^2 = E_{\boldsymbol{\nu}}||\Delta(\boldsymbol{Z}) - \boldsymbol{\nu}||_2^2 = R(I, \boldsymbol{\nu}). \tag{2.1}$$

In the last equality I represents the identity transformation. When there is no real danger of confusion, the dependence on T is suppressed. We use (2.1) later to establish an estimator $\hat{R}(T, \boldsymbol{\mu})$ for $R(T, \boldsymbol{\mu})$.

A general three steps method for estimating $\boldsymbol{\mu}$ suggests itself.

Step I: For every T, estimate $R(T, \boldsymbol{\mu})$ by $\hat{R}(T, \boldsymbol{\mu})$.

Step II: Find $\hat{T} = \text{argmin}_T \hat{R}(T, \boldsymbol{\mu})$.

Step III: Get the estimator: $\hat{\boldsymbol{\mu}} = \hat{T}^{-1}(\Delta(\hat{T}(\boldsymbol{Y}))) \equiv \Delta_{\hat{T}}(\boldsymbol{Y})$.

Note that $\hat{T}$ depends on $\boldsymbol{Y}$.

We summarize. The idea in this subsection is that by an appropriate affine transformation, that may depend on the data, we arrive to a problem that is 'easier' for the procedure Δ to handle. For example, by choosing an appropriate wavelet basis we arrive at a sparse $\boldsymbol{\nu}$, which, typically, is easier to estimate than the original vector. Moreover, by accounting for explanatory variables in a good way through a suitable transformation, the transformed variables may become (nearly) exchangeable; whence, applying a permutation invariant procedure Δ on the transformed variables becomes natural and appealing.

2.2. The parametric empirical Bayes Δ and the Fay-Herriot procedure

The purpose of this subsection is to motivate the nonparametric approach, and to give a unified treatment and presentation to the more classical Fay-Herriot approach and the nonparametric approach. We study the case where Δ is a parametric empirical Bayes that corresponds to the prior $N(0, \tau^2)$, where τ^2 is unknown. When τ^2 is known, the corresponding Bayes estimator for μ_i is $\hat{\mu}_i = [\tau^2/(\tau^2+1)]Y_i$, and its risk is $\tau^2/(\tau^2+1)$. When τ^2 is unknown, we replace τ^2 by its estimate. For our level of asymptotics all consistent estimators $\hat{\tau}^2$ induce

equivalent estimators $\hat{\mu}_i = [\hat{\tau}^2/(\hat{\tau}^2+1)]Y_i$, and the corresponding estimators are asymptotically equivalent to the James-Stein estimator up to $o(n)$, see Efron and Morris (1973). By working at this level of asymptotics, our considerations in this subsection are valid for a wide class of PEB procedures, corresponding to various consistent methods of estimating τ^2, including the J-S procedure. In particular, the risk in estimating a (deterministic) vector $\boldsymbol{\mu}$ by PEB (or James-Stein's) method is

$$\frac{n||\boldsymbol{\mu}||_2^2}{||\boldsymbol{\mu}||_2^2+n} + o(n).$$

We now examine our three-step estimation scheme, adapted for parametric Empirical Bayes (or, for a James-Stein estimator Δ). Note that, for every T and the corresponding $\boldsymbol{\nu}$ and Z_i, we have $R(I,\nu) = [n||\boldsymbol{\nu}||_2^2/(||\boldsymbol{\nu}||_2^2+n)] + o(n)$. Hence a plausible estimator for $R(T,\boldsymbol{\mu})$ is

$$\hat{R}(T,\boldsymbol{\mu}) = \hat{R}(I,\boldsymbol{\nu}) = \max\Big\{0, \frac{n(\sum Z_i^2 - n)}{(\sum Z_i^2 - n)+n}\Big\} = \max\Big\{0, \frac{n(\sum Z_i^2 - n)}{\sum Z_i^2}\Big\} \quad (2.2)$$

Our three-step adaptation scheme is the following.

Step I: For every T estimate $R(T,\boldsymbol{\mu})$ by (2.2).

Step II: Find $\hat{T} = \operatorname{argmin}_T \hat{R}(T,\boldsymbol{\mu})$.

Step III: Get the estimator: $\hat{\boldsymbol{\mu}} = \hat{T}^{-1}(\Delta(\hat{T}(\boldsymbol{Y}))) \equiv \Delta_{\hat{T}}(\boldsymbol{Y})$.

Remark 2. When $\{T\}$ corresponds to $\{\boldsymbol{b} = \boldsymbol{X}\beta : \ \beta \in R^p\}$, Step II is trivial. Minimizing the residuals $\sum Z_i^2$ is achieved for $\tilde{\boldsymbol{b}}$ which is the projection of $\boldsymbol{Y}$ on the span of the columns of $\boldsymbol{X}$, and $\hat{T}(Y) = \boldsymbol{Y} - X\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is the ordinary least squares estimator. It is then easy to see that our suggested method is that of Fay and Herriot.

2.3. A nonparametric empirical Bayes Δ

The statements and development in this sub-section are for the nonparametric empirical Bayes procedure Δ, as in Brown and Greenshtein (2009), see the appendix. A recent study in which the NPEB procedure of Jiang and Zhang (2009) is extended to handle covariates is in Jiang and Zhang (2010).

Let $Z_i \sim N(\nu_i, 1)$ be independent. Denote by $\mathcal{R}(\boldsymbol{\nu})$ the Bayes risk that corresponds to the prior defined by the empirical distribution of $\boldsymbol{\nu}$. Let $f_{\boldsymbol{\nu}} = (1/n)\sum \phi(z-\nu_i)$, where ϕ is the density of a standard normal distribution. Then

$$\mathcal{R}(\boldsymbol{\nu}) = 1 - \int \frac{(f'_{\boldsymbol{\nu}}(z))^2}{f_{\boldsymbol{\nu}}(z)} dz = 1 - E_{\boldsymbol{\nu}} \frac{(f'_{\boldsymbol{\nu}}(Z))^2}{(f_{\boldsymbol{\nu}}(Z))^2}, \quad (2.3)$$

see Bickel and Collins (1983).

The following theorem is from Brown and Greenshtein (2009). It is stated for a triangular array set-up in order to cover situations of sparse $\boldsymbol{\nu} \equiv \boldsymbol{\nu}^n$. At stage n, $Y_i \sim N(\mu_i^n, 1)$ are independent and, for any corresponding sequence T^n, $T^n \in \{T^n\}$, $Z_i \sim N(\nu_i^n, 1)$ are independent, $i = 1, \ldots, n$.

Assumption 1. *For every $\alpha > 0$ and every sequence T^n and the corresponding $\boldsymbol{\nu}^n$ we have* $\max_i(|\nu_i^n|) = o(n^\alpha)$.

Assumption 2. *For some $\alpha_0 > 0$, $n^{(1-\alpha_0)}\mathcal{R}(\boldsymbol{\nu}^n) \to \infty$ for every T^n and corresponding $\boldsymbol{\nu}^n$.*

Theorem 1. *Under Assumptions* 1 *and* 2*, for every sequence T^n,*

$$R(I, \boldsymbol{\nu}^n) = E_{\boldsymbol{\nu}^n}||\Delta(\boldsymbol{Z}) - \boldsymbol{\nu}^n||_2^2 = (1 + o(1))n\mathcal{R}(\boldsymbol{\nu}^n) \tag{2.4}$$

As explained in the Appendix, the procedure Δ in Brown and Greenshtein requires a bandwidth $h = h_n$, that approaches slowly to zero. A rate that implies the result in Theorem 1 is $h_n = 1/\log(n)$.

Given $Y_i \sim N(\mu_i, 1)$, and a transformation T, $T \in \{T\}$, let Z_i be the $i'th$ coordinate of $\boldsymbol{Z} = T(\boldsymbol{Y})$. This theorem,(2.1) and (2.3), suggest an estimator for $R(T, \boldsymbol{\mu})$,

$$\hat{R}(T, \boldsymbol{\mu}) = n - \sum \Big[\frac{(\hat{f}'_{\boldsymbol{\nu}}(Z_i))}{\hat{f}_{\boldsymbol{\nu}}(Z_i)}\Big]^2, \tag{2.5}$$

where the density $f_{\boldsymbol{\nu}}$ and its derivative are estimated, for example, by appropriate kernel estimates.

Only step I of our general three-step procedure need to be adapted, as follows.

Step I: For every T and corresponding $\boldsymbol{\nu} = \boldsymbol{\nu}(T)$, estimate $R(T, \boldsymbol{\mu})$ by (2.5).

Remark 3. Step II could be computationally complicated when the set $\{T\}$ is large. When $\{T\}$ corresponds to $\{\boldsymbol{b} = \boldsymbol{X}\beta : \ \beta \in R^p\}$, a computationally convenient choice, which is to use the least-squares residuals for $\hat{T}(\boldsymbol{Y})$, as in the PEB case. However, this can be far from optimal, as noted in Examples 3 and 4, and in the simulations section.

Note that minimizing $R(I, \boldsymbol{\nu})$ with respect to $\boldsymbol{\nu} = \boldsymbol{\nu}(T)$ is equivalent to finding the "most favorable" prior, rather than the more conventional task of finding the least favorable prior.

Remark 4. Our method that combines the NPEB method of Brown and Greenshtein (2009) with a transformation induced by covariates is termed 'NPEB with covariates' but, for simplicity, we refer to it in the sequel as NPEB.

Choosing the least squares residuals can be very inefficient, since it might cause "over smoothing" of the empirical distribution and low values can happen in $(f'_{\tilde{\boldsymbol{\nu}}})^2$ which, by (2.3), implies high risk. This can happen in transforming a sparse structure into a non-sparse one, as in Example 3, or by transforming a structure with well separated groups into a mixed structure, as in the Example 4.

Example 3. $Y_i \sim N(1,1)$, $i = 1,\dots,2m$, $2m = n$. Suppose we have only one (useless) explanatory variable $X_i = 1$ if $i \le m$ and 0 otherwise. Projecting $\boldsymbol{Y}$ on X, we get that the least squares shift is $\tilde{\boldsymbol{b}} \approx (1,\dots,1,0,\dots,0)'$ and $\boldsymbol{\nu} = \boldsymbol{\mu} - \tilde{\boldsymbol{b}} \approx (0,\dots,0,1,\dots,1)'$, which is much worse for empirical Bayes estimation than the original $\boldsymbol{\mu}$; it is easy to see that $n\mathcal{R}(\tilde{\boldsymbol{\nu}}) = O(n)$, while $n\mathcal{R}(\boldsymbol{\mu}) = 0$. From Theorem 1 we conclude that, as $n \to \infty$, the advantage of the latter (trivial) transformation compared to the least squares residuals in terms of the risk is $o(n)$ compared to $O(n)$.

Example 4. Let $Y_i \sim N(\mu_i, 1)$ be independent, where $\mu_i = \mu_1$ for $i = 1,\dots,m$, and $\mu_i = -\mu_1$ for $i = m+1,\dots,2m = n$. Suppose $X_i = (\mu_i + W_i) \sim N(\mu_i, 1)$, independent of Y_i, $i = 1,\dots,n$. Let $\tilde{\boldsymbol{\nu}} = \boldsymbol{\mu} - \tilde{\boldsymbol{b}}$ where $\tilde{\boldsymbol{b}}$ is the projection of Y on the (random) vector $\boldsymbol{X} = (X_1,\dots,X_n)'$. It is easy to check that $\tilde{\nu}_i \to \mu_i/(\mu_1^2+1) - \mu_1^2 W_i/(\mu_1^2+1)$ as $n \to \infty$. When $\mu_1 \to \infty$, the empirical distribution of $\tilde{\boldsymbol{\nu}} \equiv \boldsymbol{\nu^n}$ converges to that of a standard normal. The corresponding Bayes risk $\mathcal{R}(\tilde{\boldsymbol{\nu}}^n)$ converges to 0.5. Obviously the Bayes risk that corresponds to the trivial transformation, for which $\boldsymbol{\nu}^n = \boldsymbol{\mu}^n$, converges to zero.

2.4. Finding a good transformation

The above method is reasonable when the class $\{T^n\}$ of candidate transformations is not too large, in terms of its cardinality or its VC dimension, and the overfit effect is not significant. When $R(T^n, \boldsymbol{\mu})$ is of the order of n, an appealing condition is of the type

$$P\Big(\sup_{T^n \in \{T^n\}} |\hat{R}(T^n, \boldsymbol{\mu}) - R(T^n, \boldsymbol{\mu})| > \epsilon n\Big) \to 0. \tag{2.6}$$

We now demonstrate in the following Theorem 1 why such conditions are plausible to expect even for reasonably large sets $\{T^n\}$.

Assumption 3. *Let $m = n^\gamma$ for some $\gamma < 1$. Suppose that the cardinality of the set $\{T^n\}$ of candidate transformations at stage n, is of size* $\exp(m)$, $n = 1, 2, \dots$

Proposition 1. Under Assumptions 1 and 3 there exists a sequence of estimators $\hat{R}(T^n, \mu)$ of the form (2.5) such that

$$P\Big(\sup_{T^n \in \{T^n\}} |\hat{R}(T^n, \boldsymbol{\mu}) - R(T^n, \boldsymbol{\mu})| > \epsilon n\Big) \to 0.$$

If we specify the class of transformation, we can obtain more.

Theorem 2. *Suppose the class of transformation is* $T(Y) = Y - X\boldsymbol{\beta}$, $\boldsymbol{\beta} \in B = B^n$, *where* $X \in R^{n\times p}$, $\boldsymbol{\beta} \in R^p$, $p = p_n$, *the* l_2 *norms of the rows of* X *and of* β *are bounded by* $M = M_n$; $\sup_{\boldsymbol{\beta}\in B^n} ||X\boldsymbol{\beta}||_\infty = o(n^\alpha), \forall \alpha > 0$. *If* $p\log(M) < n^\gamma$, *for some* $\gamma < 1$, *then there is a version of* $\hat{R}$ *such that* (2.6) *is satisfied.*

The proofs are given in the Appendix.

2.5. Optimality of NPEB Δ.

Until this point the treatment has been for a concrete procedure Δ and a class $\{T\}$ of transformations. The purpose of this section is to advocate the choice of a non-parametric empirical Bayes Δ, denoted Δ_{NP}.

As noted, Step II in the non-parametric approach can be computationally intensive, so such a dominance result might not be enough to persuade one that the non-parametric approach is a good alternative to the parametric approach and to the Fay Herriot procedure. In Theorem 3 we show that for *every* two sequences $\boldsymbol{\mu}^n$ and T^n, the sequence of estimators, obtained by coupling T^n with Δ_{NP}, asymptotically dominates the sequence obtained when coupling the same T^n with any other sequence of permutation invariant procedures Δ^n.

Given a procedure Δ, a transformation T, and a mean vector $\boldsymbol{\mu}$, the corresponding risk is denoted as $R_\Delta(T, \boldsymbol{\mu}) \equiv R(T, \boldsymbol{\mu})$ as before; for the case of the nonparametric EB procedure Δ_{NP}, the corresponding risk is denoted $R_{NP}(T, \boldsymbol{\mu})$. Our asymptotic analysis is again in a triangular array setup.

Theorem 3. *Let* $\boldsymbol{\mu}^n$, Δ^n *and* T^n *be arbitrary sequences. Assume that for each* n *the procedure* Δ^n *is simple symmetric. Under Assumptions* 1 *and* 2,

$$\limsup \frac{R_{NP}(T^n, \boldsymbol{\mu}^n)}{R_{\Delta^n}(T^n, \boldsymbol{\mu}^n)} \leq 1.$$

Proof. This Follows from Brown and Greenshtein (2009) and Theorem 1. Note that the risk of the optimal simple symmetric procedure equals $n\mathcal{R}(\boldsymbol{\nu}^n)$.

Conjecture: In Theorem 3 the condition that Δ^n be simple symmetric for every n, might be replaced by the weaker condition, that Δ^n is permutation invariant for every n. This should follow by an equivalence result in the spirit of Greenshtein and Ritov (2009), though stronger. Note, the equivalence result in Greenshtein and Ritov (2009) would suffice under the assumption that $\max_i(|\nu_i^n|) = O(1)$; however, Assumption 1 allows a higher order.

2.6. Remark

Consider the case in which $\{T\}$ corresponds to $\{\boldsymbol{b} = X\beta\}$. Write $\boldsymbol{b} = (B_1, \ldots, B_n)'$. In the application we have in mind the set $\{T\}$ may be random since X_{ij} is random. When the random set of transformations is independent of $\boldsymbol{Y}$, our treatment applies by conditioning on the explanatory variables. We are interested in situations where the random set $\{T\}$ may depend on $\boldsymbol{Y}$, however we require that Y_i be independent of $X_{i1}, \ldots, X_{ip}$ for each i. Then the distribution of Z_i conditional on $X_{i1}, \ldots, X_{ip}$, is $N(\nu_i, 1)$, where $(\nu_1, \ldots, \nu_n)' = \boldsymbol{\nu} = A\boldsymbol{\mu} - \boldsymbol{b}$ as before. When the dependence of Y_i on $X_{j1}, \ldots, X_{jp}$, $j \neq i$ is not too heavy, a natural goal is still to try to approximate the best decision function for estimating ν_i among the decision functions which are simple symmetric with respect to $Z_1, \ldots, Z_n$. The conditional marginal distribution of Z_i, $i = 1, \ldots, n$, is still $N(\nu_i, 1)$; however, we may not treat these as independent observations. Thus, the rates of estimating $f_{\boldsymbol{\nu}}$ and its derivative may become slower and, for heavy dependence, Theorems 1 and 2 might not hold. Similarly rates of estimation of τ_n^2, in order to apply the PEB procedure, could be slow. However, when the dependence is not "too heavy" we may expect Theorems 1 and 3 to hold under the assumption that Y_i is independent of $X_{i_1}, \ldots, X_{ip}$ for each i.

2.7. Discussion and summary

In this subsection we summarize and compare the approach of Fay and Herriot and Jiang and Zhang, in addition to our's suggested approach. We take the liberty to follow those approaches just in "spirit". From a compound decision (non-Bayesian) perspective, all the approaches assume that $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is multivariate normal with mean zero. Given a matrix X of explanatory variables, let $\boldsymbol{\xi}'$ be the projection of $\boldsymbol{\mu}$ on the linear space spanned by the columns of X. Then we may write $\boldsymbol{\mu} = \boldsymbol{\xi}' + \boldsymbol{\xi}$, where $\boldsymbol{\xi}'$ is orthogonal to $\boldsymbol{\xi}$. If X is non-singular, there is a unique $\boldsymbol{\beta}$ such that

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\xi} + \boldsymbol{\epsilon}.$$

Under a Bayesian formulation the vector $\boldsymbol{\xi}$ is composed of i.i.d. sampled variables $\xi_i \sim G$, and G is often assumed normal. Under the Fay-Herriot approach, we estimate $\boldsymbol{\beta}$ using least squares estimator $\hat{\boldsymbol{\beta}}$, then we estimate the mean $E(\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}) \approx \boldsymbol{\xi}$ of the 'nearly' multivariate normal vector $(\boldsymbol{Y} - X\hat{\boldsymbol{\beta}})$ by a variant of the James-Stein estimator, and finally we transform back to obtain an estimator for $\boldsymbol{\mu}$. The approach of Jiang and Zhang (2010) is similar, only they prefer to estimate the mean of $\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}$ by a non-parametric mle as in Jiang and Zhang (2009). The latter method is appropriate for a general distribution G under a Bayesian approach, or for a general vector $\boldsymbol{\xi}$ under a compound decision

approach. Our approach differs from that of Jiang and Zhang in taking an estimate of the mean of the transformed vector by the NPEB estimator that was suggested by Brown and Greenshtein (2009), though the two procedures have similar performance. An important additional difference is in the method of choosing the appropriate transformation. As in Examples 3 and 4, if we use NPEB or a non-parametric mle procedure Δ, it is not at all clear that it is useful to transform the problem and estimate the mean of $\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}$; estimating the mean of $\boldsymbol{Y} - X\tilde{\boldsymbol{\beta}}$, for some $\tilde{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}$ could be far better.

Finding the appropriate alternative $\tilde{\boldsymbol{\beta}}$ could be complicated, especially computationally. However, there are examples and applications where we could find a better transformation by intelligent guessing. A general scenario is the following. Suppose we have a plausible linear model with a certain $\boldsymbol{\beta}_0$, that works fine for most cases, but does not work for a few outliers. Applying a transformation with $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ would bring us to a situation where the mean of $\boldsymbol{Y} - X\tilde{\boldsymbol{\beta}}$ is a sparse vector with many "nearly" zero components and a few components that are very different than zero; such a sparse mean vector is often "easier to estimate" compared to estimation of the mean vector of $\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}$, which could be far from sparse due to "over smoothing". In the example presented in Section 3, estimating the vector of current proportions of registered people in various areas, a useful explanatory variable is the corresponding vector of estimated proportions from the previous year. A linear relation with $\beta_0 = 1$ is appropriate for most areas, excluding areas that went through a rapid development in the last year. Indeed in our study of the census example in Section 3, we tried in addition to $\hat{\boldsymbol{\beta}}$ a few more candidates that were chosen through an "intelligent guess" and not through a numerical search. In the second part of Section 4.2 we demonstrate, in the baseball example, how to select the more appropriate among a few candidate transformations, using the estimator (2.5). Also, we present results of a 'brute force' computation and search for $argmin_{T\in\{T\}}\hat{R}(T, \boldsymbol{\mu})$ when Δ is NPEB.

3. Census Example

3.1. Preliminaries

The city of Tel Aviv-Yafo, Israel, is divided into 161 small areas called "statistical areas", each area belongs to a sub-quarter that includes about four additional statistical areas. The recent Israeli census was based on administrative records corrected by samples. Thus the proportion p_i of people who are registered in area i among those who live in area i, , $i = 1, \ldots 161$, was of interest. The estimated p_i, $i = 1, \ldots, n$ are used to adjust the administrative-registration counts and get population estimates for each area. In our example we use the parametric bootstrap concept to evaluate the performance of various estimators. In the parametric bootstrap, we use for the parameters $p_1, \ldots, p_n$ their values as

estimated in the recent census (where about 20% of the population was sampled). The mean of $p_i,\ i = 1, \ldots, 161$, is 0.75 the standard deviation is 0.13, and the histogram is roughly bell shaped.

We present a bootstrap study in which $p_i,\ i = 1, \ldots, 161$ are estimated based on samples of size m_i and the corresponding simulated independent $\tilde{Y}_i$, $\tilde{Y}_i \sim B(m_i, p_i)$. Here $\tilde{Y}_i$ is the number of people in the sample from area i, registered to area i.

In addition we simulated covariates in our parametric bootstrap. We simulated temporal variables that correspond to historical data from each area i, and spatial covariates, that correspond to samples from the neighboring areas of each area i. In the following we explore scenarios for the cases of only temporal covariates, only spatial covariates, and both temporal and spatial covariates. We compare the performance of PEB, NPEB and other methods. In all the analyzed situations, we simulated binomial observations with sample size $m_i \equiv m$, for $m = 25, 50, 100$.

In order to reduce this setup to the normal case, we applied an arcsin transformation on our binomial observations $\tilde{Y}_i,\ i = 1, \ldots, n$, as in Brown (2008). Specifically,

$$Y_i = \sqrt{4m}\arcsin\Big(\sqrt{\frac{\tilde{Y}_i + 0.25}{m + 0.5}}\ \Big). \tag{3.1}$$

Then, Y_i are distributed approximately as $N(\sqrt{4m}\arcsin(\sqrt{p_i}), 1)$. We estimated $\mu_i = E(Y_i)$, by $\hat{\mu}_i,\ i = 1, \ldots, n$, as explained in Sub-sections 2.3 and 2.3, and then let the estimate of $p_i,\ i = 1, \ldots, 161$ be,

$$\hat{p}_i = (\sin(\frac{\hat{\mu}_i}{\sqrt{4m}}))^2. \tag{3.2}$$

Similarly, we also considered the following regression estimator. We estimated $\boldsymbol{\mu}$ by $\hat{\boldsymbol{\mu}}^{Reg} = X\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the least squares, and obtained the estimator:

$$\hat{p}_i^{Reg} = \Big(\sin(\frac{\hat{\mu}_i^{Reg}}{\sqrt{4m}})\Big)^2. \tag{3.3}$$

Let $\boldsymbol{p} = (p_1, \ldots, p_n)$ and $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_n))$. We evaluated the performance of an estimator according to the risk $E_{\boldsymbol{p}}||\hat{\boldsymbol{p}} - \boldsymbol{p}||_2^2$. The risk was approximated through 1,000 simulations for each entry in the tables in the sequel. A different parametric EB approach for estimating proportions in small areas, that involves a logistic regression model, may be found in Farrell, MacGibbon, and Tomberlin (1997).

3.2. Temporal covariates

We introduce now simulated scenarios with only temporal covariates. We think of a process where each year a sample of size m is taken from each area. Suppose we use the records of the previous three years as covariates. Let $\tilde{T}_i$ be the number of people among the $3m$ that were sampled in the previous three years from area i, which were registered to the area. Although $\tilde{T}_i$ might be better modeled as a binomial mixture, we model $\tilde{T}_i$ as $B(3m, p_{it})$ for simplicity. In order to (hopefully) have a linear relation between the response and explanatory variables, we take the temporal covariates

$$T_i = \sqrt{4m}\arcsin\Big(\sqrt{\frac{\tilde{T}_i + 0.25}{3m + 0.5}}\Big). \tag{3.4}$$

Note, if there is little change from the previous three years to the current year in area i, then $p_i \approx p_{it}$ and $E(T_i) \approx E(Y_i)$.

We simulate two scenarios. One scenario is of no-change, where $p_{it} = p_i$ for $i = 1, \ldots, 161$. The other scenario is of a few abrupt changes; specifically, $p_i = p_{it},\ i = 17, \ldots, 161$, but $p_{it} = 0.3 < p_i$ for $i = 1, \ldots, 16$. Such abrupt changes could occur in areas that went in previous years through a lot of building, internal immigration and other changes.

Since the empirical distribution of $E(Y_i)$ is roughly bell-shaped, it is expected that the PEB method will work well in the no-change scenario; while under a few abrupt changes, an advantage of the NPEB procedure will be observed.

As mentioned in Section 2, the optimization step of the NPEB procedure is difficult. We try two candidate transformations $Y - \boldsymbol{b}^i$, $i = 1, 2$, coupled with the NPEB; the corresponding methods are denoted NPEB1 and NPEB2. NPEB1 corresponds to the least-squares/Fay-Herriot transformation, while NPEB2 corresponds to the transformation $Z_i = Y_i - T_i$. The latter transformation, although still sub-optimal when coupled with a NPEB Δ, could occasionally perform better than the former, as indicated by Examples 3 and 4. In addition to comparing the risks of the PEB, NPEB1, and NPEB2 methods, we also compare the risk of the naive estimator, and of the regression estimator. The regression estimator estimates $\hat{\mu}_i$ through $\hat{\boldsymbol{\mu}} = X\hat{\beta}$, does not apply an additional PEB or NPEB stage. The Naive estimator simply estimates p_i by the corresponding sample proportion.

The no-change scenario is presented in Table 1. Each entry is based on 1,000 simulated realizations. Under no-change the temporal covariate is very helpful, and even the least squares linear predictor is doing very well. Over all, the naive estimator is the worst, NPEB1, NPEB2, and Regression are about the same, while the PEB is moderately better than the other methods.

Table 1.

	Naive	Reg	NPEB1	NPEB2	PEB
$m = 25$	1.12	0.33	0.35	0.37	0.27
$m = 50$	0.56	0.17	0.18	0.18	0.14
$m = 100$	0.28	0.092	0.093	0.093	0.073

Table 2.

	Naive	Reg	NPEB1	NPEB2	PEB
$m = 25$	1.12	1.66	0.75	0.49	0.68
$m = 50$	0.56	1.64	0.46	0.22	0.42
$m = 100$	0.28	1.62	0.26	0.11	0.24

In the scenario of a few abrupt changes, the regression by itself is performing the worst, however an additional EB step is helpful. Here the NPEB2 procedure is the best, see Table 2.

3.3. Spatial covariates

We simulate a scenario with spatial covariates as follows. Tel-Aviv is divided into sub-quarters, each sub-quarter is defined by about 5 statistical areas. For every $i = 1, \ldots, 161$, we take the neighborhood of area i, to be the statistical areas *other than* area i, in the same sub-quarter.

Based on the census we have good estimates for p_{is}, the proportion of people living in the neighborhood of area i, who are registered to their areas. Those estimates are treated as the "real" values in our simulations. The correlation between p_i and p_{is}, $i = 1, \ldots, 161$ is 0.62.

For simplicity we assume that, for each i, the size of the sample from the neighborhood of area i is $4m$. Let $\tilde{S}_i$ be the number of people sampled from the neighborhood of i who are registered to their area. Although $\tilde{S}_i$ might be better modeled as a binomial mixture, we model $\tilde{S}_i$ as $\tilde{S}_i \sim B(4m, p_{is})$ for simplicity. As in the case of Temporal covariates we take the spatial covariate for area i as

$$S_i = \sqrt{4m}\,\arcsin\Big(\sqrt{\frac{\tilde{T}_i + 0.25}{4m + 0.5}}\,\Big). \tag{3.5}$$

We consider two NPEB estimates, corresponding to the projection/Fay-Herriot and to the $Z_i = Y_i - S_i$ transformations. The results of our simulations are summarized in Table 3. The advantage of the EB procedures is more noticeable for $m = 25$. The explanation is the following. Since the temporal covariate is not very strong, the mean vector of the transformed variables is not too sparse.

Table 3.

	Naive	Reg	NPEB1	NPEB2	PEB
$m = 25$	1.12	1.41	0.72	0.75	0.64
$m = 50$	0.56	1.34	0.44	0.44	0.40
$m = 100$	0.28	1.31	0.26	0.28	0.23

Table 4.

	Naive	Reg	NPEB1	NPEB2	NPEB3	NPEB4	PEB
$m = 25$	1.12	1.13	0.65	0.49	0.54	0.55	0.58
$m = 50$	0.56	1.06	0.4	0.22	0.28	0.38	0.37
$m = 100$	0.28	1.03	0.24	0.11	0.15	0.22	0.22

When m is large, under the scale induced by the variance of Z_i, the points $\nu_i,\ i = 1, \ldots, n$, may be viewed as isolated and the smoothing of the EB is hardly effective. Hence the EB methods behave roughly like the Naive estimator.

One could wonder whether the spatial covariates are helpful to the non parametric empirical Bayes, whether it is better not to transform the data at all and to apply Δ_{NP} on the original data taking $T = I$ and $\boldsymbol{\nu} = \boldsymbol{\mu}$. However this option is slightly worse than the above ones. The simulated risks that correspond to $m = 25, 50, 100$ are 0.84 , 0.5, and 0.28.

3.4. Spatial and temporal covariates

In this sub-section we study the performances of our estimators when both temporal and spatial variables are introduced. As before we apply the projection transformation for the NPEB estimator. However, we also try the transformations $Z_i = Y_i - (\alpha S_i + (1-\alpha)T_i)$, for $\alpha = 0, 0.3, 0.6$. The corresponding estimators are NPEB1 (for the projection), NPEB2, NPEB3, and NPEB4, correspondingly. For the temporal covariates we simulate the scenario of 16 abrupt changes, the spatial covariates as before. As may be expected, since the spatial covariate is weak relative to the temporal, accounting for it causes extra unnecessary smoothing. For the non-parametric EB procedure, indeed NPEB2 that corresponds to $\alpha = 0$ has the best performance, and is also the optimal among all seven methods, see Table 4.

4. Baseball Example

Following Efron and Morris (1975) and Brown (2008), one finds "The ultimate test of any empirical Bayes procedure is known to be: How well it predicts second-half-of-the-season baseball batting average", see Koenker and Mizera

(2012). In this section we analyze the Baseball data set, originally analyzed by Brown (2008) and later by Jiang and Zhang (2010). Our analysis resembles that of Jiang and Zhang. The data consists of batting records of each major league player in 2005. For each player i, denote by N_{1i} and H_{1i} the number of at bats and the number of hits he had in the first half of the season; similarly N_{2i} and H_{2i} are the corresponding quantities for the second half. In addition, for every player it is known whether he is a pitcher or a batter. For $j = 1, 2$ denote

$$R_{ji} = \frac{H_{ji}}{N_{ji}}.$$

This notation stays close to that of Brown and of Jiang and Zhang.

Our purpose is to predict the value of R_{2i} for player i, based on the data from the first half.

A reasonable model for the data is that, conditional on $N_{ji},\ i = 1, \ldots, n,\ j = 1, 2$, $H_{ji} \sim Bin(N_{ji}, p_i)$, where p_i is the (fixed in time) probability of a successful hit by player i. Thus, a reasonable approach is to estimate p_i by $\hat{p}_i$ and let our predictor for the value of R_{2i} be $\hat{R}_{2i} = \hat{p}_i,\ i = 1, \ldots, n$. So, as in previous example, we should estimate the proportions p_i.

Let $S_j = \{i | N_{ji} \geq 11\}$. The estimation of p_i is done only for players i, such that $i \in S_1$. Validation of the prediction $\hat{p}_i$ is made only for players i such that $i \in S_2$. The size of S_1 is 567, while the size of $S_1 \cap S_2$ is 499.

The criterion for the performance of a predictor $\hat{R}_{2i}, i = 1, \ldots, 499$, is based on the estimator of $E \sum_{i \in S_1 \cap S_2} (\hat{R}_{2i} - R_{2i})^2$,

$$T\hat{S}E_R = \sum_{i \in S_1 \cap S_2} \Big((\hat{R}_{2i} - R_{2i})^2 - \frac{R_{2i}(1 - R_{2i})}{N_{2i}} \Big).$$

As a benchmark for the performance of $\hat{\boldsymbol{R}}$ we take the performance of the naive estimator $\hat{R}_{2i} = R_{1i}$, specifically the value $T\hat{S}E_0 = \sum_{i \in S_1 \cap S_2} ((R_{1i} - R_{2i})^2 - R_{2i}(1 - R_{2i})/N_{2i})$. We report the results of an estimator $\hat{R}_{2i}, i = 1, \ldots, 499$, through

$$T\hat{S}E_{R^*} = \frac{T\hat{S}E_R}{T\hat{S}E_0}.$$

4.1. Covariates and transformations

Consider a few "linear models" with the following covariates. One covariate for player i is the number of trials N_{1i}. The additional covariate is an indicator of the event that the player is not a pitcher. The value of N_{1i} is potentially a

Table 5.

	WGMLEB	NPEB	Reg
Model i	0.291	0.353	0.526
Model ii	0.204	0.234	0.343
Model iii	0.175	0.186	0.214
Model iv	0.167	0.176	0.200

useful covariate, since that a high value of N_{1i} indicates that the coach perceives player i as a good batter. The response variable for player i is

$$Y_i = \sqrt{4 * N_{1i}} arcsin(\sqrt{\frac{H_{1i} + 0.25}{N_{1i} + 0.5}}).$$

Note, in the above we transformed the variables and obtained a homoscedastic model. Both Brown (2008) and Jiang and Zhang (2010) worked in a heteroscedastic setup with the variables Y_i^*, where $Y_i^* = \arcsin(\sqrt{\frac{H_{1i}+0.25}{N_{1i}+0.5}})$.

Jiang and Zhang (2010) considered the following linear models to all include intercept. Using their notation we denote the covariate N_{1i}, the number of At Bat of player i, by AB. The models they considered with respect to y_i^* were i) AB, ii) Pitcher, iii) Pitcher+AB, iv) Pitcher +AB + Pitcher*AB. In the above we used the standard notation, where the last model includes interaction of the variables AB and Pitcher. The corresponding models in terms of our response variable Y_i are: i) $AB^{0.5} + AB^{1.5}$, ii) $AB^{0.5} + AB^{0.5} * Pitcher$, iii) $AB^{0.5} + AB^{0.5} * Pitcher + AB^{1.5}$, iv)$AB^{0.5} + AB^{0.5} * Pitcher + AB^{1.5} + AB^{1.5} * Pitcher$. Note, the intercept variable is transformed to $AB^{0.5}$ when modeling with respect to Y_i.

4.2. Numerical study

We report the results of three methods applied to the baseball data. The estimators are WGMLEB, Weighted General Maximum Likelihood Empirical Bayes as studied by Jiang and Zhang; regression, where the mean $\boldsymbol{\mu}$ of $\boldsymbol{Y}$ is estimated based on the least squares $\hat{\boldsymbol{\beta}}$ and then transformed to obtain an estimator of $\boldsymbol{p}$; our NPEB method, with least squares $\hat{\boldsymbol{\beta}}$. The methods were applied in the four models, the corresponding $T\hat{S}E_R^*$ are reported in Table 5.

The results of the NPEB are slightly inferior to those achieved by the method WGMLEB of Jiang and Zhang, yet the computation of our estimator seemed significantly easier.

In Brown (2008), the covariate AB was used implicitly, through the estimation of the density f and its derivative (see Appendix) in his variant of NPEB. The density at a point y_i was estimated based on observations y_k, with N_{1k}

"close" to N_{1i}. This implicit use yielded $T\hat{S}E_R^* = 0.509$ for the corresponding NPEB. The more direct approach taken by Jiang and Zhang and by us seems beneficial.

As explained, in each case we used the transformation induced by the least squares $\hat{\boldsymbol{\beta}}$. In addition, for Model i we numerically searched for a transformation that yielded a better corresponding $T\hat{S}E_R^*$ via our NPEB method. While the least squares $\hat{\boldsymbol{\beta}} = (0.9564, 0.0006)$ gave $T\hat{S}E_R^* = 0.353$, $\tilde{\boldsymbol{\beta}} = (1.0864, 0.0003)$ yielded $T\hat{S}E_R^* = 0.316$. However, the corresponding linear predictors that used only the regression coefficients had $T\hat{S}E_R^* = 0.526$ and $T\hat{S}E_R^* = 1.565$, correspondingly to the least squares and to $\tilde{\boldsymbol{\beta}} = (1.0864, 0.0003)$. This demonstrates again that for the NPEB method, it is not necessarily better to have a transformation that works well on its own (e.g., least squares).

Obviously, in practice we do not have validation data to help us choosing the appropriate transformation for our NPEB method through $T\hat{S}E_R^*$. We can use the risk estimator of Bickel and Collins (2.5), to compare various candidate transformations. We implemented the last risk estimator by estimating the density f_ν and its derivative with a kernel density estimator, using a Normal kernel with $h = 0.4$, as described in the Appendix. We compared the estimated risk for the above $\tilde{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}}$, and for the identity transformation $\boldsymbol{\beta}^* = (0, 0)$. The NPEB that corresponds to the identity transformation that ignores the covariates has a poor performance, its $T\hat{S}E_R^* = 0.926$. The estimates (2.5) that correspond to $\hat{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$ are: 567-361=206, 567-376=191, 567-76= 491. We see that the distinction between $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ is hard to make, while $\boldsymbol{\beta}^*$ is strongly indicated as inferior.

Finally, we tried the Bickel and Collins estimator (2.5) in Model ii. The least squares is $\hat{\boldsymbol{\beta}} = (0.793, 0.297)$, the corresponding $T\hat{S}E_R^*$ for the Reg and NPEB methods are 0.343 and 0.234, as given in Table 5. We ran a brute force optimization of the Bickel and Collins estimator with $h = 0.4$ on a grid of 10,000 points, equally spaced with a 0.01 distance between neighboring points, and centered at $\hat{\beta}$. The minimizer of the Bickel Collins estimator over the grid of points is $(1.103, 0.087)$, the $T\hat{S}E_R^*$ corresponding to Reg and NPEB are 2.57 and 0.253. The last transformation is worse than the one corresponding to $\hat{\boldsymbol{\beta}}$ in terms of the $T\hat{S}E_R^*$ of both the associated NPEB and Reg. Note, that the Reg is much worse while the NPEB is only slightly so, and we have another demonstration of the limited relevance of the performance of the predictor Reg, that corresponds to a given $\boldsymbol{\beta}$, to the performance of the corresponding NPEB.

5. Summary

In this paper we studied the problem of extending Empirical Bayes methods so they can be naturally applied in situations where there are explanatory variables.

We suggested a general perspective in which the method of Fay and Herriot and our newly proposed "NPEB with covariates" method are special cases. We demonstrated through Examples 3,4, and more generally through Theorem 3 that asymptotically the NPEB method is advantageous over the method of Fay and Herriot and over a larger class of other methods. We demonstrated it also in a data example, and saw that our newly proposed method could occasionally be a good alternative to the method of Fay and Herriot in practical situations. A comparison with the recently proposed method of Jiang and Zhang (2010) was also conducted.

Some computational aspects of our newly proposed method should be further studied, but we have seen that even sub-optimal (simpler to compute) versions of the method are advantageous.

Acknowledgement

We are grateful to the referee, who suggested to add Sub-section 2.4 and Section 4, and to the associate editor for their careful reading. We are also grateful to the editor for the encouragement and the editorial suggestions. The research of Yaacov Ritov is supported in part by an ISF grant.

Appendix A: NPEB

NPEB procedure. We will the approach of Brown and Greenshtein (2009).

Assume $Z_i \sim N(\nu_i, \sigma^2)$, $i = 1, \ldots, n$, where $\nu_i \sim G$. Let

$$f(z) = \int \frac{1}{\sigma}\varphi\Big(\frac{z-\nu}{\sigma}\Big)dG(\nu).$$

It can be shown that the normal Bayes procedure, δ_N^G, satisfies

$$\delta_N^G(z) = z + \sigma^2 \, \frac{f'(z)}{f(z)}. \tag{A.1}$$

The procedure studied in Brown and Greenshtein (2009) involves an estimation of δ_N^G. By replacing f and f' in (A.1) with their kernel estimators,

$$\hat{f}_h(z) = \frac{1}{nh}\sum \phi\Big(\frac{z-Z_i}{h}\Big), \tag{A.2}$$

and

$$\hat{f}'_h(z) = \frac{1}{nh}\sum \frac{Z_i - z}{h^2} \times \phi\Big(\frac{z-Z_i}{h}\Big). \tag{A.3}$$

We obtain the decision function, $(Z_1, \ldots, Z_n) \times z \mapsto R$,

$$\delta_{N,h}(z) = z + \sigma^2 \, \frac{\hat{f}'_h(z)}{\hat{f}_h(z)}. \tag{A.4}$$

A suitable (straightforward) truncation is applied when estimating the corresponding mean of points Z_i for which $\hat{f}(Z_i)$ is too close to zero and consequently $|\delta_{N,h}(Z_i) - Z_i| > 2\log(n)$. We did not apply such truncation in our simulations. The default choice for the bandwidth $h \equiv h_n$, suggested by Brown and Greenshtein is $1/\sqrt{\log(n)}$. See also, a cross-validation method for choosing h, suggested by Brown, Greenshtein, and Ritov (2010), together with some suggested improvements of the procedure above. In our numerical studies, we chose $h = 0.4$. The procedure is not too sensitive to the choice of h.

Appendix B: Proofs of Proposition 1 and Theorem 2

We give a proof of Proposition 1 under the assumption that, for every sequence T^n and the corresponding $\boldsymbol{\nu}^n$, we have $\max_i(|\nu_i^n|) < (\log(n))^d$ for some $d > 0$. It simplifies the arguments a little compared to the weaker Assumption 1, but the same ideas work also under the weaker assumption.

Our treatment is under a Compound Decision set-up, where $Y_i \sim N(\mu_i, 1)$, $i = 1, \ldots, n$, are not identically distributed, thus we need moderate/large deviation results for sums of non-identically distributed random variables. Under an Empirical Bayes setup Y_i are i.i.d., and more familiar large deviation results for sums of i.i.d random variables may be used.

Proof of Proposition 1. We show that the claim follows for $\hat{R}(T^n, \boldsymbol{\mu}) = n - \sum(\hat{f}'_{\boldsymbol{\nu}h}(Z_i)/\hat{f}_{\boldsymbol{\nu}h}(Z_i))^2$, where $\hat{f}_{\boldsymbol{\nu}h}$ and $\hat{f}'_{\boldsymbol{\nu}h}$ are kernel estimators of $f_{\boldsymbol{\nu}}$ and $f'_{\boldsymbol{\nu}}$, $\boldsymbol{\nu} = \boldsymbol{\nu}(T^n)$, with a normal kernel and bandwidth $h = h_n = 1/\log(n)$, see (A.2) and (A.3). It may be verified, as in Brown and Greenshtein (2009), that $E\hat{f}_{\boldsymbol{\nu}h} = f_{\boldsymbol{\nu}h}$ and $E\hat{f}'_{\boldsymbol{\nu}h} = f'_{\boldsymbol{\nu}h}$, where $f_{\boldsymbol{\nu}h}$ and $f'_{\boldsymbol{\nu}h}$ are the mixture density and its derivative in an auxiliary problem, with independent observations $Z_i^* \sim N(\nu_i, 1 + h^2)$, $i = 1, \ldots, n$. Let $R^h(T, \boldsymbol{\mu})$ be the Bayes risk in the auxiliary problem. It follows from Lemma 2 in BG(2009) and its proof, that $(R^{h_n}(T^n, \boldsymbol{\mu}^n) - R(T^n, \boldsymbol{\mu}^n))/n \to 0$, also $(E\sum[f'_{\boldsymbol{\nu}h}(Z_i^*)/f_{\boldsymbol{\nu}h}(Z_i^*)]^2 - E\sum[f'_{\boldsymbol{\nu}h}(Z_i)/f_{\boldsymbol{\nu}h}(Z_i)]^2)/n \to 0$; of course those quantities would approach 0 also for sequences h_n that approach 0 faster, but we prefer to use sequences h_n that approach 0 as slowly as possible, since that the variances of $\hat{f}_{\boldsymbol{\nu}h}$ and $\hat{f}'_{\boldsymbol{\nu}h}$ are smaller when h_n is larger. In Brown and Greenshtein (2009) it is shown that h_n should approach 0 "just faster" than $1/\sqrt{\log(n)}$.

For simplicity we take $g \equiv f_{\boldsymbol{\nu}h}$, similarly we write g', $\hat{g}$ and $\hat{g}'$, the dependence on $\boldsymbol{\nu}^n \equiv \boldsymbol{\nu}(T^n)$ and h_n is being suppressed.

It is enough to show that

$$P\Big(\sup_{T^n\in\{T^n\}}\Big|\sum\Big[\frac{\hat{g}'(Z_i)}{\hat{g}(Z_i)}\Big]^2-E\sum\Big[\frac{g'(Z_i)}{g(Z_i)}\Big]^2\Big|>\epsilon n\Big)\to 0. \tag{B.1}$$

First observe that

$$P\Big(\Big|\sum\Big[\frac{g'(Z_i)}{g(Z_i)}\Big]^2-E\sum\Big[\frac{g'(Z_i)}{g(Z_i)}\Big]^2\Big|>\epsilon n\Big)=o(\exp(-m)) \tag{B.2}$$

for every $m=n^\gamma$, $\gamma<1$. This follows by moderate deviation considerations and considerations similar (yet simpler), to those in the sequel.

We show that also

$$P\Big(\Big|\sum\Big[\frac{\hat{g}'(Z_i)}{\hat{g}(Z_i)}\Big]^2-\sum\Big[\frac{g'(Z_i)}{g(Z_i)}]^2\Big|>\epsilon n\Big)=o(\exp(-m)). \tag{B.3}$$

Equation (B.1) then follows by (B.2) and (B.3) coupled with Bonferonni.

By large deviation considerations, and since we assume that that $\max|\nu_i|<\log(n)^d$, we may neglect the 'far' tail, and instead of showing (B.3), show that

$$\begin{aligned}&P\Big(\Big|\sum\Big[\frac{\hat{g}'(Z_i)}{\hat{g}(Z_i)}\Big]^2-\sum\Big[\frac{g'(Z_i)}{g(Z_i)}\Big]^2\Big|>\epsilon n\mid|Z_i|<\log(n)^\kappa,\ i=1,\dots,n\Big)\\&\quad=o(\exp(-m))\end{aligned} \tag{B.4}$$

for large enough κ. This may be seen since

$$E\sum\Big[\frac{g'(Z_i)}{g(Z_i)}\Big]^2=E\Big(\sum\Big[\frac{g'(Z_i)}{g(Z_i)}\Big]^2\mid|Z_i|<\log(n)^\kappa,\ i=1,\dots,n\Big)+o(n),$$

for large enough κ, where $o(n)$ is uniformally small. Hence, it is enough to estimate the conditional expectations and conditional probabilities, both in what follows and similarly in the above equations. This is our approach, but for simplicity the conditioning is suppressed in the notation.

Observe that conditional on $|Z_i|<\log(n)^\kappa$, $i=1,\dots,n$ for large enough κ_0, $|g(z)|$, $|g'(z)|$, $|g'(z)/g(z)|$, $|\hat{g}(z)|$, $|\hat{g}'(z)|$, and $|\hat{g}'(z)/\hat{g}(z)|$, are all bounded by $\log(n)^{\kappa_0}$ for $|z|<\log(n)^\kappa$; furthermore the conditional variances $var(\hat{g}(z))$ and $var(\hat{g}'(z))$ are bounded by $\log(n)^{\kappa_0}/n$ for $|z|<\log(n)^\kappa$.

In order to obtain (B.4), we first argue that for every κ_1

$$P\Big(\sup_{\{z\mid|z|<\log(n)^\kappa\}}|\hat{g}(z)-g(z)|>\frac{\epsilon}{\log(n)^{\kappa_1}}\Big)=o(\exp(-m)), \tag{B.5}$$

and

$$P\Big(\sup_{\{z\mid|z|<\log(n)^\kappa\}}|\hat{g}'(z)-g'(z)|>\frac{\epsilon}{\log(n)^{\kappa_1}}\Big)=o(\exp(-m)). \tag{B.6}$$

The proof of (B.5) is as follows. For every fixed z, $|z| < \log(n)^{\kappa}$, by the moderate deviations principle

$$P\Big(|\hat{g}(z) - g(z)| > \frac{\epsilon}{\log(n)^{\kappa_1}}\Big) = o(\exp(-m)), \tag{B.7}$$

see e.g., Bernstein's Inequality, p-103 in Van Der Vaart and Wellner (1996); here we use the fact that the conditional variance, $var(\hat{g}(z))$ is bounded by $\log(n)^{\kappa_0}/n$ for large enough κ_0 when $|z| < \log(n)^{\kappa}$.

For a grid $\boldsymbol{G}$ of a size which is a power of n, by Bonferroni, $P(\sup_{z\in\boldsymbol{G}} |\hat{g}(z) - g(z)| > \epsilon/\log(n)^{\kappa_1}\ ) = o(\exp(-m))$. Now, since the derivative of g and $\hat{g}$ is bounded in the relevant domain, by $\log(n)^{\kappa_0}$, we may take a dense enough grid of a size that is a suitable power of n, so that if (B.7) is satisfied for every fixed z in the grid, (B.5) is also satisfied. The proof of (B.6) is the same.

In order to obtain (B.4) we further need the following. For an appropriate κ_2, let
$A = \{i \mid g(Z_i) > \log(n)^{-\kappa_2} \cap g'(Z_i) > \log(n)^{-\kappa_2} \cap \hat{g}(Z_i) > \log(n)^{-\kappa_2} \cap \hat{g}'(Z_i) > \log(n)^{-\kappa_2}\}$. The complement of A is denoted A^c. Now

$$\begin{aligned}
&\Big|\sum \Big[\frac{\hat{g}'(Z_i)}{\hat{g}(Z_i)}\Big]^2 - \sum \Big[\frac{g'(Z_i)}{g(Z_i)}\Big]^2\Big| \\
&\quad\le \Big|\sum_{i\in A} \Big[\frac{\hat{g}'(Z_i)}{\hat{g}(Z_i)}\Big]^2 - \Big[\frac{g'(Z_i)}{g(Z_i)}\Big]^2\Big| + \Big|\sum_{i\in A^C} \Big[\frac{\hat{g}'(Z_i)}{\hat{g}(Z_i)}\Big]^2 - \Big[\frac{g'(Z_i)}{g(Z_i)}\Big]^2\Big|.
\end{aligned} \tag{B.8}$$

For a large enough κ_2, the size of the random set A^c is small enough, so that the second summand in (B.8) is smaller than $n\epsilon/2$, with probability $1 - o(\exp(-m))$; the last probability is through large deviation considerations on the size of A^c; elaborates on the fact that for a random variable $\mathcal{Z}$ with a density g such that $|\mathcal{Z}| < \log(n)^k$, $P(g(\mathcal{Z}) < \delta) < 2\delta \log(n)^k$.

The first summand of (B.8) is smaller than $n\epsilon/2$ for large enough κ_1, with probability $(1 - o(\exp(-m)))$, by (B.5) and (B.6) . Thus, we conclude (B.4) and (B.3). The claim of the theorem now follows.

Proof of Theorem 2. We emphasize the dependence of $\hat{g}$ on $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$, by writing $\hat{g}_{\boldsymbol{Z}}(z)$. As in Proposition 1 we consider the conditional setup, conditional upon $|Z_i| < \log(n)^{\kappa}$, $i = 1, \ldots, n$. It may be verified by bounding the relevant derivatives that, for $h_n = 1/\log(n)$,

$$\Big|\frac{\hat{g}'_{\boldsymbol{Z}^1}(z)}{\hat{g}'_{\boldsymbol{Z}^1}(z)} - \frac{\hat{g}'_{\boldsymbol{Z}^2}(z)}{\hat{g}'_{\boldsymbol{Z}^2}(z)}\Big| < ||\boldsymbol{Z}^1 - \boldsymbol{Z}^2||_2 \log(n)^{\kappa_3}$$

for large enough κ_3. Hence $||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2 < \epsilon/(M \log(n)^{\kappa_3})$ implies, for $\boldsymbol{Z}^i = \boldsymbol{Y} - X\boldsymbol{\beta}_i$, $i = 1, 2$,

$$\Big|\frac{\hat{g}'_{\boldsymbol{Z}^1}(z)}{\hat{g}'_{\boldsymbol{Z}^1}(z)} - \frac{\hat{g}'_{\boldsymbol{Z}^2}(z)}{\hat{g}'_{\boldsymbol{Z}^2}(z)}\Big| < \epsilon.$$

Similarly, for $h_n = 1/\log(n)$, $\boldsymbol{\nu}^i = \boldsymbol{\mu} - X\boldsymbol{\beta}_i$, $i = 1, 2$, $||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||_2 < \epsilon/(M\log(n)^{\kappa_3})$ implies

$$\Big|\frac{g'_{\boldsymbol{\nu}^1}(z)}{g'_{\boldsymbol{\nu}^1}(z)} - \frac{g'_{\boldsymbol{\nu}^2}(z)}{g'_{\boldsymbol{\nu}^2}(z)}\Big| < \epsilon;$$

here $g_{\boldsymbol{\nu}}$ is identical to $f_{\boldsymbol{\nu}h}$, as defined in Proposition 1.

By covering $B = \{\boldsymbol{\beta} \mid ||\boldsymbol{\beta}||_2 < M, ||X\boldsymbol{\beta}||_\infty = o(n^\alpha)\ \forall\alpha > 0\} \subseteq \{\boldsymbol{\beta} \mid ||\beta||_2 < M\ \}$, with m balls of radius $\epsilon/(M\log(n)^{\kappa_3})$, it is enough to consider the transformations that correspond to the centers of those m balls and apply Proposition 1. Now, that covering number should be smaller than $\exp(n^\gamma)$, for some $\gamma < 1$, by Proposition 1. The latter is satisfied if $\log(M^p/[\epsilon/M\log(n)^{\kappa_3}]^p) < n^\gamma$ for some $\gamma < 1$, or $p\log(M) < n^\gamma$ for some $\gamma < 1$.

References

Bickel, P. J. and Collins, J. R. (1983). Minimizing Fisher information over mixtures of distributions. *Sankhyā* **45**, 1-19.

Brown, L. D. (2008). In-season prediction of bating averages: A field test of simple empirical Bayes and Bayes methodologies. *Ann. App. Stat.* **2**, 113-152.

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

Brown, L. D., Greenshtein, E. and Ritov, Y. (2010). The Poisson compound decision problem revisited. Manuscript.

Chen, S. S., Donoho, D. L. and Saunders, M. A (2001). Atomic decomposition by basis pursuit. *SIAM Rev.* **43**, 129-159.

Copas, J. B. (1969). Compound decisions and empirical Bayes (with discussion). *J. Roy. Statist. Soc. Ser. B* **31**, 397-425.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

Efron, B. (2003). Robbins, Empirical Bayes, and Microarrays (invited paper). *Ann. Statist.* **31**, 364-378.

Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors- an Empirical Bayes approach. *J. Amer. Statist. Assoc.* **68**, 117-130.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70**, 311-319.

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

Farrell, P. J., MacGibbon, B. and Tomberlin, T. J. (1997). Empirical Bayes estimators of small area proportions in multistage designs. *Statist. Sinica* **7**, 1065-1083.

Greenshtein, E., Park, J. and Ritov, Y. (2008). Estimating the mean of high valued observations in high dimensions. *Journal of Stat. Theory and Pract.* **2**, 407-418.

Greenshtein, E. and Ritov, Y. (2009). Asymptotic efficiency of simple decisions for the compound decision problem. *The 3'rd Lehmann Symposium.* IMS Lecture Notes Monograph Series, 266-275.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37**, 1647-1684.

Jiang, W. and Zhang, C.-H. (2010). Empirical Bayes in-season prediction of baseball batting averages. *Borrowing Strength: Theory Powering Applications-A festschrift for L.D. Brown* **6** (Edied by J.O. Berger, T. T. Cai, I. M. Johnstone), 263-273. IMS collections.

Koenker, R. and Mizera, I. (2012). Shape constraint, compound decision and empirical Bayes rules. Manuscript.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *J. Roy. Statist. Soc. Ser. B* **34**, 1-41.

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

Robbins, H. (1951). Asymptotically subminimax solutions of compound decision problems. *Proc. Second Berkeley Symp.*, 131-148.

Robbins, H. (1955). An Empirical Bayes approach to statistics. *Proc. Third Berkeley Symp.*, 157-164.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35**, 1-20.

Van Der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods.(invited paper). *Ann. Statist.* **31**, 379-390.

Central Bureau of Statistics, Kanfei Nesharim 66, Jerusalem 95464, Israel.

E-mail: noamc@cbs.gov.il

Central Bureau of Statistics, Kanfei Nesharim 66, Jerusalem 95464, Israel.

E-mail: eitan.greenshtein@gmail.com

Department of Statistics, Hebrew University of Jerusalem, Mount Scopus 91905, Israel.

E-mail: yaacov.ritov@gmail.com

(Received March 2011; accepted May 2012)