Generalized maximum likelihood estimation of the mean of parameters of mixtures. With applications to sampling and to observational studies[*]

Eitan Greenshtein

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

Ya'acov Ritov

University of Michigan; e-mail: `yritov@umich.edu`

Abstract: Let $f(y \mid \theta)$, $\theta \in \Omega$ be a parametric family, $\eta(\theta)$ a given function, and G an unknown mixing distribution. It is desired to estimate $E_G(\eta(\theta)) \equiv \eta_G$ based on independent observations $Y_1, ..., Y_n$, where $Y_i \sim f(y \mid \theta_i)$, and $\theta_i \sim G$ are iid.

We explore the Generalized Maximum Likelihood Estimators (GMLE) for this problem. Some basic properties and representations of those estimators are shown. In particular we suggest a new perspective, of the weak convergence result by [14], with implications to a corresponding setup in which $\theta_1, ..., \theta_n$ are *fixed* parameters. We also relate the above problem, of estimating η_G, to nonparametric empirical Bayes estimation under a squared loss.

Applications of GMLE to sampling problems are presented. The performance of the GMLE is demonstrated both in simulations and through a real data example.

Keywords and phrases: GMLE, mixing distribution, nonparametric empirical Bayes, sampling.

Received November 2021.

1. Introduction

Let $f_\theta(y) \equiv f(y \mid \theta)$, $\theta \in \Omega$, be a parametric family of densities with respect to some measure μ. Let G be an unknown mixture distribution. We observe $Y_1, ..., Y_n$, which are realizations of the following process. Let $\theta_i \sim G$, $i = 1, ..., n$, be independent, where conditional on $\theta_1, ..., \theta_n$, $Y_i \sim f(y \mid \theta_i)$, $i = 1, \ldots, n$, are independent. Given a function $\eta(\theta)$, let $\eta_G \equiv \eta(G) \equiv E_G \eta(\theta)$. Based on our observations, it is desired to estimate η_G.

Consider for example the case where $Y_i \sim N(\theta_i, 1)$ and $\theta_i \sim G$ where G is supported away from 0. If $\eta(\theta) = E_\theta(Y)$, then the natural estimator for η_G is $\frac{1}{n} \sum Y_i$, which is consistent and efficient. On the other hand, deriving a

consistent estimator for η_G is less obvious if $\eta(\theta) = 1/\theta$. Yet, the Generalized Maximum Likelihood Estimator (GMLE) defined below, see (2), yields a simple consistent estimator for both functionals: If $\hat{G}$ is a GMLE estimator for G a GMLE estimator of $\eta_G \equiv E_G\eta(\theta)$ is defined by

$$\hat{\eta}_G \equiv E_{\hat{G}}\eta(\theta) \tag{1}$$

GMLE for G. Given a distribution G and a dominated family of distributions with densities $\{f(y \mid \theta) : \theta \in \Omega\}$, define

$$f_G(y) = \int f(y \mid \vartheta)dG(\vartheta).$$

Given observations $Y_1, ..., Y_n$, a GMLE $\hat{G}$ for G [14] is defined as:

$$\hat{G} = \underset{G}{\operatorname{argmax}} \ \Pi f_G(Y_i); \tag{2}$$

the maximization is with respect to all probability distributions. Note, GMLE for G is also referred in the literature as Nonparametric Maximum Likelihood Estimator (NPMLE). We use GMLE as a tribute to its originators, [14]. Traditionally, a GMLE estimator $\hat{G}$ for a mixture G is approximated via EM algorithm on a finite subset of the parameter space, see the seminal paper, [16]. [15] suggested exploitation of convex optimization techniques.

Much of the research on GMLE is about characterization of the GMLE $\hat{G}$ in terms of the size of its support, see, e.g., [17]. Asymptotics and efficiency of the estimator $\hat{\eta}_G$ for η_G is studied in semiparametric theory, see e.g., [1]. One goal of this paper is to present some new problems and applications where GMLE estimators are useful: we elaborate in particular on examples of 'sampling with no response', 'post-stratification', and 'observational studies'.

We also relate GMLE estimators $\hat{\eta}_G$ to nonparametric empirical Bayes and compound decision problems, and provide useful and insightful representations. In nonparametric empirical Bayes the goal is to estimate the specific values $\eta_i \equiv \eta(\theta_i)$, $i = 1, ..., n$, under a given loss, mainly, as in our case, a squared loss. A mixing distribution G induces a joint distribution of (Y, θ). We denote the corresponding conditional expectation of θ conditional on Y, by $E_G(\theta \mid Y)$. The nonparametric empirical Bayes approach is to estimate η_i, $i = 1, ..., n$ by the plug-in estimator $\hat{\eta}_i = E_{\hat{G}}(\eta(\theta) \mid Y_i)$, see, the general approach in the seminal papers of [22, 23, 24], see the plug-in approach in, e.g., [12, 15], and g-modeling in [6]. In subsection 3.2 we show that $\hat{\eta}_G$ may be represented as $\hat{\eta}_G = \frac{1}{n}\sum \hat{\eta}_i = \frac{1}{n}\sum E_{\hat{G}}(\eta(\theta) \mid Y_i)$. In the case of $Y_i \sim N(\theta_i, 1)$ with $\eta(\theta) = E_\theta Y$, $\theta_i \sim G$ are iid (and more generally) we show, that $\hat{\eta}_G \equiv E_{\hat{G}}\eta(\theta) = \frac{1}{n}\sum Y_i$.

Asymptotics of GMLE estimators $\hat{\eta}_G$, mainly in terms of consistency, is explored both theoretically and through simulations. There are cases where $\hat{G}$ is not unique, e.g., in situations where G is non-identifiable. When $\hat{G}$ is not unique, we consider as GMLE any $\hat{\eta}_G = E_{\hat{G}}\eta(\theta)$ that corresponds to any GMLE $\hat{G}$. In such cases, as $n \to \infty$, often, not every sequence of GMLE is consistent; we

argue that GMLE estimators are still plausible and worthwhile, and suggest "GMLE related" Confidence-Intervals for η_G.

Given realizations $Y_1, ..., Y_n$ as above, a related problem, studied in [26], is the prediction of the realized $\sum \eta(\theta_i)$, where $\theta_i \sim G$ are iid. His approach is also related to GMLE. Zhang elaborates on the difference between the estimation and the prediction problems in terms of efficiency. He also presents interesting applications where it is desired to estimate the realized $\sum \eta(\theta_i)$.

Our problem may be formalized without considering the parameters, θ_is, as random and appealing to a mixture distribution G. We consider $\theta_1, ..., \theta_n$ as unknown parameters and $Y_i \sim f(y \mid \theta_i)$, $i = 1, ..., n$, as independent (but not identically distributed) observations. The goal is to estimate $\sum \eta(\theta_i)$. Let G^n be the empirical distribution of $\theta_1, ..., \theta_n$. Let $\hat{G}^n$ be the GMLE based on $Y_1, ..., Y_n$, as defined in (2) (i.e., the GMLE pretending $\theta_1, \ldots, \theta_n$ are i.i.d. random variables sampled from a distribution G). Under suitable triangular array formulation and conditions, the sequence of signed measures $(\hat{G}^n - G^n)$, converges weakly to the zero-measure. However, a reference to such a general result is not known to us, and we give a proof in Section 3.3. This weak convergence result motivates estimating $\frac{1}{n} \sum \eta(\theta_i) = E_{G^n} \eta(\theta) = \eta_{G^n}$ by $E_{\hat{G}^n} \eta(\theta)$.

As in most of the literature, we often appeal to a mixing distribution G and random $\theta_i \sim G$, which makes the formulation more convenient. The two approaches of appealing to a mixing distribution G, versus avoiding it, are analogous to empirical Bayes versus compound decision approaches. See, e.g., [25, 2], and the fore mentioned seminal papers by Robbins. It is also related to the distinction between incidental and random nuisance parameters, see, e.g., [20]).

In Section 2 we present some motivating examples. In Section 3 we present some theoretical results concerning representations of $\hat{\eta}_G$, and some asymptotics. In Section 4 we present simulations. In Section 5 we sketch a 'GMLE related' Confidence Interval for η_G. In Section 6 we present a real data example taken from the Israeli Social Survey.

2. Examples

A main motivation for our study of GMLE are the following sampling models and problems in the context of stratification, and observational studies.

In stratified sampling and in post-stratification, it is desired to have a very fine stratification. Then, conditional on a fine stratum it appeals that missing observations are missing completely at random. Similarly in observational study (or, in convenience sampling), fine stratification makes the assumption that conditional on the strata the success or failure of a treatment are independent of the, often unknown, allocation mechanism to treatment and control. However, under very fine natural stratification many strata are likely to be randomly empty, i.e., with no sampled observations. In practice, when the number of observations is high the stratification used is fine to the level of existence of a meaningful number of empty cells. The GMLE method is a natural way of handling the resulting difficulties of empty strata.

We elaborate on the above. Suppose that it is desired to estimate the proportion of unemployed in the population. A sampled subject will respond to the survey with unknown probability. The proportion of unemployed among the responders in the survey is strictly a biased estimator of their proportion in the population, since there is a strong observed correlation between employment status and response. It is known that in the relevant surveys, subgroups with higher unemployment rates have lower response rates. We may hope, however, that if we consider small enough and homogenous stratum, the willingness to answer and the answer will be, conditionally on the stratum, practically independent. Thus, the missing responses would be practically missed completely at random (MCAR) within each stratum. See, e.g., [18].

Let K_i be the number of full observations in Stratum i (i.e., K_i were sampled and responded). Let X_i be the number of unemployed among the observations from Stratum i. For simplicity, we consider a situation of n strata with equal weights. Our observations are $Y_i = (X_i, K_i)$, $i = 1, \ldots, n$, where the conditional distribution of X_i conditional on K_i is $B(K_i, p_i)$. It is desired to estimate the population's proportion $p = n^{-1} \sum p_i$.

Similar considerations apply in observational studies. Suppose that conditional on a fine stratification, within each stratum the probability of success or failure, under treatment or control, is independent of the unknown allocation mechanism to treatment or control. Given a treatment, we thinks of X_i, the number of successes of the treatment in stratum i, and K_i, the number of times the treatment was applied in Stratum i. We observe $Y_i = (X_i, K_i)$. One approach of analyzing such observational data is to pair observations based on their propensity score. Using our method we do not need pairing, and we may handle even the extreme case, where treatment was applied in one subset of strata, while control was applied in another, disjoint, subset of strata.

2.1. Sampling models

We consider the following two realistic scenarios, where the sample size K_i is random. In light of Section 3.3, the following motivating models apply to both setups of random and fixed $\theta_1, ..., \theta_n$. We use the notations and formulation of random θ_i, $i = 1, ..., n$, for convenience.

Model (i)   *Stratified sampling with non-response.* A random sample of κ_i subjects from stratum i is sampled. The probability of a random subject from stratum i to respond is $\pi_i \leq 1$. Thus, the number of actual responses, K_i, $K_i \sim B(\kappa_i, \pi_i)$ is a random variable.

Model (ii):   *Post-Stratification.* We assume that K_i has a $Poisson(\lambda_i)$ distribution, $i = 1, 2, \ldots, n$. This scenario is reasonable whenever we have a convenience sample, observational studies, or, very low response rate.

In Model (i), $\theta_i = (\theta_{i1}, \theta_{i2}) \equiv (\pi_i, p_i)$, where π_i is the probability of response, while p_i is the proportion (of, say, unemployed) in stratum i. Conditional on θ_i, $X_i \mid K_i \sim B(K_i, p_i)$ and $K_i \sim B(\kappa_i, \pi_i)$. In Model (ii), $\theta_i = (\theta_{i1}, \theta_{i2}) \equiv (\lambda_i, p_i)$. Given θ_i, $X_i \mid K_i \sim B(K_i, p_i)$ as before, and $K_i \sim Poisson(\lambda_i)$.

Suppose we are interested in estimating $\eta_G = E_G \eta(\theta)$, for $\eta(\theta_i) = p_i$. If there are no empty strata, i.e., $K_i > 0$, $i = 1, ..., n$ then the obvious and naive estimator:

$$\hat{p}_N = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{K_i},$$

is applicable. The problem with the above estimator is that many strata may be empty, i.e., with corresponding $K_i = 0$. When there are some empty strata, a common ad-hoc approach is of 'collapsing strata' where after the data is observed some empty strata are unified with non-empty ones. An extreme collapsing is to a single stratum, which yields the Extreme Collapsing estimator:

$$\hat{p}_{EC} = \frac{\sum X_i}{\sum K_i},$$

which is, in fact, desirable when $p_1 = ... = p_n$ but not in general.

Another way of handling the difficulty of empty strata, is to assume that strata are 'missing at random', that is K_i is uncorrelated with p_i, and strata with $K_i = 0$ may be simply ignored.

Our approach, of applying GMLE, handles cases with many empty strata in a natural way, which is not ad-hoc and does not rely heavily on missing at random type of assumptions.

Finding a GMLE for a two dimensional distribution G, outside of the normal setup, only recently appears in the literature, see, e.g., [9, 19]. One reason might be its recent popularity due to Koenker and Mizera's computational methods. Our Model (ii) is presented in Feng and Dicker, however, their motivation and context is very different than ours, and is not related to sampling. In addition, they study the estimation of the the individual p_i, $i = 1, ..., n$ via non parametric empirical Bayes, while we study the estimation of $\eta_G = E_G \eta(\theta)$, where $\eta(\theta_i) = p_i$. We also provide consistency results in the estimation of η_G.

Different related sampling models are studied in [8]. In their censored-case, n individuals are attempted to be interviewed, where there are at most κ^0 attempts for each individual i, $i = 1, ..., n$. The probability of a response from individual i in any single attempt is π_i, and attempts are independent. Let $K_i \leq \kappa^0$ be the number of interviewing attempts of individual i, then K_i is a truncated geometric variable. Let Z_i be a 0-1 valued r.v., e.g., indicator of unemployment, $p_i = P(Z_i = 1)$. Suppose $\theta_i \equiv (\pi_i, p_i)$ are i.i.d distributed G. It is desired to estimate $E_G Z_i = E_G(\eta(\theta))$, where $\eta(\theta) = p$. The possible outcomes, of the attempted interviews of individual i, are $(Z_i, K_i) \in \{(1, K_i), (0, K_i), 1 \leq K_i \leq \kappa^0, \& NULL\}$, where $NULL$ indicates non-response in κ^0 attempts. This scenario is within our setup as one may think of individuals as strata of size one.

The truncated case of [8] is similar to the above censored case, except that we do not know about the existence of items that did not respond. Denote by G^t the conditional distribution of θ, conditional on response. Then: $dG^t(\theta) \propto P_\theta(K_i \leq \kappa^0)dG(\theta)$. Let $\hat{G}^t$ be the GMLE for G^t, which is based only on the

(non-truncated) observations $\{i : K_i \leq \kappa^0\}$. Then, when $\pi_i > 0$ w.p.1, we suggest the GMLE estimator

$$\hat{\eta} = \int \eta(\theta)d\hat{G}(\theta) \equiv \frac{\int \eta(\theta)P_\theta^{-1}(K_i \leq \kappa^0)d\hat{G}^t(\theta)}{\int P_\theta^{-1}(K_i \leq \kappa^0)d\hat{G}^t(\theta)}.$$

3. Asymptotic results and equivalent representations of GMLE estimators

We consider mainly the model in which $\theta_i = (\lambda_i, p_i)$, $\theta_i \sim G$, $i = 1, ...n$ are iid random variables and consider the estimation of the function $\eta(\theta_i) = p_i$. In light of subsection 3.3 below, analogous results may be derived when $\theta_1, ..., \theta_n$ are considered unknown parameters.

3.1. Consistency and asymptotics in models (i) and (ii)

We will now argue that under (ii) if $P(\lambda = 0) = 0$, then the expected value of p can be consistently estimated, while under model (i) it cannot. The difference between the two models is that on the one hand, in model (ii) K_i may get the values $0, 1, 2, \ldots$, and as $n \to \infty$, the number of observed outcomes grows to infinity. On the other hand, in model (i), κ_i is bounded by some finite κ, then the total number of outcomes is $(\kappa + 1)(\kappa + 2)/2$, which is not enough for the identification of the distribution or even the mean of of p_i. Here are the details.

3.1.1. Consistency in Model (ii)

In Model (ii) we have consistency in the estimation of η_G, when $P_G(\lambda = 0) \equiv G(\{\lambda = 0\}) = 0$, as proved in the following theorem.

It is convenient to re-parametrize the problem, as follows. Given $X_i \sim B(K_i, p_i)$ conditional on K_i, and $K_i \sim Poisson(\lambda_i)$. Denote $W_{i1} \equiv X_i$, and $W_{i2} = K_i - X_i$. Then W_{i1} and W_{i2} are independent Poissons conditional on (λ_i, p_i), with corresponding parameters $\xi_{1i} \equiv p_i\lambda_i$ and $\xi_{2i} \equiv (1 - p_i)\lambda_i$.

Before presenting the theorem, recall that G is identifiable in the model $f_G(y) = \int f_\vartheta(y)dG(\vartheta)$ if $f_G = f_{\tilde{G}}$ implies $G = \tilde{G}$.

Theorem 1. *Let G_ξ be the distribution of (ξ_1, ξ_2) and let $\hat{G}_\xi$ be the GMLE based on iid (W_{i1}, W_{i2}), $i = 1, ..., n$. then G_ξ is identifiable and $\hat{G}_\xi$ converges weakly to G_ξ.*

Proof. The proof follows from [14]. Checking the conditions is standard, the identifiability condition is verified, e.g., by [13]. □

Corollary 1. *for any function $\eta(\theta)$, such that $\eta(\theta) = \psi(\xi_1, \xi_2)$, for ψ which is continuous and bounded on the support of (ξ_1, ξ_2) under G, $E_{\hat{G}}\eta(\theta) \to E_G\eta(\theta)$. In particular, if under G_ξ, $P(\lambda > 0) = 1$, then $E_G p$ is identifiable.*

The condition $P(\lambda{>}0) = 1$ is necessary, since otherwise $P(\eta{=}0) > 0$ while the distribution of p conditioned on $\lambda = 0$ is unidentified, since there are no observations on p from the atom at $\{\lambda = 0\}$. On the other hand, if $P(\lambda{>}0) = 1$ than $\eta_\epsilon(\theta) = \psi_\epsilon(\xi_1, \xi_2) \equiv \frac{\xi_1}{\xi_1+\xi_2+\varepsilon}1\{\xi_1 + \xi_2{>}0\}$ is bounded and continuous, $E_{\hat{G}}\eta_\varepsilon(\theta) \to E_G\eta_\varepsilon(\theta)$ for any $\epsilon > 0$. But $\lim_{\epsilon\to 0} E_G\eta_\epsilon = E_G\eta(\theta)$, and hence $E_{\hat{G}}\eta(\theta) \to E_G\eta(\theta)$.

We conclude that when confining to the class of distributions in $\Gamma = \{G \mid G(\{\lambda = 0\}) = 0\}$, we have consistency of the estimator $E_{\hat{G}}\eta(\theta)$ for $E_G\eta(\theta)$ for any $G \in \Gamma$. However, the convergence may be arbitrarily slow and depends heavily on the probability concentration of G in the neighborhood of 0.

Consider the following example. $\lambda \sim G_\lambda$, $p = p_0 + \delta 1(\lambda < \lambda_0)$ for some λ_0. To get a bound on the rate, suppose we know all parameters except for δ and we are told when $\lambda < \lambda_0$. Then the only sample relevant to the information is the sample of size $O_p\big(nG_\lambda(\lambda_0)\big)$ coming from $\lambda < \lambda_0$. Given the sample, $\sum X_i 1(\lambda < \lambda_0)$ is binomial, hence we can estimate δ with accuracy of $\big(n\lambda_0 G_\lambda(\lambda_0)\big)^{-1/2}$. This error has a contribution of $G_\lambda(\lambda_0)\big(n\lambda_0 G_\lambda(\lambda_0)\big)^{-1/2}$. Since λ_0 can be arbitrarily small $G_\lambda(\lambda)$ can converge to 0 as slow as we want, the estimating error of δ has as slow rate as we want.

3.1.2. Inconsistency in Model (i)

Again, we consider the function $\eta(\theta_i) = p_i$. Now $\theta_i = (\pi_i, p_i)$.

In Model (i) there is no consistency in estimating G for every G, neither a consistency in estimating $E_G(\eta(\theta))$. This is a result of the non-identifiability, and it is demonstrated in the following example for $\kappa_i \equiv 1$.

Example 1. *When $\kappa_i \equiv 1$ there are $M = 3$ possible outcomes of the i'th observation, we list them as: $X_i = 1$, $X_i = 0$ (while $K_i = 1$), and $K_i = 0$; the corresponding probabilities are: $(\pi_i p_i, \pi_i(1 - p_i), (1 - \pi_i))$. Suppose the outcomes of n realization have n_1 occurrences of $X_i = 1$, n_2 occurrences of $X_i = 0$, and n_3 occurrences of $K_i = 0$. Note (n_1, n_2, n_3) is multinomial. Suppose $\frac{1}{n}(n_1, n_2, n_3) = (0.25, 0.25, 0.5)$, obviously the following $\hat{G}_1$ and $\hat{G}_2$ are both GMLE. Let $\hat{G}_1$ be degenerate at $(\pi, p) = (0.5, 0.5)$. Let $\hat{G}_2$ be the distribution whose support is $(0, 1), (1, 0), (1, 1)$ with corresponding probabilities 0.5, 0.25, 0.25. Then, obviously both $\hat{G}_1$ and $\hat{G}_2$ are GMLE, while $E_{\hat{G}_1}\eta(\theta) = 0.5 \neq E_{\hat{G}_2}\eta(\theta) = 0.75$.*

More generally, if $\kappa_i \in \mathcal{K}$, $\mathcal{K}$ a bounded set, then (κ_i, K_i, X_i) may get at most M values, and hence the distribution G is not identified (as long as it is not constraint to have a finite support with cardinality less the above number of possible values). In fact, one can estimate the expectations

$$g_G(k, x) = E_G \pi^k (1 - \pi)^{\kappa_i - k} p^x (1 - p)^{k-x}, \quad 0 \le x \le k \le \kappa_i \in \mathcal{K}. \qquad (3)$$

Note that $\eta(\pi, p) \equiv p$ is not the linear span of the functions under the expecta-

tion in right hand side of (3). Consider now the system of equations:

$$g_{G_0}(k,x) = E_G \pi^k (1-\pi)^{\kappa_i - k} p^x (1-p)^{k-x}, \quad 0 \le x \le k \le \kappa_i \in \mathcal{K}$$
$$\eta = E_G p \tag{4}$$

The set of equations (4) is a linear system of M, say, linearly independent equation in G. In fact, fix any κ_i and k, the span of the subset of functions with κ_i and k is $\pi^k (1-\pi)^{\kappa_i - k} p^i$, $i = 0, 1, \ldots, k$. Similarly $\text{span}\{\pi^k (1-\pi)^{\kappa_i - k} : k = 1, \ldots, \kappa_i\} = \text{span}\{\pi^j : j = 1, \ldots, \kappa_i\}$. Thus, span of the functions on the top line of (4) is span $\{\pi^j p^i : i \le j \le \max \mathcal{K}\}$. In particular, p is not in this span, and hence (4) is a set of linearly independent equations. Suppose, for simplicity, that the true distribution G_0 has M support points. Then, for any η, (4) has as a solution with the same support as G_0. It is a distribution function, since 1 is in the span of the functions in the right hand side of the top line of (4) and hence $\int dG = \int dG_0 = 1$. However, it is not necessarily positive. But, by the inverse function theorem, the solution is continuous in η and hence for any η in some neighborhood of $\eta_0 = E_{G_0} p$ there is a positive solution G which is a probability distribution function.

This argument fails for model (ii), where $\mathcal{K} = \{0, 1, 2, \ldots\}$, since p is now in the closed linear span of the probability functions. However, this is another explanation for the slow potential convergence. We need to observe the rare large values of κ_i to estimate the functional.

3.2. The GMLE and the mean of the empirical Bayes estimates

Given a function $\eta(\theta)$, suppose it is desired to estimate $\eta_G = E_G \eta(\theta)$. Again, for convenience we consider the notations of random parameters $\theta_1, ..., \theta_n$, but, in light of Section 3.3 the results apply also to the setup of fixed $\theta_1, ..., \theta_n$. By the following theorem, the estimator $\hat{\eta}_G = E_{\hat{G}} \eta(\theta)$ equals to the average of $E_{\hat{G}}(\eta(\theta) \mid Y_i)$, $i = 1, ..., n$. The appeal of this fact is that if we estimate the values of the individual $\eta(\theta_i)$ via nonparametric empirical Bayes under squared loss, specifically by $E_{\hat{G}}(\eta(\theta) \mid Y_i)$, there is a consistency and agreement between the estimates of the individual parameters and the estimate of their total, or, of their average. In [26], the problem of estimating random sums involving a latent variable is explored. One approach in [26] is to estimate the random sum, by the sum of the estimated conditional expectations of the summands. This is analogous to estimate $\sum_i \eta(\theta_i)$ by $\sum_i E_{\hat{G}}(\eta(\theta) \mid Y_i)$, the last term equals to $n E_{\hat{G}} \eta(\theta)$ by the following theorem.

Theorem 2. *Assume $\eta(\theta)$ is a bounded function, then*

$$\hat{\eta}_G \equiv E_{\hat{G}} \eta(\theta) = \frac{1}{n} \sum_i E_{\hat{G}}(\eta(\theta) \mid Y_i). \tag{5}$$

Proof. Let

$$d\hat{G}_t(\theta) = \big(1 + t(\eta(\theta) - \hat{\eta}_G)\big) d\hat{G}(\theta).$$

Since $\eta(\theta)$ is bounded, it follows that for $\varepsilon > 0$ small enough, $\mathcal{G} = \{G_t, \ t \in (-\varepsilon, \varepsilon)\}$ is a curve of cdf's (i.e., G_t is a positive measure with total mass 1).

Since $\hat{G}$ is a GMLE, it maximizes the likelihood within $\mathcal{G}$. Hence:

$$
\begin{aligned}
0 &= \frac{1}{n}\frac{d}{dt}\sum_i \log\left(\int f(Y_i \mid \theta) d\hat{G}_t(\theta)\right)\big|_{t=0} \\
&= \frac{1}{n}\sum_i \frac{\int \left(\eta(\theta) - \hat{\eta}_G\right) f(Y_i \mid \theta) d\hat{G}(\theta)}{\int f(Y_i \mid \theta) d\hat{G}(\theta)} = \frac{1}{n}\sum_i \int \eta(\theta) d\hat{G}(\theta \mid Y_i) - \hat{\eta}_G.
\end{aligned}
$$

This concludes the proof of the theorem. □

Theorem 2 represent the GMLE estimator, $\hat{\eta}_G$ as an average of what may look like of almost i.i.d. random variables, $n^{-1}\sum E_G\left(\eta(\theta) \mid Y_i\right)$. The latter is $\sqrt{n}$ consistent and asymptotically normal. However, this is misleading. All terms in the sum in (5) depend on all other through $\hat{G}$ and $E_{\hat{G}}(\eta(\theta) \mid Y_i)$ is possibly a biased estimator for $E_G(\eta(\theta) \mid Y_i)$. In fact, as we argue, $\eta_{\hat{G}}(\theta)$ may be inconsistent or consistent but converges in a very slow rate. On the other hand, in many other models, the GMLE is actually consistent, see the discussion in Section 7.8 of [1]. One such a case is described in Section 3.2.1 below. See further discussion there.

3.2.1. GMLE for the mean of exponential mixtures

Let $Y_i \sim N(\theta_i, 1)$, $i = 1, ..., n$ be independent observations, the obvious estimator for $\sum \theta_i$ is $\sum Y_i$. One may wonder about a comparison between the trivial estimator $\sum Y_i$ and the estimator $nE_{\hat{G}}\theta = \sum E_{\hat{G}}(\theta \mid Y_i)$. The nonparametric empirical Bayes $E_{\hat{G}}(\theta \mid Y_i)$ is strictly better than Y_i as a point-wise estimator of θ_i. c.f., [2, 12, 15]. Similarly one may wonder about the similar model, estimating $\sum \lambda_i$ when $Y_i \sim Poisson(\lambda_i)$. In the following we will show that the two estimators are in fact equal. Thus, GMLE does not improve over the trivial estimator, yet, it does not harm.

In the following theorem the notations are for a one dimensional exponential family, but it applies for a general exponential family.

Theorem 3. *Let $Y_1, ..., Y_n$, be independent observations, $Y_i \sim f_{\theta_i}(y)$. Suppose that the density of Y is of the form $f_\theta(y) \equiv f(y \mid \theta) = \exp(\theta y - \psi(\theta))$, $\theta \in \Omega$, with respect to some dominating measure μ. Let $\hat{G}$ be a GMLE supported on the interior of the parameter set Ω. Let $\eta(\theta) = E_\theta Y$. Then:*

$$
nE_{\hat{G}}\eta(\theta) = \sum Y_i.
$$

Proof. Given the observations $Y_1, ..., Y_n$, let $\hat{G}$ be a GMLE. Define the translation $\hat{G}_\Delta(\,.\,) = \hat{G}(\,.\, - \Delta)$.

Note:

$$
\int \exp(\theta y - \psi(\theta)) d\hat{G}_\Delta(\theta) = \int \exp((\theta - \Delta)y - \psi(\theta - \Delta)) d\hat{G}(\theta).
$$

Since $\hat{G}$ is GMLE,

$$
\begin{aligned}
0 &= \frac{d}{d\Delta} \sum \log(\int \exp(\theta Y_i - \psi(\theta))d\hat{G}_\Delta(\theta)) \mid_{\Delta=0} \\
&= \frac{d}{d\Delta} \sum \log(\int \exp((\theta - \Delta)Y_i - \psi(\theta - \Delta))d\hat{G}(\theta))|_{\Delta=0} \\
&= \sum \frac{\int(-Y_i + \psi'(\theta)) \exp(\theta Y_i - \psi(\theta))d\hat{G}(\theta)}{\int \exp(\theta Y_i - \psi(\theta))d\hat{G}(\theta)} \\
&= -\sum Y_i + \sum E_{\hat{G}}(\psi'(\theta) \mid Y_i) \\
&= -\sum Y_i + nE_{\hat{G}}\eta(\theta)
\end{aligned}
$$

The last equality follows by Theorem 2, and since for exponential family $\eta(\theta) = \psi'(\theta) = E_\theta Y$. □

The above theorem implies that under an exponential family setup, $E_{\hat{G}}\eta(\theta)$ is unique for $\eta(\theta) = E_\theta Y$ even if $\hat{G}$ is not. See the following example.

Example 2. *Let $Y_i \sim Bernoulli(p_i)$ be independent where $p_i \sim G$ are independent, $i = 1, ..., n$. Suppose $Y_1 = ... = Y_{\frac{n}{2}} = 0$ and $Y_{\frac{n}{2}+1} = ... = Y_n = 1$. Then, obviously both $\hat{G}_1$ and $\hat{G}_2$ are GMLE, where $\hat{G}_1$ has half of its mass at $p = 0.1$ and the other half at $p = 0.9$, while $\hat{G}_2$ is degenerate at $p = 0.5$. Obviously $\hat{G}_1 \neq \hat{G}_2$, however, $E_{\hat{G}_1}p = E_{\hat{G}_2}p = 0.5$.*

In the normal model, $Y_i \sim N(\theta_i, 1)$ all finite cumulants of θ can be estimated in the $\sqrt{n}$ rate by the corresponding cumulants of Y. For example, estimate $E_G\theta$ by $n^{-1}\sum Y_i$. That these estimators are equivalent to the GMLE can be argued by a similar argument as in the proof of Theorem 3 by taking further derivatives. However, other functions cannot be estimated in such a rate, e.g., the deconvolution problem of estimating $G(\theta_0) = E_G 1(\theta \leq \theta_0)$ has a logarithmic rate see [7]. Even a very smooth function like $\eta(\theta) = e^{-\alpha^2\theta^2/2}$, $\alpha \geq 1$, cannot be estimated in the parametric rate, see [5]. However, estimating functions of finite cumulants is efficient, as the tangent space is saturated, and any mean is an efficient estimate of its expectations. An example of a less trivial function that can be estimated efficiently at the $\sqrt{n}$ is the density at 0 of $G * N(0, \alpha^{-2})$, $\alpha^2 < 1$, which in our framework is $E_G\eta(\theta)$, where $\eta(\theta) = \sqrt{\alpha^2/2\pi}e^{-\alpha^2\theta^2/2}$. This is also, the density at 0 of $f_G * N(0, (1 - \alpha^2)/\alpha^2)$, and hence can be estimated with the unbiased kernel estimator: $n^{-1}\sqrt{(1 - \alpha^2)/2\pi\alpha^2}\sum e^{-\alpha^2 Y_i/2(1-\alpha^2)}$. See Bickel et al. Section 4.5 for the tangent space and for the argument that it is efficient.

3.3. The compound decision model and weak convergence

The seminal paper of [14], appeals to a completely unknown distribution G, assuming that $\theta_i \sim G$, $i = 1, 2, ...$ are iid. They prove weak convergence of $\hat{G}$ to G. It may be seen in the introduction of that paper, that the authors are somewhat

uncomfortable with this (very weak) assumption. The weak convergence in [14] is explored from a broader perspective in [3].

In this subsection we attempt to present a weak convergence result when $\theta_1, ..., \theta_n$ are considered as fixed unknown parameters that may depend on n. That is, we consider a general triangular array where at stage n the parameters are some $(\theta_1^n, ..., \theta_n^n)$, as also elaborated in the sequel.

Recently there are results about rates of convergence of $\hat{G}$ and other estimators to G. See, e.g., [21] for the multivariate Gaussian case, [10] for finite mixtures. Those papers do not cover the weak convergence for triangular arrays presented in this subsection. In [10] a non-GMLE estimator, $\tilde{G}$ for G, is studied and they provide minmax rates for the Wasserstein distance between $\tilde{G}$ and G, for mixtures with m components. However, those rates do not imply our weak convergence result in the triangular array formulation. Specifically, in our triangular array setup, as the number of observations n increases, the number of components m in the mixture is also increased since we allow $m = n$.

Given any triangular array sequence of sets $\{\theta_1^n, ..., \theta_n^n\}$, $n = 1, 2, ...$, denote by G^n its corresponding empirical distribution. Given independent observations $Y_1^n, ..., Y_n^n$, $Y_i^n \sim f(y \mid \theta_i^n)$, we denote the corresponding GMLE, as defined above, by $\hat{G}^n$. That is, $\hat{G}^n$ is the GMLE ignoring the fact that the observations are not i.i.d. from the mixture. In the sequel we may omit the superscript n and write θ_i, Y_i.

Note that, in our current setup the true joint likelihood $f_G^n(Y_1, ..., Y_n)$ does *not* satisfy $f_G^n(Y_1, ..., Y_n) = \Pi_i f_G(Y_i)$, since Y_i are *not* independent.

Theorem 4. *Assume A1-A5 below, then for any bounded and continuous function f, $\int f d\hat{G}^n - \int f dG^n \to_{a.s.} 0$.*

Assumptions

A1 The parameter space Ω is compact.
   Let
$$h(y) = \sup_\theta |\log(f(y \mid \theta))|,$$

A2 $\sup_{\theta \in \Omega} E_\theta h^4(Y) < \infty$.
A3 The functions $f(y_0 \mid \theta)$ are uniformly bounded and continuous in θ for every y_0.
A4 The class of densities $\{f_G(y)\}$, that correspond to the set of all possible mixtures $\{G\}$, is identifiable.
A5 For every $G_0 \in \{G\}$, every $\epsilon > 0$, and every $M > 0$, the set

$$\{G \mid \|\log(f_G^{(M)}) - \log(f_{G_0}^{(M)})\|_\infty < \epsilon\},$$

   is an open set under the weak convergence topology, where $f_G^{(M)}(Y) = [f_G(Y)]_M$ and $[x]_M = \max(-M, \min(X, M))$.

Condition A1 is used to avoid situations where mass can escape. The theorem fails when $(\theta_1, \ldots, \theta_n) = (n, n + 1/n, \ldots, n + (n-1)/n)$ since we can consider a bounded and continuous f which is not uniform continuous, for example,

$f(\theta) = \sin(\theta^4)$. This may be considered as a technical issue, as the GMLE is translation equivariant, and in this case we do have convergence to uniform, as could be obtained by restricting to a smaller set of test functions, e.g., to bounded functions with a bounded first derivative. More problematic example is with the parameters at sample size n being $(1, 2, \ldots, n)$. In this case the convergence fails even when $f(\theta) = e^{-\theta^2}$. Assumption A1, however, can be replaced by a weaker one that ensures compactness of the measures such as strong moments conditions. We preferred to keep it simple, but strong enough for our models (although it does not cover, for example, a Gaussian mixing distribution).

The purpose of A5 is to ensure, that given an open cover for $\{G\}$, a finite covering subset may be extracted by compactness. The assumption is implied by A1–A4 if, in addition, it is assumed that for every $M > 0$, the functions $\log((f_\theta^{(M)}(y))$, $\theta \in \Omega$ are uniformly continuous.

Proof. By A1 we trivially have tightness, and hence any sequence G^n of measures, has a weakly converging sub-sequence. Our plan is to show that every sub-sequence of $a_n \equiv \int f d\hat{G}^n - \int f dG^n$ has a further sub-sequence that converges to zero w.p.1. The later implies the assertion of the theorem.

Given any sub-sequence of a_n, take a corresponding further sub-sequence n_j, so that $G^{n_j} \Rightarrow_w G^0$ and $\hat{G}^{n_j} \Rightarrow_w \hat{G}^0$, for some $\hat{G}^0$ and G^0.

Denote by G_1^n the joint distribution of $\theta_1, \ldots, \theta_n$ when the parameters are sampled independently from G^n, as if the parameters are sampled with replacement from the set $\{\theta_1, \ldots, \theta_n\}$. We denote by G_2^n the joint distribution of $\theta_1, \ldots, \theta_n$ when the sampling is done without replacement. Note that, for any function ψ,

$$E_{G_1^n} \sum_i \psi(Y_i) = E_{G_2^n} \sum_i \psi(Y_i). \tag{6}$$

We keep the notations that for any function ψ, $E_G \psi(Y)$ is the expectation when $\theta \sim G$ and $Y \mid \theta \sim f(y \mid \theta)$.

By Lemma 1 in the Appendix, under both sequences of joint measures $G_1^{n_j}$ and $G_2^{n_j}$,

$$\frac{1}{n_j} \sum_{i=1}^{n_j} \log(f_{\hat{G}^{n_j}}(Y_i)) \to_{a.s.} E_{G^0} \log(f_{\hat{G}^0}(Y)). \tag{7}$$

Similarly, under both sequences of joint measures $G_2^{n_j}$, and $G_1^{n_j}$

$$\frac{1}{n_j} \sum_{i=1}^{n_j} \log(f_{G^{n_j}}(Y_i)) \to_{a.s.} E_{G^0} \log(f_{G^0}(Y)). \tag{8}$$

By definition of $\hat{G}^{n_j}$ as GMLE, $\frac{1}{n_j} \sum_i \log(f_{\hat{G}^{n_j}}(Y_i)) \geq \frac{1}{n_j} \sum_i \log(f_{G^{n_j}}(Y_i))$,

thus under the sequence of measures $G_2^{n_j}$:

$$
\begin{aligned}
E_{G^0} \log(f_{\hat{G}^0}(Y)) &= \lim \frac{1}{n_j} \sum_{i=1}^{n_j} \log(f_{\hat{G}^{n_j}}(Y_i)) \\
&\geq \lim \frac{1}{n_j} \sum_{i=1}^{n_j} \log(f_{G^{n_j}}(Y_i)) \\
&= E_{G^0} \log(f_{G^0}(Y)),
\end{aligned} \tag{9}
$$

Since $f_{G^0} = \operatorname{argmax}_{f_G} E_{G^0} \log(f_G(Y))$, we obtain that:

$$
E_{G^0} \log(f_{\hat{G}^0}(Y)) \leq E_{G^0} \log(f_{G^0}(Y)). \tag{10}
$$

From the last two inequalities we obtain:

$$
E_{G^0} \log(f_{G^0}(Y)) = E_{G^0} \log(f_{\hat{G}^0}(Y)).
$$

By the identifiability assumption A4, and by concavity of the log function, $\operatorname{argmax}_{\{f_G\}} E_{G^0} \log(f_G(Y))$ is unique, and $\hat{G}^0 = G^0$.
This concludes our proof.   □

Remark. In *compound decision*, where $\theta_1, ..., \theta_n$ are fixed, it is desired to estimate $\eta_i = \eta(\theta_i)$. Then, the goal is to approximate the optimal separable estimator, some times also termed optimal simple symmetric estimator. Under a squared loss, the later estimator for η_i, given an observation Y_i, is $E_{G^n}(\eta(\theta_i) \mid Y_i)$, see, e.g., [25, 2]. In our triangular array setup, our weak convergence result, implies under suitable conditions $[E_{G^n}(\eta(\theta_i) \mid Y_i) - E_{\hat{G}^n}(\eta(\theta_i) \mid Y_i)] \to 0$. A simple such possible additional condition in our triangular array setup, is that $\liminf f_{G^n}(Y_i^n) > 0$.

4. Simulations

In this section, we simulate the estimation of $\eta_G = E_G \eta(\theta)$, where $\eta(\theta_i) = p_i$, under both models (i) and (ii). In all of the following simulations, for convenience, η_G equals 0.5. For any parameter configuration, the number of repetitions is 50.

We compute the GMLE $\hat{G}$ via EM algorithm on a grid. As suggested by [15], we search for a "confined GMLE" where we confine our search to distributions with a finite and specific support. Our choice of a specific support is ad-hoc, see some rigorous treatment in [4]. The search for GMLE among distributions on the specific grid via EM-algorithm, is also the approach in [19], see further elaboration there.

The grids for $(\theta_{i1}, \theta_{i2})$ contain $40 \times 40 = 1600$ equally spaced grid points in a range that fits the relevant problem. The parametrization in Model (ii) is via a two dimensional Poisson, as explained in Section 3.1.1. The EM algorithm started with G uniform on the grid and ran for 1000 iterations.

TABLE 1

Poisson Simulation discrete G. The estimated mean and standard deviations of two estimators.

Support points of $(\lambda, p))$	The naive estimator	The GMLE
$(2, 0.4), (1, 0.6)$	**0.486**, (0.014)	**0.503**, (0.020)
$(2, 0.2), (1, 0.8)$	**0.453**, (0.015)	**0.496**, (0.018)
$(2, 0.2), (0.5, 0.8)$	**0.385**, (0.013)	**0.505**, (0.022)

TABLE 2

Poisson simulation with continuous G. The distribution of λ is $U(0.5, 1)$, for the first 500 strata, and $U(0.5, 2)$ for the remaining strata. The binary probabilities p^I corresponding to the first 500 strata are given the table. The binary probabilities of in the rest of the strata are $1 - p^I$. The table gives estimates of the mean and standard deviation of the two estimators.

p^I	Naive	GMLE
0.4	**0.513**, (0.015)	**0.500**, (0.023)
0.3	**0.529**, (0.017)	**0.501**, (0.027)
0.2	**0.538**, (0.016)	**0.491**, (0.027)

4.1. Poisson sample sizes

Table 1 presents the results of Model (ii) with 2-points support distribution G. There are 500 strata corresponding to each of the two points in the support of G. The table presents the mean and the sample standard deviation of the naive and the GMLE estimators for η_G.

In Table 2 we report on one such set of simulations. In the three cases presented in Table 2 there are again 500 strata of two types. The probability p^I of first 500 strata is fixed while for the rest of the strata the probability is $p^{II} = (1 - p^I)$. The Poisson parameter λ is continuous in all the three cases and is chosen from a uniform distribution, $U(0.5, 1)$ in the first 500 strata and from $U(0.5, 2)$ in each of the remaining 500 strata.

4.2. Binomial sample sizes

We next study Model (i), where K_i, the realized sample size from stratum i, is distributed $B(\kappa_i, \pi_i)$. Again, our simulated populations have two types of strata, 500 of each type. In the simulations reported in Table 3, $\kappa_i = 4$ for all of the 1000 strata, and $\eta_G = 0.5$ throughout the three simulations summarized in the table. In 500 strata $\pi_i = p_i = 0.5 - \delta$ while in the other 500 strata, $\pi_i = p_i = 0.5 + \delta$.

The final reported simulations are summarized in Table 4. We study Binomial sampling with various values of κ, fixed at $\kappa = 1, \ldots, 5$. Again, there are two types of strata, 500 strata for each of the two types; for 500 strata p_i and π_i are sampled independently from $U(0.1, 0.6)$, while for the rest of the strata they are i.i.d. from $U(0.4, 0.9)$.

It is surprising how well the GMLE is doing already for $\kappa = 2, 3$, in spite of the non-identifiability of G and the inconsistency of the GMLE.

TABLE 3
The mean and standard deviation of two estimators. Binomial Simulation. $\kappa \equiv 4$ and $\pi_i = p_i = 0.5 \pm \delta$.

δ	Naive	GMLE
0.3	**0.559**, (0.012)	**0.502**, (0.014)
0.2	**0.522**, (0.011)	**0.504**, (0.012)
0.1	**0.504**, (0.010)	**0.501**, (0.010)

TABLE 4
Binomial Simulations with continuous G. κ=1,2,3,4,5.

κ	Naive	GMLE
1	**0.544**, (0.019)	**0.530**, (0.015)
2	**0.528**, (0.014)	**0.502**, (0.021)
3	**0.522**, (0.014)	**0.498**, (0.022)
4	**0.517**, (0.012)	**0.499**, (0.020)
5	**0.512**, (0.009)	**0.501**, (0.013)

5. Confidence Interval for η_G

Although the GMLE $\hat{\eta}_G$ is an appealing estimator, we do not know how good is its performance, beyond the consistency results we established. This is especially under non-identifiability where consistency is not implied. In the following we suggest an asymptotically level-$(1-\alpha)$ conservative confidence interval for $\eta_G = E_G \eta(\theta)$. We only elaborate on defining the corresponding convex optimization problem. Implementing and solving the sketched convex optimization problem, could be challenging.

Let $Z = Z(Y)$ be a random variable with M possible outcomes, z_j, $j = 1, ..., M$. One may think of M "chosen cells" in a goodness of fit test. The considerations for the choice and for the number of cells is beyond the scope of this section, in particular $M(n) \equiv M$ is fixed. Then the densities $f_Y(y \mid \theta)$ induce densities $f_Z(z \mid \theta)$, $\theta \in \Omega$. Denote

$$p_G^j = P_G(Z = z_j) = \int f_Z(z \mid \vartheta) dG(\vartheta);$$

denote $\hat{p}^j = \frac{\{\#i|Z_i=z_j\}}{n} \equiv \frac{n_j}{n}$, $j = 1, ..., M$, where n_j is implicitly defined. Recall that:

$$2(\sum_j n_j \log(\hat{p}^j) - \sum_j n_j \log(p_G^j)) \Rightarrow_G \chi^2_{(M-1)}.$$

Let $\chi^2_{(M-1),1-\alpha}$, be the $(1-\alpha)$ quantile of a $\chi^2_{(M-1)}$ distribution. Define

$$\Gamma_\alpha = \{G \ : \ \sum_j n_j \log(p_G^j) \geq \sum_j n_j \log(\hat{p}^j) - \frac{1}{2}\chi^2_{(M-1),1-\alpha}\}.$$

Observe that Γ_α is a convex set of distributions.

The functional $\eta(G) = E_G \eta(\theta)$, is a linear functional.

Let

$$\eta^L = \min_{G \in \Gamma_\alpha} \eta(G); \; \eta^U = \max_{G \in \Gamma_\alpha} \eta(G).$$

By the above, finding (η^L, η^U) is a convex problem, and in addition the later interval is a $(1 - \alpha)$ conservative confidence interval for $E_G \eta(\theta)$.

The above set Γ_α is in the spirit of the F-localiztion in [11]. While our suggestion for the choice of $Z(Y)$ is ad-hoc, as often done in goodness of fit tests, they carefully elaborate on optimal choices of 'affine estimators'.

It may be shown that the above is also a conservative $(1 - \alpha)$-level CI for $\sum \eta(\theta_i)$ for fixed $\theta_1, ..., \theta_n$, and also a $(1 - \alpha)$-level conservative credible set for $\sum \eta(\theta_i)$ when $\theta_1, ..., \theta_n$, are independent realizations $\theta_i \sim G$.

6. Real data example

Our example is based on data from the Social-Survey conducted yearly by Israel Census Bureau (ICB). A random sample representing a $1/1000$ fraction of the age 20 or older individuals in the registry is drawn. The home addresses of those in the sampled are verified and they are interviewed in person.

We study here the data accumulated for Tel-Aviv, in the surveys collected during 2015–2017. The total sample size in those three years is 1256. There are 156 'statistical-areas' in Tel-Aviv, very roughly of equal size, about 3000 individuals in each. This means that around three individuals are sample from each statistical area. Statistical-areas are considered homogeneous in many respects, and we take them as our strata.

Let K_i be the sample size in stratum i, $i = 1, ...156$, then our data satisfy $K_i > 0$, $i = 1, ..., 156$. (Admittedly, we neglected a few small statistical-areas that actually had zero sample sizes). In our analysis, p_i is the proportion of individuals in stratum i that own their living place (or, it is owned by a member of their household). The goal is to estimate $n^{-1} \sum p_i$, roughly the proportion of individuals that own their living-place.

The naive estimator is applicable since $K_i > 0$, $i = 1, ..., 156$. The estimated proportion obtained by the naive estimator is:

$$\frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{K_i} = \mathbf{0.434}.$$

On the other hand, the 'extreme collapse' estimator, satisfy:

$$\frac{\sum X_i}{\sum K_i} = \mathbf{0.488}.$$

The significant difference between the two estimates has to do *also* with the fact that strata are, in fact, not of equal size. But, more importantly for us, the 'extreme collapse' seems to over estimate the proportion, since owners are over represented in the sample. One reason is that their address in the registry

TABLE 5
Home ownership in Tel-Aviv. Performance of the two estimators as function of sampling rate.

γ	Naive	GMLE
0.1	0.471	0.457
0.2	0.467	0.443
0.25	0.447	0.434

is more accurate and thus it is easier to find them. In other words, individuals are not MCAR (missing completely at random). This phenomena is partially corrected by the stratification, when MAR (missing at random) conditional on the fine strata is approximately right.

In Table 5 we compare the GMLE and the naive estimators in semi-real simulated scenarios in which only a randomly sub-sampled with probability γ, of the described sample, is retained. The average estimates based on 25 simulations applied on the real data with $\gamma = 0.1, 0.2, 0.25$, are presented in Table 5. The (random) number of simulated strata with zero sample sizes, corresponding to $\gamma = 0.1, 0.2, 0.25$, are around 70, 40, and 30, correspondingly.

It is reasonable to assume that the number of observations in the strata are Poisson random variables. Thus, the sample sizes K_i in the simulated sub-sample are $Poisson(\lambda_i)$ too, $i = 1, ..., 156$. The naive estimate based on the entire data equals **0.434**, which is a reasonable benchmark, since that based on the entire data we have no empty strata.

In the semi-real simulations the GMLE seems to perform much better than the naive estimator.

7. Appendix

Lemma 1. *Assume the setup of Section 3.3, Assumptions A1–A5, and the notion of the proof of Theorem 4. Then, under both $G_1^{n_j}$ and $G_2^{n_j}$,*
  i)

$$\frac{1}{n_j} \sum_{i=1}^{n_j} \log(f_{\hat{G}^{n_j}}(Y_i)) \to_{a.s.} E_{G^0} \log(f_{\hat{G}^0}(Y)). \tag{11}$$

  ii)

$$\frac{1}{n_j} \sum_{i=1}^{n_j} \log(f_{G^{n_j}}(Y_i)) \to_{a.s.} E_{G^0} \log(f_{G^0}(Y)). \tag{12}$$

Proof. We prove part i).
  First we show that:

$$\lim E_{G^{n_j}} \log(f_{\hat{G}^{n_j}}(Y)) = E_{G^0} \log(f_{\hat{G}^0}(Y)). \tag{13}$$

Note that by A2 for every $\epsilon > 0$, there exists large enough $m > 0$, such that for every $G', G'' \in \{G\}$,

$$\left| E_{G'} \log f_{G''}(Y) \times I(h(Y) \le m) - E_{G'} \log f_{G''}(Y) \right| < \epsilon,$$

here I is an indicator function.

Hence, (13) is implied by the following,

$$\lim_{n_j} E_{G^{n_j}} \log(f_{\hat{G}^{n_j}}(Y)) \times I(h(Y) \le m)$$

$$= \lim_{n_j} E_{G^0} \log(f_{\hat{G}^{n_j}}(Y)) \times I(h(Y) \le m) \times \frac{f_{G^{n_j}}(Y)}{f_{G^0}(Y)}$$

$$= E_{G^0} \lim_{n_j} \log(f_{\hat{G}^{n_j}}(Y)) \times I(h(Y) \le m) \times \frac{f_{G^{n_j}}(Y)}{f_{G^0}(Y)}$$

$$= E_{G^0} \log(f_{\hat{G}^0}(Y)) \times I(h(Y) \le m).$$

The above is by Lebesgue dominating convergence theorem, applying A2. Note that, the likelihood ratio $\frac{f_{G^{n_j}}(Y)}{f_{G^0}(Y)}$ is bounded when $h(Y) \le m$. Equation (13) follows by letting $m \to \infty$ and utilizing A2.

Under $G_1^{n_j}$, the proof will follow, by Borel Cantelli, if we show that for every $\epsilon > 0$:

$$\sum_{n_j} P_{G_1^{n_j}} \left(\left| \frac{1}{n_j} \sum_{i=1}^{n_j} \log(f_{\hat{G}^{n_j}}(Y_i)) - E_{G^{n_j}} \log(f_{\hat{G}^{n_j}}(Y)) \right| > \epsilon \right) < \infty. \qquad (14)$$

The complication in the above is that in the terms $f_{\hat{G}^{n_j}}(Y_i)$, the random function $f_{\hat{G}^{n_j}}$ depends on its argument Y_i. We will prove (14) through the following steps.

In fact, the convergence in equation (14) is of a sub-series n_j, but also the original series converges. Indeed, in the sequel we consider the original series with a general index n. Let $\log(\Psi) \in \{\log(f_G), G \in \{G\}\}$, be a *fixed* function, in particular, independent of the observations. Then for every $\epsilon > 0$

$$P_{G_1^n} \left(\left| \frac{1}{n} \sum_{i=1}^{n} \log(\Psi(Y_i)) - E_{G^n} \log(\Psi(Y)) \right| > \epsilon \right) < \kappa/n^2,$$

for a suitable large enough κ. The above is by is by Markov inequality, utilizing the bounded fourth moment assumption A2, and the fact that under G_1^n (sampling with replacement), Y_i, $i = 1, ..., n$, are independent.

Note, that by tightness, $\{G\}$ is compact under the weak convergence topology, since that every sequence G^k has a converging sub-sequence. For every $\epsilon > 0$ and $M > 0$, by A5 and compactness of $\{G\}$, we may find a finite cover of L open sets, and corresponding functions $\Psi_1, ..., \Psi_L$, $\Psi_i \in \{f_G\}$, such that for every $f_0 \in \{f_G\}$

$$\inf_{\Psi \in \{\Psi_1, ..., \Psi_L\}} \| \log(f_0^{(M)}) - \log(\Psi^{(M)}) \|_\infty < \epsilon. \qquad (15)$$

Next, since L is finite, the above considerations coupled with Bonferroni, imply that:

$$P_{G_1^n} \left(\max_{\Psi \in \{\Psi_1, ..., \Psi_L\}} \left| \frac{1}{n} \sum_{i=1}^{n} \log(\Psi(Y_i))p - E_{G^n} \log(\Psi(Y)) \right| > \epsilon \right) < \kappa/n^2, \qquad (16)$$

for a suitable large enough κ.

Now we show that in Equation (14), for large enough M, we may neglect the terms where $|\log(f_{G^{n_j}}(Y_i))| > M$. For every $\epsilon > 0$, for large enough M, a.s.,

$$\limsup \frac{1}{n} \sum_{i:\ |\log(f_{\hat{G}^n}(Y_i))| \geq M} |\log(f_{\hat{G}^n}(Y_i))| < \epsilon.$$

This follows since $|\log(f_{\hat{G}^n}(Y_i))| \leq h(Y_i)$, and by A2, coupled with Borel Cantelli. Hence, in the following, considering large M, we may neglect

$$\frac{1}{n} \sum_{i:\ |\log(f_{\hat{G}^n}(Y_i))| \geq M} |\log(f_{\hat{G}^n}(Y_i))|.$$

By (15) we may conclude, by taking large enough M, that for every $\epsilon > 0$,

$$P_{G_1^n}\Big(\sup_{\Psi \in \{\ f_G\}} \Big| \frac{1}{n} \sum_{i=1}^n \log(\Psi(Y_i)) - E_{G^n} \log(\Psi(Y)) \Big| > \epsilon \Big) < \kappa/n^2, \qquad (17)$$

for a suitable large enough κ.

Equation (14) follows by the last equation and the above.

The Lemma is now implied by Borel Cantelli, for the case $G_1^{n_j}$, of sampling with replacement.

The convergence under $G_2^{n_j}$ is obtained by comparing sampling with replacement to sampling without replacement. Let $S_{n_j} = \sum_{i=1}^{n_j} \log(\Psi(Y_i))$. Then, $E_{G_1^{n_j}} S_{n_j}^4 \geq E_{G_2^{n_j}} S_{n_j}^4$. It may be seen by applying Rao-Blackwell on the convex function $g(S_{n_j}) = S_{n_j}^4$, when conditioning on $k_i,\ i = 1, ..., n_j$, where k_i is the number of times θ_i was sampled in the sampling with replacement process, that defines $G_1^{n_j}$. The implied bounded 4'th moment of S_{n_j} under $G_2^{n_j}$, coupled with (6), implies (14), along the lines of the above, under $G_2^{n_j}$. This concludes our proof. □

References

[1] P.J. Bickel, C. A. Klassen, Y.Ritov and J.A. Wellner (1998). *Efficient and Adaptive Estimation For Semiparametric Models.* Springer Verlag NY. MR1623559

[2] 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**, No 4, 1685–1704. MR2533468

[3] Chen, J. (2017). "Consistency of the MLE under Mixture Models." *Statist. Sci.* **32** (1) 47–63. MR3634306

[4] Dicker, L. and Zhao, S. (2014). Nonparametric empirical Bayes and maximum likelihood estimation for high-dimensional data analysis. ArXiv preprint arXiv:1 MR3465819

[5] Donoho, D. L. and Low, M. G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Stat.* **20**, No 2, 944–970. MR1165601

[6] B. Efron. (2014). Two modeling strategies for empirical Bayes estimation. Statistical science: a review. *Journal of the Institute of Mathematical Statistics*, **29**:285. MR3264543

[7] Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Stat.* **19** No.3, 1257–1272. MR1126324

[8] Greenshtein, E. and Itskov, T (2018), Application of Non-Parametric Empirical Bayes to Treatment of Non-Response. *Statistica Sinica* **28**, 2189–2208. MR3839010

[9] Gu, J. and Koenker, R. (2017). Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective. *Journal of Business and Economics Statistics.* **35** MR3591533

[10] Heinrich Philippe and Jonas Kahn(2018). "Strong identifiability and optimal minimax rates for finite mixture estimation." *Annals of Statistics* **46.6A** 2844–2870. MR3851757

[11] Ignatiadis Nikolaos and Wager Stefan (2021). Confidence Intervals for Nonparametric Empirical Bayes Analysis. *JASA*, to appear. MR4480697

[12] Jiang, W. and Zhang, C.-H., (2009), General maximum likelihood empirical Bayes estimation of normal means. *Ann.Stat..* **37**, 1647–1684. MR2533467

[13] Karlis, D. and Xekalaki, E. (2005), Mixed Poisson Distributions. *International Statistical Review*, **73**, 1, 35–58.

[14] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann.Math.Stat.* **27**, 887–906. MR0086464

[15] Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions and empirical Bayes rules. *JASA* **109**, 674–685. MR3223742

[16] Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *JASA* **78**, 805–811. MR0521328

[17] Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications.* IMS, Hayward.

[18] Little, R.J.A and Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* Wiley, New York MR1925014

[19] Long Feng and Lee, H. Dicker (2018). Approximate nonparametric maximum likelihood for mixture models: A convex optimization approach to fitting arbitrary multivariate mixing distributions. *Computational Statistics & Data Analysis* **122** 80–91. MR3765816

[20] J.Pfanzagl (1993). Incidental Versus Random Nuisance Parameters. *Ann Stat*, **21** 1663–1691. MR1245763

[21] Sujayam Saha. Adityanand Guntuboyina (2020) "On the nonparametric maximum likelihood estimator for Gaussian location mixture densities with application to Gaussian denoising." *Ann. Statist.* **48** 738–762 MR4102674

[22] Robbins, H. (1951). Asymptotically subminimax solutions of compound decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1950 131–148. Univ. California,

Berkeley. MR0044803

[23] ROBBINS, H. (1956). An empirical Bayes approach to statistics. *In Proc. Third Berkeley Symp.* 157–164. Univ. California Press, Berkeley. MR0084919

[24] ROBBINS, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.* **35** 1–20. MR0163407

[25] ZHANG, C.-H. (1997). Empirical Bayes and compound estimation of a normal mean. Statist. *Sinica* **7** 181–193. MR1441153

[26] ZHANG, C-H. (2005). Estimation of sums of random variables: Examples and information bounds. *Ann. Stat.* **33**, No.5. 2022–2041. MR2211078