Statistica Sinica Preprint No: SS-2016-0270	
Title	Application of non-parametric empirical Bayes to treatment of non-response
Manuscript ID	SS-2016-0270
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0270
Complete List of Authors	Eitan Greenshtein and Theodor Itskov
Corresponding Author	Eitan Greenshtein
E-mail	eitan.greenshtein@gmail.com

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

Application of non-parametric empirical Bayes to treatment of non-response.

Eitan Greenshtein

Israel Central Bureau of Statistics; e-mail: eitan.greenshtein@gmail.com

and

Theodor Itskov

Israel Central Bureau of Statistics ; e-mail: itsmatis@gmail.com

Dedicated to Larry Brown

Abstract

Let (Y_i, θ_i), $i = 1, ..., n$, be independent random vectors distributed as $(Y, \theta) \sim G^*$, where the marginal distribution of θ is completely unknown, and the conditional distribution of Y conditional on θ is known. It is desired to estimate G^*, as well as $E_{G^*}h(Y, \theta)$ for a given h, based on the observed $Y_1, ..., Y_n$.

In this paper we suggest a method for these problems and discuss some of its applications. The method involves a quadratic programming step. It is computationally efficient and may handle large data sets, where the popular method that uses EM-algorithm is impractical.

The general approach of empirical Bayes, together with our computational method, is demonstrated and applied to problems of treating non-response. Our approach is nonstandard and does not involve missing at random type

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

of assumptions. We present simulations, as well as an analysis of a data set from the Labor Force Survey in Israel.

We also suggest a method, that involves convex optimization for constructing confidence intervals for $E_{G^*}h$ under the above setup.

1. Introduction and Preliminaries.

Consider a general empirical Bayes setup, where (Y_i, θ_i), are i.i.d., $i = 1, ..., n$, distributed as $(Y, \theta) \sim G^*$, and the conditional distribution of Y conditionally on θ is F_θ, $\theta \in \Theta$. The marginal distribution of θ under G^* is denoted G. Suppose that $\{F_\theta\}$ are known, while G is unknown. It is desired to estimate $\eta = E_{G^*}h(Y, \theta)$ for a given h, based on the observed $Y_1, ..., Y_n$. Our approach is to estimate η by $\hat{\eta} = E_{\hat{G}^*}h(Y, \theta)$ for a suitable estimator $\hat{G}^*$ of G^*. To clarify terminology, when necessary we refer to $\{(Y_i, \theta_i),\ i = 1, ..., n\}$ as the "aggregate sample" and to $\{Y_i,\ i = 1, ..., n\}$ as the "observed sample".

We concentrate on the non-parametric empirical Bayes setup where G is completely unknown.

There are two main contributions to this paper. One, is suggesting a method of estimating G. The method is based on quadratic programming. It is computationally much more efficient than the common approach of EM-algorithm, in addition, it naturally incorporates calibration constraints when available. An estimator $\hat{G}$ for G induces a corresponding estimator $\hat{G}^*$ for G^*, through $d\hat{G}^*(y, s) = dF_s(y)d\hat{G}(s) \equiv dG^*(y|\theta = s)d\hat{G}(s)$.

The other main contribution is a nonstandard application of empirical Bayes and estimators $\hat{\eta}$ as above to the problem of treating non-response. The suggested treatment does not involve Missing At Random (MAR) type

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

of assumptions, see, e.g., Little and Rubin (2002) and Lohr (1999). Instead, it uses, often available, information about the 'effort' invested in getting responses.

Capture Recapture Example, and Relation to Causal Inference.

We briefly explain our idea and its relation to approaches in *causal inference*. This is done in light of the familiar 'capture re-capture' example. Suppose it is desired to estimate N–the population's size of fish in a lake. For this purpose there are M capturing attempts, in each attempt, captured fish are tagged and released. Suppose n (different) fish were captured in the M attempts. For each fish among the n that were captured at least once, there is a record of the number of times it was captured. For each i, $i = 1, ..., N$, let Y_i be the number of times the corresponding fish was captured. Suppose $Y_i \sim B(M, \pi_i)$, where $p_i = p_i(\pi_i) = P_{\pi_i}(Y_i > 0)$ and $\pi_i = \pi_i(p_i)$.

Given n captured fish, if their corresponding p_i were known, then the Horvitz–Thompson estimator $\hat{N} = \sum_{i=1}^{n} \frac{1}{p_i} \equiv \sum_{i=1}^{n} h(p_i)$ could be applied. Since the p_i are unknown, a common way to simplify is to assume that $p_i \equiv p$, estimate p, e.g., by the mle $\hat{p}$, and get the estimator $\tilde{N} = \frac{n}{\hat{p}}$. Less restrictive assumptions are used in causal-inference, in related problems, as briefly discussed in what follows.

In causal-inference, when it is desired to estimate Average Treatment Effect, a similar task should be carried out, where the analogous of the unknown p_i, that corresponds to subject i, is its probability to be assigned to a certain treatment. The common approaches use estimates of $1/p_i$ in terms of propensity score see, e.g., Rosenbaum and Rubin (1983). A related approach is that

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

of Robbins et. al. (1994), in which some data points are fully observed while for some other data points there are missing covariates. The aim is to find the appropriate weights for fully observed and partially observed data points. The weights are based on an analogue of the estimated inverse probabilities of the data points to be fully observed. Those approaches involve estimation of p_i based on some covariates and some parametric model (typically logistic), under which one may *consistently* estimate the *individual* values p_i. For example, in the capture re-capture case, assume a logistic model for p_i that uses covariates like the length, weight of the fish, etc.

In the basic capture-recapture setup, with no covariates and no logistic (or other) model, applying the estimator $\sum \frac{1}{\hat{p}_i}$, where $\hat{p}_i$ is the point-wise mle estimator for p_i based on Y_i, yields a grossly biased and inefficient estimator, since $\hat{p}_i$ are not consistent for p_i, $i = 1, ..., n$.

Our approach here is the following. Let G be the distribution of p_i, $i = 1, ..., N$. Given $p \sim G$, define a two dimensional distribution G^* by introducing a variable Y, whose conditional distribution conditionally on p is $B(M, \pi)$, $\pi = \pi(p)$. As elaborated in the sequel, a pseudo Horvitz-Thompon type estimator for N is $nE_{G^*}(\frac{1}{p}|Y > 0) \equiv nE_{G^*}(h(p)|Y > 0)$. The term 'pseudo' is added since G^* is unknown. Denote by G^{*t} the conditional distribution of p condition on $Y > 0$. We suggest an NPMLE estimator $\hat{G}^{*t}$ for G^{*t} that will yield a corresponding estimator $nE_{\hat{G}^{*t}}\frac{1}{p}$.

Our approach involves fewer model assumptions than is common in causal-inference. We do not assume models that imply consistent estimation of the individual p_i. Consequently, the corresponding theoretical properties are weaker; at the same time, one may feel more comfortable with an analysis

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

that is based on weaker assumptions.

Non Parametric Maximum Likelihood Estimators (NPMLE).

Suppose $(Y_i, \theta_i) \sim G^*$, $i = 1, ..., n$ are i.i.d. as above.

Given the distributions F_θ and their corresponding densities f_θ, $\theta \in \Theta$, for every distribution G on Θ let

$$P_G(y) = \int f_\theta(y) dG(\theta).$$

An NPMLE $\hat{G}$ for G, based on the observations $Y_1, ..., Y_n$, is any $\hat{G}$ that satisfies

$$\hat{G} = argmax_G \, \Pi_{i=1}^n P_G(Y_i).$$

In the literature, NPMLE is also termed GMLE, Generalized Maximum Likelihood Estimator.

This estimator was suggested by Kiefer and Wolfowitz (1956), who gave conditions, that imply weak convergence $\hat{G} \Rightarrow G$.

The common way for computing and approximating $\hat{G}$ is through the EM-algorithm. In Section 3 we will suggest a quadratic programming-based approach. Our approach is related to and affected by the convex optimization approach that was suggested by Koenker and Mizera (2014), and the formulation in Efron (2014).

It was pointed out to us that our quadratic programming approach is very close to that of Wager (2014). Our development was done independently and about the same time, see Greenshtein and Itskov (2013) arXiv. In spite of

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

the similarity, our setup is slightly more general in order to allow various interesting non-response scenarios. Under the setup in Wager (2014), $Y_i = \theta_i + \epsilon_i$ where θ_i and ϵ_i are independent, $i = 1, ..., n$, ϵ_i are i.i.d., $i = 1, ..., n$, or *additive* noise. The later setup does not include, e.g., our example where θ_i are i.i.d, and Y_i are censored variables distributed $Geometric(\theta_i)$. Our approach is also flexible enough to exploit covariates as demonstrated later.

The estimation of G is called de-mixing or identifying mixtures. On mixture models, see, e.g., Lindsay (1995).

Motivating NPMLE-type estimators for $\mathbf{E_G h}$. Consider a function $h = h(\theta)$. A naive way to estimate $\eta = E_G h$, is to plug in a point-wise estimator of θ_i, e.g., the mle, resulting in $\tilde{\eta} = \frac{1}{n}\sum h(\hat{\theta}_i)$. The estimator $\tilde{\eta}$ is typically biased and not consistent. On the other hand, in situations where $\hat{G} \Rightarrow G$, consistency of $\hat{\eta} = E_{\hat{G}} h$, as an estimator for $\eta = E_G h$, is implied for continuous functions h.

Example 1: Suppose $Y_i \sim N(\theta_i, 1)$ are independent, $\theta_i \sim G$, $i = 1, ..., n$. Consider θ_i, such that $\theta_i \geq 1$, as "meaningful" signals. Then we might be interested in estimating the proportion of meaningful signals in our sample. For large n, that proportion is close to $\eta = E_G h$, for the function $h(\theta) = I(\theta_i \geq 1)$–the indicator of the event $\theta \geq 1$. The estimator $\tilde{\eta} = \frac{1}{n}\sum h(\hat{\theta}_i) = \frac{1}{n}\sum I(Y_i \geq 1)$ is not consistent, while $\hat{\eta} = E_{\hat{G}} h$ is consistent if G is continuous at 1. For example, when G is degenerate at 0, $\eta = E_G h = 0$ but $\tilde{\eta} \to 1 - \Phi(1) > 0$. On the other hand, by the results of Kiefer and Wolfowitz, $\hat{G} \Rightarrow G$ and thus $\hat{\eta} = E_{\hat{G}} h \to 0 = \eta$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

In cases where $\hat{G} \Rightarrow G$, we demonstrate the potential advantage of our NPMLE estimator compared to the plug in point-wise mle estimator. This advantage motivates us to apply such estimators also when there is no consistency. In particular, such estimators are applied in our main application example, presented in the next section. In cases with no consistency, the Confidence Interval method, presented in Section 4, could reassure or indicate whether such estimators are indeed worthwhile in particular cases.

There are many other applications involving the estimation of $E_G h$, for appropriate h, in various setups. See for example Zhang (2005) for a related empirical Bayes approach and more examples. The formulation in Zhang is of estimating $\sum_{i=1}^{n} h(Y_i, \theta_i)$, for latent/unobserved parameters θ_i. In that paper some efficiency results, of empirical Bayes approach related to ours, are obtained under appropriate conditions. A notable related early work, on predicting random sums, is Robbins (1977); see also Greenshtein et.al., (2008).

Our main application of treating non-response is described in the next section. In Section 3 we describe a general method to compute the NPMLE. Section 4 discusses the construction of confidence intervals for functionals $E_G h$ for a given h and an unknown G. Section 5 presents simulation results. Section 6 presents a data example involving the Israeli labor force survey, analyzed by our approach.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

2. Non-response and Empirical Bayes type Horvitz Thompson estimators.

2.1. Repeated interviewing attempts

To motivate our notation, we introduce a realistic sampling scheme from a large population. A subject from the population is randomly sampled. Then there are repeated interviewing attempts of this subject until either a response is obtained, or until M unsuccessful attempts are made. The value of M is known and it is part of the design of the survey. We model Z_i, the number of attempts until a response from subject i, as a Geometric random variable with success probability π_i. The value of π_i is unknown. The corresponding probability of response before the cutoff of M unsuccessful attempts, is denoted p_i, where

$$p_i = p(\pi_i) = P(Z_i \leq M) = 1 - (1 - \pi_i)^M.$$

The quantity of interest in each subject i is denoted X_i, observed if and only if a response is received. Suppose that N subjects are sampled from the population. We consider the "*initial* aggregate sample" (X_i, Z_i, p_i), $i = 1, ..., N$, as N realizations of independent random variables $(X, Z, p) \sim G^*$.

There are two related scenarios, we label them truncated and censored. The two scenarios induce two types of aggregate samples and observed samples.

Truncated scenario. Here we observe only the $n \leq N$ observations that correspond to responses, where the corresponding Z_i satisfy $Z_i \leq M$. We re-index, and those n points in the *initial* aggregate sample become our aggregate sample (X_i, Z_i, p_i), $i = 1, ..., n$. The corresponding observed sample

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

is $(X_i, Z_i),\ i = 1, ..., n$. The other $(N - n)$ observations are truncated and we do not know about their existence.

As an example for truncated observations whose "'existence"' is unknown, consider the capture-recapture example presented in the Introduction. The fish with zero captures are truncated. In particular we have no knowledge of how many such cases exist, if any.

Censored scenario. For responded subjects we observe (X_i, Z_i). For non-responded subjects we do not observe the value of interest X_i. However, we do get the censored information that for the corresponding i, $Z_i > M$.

It will be seen in our simulation section that the seemingly minor extra censored information can be very helpful for the estimation of $E_{G^*}X$, compared with the truncated scenario. It is also demonstrated in the following trivialized example.

Example 2: Suppose there are $M = 1$ "'repeated attempts"'. There are n observations with corresponding $Z_i = 1$, $i = 1, ..., n$. It is desired to estimate $E_{G^*}X$. In the truncated case, where the number $N - n$ is unknown, there is no way to consistently estimate G^* or $E_{G^*}X$.

Consider a censored scenario where it is known that $N - n = 0$. Then, asymptotically as $n \to \infty$, the NPMLE $\hat{G}$ converges to a degenerate distribution under which $p = 1$ almost surely. The corresponding estimator for $E_{G^*}X$ converges to the sample average $\bar{X}$, i.e., $\bar{X} - E_{\hat{G}^*}X \to_p 0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

2.2. General formulation.

The formulation in this subsection is in light of this example, but it is more general. It includes more modeling situations, such as that in our data example in Section 6, and beyond.

Let $(X_i, Z_i, I_i, p_i) \sim G^*$, $i = 1, ..., N$, be independent distributed as (X, Z, I, p). The variable X is the variable of interest and it is desired to estimate its expectation under G^*. In light of the example of repeated interviewing attempts, the variable I_i is an indicator of the event that subject i responded. The variables X_i and Z_i are observed if and only if $I_i = 1$. In terms of our example of repeated interviewing attempts, I_i is a function of Z_i and thus redundant, but we set here a general formulation. The unobserved p_i is abstract in the current general setup, but in our main example it is the probability of response; Z_i is some covariate.

Let

$$Y_i = Y(X_i, Z_i, I_i)$$

be (functions of) the observed sample, $i = 1, ..., n$. Again, $n \leq N$ in the truncated scenario and $n = N$ in the censored scenario. Let

$$\theta_i \equiv (X_i, p_i).$$

Suppose the conditional distribution of Y conditional on $\theta \equiv (X, p)$, denoted by $F_\theta(y)$, is known.

It is desired to estimate

$$\eta = E_{G^*} h(\theta) \equiv E_{G^*} X, \text{ for } h(\theta) = X.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

This setup may be readily generalized to a general function $h(\theta, Y)$ and its corresponding $\eta = E_{G^*} h(Y, \theta)$.

Here, the X-part of the parameter θ is observed under response, which is non-conventional. Our main interest is in the distribution of X in terms of $E_{G^*}X$, while the value of the unobserved/latent p is of a secondary importance. By incorporating the value of X in θ and estimating the population distribution G of θ, we learn about the desired population's distribution of X.

In addition, there are cases where we have some partial knowledge regarding the distribution of X in the population in terms of calibration constraints. Those constraints are conveniently expressed in terms of θ in our NPMLE method, see sub-section 3.2.

2.3. Censored version.

Let G be the marginal distribution of θ under G^*. Note that $E_{G^*}X = E_G X$. Thus, once an estimator $\hat{G}$ for G is available, an induced estimator for $\eta = E_G X$ is defined. Specifically we have the estimator:

$$\hat{\eta}^c = E_{\hat{G}} X = E_{\hat{G}^*} X. \tag{1}$$

The last equation and reasoning applies under the general and abstract setup of the previous sub-section. In the sequel we present helpful representations of (1), motivated by our main example of repeated interviewing attempts.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

In the rest of this subsection we are oriented toward our main example, through assumption (2) in the sequel.

Let

$$\theta_i = (X_i, p_i),$$

and let

$$Y(X_i, Z_i, I_i) = \begin{cases} (X_i, Z_i) & I_i = 1 \\ NR & I_i = 0 \end{cases}$$

Here, "'NR"' is a formal value, expressing the fact that there was No-Response.

Assume

$$p = P_{G^*}(I = 1|p) = P_{G^*}(I = 1|X, p), \tag{2}$$

$$E_{G^*}(p|X) > 0 \text{ w.p.1, under } G^*. \tag{3}$$

Under the censoring scenario the event $I_i = 0$ provides some information, e.g., the mere fact that $I_i = 0$. Thus, $n = N$.

In our main example, this is formally expressed through $F_\theta(y) = P_{G^*}(Y = y|\theta)$. For $\pi = \pi(p)$:

$$P_{G^*}(Y = y|\theta = (x_0, p)) = \begin{cases} (1-\pi)^{z-1}\pi & y = (x_0, z),\ z = 1, ..., M \\ (1-\pi)^M & y = NR \\ 0 & otherwise. \end{cases}$$

Under (2) and (3) we obtain

$$E_{G^*}\frac{X}{E_{G^*}(p|X)}I = EE_{G^*}(\frac{X}{E_{G^*}(p|X)}I|X) = EX\int E(\frac{I}{E_{G^*}(p|X)}|X, p)dG^*(p|X) = E_{G^*}X, \tag{4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

$$E_{G^*}\frac{1}{N}\sum^{N}\frac{X_i}{E_{G^*}(p|X_i)}I_i = E_{G^*}X. \tag{5}$$

When replacing G^* by $\hat{G}^*$, (5) induces the estimator:

$$\hat{\eta}^{cA} = \frac{1}{N}\sum^{N}\frac{X_i}{E_{\hat{G}^*}(p|X_i)}I_i = \frac{1}{N}\sum^{N}\frac{X_i}{E_{\hat{G}}(p|X_i)}I_i \tag{6}$$

for $E_{G^*}X$.

The estimator $\hat{\eta}^{cA}$ allows the flexibility of estimating $E_{G^*}(p|X=x)$ based on a possible source/sample other than the observed sample $(X_i, Z_i),\ i = 1, ..., N$, as done in Section 6. The estimator is an empirical Bayes variant of Horvitz-Thompson estimator, see the next subsection.

2.4. Truncated version.

The initial aggregate sample is (X_i, Z_i, I_i, p_i), $i = 1, ..., N$; here (X_i, Z_i, I_i, p_i) are i.i.d distributed like $(X, Z, p, I) \sim G^*$. Under the truncated version, the aggregate sample (X_j, Z_j, p_j, I_j) consists of the $n \leq N$ sample points for which $I_j = 1$. We re-index those sample point as (X_i, Z_i, p_i, I_i), $i = 1, ..., n$.

Denote by G^{*t}, the conditional distribution of (X, Z, I, p) conditional on $I = 1$. Let

$$Y_i = (X_i, Z_i), \quad \theta_i = (X_i, p_i),\ i = 1, ..., n,$$

and denote the marginal of G^{*t} on $\theta \equiv (X, p)$ by G^t.

In the current truncated setup we can attempt to estimate only G^{*t} and its functionals. Indeed in this section we present $E_{G^*}X$ as a functional of G^{*t}, and our estimators are presented through expectations under its estimators $\hat{G}^{*t}$. In particular our estimators are not functions of the unobserved N.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

Assume (2), coupled with

$$p > 0, \text{ and } E_{G^*}(\frac{1}{p}|X) < \infty, \text{ w.p.1 under } G^*. \tag{7}$$

In our main example, for $\pi = \pi(p)$, the expression for $F^t(y|\theta) \equiv P_{G^*}(y|\theta, I = 1) = P_{G^t}(y|\theta) =$ is

$$P_{G^t}(Y = y|\theta = (x_0, p)) = \begin{cases} \frac{(1-\pi)^{z-1}\pi}{1-(1-\pi)^M} & y = (x_0, z),\ z = 1, ..., M \\ 0 & otherwise. \end{cases}$$

Empirical Bayes Horvitz-Thompson estimators. When we condition on $(X_i, p_i),\ i = 1, ..., N$, the only remaining randomness is in I_i. By (2)

$$E(\ I_i \mid (X_j, p_j),\ j = 1, ..., N\ ) = p_i.$$

Thus, by (7), just as in the derivation of Horvitz-Thompson,

$$E_{G^*}(\sum \frac{X_i}{p_i} I_i | (X_j, p_j),\ j = 1, ..., N) = \sum_{j=1}^{N} X_i;$$

by taking expectation of the conditional expectation we obtain

$$E_{G^*}(\frac{1}{N}\sum_{i=1}^{N} \frac{X_i}{p_i} I_i) = E_{G^*} X. \tag{8}$$

Thus,

$$\hat{\eta}^{oracle} = \frac{1}{N}\sum_{i=1}^{N} \frac{X_i}{p_i} I_i \tag{9}$$

is an unbiased pseudo-estimator for $E_{G^*}X$, that could be used by an oracle who knows both p_i, $i = 1, ..., n$ and N. We later use this pseudo estimator as a benchmark for the performance of our estimators.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

We now use this idea to derive our "legitimate" estimator for the truncated version. Legitimate is in the sense that it depends only on the observed portion of the aggregate sample. It is convenient to write the argument for a discrete X. In any case our general technique, described in Section 3, is for discrete (or discretized) parameters.

Suppose the support of X is $\{x_1, ..., x_L\}$. We first estimate $P_{G^*}(X = x_l),\ l = 1, ..., L$. Let χ_x be the indicator $\chi_x \equiv I(X = x)$.

Then by (2) and (7):

$$P_{G^*}(X = x) = E_{G^*}(\frac{\chi_x}{p} I) = P_{G^*}(I = 1) E_{G^*}(\frac{\chi_x}{p} | I = 1) = P_{G^*}(I = 1) E_{G^t} \frac{\chi_x}{p},$$

with the first equality obtained similarly to (8).

From this we get:

$$P_{G^*}(X = x_k) = \frac{P_{G^*}(X = x_k)}{\sum_l P_{G^*}(X = x_l)} = \frac{E_{G^t} \frac{\chi_{x_k}}{p}}{\sum_l E_{G^t} \frac{\chi_{x_l}}{p}} = \frac{E_{G^t} \frac{\chi_{x_k}}{p}}{E_{G^t} \frac{1}{p}}. \tag{10}$$

After obtaining an estimator $\hat{G}^t$ for G^t, through our general method described in Section 3, we arrive at an estimator $\hat{P}^t_{G^*}(X = x_k)$ for $P_{G^*}(X = x_k)$:

$$\hat{P}^t_{G^*}(X = x_k) = \frac{E_{\hat{G}^t} \frac{\chi_{x_k}}{p}}{\sum_l E_{\hat{G}^t} \frac{\chi_{x_l}}{p}} = \frac{E_{\hat{G}^t} \frac{\chi_{x_k}}{p}}{E_{\hat{G}^t} \frac{1}{p}}. \tag{11}$$

Thus, we obtain the estimator for $\eta = E_{G^*} X$,

$$\hat{\eta}^t = \sum_l x_l \hat{P}^t_{G^*}(X = x_l). \tag{12}$$

A related representation alternative to (11) is given at (13). It suits better in some cases, e.g., the example in Section 6. A simple conditioning argument,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

together with Horvitz-Thompson reasoning, implies by (2) and (7) that

$$\begin{aligned} E_{G^*}\sum_{i=1}^{N}\chi_{x_k} = E_{G^*}\sum^{N}\chi_{x_k}\frac{I_i}{p} &= E_{G^*}\sum_{i=1}^{N}\chi_{x_k}I_iE_{G^*}(\frac{1}{p}|I_i=1,X=x_k) \\ &= E_{G^*}\sum_{i=1}^{N}\chi_{x_k}I_iE_{G^t}(\frac{1}{p}|X=x_k) \end{aligned}$$

Let n_l be the number of items that had the value x_l in our observed sample. Let $\hat{G}^t$ be an estimator for G^t. Then, since $P_{G^*}(X=x_k)=\frac{E_{G^*}\sum_{i=1}^{N}\chi_{x_k}}{\sum_l E_{G^*}\sum_{i=1}^{N}\chi_{x_l}}$, an alternative estimator for $P_{G^*}(X=x_k)$ is:

$$\hat{P}^{tA}_{G^*}(X=x_k)=\frac{n_kE_{\hat{G}^t}(\frac{1}{p}|X=x_k)}{\sum_l n_lE_{\hat{G}^t}(\frac{1}{p}|X=x_l)}. \tag{13}$$

The last representation defines a more flexible estimator, where the estimator $\hat{G}^t$ and the counts $n_l,\ l=1,...,L$, may come from different sources. The example in Section 6 exploits this flexibility.

Here, the alternative estimator for η under truncation is

$$\hat{\eta}^{tA}=\sum_l x_l\hat{P}^{tA}_{G^*}(X=x_l). \tag{14}$$

Remark: In our simulations and data analysis, we assume that $0<p_0\leq p$, w.p.1, for a suitable p_0. If such an assumption cannot be reasonably made, then, especially in the truncated scenario, the estimators may be very unstable.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

3. Estimation of the mixing distribution

3.1. Approximate NPMLE.

This section presents a general NPMLE method for a discrete setup. In particular, the censored and truncated scenarios of the previous section, may be obtained as special cases.

Consider a standard empirical Bayes setup, as described in the introduction, where $(Y_i, \theta_i) \sim G^*$, are i.i.d., $i = 1, ..., n$. We assume discrete distributions, in particular the F_θ, $\theta \in \Theta$, are discrete with a common finite support denoted $\{y_1, ..., y_J\}$, and that G is discrete with a given finite support $\{s_1, ..., s_K\}$. The treatment of continuous cases may be done through discretization: in light of our examples in the previous section, take $\Theta = \{x_1, ..., x_L\} \times \{p_1, ..., p_\kappa\}$, with $K = L\kappa$ points; here $p_1, ..., p_\kappa$, is a dense grid in the interval $[p_0, 1]$ for a suitable $0 < p_0$, of the possible response probabilities.

Our observations Y_i, $i = 1, ..., n$, are independent and identically distributed like a random variable Y. Denote their discrete density by $\boldsymbol{f} = (f_1, ..., f_J)'$, where

$$f_j = P(Y = y_j),\ j = 1, ..., J.$$

Denote

$$p_{jk} = P(Y = y_j | \theta = s_k), j = 1, ..., J;\ k = 1, ..., K,$$

and denote the density of the discrete distribution G by $\boldsymbol{g}^0 = (g_1^0, ..., g_K^0)'$, where

$$g_k^0 = P_G(\theta = s_k),\ k = 1, ..., K.$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

Denote by P the $J \times K$ matrix $P = (p_{jk})$.

Then:

$$\boldsymbol{f} = P\boldsymbol{g^0}. \tag{15}$$

This formulation and (15) are given in Efron (2014), but the proposed solution at (16) differs from his suggestion.

Recall, the support of G is known (or, practically approximated by a dense grid $\{s_1, ..., s_K\}$); it is the density $\boldsymbol{g}^0$ that should be estimated. We reduce the problem through sufficiency. A sufficient statistic is $\hat{\boldsymbol{f}} = (\hat{f}_1, ..., \hat{f}_J)'$, where $\hat{f}_j$ is the proportion of observations among $Y_1, ..., Y_n$, that took the value y_j, $j = 1, ..., J$. Now, $\hat{\boldsymbol{f}}$ is a scaled multinomial vector with mean $\boldsymbol{f}$ and a corresponding covariance matrix $\Sigma_{\boldsymbol{f}}/n$. Its distribution is asymptotically multivariate normal. As there is a linear dependence, the corresponding covariance matrix $\Sigma_{\boldsymbol{f}}^{-1}$ does not exist. We can replace $\hat{\boldsymbol{f}}$ by the sufficient statistic $\hat{\boldsymbol{f}}^{\boldsymbol{*}} = (\hat{f}_1, ..., \hat{f}_{J-1})'$, whose corresponding covariance matrix is Σ^*/n. The mean of $\hat{\boldsymbol{f}}^{\boldsymbol{*}}$ is $P^*\boldsymbol{g^0}$, where $P^*_{(J-1)\times K}$ is obtained from P by deleting its last row. Assume that Σ^* is non-singular. Since the distribution of $\hat{\boldsymbol{f}}^{\boldsymbol{*}}$ is asymptotically multivariate normal, a solution $\hat{\boldsymbol{g}}$ to

$$\min_{\boldsymbol{g}} (\hat{\boldsymbol{f}}^{\boldsymbol{*}} - P^*\boldsymbol{g})'\Sigma^{*-1}(\hat{\boldsymbol{f}}^{\boldsymbol{*}} - P^*\boldsymbol{g}), \tag{16}$$
$$\text{s.t. } 0 \le g_k, \ \sum g_k = 1,$$

is asymptotically an mle estimator for $\boldsymbol{g^0}$.

Practically, Σ^* is replaced by its estimate, which is obtained by utilizing the multinomial distribution of $n\hat{\boldsymbol{f}}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

We write 'an mle' rather than 'the mle', since a solution and an mle are not necessarily unique. A solution of (16) is unique if P^* is full rank.

The numerical work in this paper was done by applying the quadratic programming function *ipop*, from the R-package *kernlab*, Karatzoglou, et. al. (2004).

3.2. Covariates and Calibration.

Our formulation can accommodate covariates. Suppose the data includes a variable, denoted X. Here X may be any observed covariate, on which there are available external additional known constraints. Let $\theta_i = (X_i, p_i)$, where p_i is unobserved/latent parameter. Suppose that X is discrete and again, let $s_1, ..., s_K$ be the (approximated) discrete support of θ. For simplicity, assume that $X = 0$ or 1, indicating, e.g., whether the subject is a male or a female. Suppose, it is known that $P_{G^*}(X = 1) = 0.5$. Take $\psi(\theta) = 1 \iff X = 1$. As before $P_{G^*}(y|\theta)$, is known.

In such a case we can add to the quadratic programming problem (16) the linear 'calibration' constraint

$$\sum g_k \psi(s_k) = c \equiv 0.5.$$

Similarly, more generally, when there are a few such functions $\psi_1, ..., \psi_m$, and corresponding constants $c_1, ..., c_m$.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

4. Confidence intervals and linear optimization.

Lack of identifiability might yield very poor, non-unique and inconsistent NPMLE estimators. Nevertheless, even in a non-identifiable setup, under a specific configuration determined by G and possibly additional (calibration) constraints, one might still obtain reliable NPMLE estimates, as demonstrated in Example 2 and in Example 3 in the sequel.

In this section we suggest a confidence interval method, that could indicate whether the obtained estimator is reliable. In cases where the CI is non-informative, one might want to turn to models with further assumptions, or gather more data.

We consider the setup of Section 3, where the aggregate sample consists of i.i.d $(Y_i, \theta_i) \sim G^*$. Suppose it is desired to estimate

$$\eta = E_{G^*} h(\theta) = \sum_k h(s_k) g_k^0,$$

where $\boldsymbol{g}^0 = (g_1^0, ..., g_K^0)'$ is the discrete density of G. The truncated setup of Section 3 is obtained, when letting $G^* = G^{*t}$ and $G = G^t$.

Let $\hat{\mathbf{f}}^*$ and Σ^* be as before. Suppose that Σ^* is non-singular. Let $\hat{\Sigma}^*$ be the empirical covariance matrix. Then, as the sample size approaches infinity, $\hat{\Sigma}^{*-1}$ approaches Σ^{*-1} in probability. Furthermore, the distribution of $\sqrt{n}(\hat{\mathbf{f}}^* - P^* \boldsymbol{g}^{\mathbf{0}})$ converges weakly to a multivariate normal distribution with a zero mean vector, and covariance matrix Σ^*.

Consider the solution of the following problem of linear optimization under

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

convex constraints:

$$\hat{\eta}_U = \max_g \sum_k h(s_k) g_k \tag{17}$$
$$\hat{\eta}_L = \min_g \sum_k h(s_k) g_k$$

s.t.

$$n(\hat{\boldsymbol{f}}^* - P^* \boldsymbol{g})' \hat{\Sigma}^{*-1} (\hat{\boldsymbol{f}}^* - P^* \boldsymbol{g}) < \chi^2_{(J-1),1-\alpha},$$

$$0 \leq g_k, \quad \sum_k g_k = 1.$$

Here $\chi^2_{(J-1),1-\alpha}$ is the critical value of the appropriate α-level χ^2 test with $J-1$ degrees of freedom, the size of the discrete support of Y being J.

Theorem 1: If Σ^* is non-singular, then $(\hat{\eta}_L, \hat{\eta}_U)$ is a conservative $(1-\alpha)$ level confidence interval for η, asymptotically as $n \to \infty$.

The generalization for $h = h(Y, \theta)$ and $\eta = E_{G^*} h(Y, \theta)$ is straightforward.

Calibration. As in the previous section, additional calibration constraints of the form: $\sum_k \psi_j(s_k) = c_j$, $j = 1, ..., m$, can be added to the above convex optimization problem when available.

Example 3: Consider Example 2 under censoring with $M = 1$, and let $\eta = E_G(\frac{1}{p})$. Under this trivial setup, suppose that the n observed Z_i are $Z_1 = ... = Z_n = 1$. Consider the censored scenario, where the known number of censored observations is $N - n = n$. Due to lack of identifiability, any

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

$\hat{G}$ that satisfies $\int p d\hat{G}(p) = 0.5$ is an NPMLE. Under no further constraints (17) yields $\eta \in (1, \infty)$. Suppose that under G, it is known that $0 < p_0 \leq p$ a.s. Then the corresponding convex optimization is an exercise, with a non-trivial solution, as shown in the following.

By letting $n \to \infty$, the following is asymptotically valid for any $(1-\alpha)$ level CI. Asymptotically, as $n \to \infty$, $\hat{\eta}_U$ is obtained for $\hat{G}_U$ which has its support at the points p_0 and 1, with maximal possible weight assigned to p_0. The corresponding weights are $g_1^U = \frac{1}{2(1-p_0)}$ and $g_2^U = 1 - g_1^U$. The corresponding $\hat{\eta}_U = \frac{1+p_0-2p_0^2}{2p_0(1-p_0)}$. Here $\hat{G}_U$ satisfies the chi-square constraint with chi-square value equal to zero.

If $p_0 \leq 0.5$, then $\hat{\eta}_L$ is obtained for $\hat{G}_L$, which has all its mass at $p = 0.5$. The corresponding $\hat{\eta}_L = 2$. Again, $\hat{G}_L$ satisfy the constraint with a chi-square value of zero. If $p_0 = 0.5$, then asymptotically $\hat{\eta}_U = \hat{\eta}_L = 2$.

5. Simulations

In this section we report on simulation results for the repeated interviewing attempts example, as described in Section 2.1.

The variable of interest X is binary. We simulated under various choices of $G^* = 0.5G^{*0} + 0.5G^{*1}$, and under $M = 4, 6, 8$; under G^{*0}, $X = 0$ w.p.1, while $p = 1 - (1-\pi)^M$, where $\pi \sim \Gamma$, and $Z \sim Geometric(\pi)$; under G^{*1}, $X = 1$ w.p.1., while $p = 1 - (1-\pi)^M$, where $\pi \sim \Gamma_\gamma$ and $Z \sim Geometric(\pi)$.

The description of the various choices of Γ and Γ_γ follows:

Two Points. The distribution Γ has two-point support, at the points 0.5 and 0.9, with probability mass 0.5 at each. The distribution Γ_γ is a $(-\gamma)$ translation of Γ.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

Uniform. The distribution Γ is uniform on the interval $(0.1, 1)$. The distribution Γ_γ is a mixture of Γ and a point mass at 0.1, with mixing weights of $(1-\gamma)$ and γ.

Normal. The distribution Γ is $N(0.5, 0.1)$, rounded up to 0.1 and rounded down to 1. The distribution Γ_γ is $N(0.5-\gamma, 0.1)$ rounded up to 0.1 and rounded down to 1.

In all these cases, we used $\gamma = 0.1, 0.2, 0.3, 0.4$. As γ increases, subjects with $X = 1$ are less likely to respond compared to subjects with $X = 0$. We are estimating $\eta = E_{G^*}X = P_{G^*}(X = 1) = 0.5$. The naive estimator defined as the sample average is a biased estimator and the bias is increased with γ.

We considered the performances of $\hat{\eta}^c$ and $\hat{\eta}^t$ that correspond to the estimators based on censored and truncated observations, as explained in Sections 2.3 and 2.4. As a benchmark we also considered the simulated results for the estimator $\hat{\eta}^{oracle} = \frac{1}{N}\sum_{i=1}^{N}\frac{X_i}{p_i}I_i$, see (9), which could be used by an 'oracle' who knows the values of (the unobserved) p_i, and the value of N.

Table 1 reports simulations that correspond to $N = 1000$. Applying our method to compute $\hat{G}$, we discretized the parameter space so that $p_i \in \{1-(1-\pi_j)^M,\ j = 1, ..., 91\}$, where $\pi_1 = 0.1, \pi_2 = 0.11, ..., \pi_{91} = 1$. The columns "m- " correspond to the simulated mean of the corresponding estimator. The columns "S- " correspond to the square-root of the simulated mean of the MSE of the corresponding estimator. The "'naive"' estimator, is the estimator that estimates the population mean by the sample average. The number of simulations in each of the configurations is 1000.

Due to the non-identifiability, the NPMLE is not unique. Our choice of

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

$\hat{G}$ in the simulations was the one suggested by the quadratic programming routine.

We found the following.

i) The seemingly minor extra censored information is very helpful and $\hat{\eta}^c$ is significantly better than $\hat{\eta}^t$.

ii) The performance of $\hat{\eta}^c$ is comparable to that of the oracle and in a few cases, such as the two-point G^* with $M = 8$, the performances are virtually the same.

iii) The advantage in increasing M, in terms of the reduction of the MSE, is greater for our EB type estimators $\hat{\eta}^t$ and $\hat{\eta}^c$, in comparison with the naive estimator. This should further encourage the effort to get a response, when using such estimators.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

TABLE 1
Simulation Results

Γ	M	γ	m-naive	m-$\hat{\eta}^t$	m-$\hat{\eta}^c$	S-naive	S-$\hat{\eta}^t$	S-$\hat{\eta}^c$	S-oracle
TwoPts	4	0.1	0.4909	0.4206	0.4963	0.0184	0.0868	0.0168	0.0161
TwoPts	4	0.2	0.4743	0.4094	0.4891	0.0304	0.0996	0.0250	0.0165
TwoPts	4	0.3	0.4470	0.3867	0.4766	0.0556	0.1230	0.0405	0.0173
TwoPts	4	0.4	0.3978	0.3456	0.4532	0.1035	0.1618	0.0668	0.0192
TwoPts	6	0.1	0.4966	0.4815	0.4995	0.0164	0.0300	0.0164	0.0161
TwoPts	6	0.2	0.4872	0.4823	0.4978	0.0208	0.0335	0.0173	0.0165
TwoPts	6	0.3	0.4663	0.4726	0.4937	0.0373	0.0417	0.0203	0.0162
TwoPts	6	0.4	0.4221	0.4358	0.4846	0.0796	0.0731	0.0274	0.0178
TwoPts	8	0.1	0.4978	0.4975	0.4992	0.0156	0.0172	0.0155	0.0154
TwoPts	8	0.2	0.4933	0.5007	0.4996	0.0173	0.0200	0.0160	0.0160
TwoPts	8	0.3	0.4788	0.5022	0.4990	0.0268	0.0245	0.0169	0.0165
TwoPts	8	0.4	0.4394	0.4762	0.4948	0.0629	0.0349	0.0185	0.0178
Uniform	4	0.1	0.4855	0.3739	0.4921	0.0224	0.1335	0.0446	0.0181
Uniform	4	0.2	0.4682	0.3638	0.4816	0.0360	0.1435	0.0548	0.0184
Uniform	4	0.3	0.4504	0.3562	0.4777	0.0530	0.1516	0.0609	0.0201
Uniform	4	0.4	0.4301	0.3509	0.4710	0.0720	0.1571	0.0664	0.0197
Uniform	6	0.1	0.4882	0.4441	0.4952	0.0205	0.0629	0.0287	0.0174
Uniform	6	0.2	0.4738	0.4399	0.4893	0.0312	0.0679	0.0340	0.0174
Uniform	6	0.3	0.4597	0.4347	0.4860	0.0438	0.0735	0.0371	0.0176
Uniform	6	0.4	0.4457	0.4314	0.4858	0.0570	0.0770	0.0388	0.0183
Uniform	8	0.1	0.4908	0.4757	0.4973	0.0189	0.0340	0.0224	0.0166
Uniform	8	0.2	0.4794	0.4709	0.4941	0.0261	0.0373	0.0238	0.0162
Uniform	8	0.3	0.4679	0.4687	0.4937	0.0362	0.0408	0.0256	0.0172
Uniform	8	0.4	0.4555	0.4634	0.4913	0.0476	0.0449	0.0255	0.0173
Normal	4	0.1	0.4792	0.3570	0.4966	0.0267	0.1492	0.0227	0.0168
Normal	4	0.2	0.4422	0.3485	0.4917	0.0602	0.1594	0.0295	0.0176
Normal	4	0.3	0.3859	0.3414	0.4863	0.1156	0.1679	0.0404	0.0199
Normal	4	0.4	0.3231	0.3332	0.4833	0.1778	0.1738	0.0471	0.0211
Normal	6	0.1	0.4902	0.4571	0.4989	0.0195	0.0523	0.0184	0.0169
Normal	6	0.2	0.4664	0.4489	0.4955	0.0375	0.0631	0.0214	0.0169
Normal	6	0.3	0.4223	0.4380	0.4920	0.0796	0.0744	0.0257	0.0180
Normal	6	0.4	0.3691	0.4333	0.4919	0.1321	0.0782	0.0272	0.0191
Normal	8	0.1	0.4945	0.4899	0.4987	0.0169	0.0232	0.0168	0.0160
Normal	8	0.2	0.4777	0.4875	0.4968	0.0277	0.0277	0.0178	0.0166
Normal	8	0.3	0.4461	0.4813	0.4964	0.0564	0.0350	0.0187	0.0170
Normal	8	0.4	0.4016	0.4762	0.4962	0.0999	0.0401	0.0196	0.0183

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

6. Data Example.

In this section we report on the application of our method to a data set from the Labor Force Survey, conducted by the Israel Central Bureau of Statistics. The sampling method is 4-8-4 rotating panels, where each panel has 4 consecutive investigations in 4 months, then 8 months break and finally another 4 consecutive investigations in additional 4 months. In our analysis we treated the the two panels with their i'th and $(4+i)'th$ investigation, $i = 1, 2, 3, 4$, as one. Thus, for our analysis the rotation method may be equivalently treated as a 4-in rotation, as described in the following.

The survey is given to four panels, where each panel is investigated four times in four consecutive months. Each month one panel finishes its fourth investigation and in the next month it is replaced by a new panel that remains for four months. The main purpose of the survey is to estimate the proportion of 'Unemployment'$\equiv x_1$, 'Employment'$\equiv x_2$, and those who are 'Not in Working Force (NWF)' $\equiv x_3$; the last category is of those who do not have a job, nor are they looking for one. Under the sampling method, persons are equally likely to be sampled. The monthly sample size is about 20000. The response rate is around 80%.

Our observed sample each month consists of $Y_i = (X_i, Z_i)$ that correspond to the people who respond that month, i.e., their corresponding $I_i = 1$. Our variable of interest is X-'working status'. For each person i, Z_i is the corresponding number of responses he has made so far. For people in a panel who had so far B scheduled investigations and belong to the observed sample of the relevant month, we model Z_i as $Z_i = 1 + W_i$, where W_i is a Binomial ran-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

dom variable $W_i \sim Binomial(B-1, p_i),\ B = 1, 2, 3, 4$, p_i is the probability of a response from person i in a single month; p_i is assumed to be fixed for the same person in different months. The estimation of G^t is done based on data only from the panel with its fourth investigation that month. Incorporating the information from the other panels whose $Z_i \leq 3$, does not add much to the estimation accuracy, while undesirably complicating the analysis.

Let n_l, $l = 1, 2, 3$, be the number of occurrences of $X = x_l$ in our observed sample. We emphasize, the n_l counts are from *all* the four panels; it is the estimation of G^t, which is based only on one panel. We estimate $P(X = x_l),\ l = 1, 2, 3$, by (13), the estimator is denoted $\hat{P}$. Our tables below are based on the last estimator.

As in the previous section, the particular (non-unique) approximate NPMLE $\hat{G}$ that we chose, was simply the one suggested by the quadratic programming routine.

Remark: The reason that we use the truncated version and not the censored one, is that under non-response it could be that the corresponding apartment is simply unoccupied and thus we are not sure about the effective size N of the initial aggregate sample, where N corresponds to the number of occupied households.

Estimation of the (known) proportions of sex and age categories.

Since the true proportions of the various working statuses are unknown, we first demonstrate the performance of our estimation method in estimating the following *known* true proportions, based on the responses in a given month.

In one case we estimate the proportion of males in the population, our X

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

variable is an indicator of the event 'the person is a male', the proportion of males in the population is known to be 0.485. The proportion of males in the survey among responders in our observed sample, is about one percent lower.

In another example we estimate the proportion of the group age 20-39. The known proportion of this age group in the population is 0.397, while their response rate is particularly low; their proportion among responders is nearly 3 percent lower.

Each of Tables 2 and 3 has three lines that correspond to the data obtained in Aug/2012, Dec/2012, and April/2013. We took periods that are four months apart in order not to have overlapping panels. The general picture persist in other months.

The columns True, Naive, and $\hat{P}$, correspond to the true population's proportion, the sample proportion among responders, and our estimator $\hat{P}$. In each case one sees that $\hat{P}$ corrects the naive estimator in the right direction.

TABLE 2
Comparison of estimates of male's proportion.

	True	Naive	$\hat{P}$
Male	0.4853	0.4752	0.4822
	0.4853	0.4751	0.4819
	0.4853	0.4776	0.4842

TABLE 3
Comparison of estimates of proportion of 20-39 age group.

	True	Naive	$\hat{P}$
Age 20-39	0.3970	0.3664	0.3815
	0.3970	0.3631	0.3984
	0.3970	0.3598	0.3842

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

Estimation of the proportion of Employment statuses

After gaining some confidence in $\hat{P}$, we now examine its estimates in the estimation of the proportion of 'Unemployed', 'Employed' and those 'Not in Working Force' (NWF). In Table 4 the columns Naive and $\hat{P}$ are as before. The column Bureau gives the estimates of the Israel Central Bureau of Statistics, for the three categories of working statuses. The estimator of the bureau is obtained through a method that involves calibration in a 'post-stratification manner' (the final estimate involves additional seasonal adjustment that we neglect). The three parts of the table refer to the three working statuses. The three lines in each part refer to the three months as before. The Bureau and the $\hat{P}$ estimators 'correct' the naive estimator for Employment and NWF, in opposite directions. It seems that the correction of the naive estimates by the Bureau estimates, in the cases 'Employment' and the 'NWF' are in the wrong direction. This is suggested also by incorporating revealed working status of non-responders as revealed through their "future" investigations. On the other hand, both the Bureau and $\hat{P}^t$ correct the unemployment naive estimate by increasing it.

TABLE 4
Comparison of unemployment estimates.

	Bureau	Naive	$\hat{P}$
Emp	0.6104	0.5931	0.5761
	0.6081	0.5992	0.5910
	0.6089	0.5986	0.5881
NWF	0.3416	0.3594	0.3748
	0.3465	0.3576	0.3605
	0.3491	0.3621	0.3720
UnEmp	0.0479	0.0475	0.0492
	0.0454	0.0431	0.0484
	0.0420	0.0392	0.0399

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

Acknowledgment.

We are indebted to an associate editor and the referee for their careful reading, corrections, and their especially helpful comments.

Larry Brown was closely involved in this project since its initial stage. Yet, he did not want to be listed as a co-author. The paper is dedicated to him.

References

Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statistical Science* **29** 285-301.

Greenshtein, E., Park, J., and Ritov, Y. (2008). Estimating the mean of high valued observations in high dimensions. *JSTP* **2** No. 3 407-418.

Greenshtein, E., and Itskov, T. (2013). Deconvolution with application to estimation of sampling probabilities and the Horvitz-Thompson estimator. arXiv: 1309.2136V3

Karatzoglou, A., Smola, A., Hornik, K., and Zeleis, A., (2004). An S4 package for kernel methods in R. *Journal of Statistical Software* **11**, No. 9, 1-20.

Kiefer and Wolfowitz (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann.Math.Stat.* **27** No. 4, 887-906.

Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions and empirical Bayes rules. *JASA* **109**, 674-685.

Lindsay, B. G. (1995). Mixture Models: Theory, Geometry and Applica-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0270

tions. Hayward, CA, IMS.

Little, R.J.A and Rubin, D.B. (2002). Statistical Analysis with Missing Data. New York: Wiley.

Robbins, H. (1977). Predictions and estimation for the compound Poisson distribution. *PNAS* , **74**, No 7, 2670-2671.

Robins, J.M., A Rotnitzky, and Zhao, L.P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *JASA*, **89**, 846-866.

Rosenbaum, P.R., and Rubin, D.B (1983). The central role of propensity score in observational studies for causal effects. *Biometrika*, **70**, 1, 41-55.

Sharon L. Lohr (1999). *Sampling Design and Analysis*. Brooks/Cole publishing company.

Wager, S, (2014). A geometric approach to density estimation with additive noise. *Stat. Sinica* **24**, 533-554.

Zhang, C-H. (2005). Estimation of sums of random variables: Examples and information bounds. *Ann. Stat.* **33** No.5. 2022-2041.