IMS Collections
Borrowing Strength: Theory Powering Applications –
A Festschrift for Lawrence D. Brown
Vol. 6 (2010) 263–273
© Institute of Mathematical Statistics, 2010
DOI: 10.1214/10-IMSCOLL618

Empirical Bayes in-season prediction of baseball batting averages

Wenhua Jiang[1] **and Cun-Hui Zhang**[2,*]

National Heart, Lung, and Blood Institute and Rutgers University

Abstract: The performance of a number of empirical Bayes methods are examined for the in-season prediction of batting averages with the 2005 Major League baseball data. Among the methodologies considered are new general empirical Bayes estimators in homoscedastic and heteroscedastic partial linear models.

1. Introduction

This paper is motivated by our desire to extend advances in empirical Bayes to more general settings, including homoscedastic and heteroscedastic partial linear models, and by Brown's [3] field test of empirical Bayes with the 2005 Major League baseball data.

The main thrust of empirical Bayes is the possibility of substantial reduction of a compound risk in a collection of statistical decision problems by making decisions in each problem with observations from all problems to be combined [14, 15, 17, 10, 1, 2, 18, 19, 7, 8, 21]. In the classic compound decision theory, the significance of this phenomenon is proven for similar decision problems involving independent observations and different unknown parameters. Baseball is a favorite example for empirical Bayes since it raises problems involving many players in well understood statistical models for counting data.

Before the benefits and importance of empirical Bayes became well understood, there were heated debates about its usefulness, especially about what collection of decision problems should be combined in actual statistical practice. In the midst of such a debate, Efron and Morris [8] used predicting the baseball batting average (ratio of base hits to the number of times at bat for an individual player) as a primary example of a proper collection of decision problems to combine. In that example, the James-Stein estimator is applied to the arcsin-root transformation of the batting averages of 14 players for the first 45 at bats in the 1970 season to predict their batting averages in the rest of the season. In fact, the 14 players were selected as Roberto Clemente and all those having the same number of 45 at bats as him according to a certain 1970 issue of the New York Times. The selection of Roberto Clemente, one of the most successful hitters in the preceding

[*] Supported in part by NSF Grants DMS 0804626, DMS 0906420 and NSA Grant H98230-09-1-0006.

[1] Office of Biostatistics, DECA, National Heart, Lung, and Blood Institute, 2 Rockledge Center, Bethesda, Maryland, USA, e-mail: jiangw3@nhlbi.nih.gov

[2] Department of Statistics and Biostatistics, Rutgers University, Hill Center, Busch Campus, Piscataway, NJ 08854, USA, e-mail: czhang@stat.rutgers.edu

Keywords and phrases: empirical Bayes, compound decisions, partial linear model, nonparametric estimation, semiparametric estimation, sports, hitting, batting average.

AMS 2000 subject classifications: Primary 62J05, 62J07; secondary 62H12, 62H25.

seasons, was designed “to embarrass the Jame-Stein rule” [8], since the method does not favor heterogeneity with the unknown means. Incidentally, the selection of these 14 players also provided homoscedasticity since the variance of the arcsinroot transformed success rate in Bernoulli trials is approximately a quarter of the reciprocal of the number of trials.

Brown [3] examined the performance of seven statistical methods for the in-season prediction of batting averages with the 2005 Major League data. He used the data from all players with at least 11 at bats in the first half season to predict the batting averages in the second half season for all players with at lease 11 at bats in both half seasons. This involves about 500 players. Since different players have different at bats in the first half season, he raised an interesting heteroscedastic compound decision problem and studied extensions of empirical Bayes methods for such data. He observed a moderate correlation in the 2005 data between at bats and batting averages, and significantly improved the performance of empirical Bayes predictors by applying them separately to the groups of pitchers and nonpitchers.

Brown’s [3] results motivated us to consider a heteroscedastic partial linear model. In this model, the unknown means are partially explained by a certain linear effect but not completely, and the errors are normal variables with zero mean and different known variances. In the baseball application, the linear component could include pitcher-nonpitcher group and at bats effects. We extend a number of empirical Bayes methods to this heteroscedastic partial linear model and examine their performance with the 2005 baseball data.

The rest of the paper is organized in four sections to cover data description and performance measures, partial linear models, extensions of empirical Bayes methods, and experiments with the 2005 baseball data.

2. Data description and prediction loss functions

We shall keep most of Brown’s [3] notation and consider the same set of loss functions for the prediction of batting averages.

The 2005 baseball data in [3] provide the name, status as a pitcher or not, at bats for the entire 2005 regular season and for the months April through September, and monthly total base hits for each Major League player. Post season records are excluded and the records of a few October games are merges into September. For convenience, we treat April, May and June as the first half season and the rest of the regular season as the second half [3].

Let $\mathcal{S}$ denote the set of all players. For each $i \in \mathcal{S}$, let N_{ji} and H_{ji} denote the number of at bats and number of hits for the first (second) half season for $j = 1$ ($j = 2$). The corresponding batting averages are then

$$R_{ji} = H_{ji}/N_{ji}. \tag{1}$$

Let $\mathcal{S}_j = \{i : N_{ji} \geq 11\}$ be the set of all players with at least 11 at bats for each half season. We consider the prediction of $\{R_{2i}, i \in \mathcal{S}_1 \cap \mathcal{S}_2\}$ based on the data $\{\Delta_i, N_{1i}, H_{1i}, i \in \mathcal{S}_1\}$, where $\Delta_i = 1$ if the i-th player is a pitcher and $\Delta_i = 0$ otherwise.

A reasonable model for the data assumes that the number of hits H_{1i} and H_{2i} are conditionally independent with the binomial distribution

$$H_{ji}|(\Delta_i, N_{ji}, p_i) \sim \mathrm{Bin}(N_{ji}, p_i),\ j = 1, 2, \tag{2}$$

where p_i is the batting probability of the i-th player. The standard variance stabilizing transformation for H ~Bin(N,p) is $\arcsin\sqrt{H/N}$. Brown [3] suggested a finer version of this arcsin-root transformation as

$$X = \arcsin\sqrt{\frac{H+1/4}{N+1/2}} \approx N\Big(\arcsin\sqrt{p}, \frac{1}{4N}\Big)$$

to minimize the approximation error for the mean to the order of N^{-2}. Under the binomial model (2), the batting averages are transformed into

$$X_{ji} = \arcsin\sqrt{\frac{H_{ji}+1/4}{N_{ji}+1/2}} \approx N\big(\theta_i, \sigma^2_{ji}\big), \tag{3}$$

where $\theta_i = \arcsin\sqrt{p_i}$ and $\sigma^2_{ji} = 1/(4N_{ji})$.

Let $\boldsymbol{R}_1 = (R_{1i}, i \in \mathcal{S}_1)'$ and $\boldsymbol{R}_2 = (R_{2i}, i \in \mathcal{S}_1 \cap \mathcal{S}_2)'$ with $\mathcal{S}_j = \{i : N_{ji} \geq 11\}$. Define vectors $\boldsymbol{\Delta}$, $\boldsymbol{X}_j$, $\boldsymbol{\theta}_j$ and $\boldsymbol{\sigma}_j = 1/\sqrt{4\boldsymbol{N}_j}$ analogously via (3). The problem is to predict $\boldsymbol{R}_2$ or $\boldsymbol{X}_2$ based on $(\boldsymbol{\Delta}, \boldsymbol{X}_1, \boldsymbol{\sigma}_1)$. Thus, a predictor is a Borel map $\boldsymbol{\delta}(\boldsymbol{\Delta}, \boldsymbol{X}_1, \boldsymbol{\sigma}_1) \in \mathbb{R}^{\mathcal{S}_2}$.

The data $\boldsymbol{X}_2$ and $\boldsymbol{\sigma}_2$ are used to validate the performance of predictors. As in [3], we use error measures

$$\widehat{TSE}^*[\boldsymbol{\delta}] = \frac{\widehat{TSE}[\boldsymbol{\delta}]}{\widehat{TSE}[\boldsymbol{\delta}_0]} \quad \text{with} \quad \widehat{TSE}[\boldsymbol{\delta}] = \sum_{i\in\mathcal{S}_1\cap\mathcal{S}_2} \big\{(X_{2i}-\delta_i)^2 - \sigma^2_{2i}\big\}$$

and its weighted version

$$\widehat{TWSE}^*[\boldsymbol{\delta}] = \frac{\widehat{TWSE}[\boldsymbol{\delta}]}{\widehat{TWSE}[\boldsymbol{\delta}_0]} \quad \text{with} \quad \widehat{TWSE}[\boldsymbol{\delta}] = \sum_{i\in\mathcal{S}_1\cap\mathcal{S}_2} \frac{(X_{2i}-\delta_i)^2 - \sigma^2_{2i}}{4\sigma^2_{1i}}.$$

These error measures compare predictors $\boldsymbol{\delta}$ of $\boldsymbol{X}_2$ with the naive $\boldsymbol{\delta}_0 = (X_{1i}, i \in \mathcal{S}_1 \cap \mathcal{S}_2)$. They can be viewed as approximations for the (weighted) MSE for the estimation of $\boldsymbol{\theta}$, since

$$E\Big(\widehat{TSE}[\boldsymbol{\delta}]\Big) = E\sum_{i\in\mathcal{S}_1\cap\mathcal{S}_2}(\delta_i-\theta_i)^2,\ E\Big(\widehat{TWSE}[\boldsymbol{\delta}]\Big) = E\sum_{i\in\mathcal{S}_1\cap\mathcal{S}_2}\frac{(\delta_i-\theta_i)^2}{4\sigma^2_{1i}},$$

in the normal model. For the prediction of $\boldsymbol{R}_2$ we use the error measure

$$\widehat{TSE}^*_R[\widehat{\boldsymbol{R}}] = \frac{\widehat{TSE}_R[\widehat{\boldsymbol{R}}]}{\widehat{TSE}_R[\widehat{\boldsymbol{R}}_0]},$$

where $\widehat{\boldsymbol{R}} = (\widehat{R}_i, i \in \mathcal{S}_1 \cap \mathcal{S}_2)$ with $\widehat{R}_i = \sin^2(\delta_i)$ for any predictor $\boldsymbol{\delta}$ of $\boldsymbol{X}_2$, $\widehat{TSE}_R[\widehat{\boldsymbol{R}}] = \sum_{i\in\mathcal{S}_1\cap\mathcal{S}_2}\big\{(\widehat{R}_i - R_{2i})^2 - R_{2i}(1-R_{2i})/N_{2i}\big\}$ and $\widehat{\boldsymbol{R}}_0 = \sin^2(\boldsymbol{\delta}_0)$.

3. A partial linear model

Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be a response vector and $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)' = (z_{ij})_{n\times p}$ be a covariate matrix. In a partial linear model, the mean of $\boldsymbol{y}$ is partially explained by a linear function of $\boldsymbol{Z}$ but not completely. This can be written as

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\xi} + \boldsymbol{\varepsilon}, \tag{4}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of unknown deterministic regression coefficients, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ is an unknown vector not dependent on known covariates, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is a vector of independent random errors with $E\varepsilon_i = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma_i^2$. The partial linear model is homoscedastic if σ_i are all equal and heteroscedastic otherwise. In the homoscedastic model, the common variance σ^2 can be estimated with the data in (4) if ξ_i are known to be sparse, e.g. by the $\widehat{\sigma}$ in (11). In the heteroscedastic model, the estimation of $\boldsymbol{\xi}$ requires known $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)'$ or the availability of consistent or χ^2 type estimates of σ_i^2.

The partial linear model (4) is different from the partial (semi, semiparametric) linear regression model in which ξ_i is assumed to be a smooth (possibly nonlinear) function of some additional covariates.

From a likelihood point of view, the linear effect $\boldsymbol{Z\beta}$ in (4) may not help since a sum of the linear effect and a general unknown effect is still a general unknown effect. However, after removing the linear effect, the benefits of the compound estimation of $\boldsymbol{\xi}$ could be far greater than that of the direct compound estimation of the mean of $\boldsymbol{y}$. This is also clear from an empirical Bayes point of view since the sum of $\boldsymbol{Z\beta}$ and a vector $\boldsymbol{\xi}$ with i.i.d. components is typically not a vector with i.i.d. components.

The partial linear model is closely related to the parametric random effects model

$$y_{ik} = \boldsymbol{\beta}'\boldsymbol{z}_{ik} + \xi_i + \varepsilon_{ik}, \ k = 1, \ldots, K_i, \tag{5}$$

where ξ_i are assumed to be i.i.d. variables from a distribution depending on an unknown parameter τ (e.g. $N(0, \tau^2)$). Existing work on random effects models typically focuses on the estimation of the parameters $\boldsymbol{\beta}$ and τ in the case where the identifiability of τ is ensured by the multiplicity $K_i > 1$. Parametric statistical inference for $\{\boldsymbol{\beta}, \tau\}$ and the noise level in (5) is well understood.

For the baseball data, (4) provides an interpretation of the data in (3) with

$$\boldsymbol{X}_1 = \boldsymbol{y}, \ \boldsymbol{\theta} = \boldsymbol{Z\beta} + \boldsymbol{\xi}, \ \frac{1}{4\boldsymbol{N}_1} = \boldsymbol{\sigma}^2, \ |\mathcal{S}_1| = n. \tag{6}$$

We use the least squares method (LSE) $\widehat{\boldsymbol{\theta}} = (\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i')^{-1}\sum_{i=1}^n \boldsymbol{z}_i y_i$ (heteroscedasticity ignored) and weighted least squares estimator (WLSE) $\widehat{\widehat{\boldsymbol{\theta}}} = (\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i'/\sigma_i^2)^{-1}\sum_{i=1}^n \boldsymbol{z}_i y_i/\sigma_i^2$ to carry out a preliminary examination of possible choices of $\boldsymbol{Z}$ for the linear component. Brown's [3] results suggest a strong group effect with $\boldsymbol{\Delta}$ (Pitcher) and a moderate at bats $\boldsymbol{N}_1$ (AB) effect. This is confirmed in Table 1. It turns out that the regression analysis exhibits no Pitcher-AB interaction. The scatterplots in Figure 1 demonstrate a moderate AB effect for nonpitchers and a very small AB effect for pitchers. Since most data points are within $\pm 1.96\sigma_i$ from the regression line, the model for the noise level $\sigma_i^2 = 1/(4N_{1i})$ explains a vast majority of the residuals in the regression analysis, suggesting a sparse $\boldsymbol{\xi}$.

Table 1
Squared multiple correlation coefficients, R^2

Models	LSE	WLSE
Pitcher	0.256	0.190
AB	0.247	0.208
Pitcher+*AB*	0.342	0.292
Pitcher**AB*	0.342	0.293

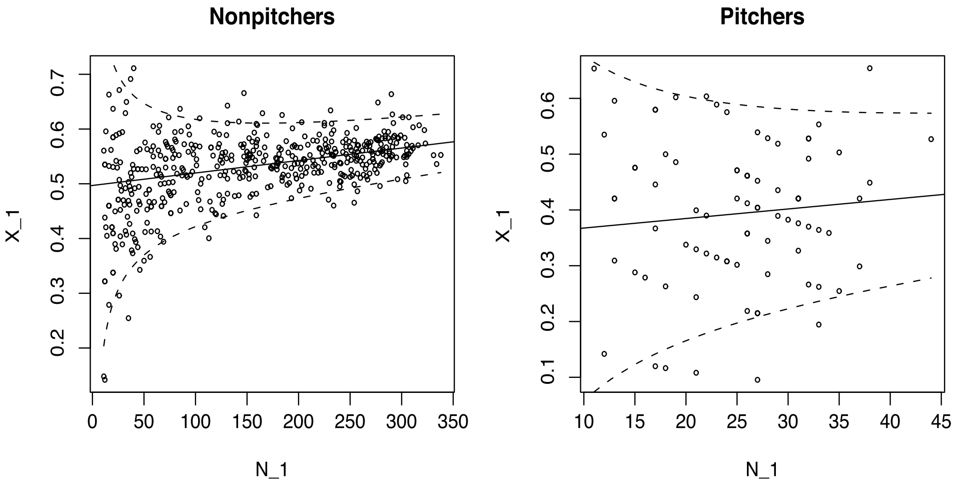

FIG 1. *Scatterplots of the transformed batting average (X_1) vs at bats (N_1) with weighted least squares regression line* $\pm 1.96\sqrt{1/(4N_1)}$*;* $R^2 = 0.151$ *for nonpitchers and* $R^2 = 0.009$ *for pitchers.*

4. Empirical Bayes methods in partial linear models

We consider in separate subsections the general (nonparametric) empirical Bayes, parametric (linear) empirical Bayes, and the James-Stein methods.

4.1. General empirical Bayes

Suppose that for a deterministic or given unknown vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$, the observation $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is distributed as independent $y_i \sim f(y|\theta_i)\nu(dy)$ with a known $f(\cdot|\cdot)$. Given a loss function $L(a, \theta)$ for the statistical inference about θ_i, the compound risk for a statistical procedure $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_n)'$ is

$$E_{\boldsymbol{\theta}} \sum_{i=1}^{n} \frac{L(\widehat{\theta}_i, \theta_i)}{n}. \tag{7}$$

For any separable statistical procedure of the form $\widehat{\theta}_i = t(y_i)$, the compound risk is identical to the Bayes risk

$$E_{\boldsymbol{\theta}} \sum_{i=1}^{n} \frac{L(\widehat{\theta}_i, \theta_i)}{n} = \int\int L(t(y), \theta) f(y|\theta)\nu(dy)\, dG_n(\theta) \tag{8}$$

with respect to the unknown prior $G_n(A) = \sum_{i=1}^{n} I\{\theta_i \in A\}/n$. Thus, the risk of separable rules is minimized at the ideal Bayes rule

$$t^*_{G_n}(y) = \arg\min_{a} \int L(a, \theta) f(y|\theta)\, dG_n(\theta). \tag{9}$$

In the original formulation of Robbins [14, 15], empirical Bayes seeks statistical procedures which approximate the ideal Bayes rule $t^*_{G_n}(\cdot)$ or its performance. Robbins [16] referred to his approach as general empirical Bayes.

From a methodological point of view, there is minimum difference in the general empirical Bayes approach between the cases of deterministic $\boldsymbol{\theta}$ (the compound setting) and i.i.d. $\theta_i \sim G$ (the empirical Bayes setting), since one seeks an approximation of a (nominal) Bayes rule t^*_G for some (nearly) completely unknown G anyway. However, theoretical results in the compound setting typically have broader impact (e.g. on minimaxity and adaptive estimation). The use of (9) as a benchmark, originally designed to "beat" standard (minimax, invariant, or maximum likelihood) procedures, has been further justified via the minimax theory for the estimation of sparse vectors [6] and the approximate risk equivalence between exchangeable and separable rules [9].

Jiang and Zhang [11] developed and studied a general maximum likelihood empirical Bayes (GMLEB) method for the approximation of $t^*_{G_n}$ in (9):

$$\widehat{\boldsymbol{\theta}} = t^*_{\widehat{G}}(\boldsymbol{y}), \quad \widehat{G} = \arg\max_{G\in\mathscr{G}} \prod_{i=1}^{n} \int f(y_i|\theta)G(d\theta), \tag{10}$$

where $\mathscr{G}$ is the class of all distribution functions. This procedure estimates $t^*_{G_n}$ with the generalized (nonparametric) maximum likelihood method [12, 22]. Different kernel estimates of the ideal Bayes rule for the estimation of $\boldsymbol{\theta}$ have been developed in [20] and [4].

Nothing in the general empirical Bayes formulation prevents adding a parameter $\boldsymbol{\beta}$ as long as the link (8) between the compound and empirical Bayes settings holds and the corresponding ideal Bayes rule $t^*_{G_n,\boldsymbol{\beta}}$ can be estimated.

In the homoscedastic partial linear model (4) where $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 \boldsymbol{I}_n)$, the GMLEB can be directly extended with

$$\widehat{\theta}_i = t^*_{\widehat{G},\widehat{\sigma}}\big(y_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i\big), \ t^*_{G,\sigma}(y) = \arg\min_a \int L(a,\theta)\varphi(y/\sigma)\,dG_n(\theta), \tag{11}$$

where $\widehat{\sigma} = \sigma$ for known σ and $\widehat{\sigma}$ =median($|y_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i|$)/median($|N(0,1)|$) otherwise, $\varphi(x) = e^{-x^2/2}/\sqrt{2\pi}$ is the $N(0,1)$ density and

$$\{\widehat{\boldsymbol{\beta}}, \widehat{G}\} = \arg\max_{\boldsymbol{\beta}, G\in\mathscr{G}} \prod_{i=1}^{n} \int \varphi\Big(\frac{y_i - \boldsymbol{\beta}'\boldsymbol{z}_i - \xi}{\widehat{\sigma}}\Big)G(d\xi). \tag{12}$$

It follows from (8) that for known $\{\boldsymbol{\beta}, \sigma\}$, the compound risk for rules of the form $t(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})$ is minimized by the ideal Bayes rule $t^*_{G_n,\sigma}(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})$ with $G_n(A) = \sum_{i=1}^n I\{\xi_i \in A\}/n$. Thus, for known σ, (11) simply replaces the unknown $\{\boldsymbol{\beta}, G_n\}$ in the ideal Bayes rule with their joint generalized maximum likelihood estimator. For unknown σ, we compute $\{\widehat{\boldsymbol{\beta}}, \widehat{G}, \widehat{\sigma}\}$ by iterating the estimating equations for $\boldsymbol{\beta}$, G_n and σ.

The direct link (8) between the compound and empirical Bayes settings breaks down in the heteroscedastic partial linear model with known σ_i, since it is unreasonable to use a prior to mix known quantities σ_i. However, the GMLEB still makes sense in the empirical Bayes setting where ξ_i are (nominally) treated as i.i.d. variables from an unknown distribution G. For $\varepsilon_i \sim N(0, \sigma_i^2)$ in (4) and the squared loss $L(a,\theta) = (a-\theta)^2$,

$$\widehat{\theta}_i = \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i + \frac{\int \xi\varphi\big((y_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i - \xi)/\sigma_i\big)\widehat{G}(d\xi)}{\int \varphi\big((y_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i - \xi)/\sigma_i\big)\widehat{G}(d\xi)}, \tag{13}$$

where $\widehat{\boldsymbol{\beta}}$ and $\widehat{G}$ could be solved by iterating

$$\text{(14)}\qquad \begin{cases} \widehat{\boldsymbol{\beta}} = \arg\max\limits_{\boldsymbol{b}} \prod_{i=1}^{n} \int \sigma_i^{-1}\varphi((y_i - \boldsymbol{b}'\boldsymbol{z}_i - \xi)/\sigma_i)\widehat{G}(d\xi), \\ \widehat{G} = \arg\max\limits_{G\in\mathscr{G}} \prod_{i=1}^{n} \int \sigma_i^{-1}\varphi((y_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i - \xi)/\sigma_i)G(d\xi). \end{cases}$$

Another general empirical Bayes method in the heteroscedastic partial linear model first rescales the data to the unit variance and then applies the GMLEB (11) in the homoscedastic partial linear model. This effectively treats G as a nominal prior for $\zeta_i = \xi_i/\sigma_i$. We call this estimator the weighted general maximum likelihood empirical Bayes (WGMLEB) since the approach is parallel to the extension of LSE to WLSE. Explicitly, the WGMLEB is

$$\text{(15)}\qquad \widehat{\theta}_i = \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i + \sigma_i \frac{\int \zeta\varphi(y_i/\sigma_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i/\sigma_i - \zeta)\widehat{G}(d\zeta)}{\int \varphi(y_i/\sigma_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i/\sigma_i - \zeta)\widehat{G}(d\zeta)}$$

for $L(a,\theta) = (a-\theta)^2$, where $\widehat{\boldsymbol{\beta}}$ and $\widehat{G}$ could be solved by iterating

$$\text{(16)}\qquad \begin{cases} \widehat{\boldsymbol{\beta}} = \arg\max\limits_{\boldsymbol{b}} \prod_{i=1}^{n} \int \varphi(y_i/\sigma_i - \boldsymbol{b}'\boldsymbol{z}_i/\sigma_i - \zeta)\widehat{G}(d\zeta), \\ \widehat{G} = \arg\max\limits_{G\in\mathscr{G}} \prod_{i=1}^{n} \int \varphi(y_i/\sigma_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i/\sigma_i - \zeta)G(d\zeta). \end{cases}$$

The WGMLEB maintains the link (8) between the compound estimation and empirical Bayes under the weighted compound risk $E_{\boldsymbol{\theta}} \sum_{i=1}^{n} (\widehat{\theta}_i - \theta_i)^2/(\sigma_i^2 n)$.

4.2. Parametric empirical Bayes

A parametric empirical Bayes method, as defined in [13], approximates

$$\text{(17)}\qquad t_{\boldsymbol{\tau}}^{*}(y) = \arg\min_{a} \int L(a,\theta) f(y|\theta)\, dG(\theta|\boldsymbol{\tau}),$$

where $\boldsymbol{\tau}$ is an unknown parameter and $G(\cdot|\cdot)$ is a known family of distributions.

Brown [3] extended parametric empirical Bayes methods to the heteroscedastic model

$$y_i = \xi_i + \varepsilon_i, \quad \xi_i \sim N(\mu,\tau),\ \varepsilon_i \sim N(0,\sigma_i^2)$$

for the estimation of ξ_i under the squared loss. Specifically, he considered the maximum likelihood (ML) or method of moments (MM) estimates of (μ,τ). Direct extension of these methods to the partial linear model yields

$$\text{(18)}\qquad \widehat{\theta}_i = \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i + \frac{\widehat{\tau}^2(y_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i)}{\sigma_i^2 + \widehat{\tau}^2},$$

where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\tau}$ are computed by iteratively solving

$$\text{(19)}\qquad \begin{cases} \widehat{\boldsymbol{\beta}} = \left\{\sum_{i=1}^{n} \boldsymbol{z}_i\boldsymbol{z}_i'/(\widehat{\tau}^2 + \sigma_i^2)\right\}^{-1} \sum_{i=1}^{n} \boldsymbol{z}_i y_i/(\widehat{\tau}^2 + \sigma_i^2), \\ \sum_{i=1}^{n} (y_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i)^2/(\sigma_i^2 + \widehat{\tau}^2)^2 = \sum_{i=1}^{n} 1/(\sigma_i^2 + \widehat{\tau}^2) & \text{(ML)}, \\ \widehat{\tau}^2 = \left\{\sum_{i=1}^{n} (y_i - \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_i)^2/(n-p) - \sum_{i=1}^{n} \sigma_i^2/n\right\}_+ & \text{(MM)}. \end{cases}$$

4.3. The James-Stein estimator

A distinct feature of the James-Stein estimator is its dominance of a standard (e.g. minimum risk unbiased or invariant) estimator under the squared loss, although it has an empirical Bayes interpretation [7, 8]. In this spirit, the name James-Stein estimator is typically reserved for its extensions with this dominance feature to explicitly allow their applications to small samples.

In the homoscedastic partial linear model (4) with $\varepsilon_i \sim N(0, \sigma^2)$, one way of achieving this dominance is to apply the James-Stein estimator separately to the projection of $\boldsymbol{y}$ to the linear span of the columns of $\boldsymbol{Z}$ and its orthogonal complement.

In the heteroscedastic partial linear model (4) with known σ_i, one may first scale to unit variance and then apply the James-Stein estimator in the homoscedastic partial linear model to achieve dominance of the naive estimator $\boldsymbol{y}$ under the σ_i^{-2} weighted mean squared error. This was done in [3] for the common mean model ($\boldsymbol{z}_i = 1, \ \forall i$). In partial linear models with general $\boldsymbol{z}_i \in \mathbb{R}^p$, this James-Stein estimator is

$$(20) \quad \boldsymbol{\theta} = \Big(1 - \frac{p-2}{\sum_{i=1}^n (\widehat{\boldsymbol{\beta}}' \boldsymbol{z}_i)^2/\sigma_i^2}\Big)_+ \boldsymbol{Z}\widehat{\boldsymbol{\beta}} + \Big(1 - \frac{n-p-2}{\sum_{i=1}^n (y_i - \widehat{\boldsymbol{\beta}}' \boldsymbol{z}_i)^2/\sigma_i^2}\Big)_+ (\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}})$$

with the WLSE $\widehat{\boldsymbol{\beta}} = \{\sum_{i=1}^n \boldsymbol{z}_i \boldsymbol{z}_i'/\sigma_i^2\}^{-1} \sum_{i=1}^n \boldsymbol{z}_i y_i/\sigma_i^2$. Brown and Zhao [5] proposed further shrinkage of (20) to reduce the weighted mean squared error.

5. Empirical Bayes prediction with the baseball data

We report applications of the GMLEB in (13)-(14), WGMLEB in (15)-(16) and the Jame-Stein estimator in (20) to the 2005 baseball data described in Section 2, and compare them with the LSE, WLSE and Brown's [3] results. The data map is given in (6) and the predictors are

$$\widehat{X}_{2i} = \widehat{\theta}_i, \quad \widehat{R}_{2i} = \sin^2\big(\widehat{\theta}_i\big), \ i \in \mathcal{S}_1 \cap \mathcal{S}_2,$$

for any estimators under consideration. The models always include the intercept and are given in parentheses, with (Pitcher*AB) as the model (Pitcher + AB + interaction). For example, James-Stein(Null) means the estimator (20) with $\boldsymbol{z}_i = 1$ as in Brown [3]. Since, p is small and n is large in all cases, the shrinkage of the regression component $\boldsymbol{Z}\widehat{\boldsymbol{\beta}}$ and positive part in (20) were not implemented with the James-Stein estimator in our experiments.

Table 2 reports results based on data from all players with at least 11 at bats in the first half season for the prediction of the batting averages of all players with at least 11 at bats in both half seasons. Brown's [3] results are included in the first block of rows as the parametric EB(MM) and EB(ML) in (18)-(19) with $\boldsymbol{z}_i = 1$, a modified nonparametric empirical Bayes estimator (NPEB) of Brown and Greenshtein's [4], and a Bayes estimator with a harmonic prior (HP; [18]). We refer to Brown [3] for detailed description and further discussion of these methods. In the next 5 blocks of rows, we report results in the (Null), (AB), (Pitcher), (AB+Picher) and (AB*Pitcher) models for the LSE, WLSE, James-Stein, GMLEB and WGMLEB methods.

Table 3 reports results for separate applications of the predictors to the non-pitcher and pitcher groups. Again, Brown's [3] results are included in the first block of rows, followed by our results in the (Null) and (AB) models.

Table 2
Midseason prediction for all batters, $(|\mathcal{S}_1|, |\mathcal{S}_1 \cap \mathcal{S}_2|) = (567, 499)$

	$\widehat{\boldsymbol{TSE}}^*$	$\widehat{\boldsymbol{TSE}}^*_{\boldsymbol{R}}$	$\widehat{\boldsymbol{TWSE}}^*$
Naive	1	1	1
EB(MM)	0.593	0.606	0.626
EB(ML)	0.902	0.925	0.607
NPEB	0.508	0.509	0.560
HP	0.884	0.905	0.600
LSE(Null)	0.853	0.897	1.116
WLSE(Null)	1.074	1.129	0.742
James-Stein(Null)	0.535	0.540	0.502
GMLEB(Null)	0.663	0.671	0.547
WGMLEB(Null)	0.306	0.298	0.427
LSE(AB)	0.518	0.535	0.686
WLSE(AB)	0.537	0.527	0.545
James-Stein(AB)	0.370	0.352	0.443
GMLEB(AB)	0.410	0.397	0.455
WGMLEB(AB)	0.301	0.291	0.424
LSE(Pitcher)	0.272	0.283	0.559
WLSE(Pitcher)	0.324	0.343	0.519
James-Stein(Pitcher)	0.243	0.244	0.427
GMLEB(Pitcher)	0.259	0.266	0.429
WGMLEB(Pitcher)	0.208	0.204	0.401
LSE(Pitcher+AB)	0.242	0.246	0.477
WLSE(Pitcher+AB)	0.219	0.215	0.435
James-Stein(Pitcher+AB)	0.184	0.175	0.391
GMLEB(Pitcher+AB)	0.191	0.183	0.387
WGMLEB(Pitcher+AB)	0.184	0.175	0.385
LSE(Pitcher*AB)	0.240	0.244	0.476
WLSE(Pitcher*AB)	0.204	0.201	0.429
James-Stein(Pitcher*AB)	0.171	0.162	0.386
GMLEB(Pitcher*AB)	0.178	0.170	0.382
WGMLEB(Pitcher*AB)	0.177	0.167	0.382

Table 3
Midseason prediction for nonpitchers and pitchers, $(|\mathcal{S}_1|, |\mathcal{S}_1 \cap \mathcal{S}_2|) = (486, 435)$ *for nonpitchers* $(|\mathcal{S}_1|, |\mathcal{S}_1 \cap \mathcal{S}_2|) = (81, 64)$ *for pitchers*

	Nonpitchers		**Pitchers**	
	widehatTSE*	$\widehat{\boldsymbol{TWSE}}^*$	$\widehat{\boldsymbol{TSE}}^*$	$\widehat{\boldsymbol{TWSE}}^*$
Naive	1	1	1	1
EB(MM)	0.387	0.494	0.129	0.191
EB(ML)	0.398	0.477	0.117	0.180
NPEB	0.372	0.527	0.212	0.266
HP	0.391	0.473	0.128	0.190
LSE(Null)	0.378	0.606	0.127	0.235
WLSE(Null)	0.468	0.561	0.127	0.234
James-Stein(Null)	0.348	0.469	0.165	0.202
GMLEB(Null)	0.378	0.465	0.134	0.178
WGMLEB(Null)	0.326	0.446	0.173	0.212
LSE(AB)	0.333	0.514	0.115	0.218
WLSE(AB)	0.290	0.465	0.087	0.182
James-Stein(AB)	0.262	0.436	0.142	0.177
GMLEB(AB)	0.257	0.415	0.111	0.154
WGMLEB(AB)	0.261	0.423	0.141	0.178

Take-away messages

The empirical Bayes in-season prediction results with the 2005 baseball data suggest the following four messages:

I. Empirical Bayes methods may substantially improve upon the least squares predictor even when the linear model assumption seems to hold well: Empirical Bayes predictors outperform lease squares predictors in all models, groups and loss functions studied with the exception of $\widehat{TSE}^*$ for pitchers in Table 3. The exception is probably due to the smaller sample size and the mismatch between the loss function and the weighted empirical Bayes methods. The regression analysis in Section 3 exhibits satisfactory fit in both the mean and variance.

II. Empirical Bayes methods may capture a great portion of the effects of missing covariables in the linear model: The phenomenon is clear in comparisons between the empirical Bayes predictors in smaller models and the lease squares predictors in larger models. It may have significant implications in dealing with latent effects and in semilinear regression when a nonparametric component suffers the curse of dimensionality.

III. The GMLEB is highly competitive compared with the James-Stein predictor with moderately large samples: In fact, the WGMLEB outperforms the (also weighted) James-Stein predictor in a great majority of combinations of models, groups and loss functions.

IV. The heteroscedastic partial linear model is tricky: The unweighted GMLEB may not handle the correlation between the mean and variance well compared with the weighted empirical Bayes methods, as the (Null) model in Table 2 demonstrates. Still, the unweighted GMLEB significantly outperforms both least squares methods in the case.

References

[1] Brown, L. D. (1966). On the admissibility of invariant estimators of one or more location parameters. *Ann. Math. Statist.* **37** 1087–1136.

[2] Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Statist.* **42** 855–903.

[3] Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *Ann. Apply. Statist.* **2** 113–152.

[4] Brown, L. D. and Greenshtein, E. (2009). Empirical Bayes and compound decision approaches for estimation of a high dimensional vector of normal means. *Ann. Statist.* **37** 1685–1704.

[5] Brown, L. D. and Zhao, L. H. (2009). Estimators for Gaussian models having a block-wise structure. *Statist. Sinica* **19** 885–903.

[6] Donoho, D. L. and Johnstone, I. M. (1994). Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Related Fields* **99** 277–303.

[7] Efron, B. and Morris, C. (1972). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika* **59** 335–347.

[8] Efron, B. and Morris, C. (1973). Combining possibly related estimation problems (with discussion). *J. Roy. Statist. Soc. Ser. B* **35** 379–421.

[9] Greenshtein, E. and Ritov, Y. (2009). Asymptotic efficiency of simple decisions for the compound decision problem. In *Optimality: The Third Erich L. Lehmann Symposium* (J. Rojo, ed.). *IMS Lecture Notes—Monograph Series* **57** 266–275.

[10] James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1** 361–379. Univ. California Press, Berkeley.

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

[12] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887–906.

[13] Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–55.

[14] Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proc. Second Berkeley Symp. Math. Statist. Probab.* **1** 131–148. Univ. California Press, Berkeley.

[15] Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 157–163. Univ. California Press, Berkeley.

[16] Robbins, H. (1983). Some thoughts on empirical Bayes estimation. *Ann. Statist.* **11** 713–723.

[17] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 157–163. Univ. California Press, Berkeley.

[18] Strawderman, W. E. (1971). Proper Bayes estimators of the multivariate normal mean. *Ann. Math. Statist.* **42** 385–388. MR0397939

[19] Strawderman, W. E. (1973). Proper Bayes minimax estimators of the multivariate normal mean for the case of common unknown variances. *Ann. Math. Statist.* **44** 1189–1194. MR0365806

[20] Zhang, C.-H. (1997). Empirical Bayes and compound estimation of normal means. *Statist. Sinica* **7** 181–193.

[21] Zhang, C.-H. (2003). Compound decision theory and empirical Bayes method. *Ann. Statist.* **33** 379–390.

[22] Zhang, C.-H. (2009). Generalized maximum likelihood estimation of normal mixture densities. *Statist. Sinica* **19** 1297–1318.