# Distribution-agnostic Linear Unbiased Estimation with Saturated Weights for Heterogeneous Data

Francesco Grassi and Angelo Coluccia, *Senior Member, IEEE*

***Abstract*—The challenging problem of distribution-agnostic linear (weighted) unbiased estimation of a global parameter from heterogeneous and unbalanced data is addressed. This setup may originate in different signal processing contexts involving the joint processing of non-homogeneous groups of data whose statistical distribution is unknown, with (possibly highly) diverse sample sizes. Since sample estimators of the local variances are inaccurate in the low-sample regime, suitable weighting schemes are required. For this problem, we study a family of estimators based on the idea of trimmed weights, i.e., proportional to the sample size but with a proper saturation. Such an approach is theoretically analyzed, showing that it can be linked to the Maximum Entropy principle under uncertainty on the data generative model (as well as to a broader class of cost functions). Different criteria for setting the "cut-off" threshold between the linear and saturated regions are analyzed, also obtaining a reduced-complexity approximation of the optimal minimum-variance estimator for a generalized mixed-effect model. To this aim, a further contribution is that several estimators of an hyperparameter are derived and analyzed. The proposed approach is analyzed theoretically and its performance are assessed against state-of-the-art estimators. An illustrative application to real-world COVID-19 data is also finally developed.**

***Index Terms*—linear unbiased estimator, robust estimation, unbalanced sample size, heteroscedasticity, trimmed weights**

## I. Introduction and Motivations

THE estimation of an unknown parameter from a set of noisy observations is a recurrent problem in signal processing. A classical instance is the estimation of the DC component of a signal embedded in (homogeneous) white noise, typically modeled as Gaussian [1]. Several examples are found in multi-sensor contexts, e.g. wireless sensor networks (WSN), where the goal is to estimate a common parameter by fusing data from different sources (sensors) [2], [3]. Assumptions about the noise model are often introduced for more specific applications: besides the widely-adopted Gaussian assumption, non-Gaussian models can be adopted, e.g., a distribution with positive support is of interest in "location estimation" problems involving arrival times of radio or sound waves [4]. However, heterogeneous (also called non-uniform) noise, in which local variances can be different, is frequently found in practice [5], [6], as noise levels realistically vary across data samples in several signal processing problems. Besides WSN, other examples include direction of arrival estimation [7], [8], spectrum sensing for cognitive radio [9], and time-delay multistatic target localization [10].

In general, data heterogeneity may arise from both heterogeneous noise in each source or group of data, due to local causes [11], [12] (including the presence of outliers [13]), as well as unbalanced sample sizes. Robustness to "wild" variability of sample sizes among data groups, a different issue compared to the classical definition of robustness to data distributional assumptions [14] and other forms of mismatches (due to e.g., non-idealities, errors, or missing data [15]–[17]), is of significant practical interest.

A particularly challenging situation occurs when unbalanced sample sizes — with possibly wild variability across groups and a significant number of low-sample groups[1] — and data heterogeneity combine with lack of knowledge about the statistical distribution of the data. This paper addresses the problem of linear unbiased estimation in such a setup, for which, thus, distribution-agnostic estimators are needed.[2] Indeed, maximum likelihood (ML) estimation theory (as well as Bayesian MAP/MMSE) is inapplicable when the probability distribution of the noise, and thus the data, is unknown. Relevant state-of-the-art solutions typically rely on the least squares (LS) paradigm under different assumptions on the data model, with varying levels of heterogeneity ranging from the purely homogeneous case of the Grand Mean (GM) [21] to the best linear unbiased estimator (BLUE) [1], [5], [6], analysis-of-variance (ANOVA) [21], and minimum-variance linear unbiased estimator (MVLUE) [22] for generalized *random-effect model*, which will be all reviewed in Sec. II. The original contribution of this work is instead to investigate an alternative estimation approach that is generally-applicable yet simple and does not introduce specific hypotheses, so that it can cope with uncertainty about the data model.

### A. Statement of contributions

More specifically, we consider a family of estimators which we refer to as linear unbiased estimators with saturated sample-size based weights (LUE-S). This family exhibits a weight profile linear in the sample size but with a saturation, depending on a single parameter — the "knee-point" of the resulting piecewise-linear profile, or equivalently the "cut-off" saturation level. This is motivated by several arguments, as specified in Sec. I-B. We provide the following contributions:

*i)* We study the LUE-S family for the problem of estimating a scalar deterministic parameter from unbalanced, possibly

F. Grassi and A. Coluccia are with the Dipartimento di Ingegneria dell'Innovazione, Università del Salento, via Monteroni, 73100 Lecce, Italy. Phone: +39 0832 297 206. E-mail: angelo.coluccia@unisalento.it.

[1] In other words, severely unbalanced data groups may mostly be low-sample but for a few much larger ones, which thus qualify as outliers with respect to the sample size distribution — i.e., the latter may be heavy-tailed.

[2] Early work addressed instead the problem by assuming specific statistical models suitable for particular application contexts, e.g. [18]–[20].

low-sample heterogeneous data; theoretical contributions include derivation of the probability distribution of the weights, study of the minimum-variance setting of the saturation parameter, investigation of relationships with other estimators, and framing within the Maximum Entropy rationale.

*ii)* We derive suitable estimators for the hyperparameter needed in the state-of-the-art estimator MVLUE, of which the proposed saturated approach can be seen as an approximation for a certain (data-dependent) choice of the saturation level; the resulting estimator is easily implementable, making it particularly suitable for practical use since, compared to the MVLUE, it has lower computational and conceptual complexity.

*iii)* We provide an insightful performance assessment of the proposed approach supported by simulations, showing that the proposed approach can outperform state-of-the-art estimators; an illustrative application to COVID-19 mortality rate estimation is also provided using a real-world dataset.

### *B. Specific motivations and related work*

The idea of saturating large weights is heuristically adopted in practice, often under the term “trimmed weights” or “weight trimming”, in contexts where there is uncertainty on the weight profile. For instance, in estimation with stratified populations, samples are weighted inversely by the probability of inclusion, e.g., the classical Horvitz-Thompson estimator [23] or other inverse probability weightings [24], to account for over- or under-sampling in representing the sampled population; however, if the resulting weights are highly dispersed, the final population mean or linear regression estimates may exhibit a large variance [25]–[27]. Weight trimming or winsorizing is often introduced to reduce the weight variability and hence improve the efficiency of survey estimates [28]–[30].[3] The same applies to other statistical settings, such as propensity score methods in which misspecified models may cause extreme weights: a possible solution is again to reduce their impact through trimming [34]–[36]. We consider this idea for the signal processing problem addressed in the paper, which has different peculiarities compared to survey scenarios.

There are further theoretical reasons for studying linear unbiased estimation based on saturated weights, that is the LUE-S family. Winsorizing and trimming are known for their capability to promote robustness to outliers in the observations in a distribution-agnostic way. The rationale of this work is to apply the same idea, which is somewhat reminiscent of the L-estimation approach[4] to sample-size based weights rather than on data values (observations), so leading to a weight profile proportional to the sample size but with a proper saturation.

It is also worth noting that the resulting two-region piecewise-linear profile (linearly increasing until a cut-off value, then constant at the saturation level) shares, up to mirror transformations, the basic non-linear shape adopted in signal processing and statistical learning to reduce the impact of a range of values. Examples are the popular rectified linear unit (ReLU) activation function used in neural networks [33] and the hinge loss adopted for maximum-margin classification, in particular in support vector machines (SVM) [33]. These and other functions indeed share with trimmed weights the functional form involving a maximum or minimum between the independent variable and a chosen value (ref. eq. (13)).

Furthermore, one of the contributions of the present paper is to provide a principled theoretical justification for saturated weights, which can be found in the Maximum Entropy formulation of the distribution-agnostic estimation problem — and actually in a broader class of cost functions (ref. Sec. III-A).

### *C. Notation*

Boldface letters are used for vectors, with $^\intercal$ indicating transpose. $\mathbb{R}^n$ denotes real-valued $n$-dimensional vectors, whereas $\mathbb{R}_+$ (respectively, $\mathbb{R}_{++}$) is the set of non-negative (positive) real numbers. $\|\boldsymbol{x}\|_2$ is the $\ell_2$ (Euclidean) norm of $\boldsymbol{x}$. $\mathbb{1}_{\{\cdot\}}$ is the indicator function. $\mathrm{E}[\cdot]$ and $\mathrm{VAR}[\cdot]$ denote statistical expectation and variance of a random variable, respectively. $X \sim D(p)$ indicates that the random variable $X$ has distribution $D$ with parameters $p$. Finally, $\min(x, y)$ (respectively, $\max(x, y)$) denotes the minimum (maximum) between scalars $x$ and $y$, whereas $\min(\boldsymbol{x})$ (respectively, $\max(\boldsymbol{x})$) is the minimum (maximum) among the elements of $\boldsymbol{x}$.

## II. Problem formulation, background and state-of-the-art approaches

In this section we formulate the problem and review different state-of-the-art estimation approaches that assume increasing levels of heterogeneity.

The addressed problem can be abstractly formalized by considering $N$ data sources, hereafter referred to as groups, each with a variable number $n_i \geq 1$ of independent measurements $x_{i,m} \in \mathbb{R}$, $m = 1, \ldots, n_i$, $i = 1, \ldots, N$. The goal is to estimate a common (global) deterministic parameter $\mu$ based on the independent but *non-identically* distributed local means

$$\hat{\theta}_i \overset{\text{def}}{=} \frac{1}{n_i} \sum_{m=1}^{n_i} x_{i,m} \tag{1}$$

$$x_{i,m} = \mu + e_{i,m} \tag{2}$$

through an unbiased linear (weighted) estimator, i.e.

$$\hat{\mu} = \sum_{i=1}^{N} w_i \hat{\theta}_i = \boldsymbol{w}^\intercal \hat{\boldsymbol{\theta}} \tag{3}$$

with $e_{i,m}$ independent noise terms, possibly heterogeneous across groups $i = 1, \ldots, N$. The sample-size vector $\boldsymbol{n} = [n_1 \cdots n_N]^\intercal$ is given in input to the problem[5] and $\hat{\boldsymbol{\theta}} =$

[3] Variance reduction comes at the expense of introducing bias, but if the variance reduction is larger than the squared bias increase (as a result of a suitable choice of the tuning parameter) then the net result is an overall decrease in mean square error (MSE) [31], [32] (the same motivation supports the use of $\ell_2$-norm regularization in machine learning, see [33, Sec. 3.8]).

[4] L-estimators are linear combinations of order statistics [37], hence stem for their simplicity and robustness against outliers while being distribution-agnostic [38]. They have found application in many signal processing problems, e.g. [39], [40]. Popular L-estimators (besides trivial single-point cases involving percentiles, e.g., median, maximum/minimum) are the truncated or $\alpha$-trimmed mean, where $\alpha$% of the largest values are excluded, and the winsorized mean, where conversely largest values are clipped to a constant value [41], as we propose here for the sample size values in the weight profile.

[5] In fact, in many scenarios it is not possible to obtain an arbitrary amount of data, i.e., $\boldsymbol{n}$ is known but values cannot be chosen. For completeness, we mention that there might exist situations in which only the local means $\hat{\theta}_i$s (1) are provided, while the $n_i$s are not available or may be (severely) quantized.

$[\hat{\theta}_1 \ \cdots \ \hat{\theta}_N]^\mathsf{T}$ are the observed data, whose heterogeneity thus comes from both the (possibly wildly) different sample sizes $\boldsymbol{n}$ and the non-identical distribution of the noise terms (as also discussed in Sec. I), while $\boldsymbol{w} \in \mathbb{R}_+^N$ are normalized non-negative weights to be determined.

Several choices of weights $\boldsymbol{w}$ are possible, leading to different properties of the corresponding estimators. Since a key aspect of the problem setup is that the statistical distribution (pdf) $f_{X_i}$ of the measurements $x_{i,m}$ is unknown, (weighted) LS approaches are typically adopted in the literature for this type of problems, as mentioned in Sec. I and reviewed below under different assumptions on the heterogeneous noise. Only essential aspects are highlighted here, while more details can be found in Appendix A.

### A. Homogeneous model

In general, the LS approach requires that at least the first two moments of $f_{X_i}$ exist. If all measurements are assumed to be homogeneously affected by independent white noise (ref. Sec. A-A for details), the optimal estimator in the LS sense (optimization problem (A.1)) is the Grand Mean (GM) [21]

$$\hat{\mu}^{\text{GM}} \stackrel{\text{def}}{=} \sum_{i=1}^{N} w_i^{\text{GM}} \hat{\theta}_i = \frac{\sum_{i=1}^{N} n_i \hat{\theta}_i}{\sum_{j=1}^{N} n_j}. \tag{4}$$

A different estimator is instead obtained under the assumption of homogeneity on local averages rather than on raw measurements (ref. Sec. A-B for details); in that case the LS estimator (optimization problem (A.2)) is the so-called Mean of Group Means (MGM) [21] with constant weights $w_i^{\text{MGM}} \stackrel{\text{def}}{=} 1/N$:

$$\hat{\mu}^{\text{MGM}} \stackrel{\text{def}}{=} \sum_{i=1}^{N} w_i^{\text{MGM}} \hat{\theta}_i = \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}_i. \tag{5}$$

### B. Non-homogeneous model

To cope with the more general non-homogeneous case, assuming each group of data has *known* variance $\bar{\sigma}_i^2$ (ref. Sec. A-C for details), the optimal estimator (optimization problem (A.4)) is the best linear unbiased estimator (BLUE) [1]

$$\hat{\mu}^{\text{BLUE}} = \frac{\sum_{i=1}^{N} \frac{n_i}{\bar{\sigma}_i^2} \hat{\theta}_i}{\sum_{i=1}^{N} \frac{n_i}{\bar{\sigma}_i^2}} \tag{6}$$

as in fact $\text{VAR}[\hat{\theta}_i] = \bar{\sigma}_i^2 / n_i$. Since group variances are often unknown in practice, sample estimators such as

$$\hat{\bar{\sigma}}_i^2 = \frac{1}{n_i} \sum_{m=1}^{n_i} (x_{i,m} - \hat{\theta}_i)^2 \tag{7}$$

are typically used in place of the true $\bar{\sigma}_i^2$, though estimator (6) with (7) plugged in is no longer the BLUE.

### C. Non-homogeneous model with random-effect

The non-homogeneous model can be further generalized as the so-called *random-effect model*. This is a well-established way to take into account heterogeneous conditions in the data [21], and consists in regarding the heterogeneous noise as the sum of two stochastic components, i.e., homogeneous noise plus a random term that represents the local contribution to non-homogeneity (ref. Sec. A-D for details). This results in $\text{VAR}[\hat{\theta}_i] = \sigma_\theta^2 + \sigma_i^2 / n_i$ to be used in place of the BLUE's $\bar{\sigma}_i^2 / n_i$ to inversely weight the $\hat{\theta}_i$s. The implementation requires however knowledge of the group variances $\sigma_i^2$s as well as of the additional variance $\sigma_\theta^2$ accounting for random effects; again, sample estimators are often used in practice instead of the true variances, which leads to the well-known analysis-of-variance (ANOVA) approach [21]:

$$\hat{\mu}^{\text{ANOVA}} = \frac{\sum_{i=1}^{N} \frac{n_i}{n_i \hat{\sigma}_\theta^2 + \hat{\sigma}_i^2} \hat{\theta}_i}{\sum_{i=1}^{N} \frac{n_i}{n_i \hat{\sigma}_\theta^2 + \hat{\sigma}_i^2}}. \tag{8}$$

A limitation of this estimator is that it requires large values of the sample sizes $n_i$s to properly estimate the unknown variances. To cope with the challenging scenario where $n_i$ can be very small and with wild variability (heavy-tailed distribution), for which the ANOVA approach is very inaccurate, in [22] not only the $\theta_i$s but also the $\sigma_i^2$s are modeled as random variables, instead of deterministic parameters (ref. Sec. A-E for details). In doing so, the classical random-effect model is generalized, and the issue with the inaccuracy of sample variance estimates overcome. Assuming existence of the mean of the (unknown) distribution of $\sigma_i^2$s, and denoting it by

$$\mathcal{E} \stackrel{\text{def}}{=} \text{E}[\sigma_i^2] \in \mathbb{R}_+, \tag{9}$$

the minimum-variance linear unbiased estimator (MVLUE, optimization problem (A.14)) is obtained as [22]

$$\hat{\mu}^{\text{MVLUE}} = \frac{\sum_{i=1}^{N} \frac{n_i}{n_i + \gamma} \hat{\theta}_i}{\sum_{j=1}^{N} \frac{n_j}{n_j + \gamma}} \tag{10}$$

where the hyperparameter

$$\gamma \stackrel{\text{def}}{=} \frac{\mathcal{E}}{\sigma_\theta^2} \in \mathbb{R}_+ \tag{11}$$

is the ratio of the statistics describing local and global variability. It is straightforward to show that

$$\lim_{\gamma \to 0} \hat{\mu}^{\text{MVLUE}} = \hat{\mu}^{\text{MGM}}, \quad \lim_{\gamma \to \infty} \hat{\mu}^{\text{MVLUE}} = \hat{\mu}^{\text{GM}}, \tag{12}$$

as also apparent in Fig. 1 where the weight profiles of the MVLUE, GM and MGM are graphically compared with the saturated (trimmed weights) estimators that will be derived in the following, namely LUE-S, MVLUE-S, and ELUE-S. For a discussion on the interpretation of $\gamma$ please see Sec. A-D and, for more details, [22].

We remark that the LUE-S family we study takes a different path compared to the approaches reviewed above, as it adopts a generally-applicable yet simple weight profile that does not exploit any of the assumptions of the state-of-the-art approaches; as a consequence, it can naturally cope with uncertainty about the most appropriate model to describe the data. Still, the estimators reviewed in this section can serve as reference: specifically, despite they are derived under a different setup and without imposing constraints on the weights, both GM and MGM are unbiased estimators of $\mu$ in general, since their weights sum up to one. Thus they have general-purpose

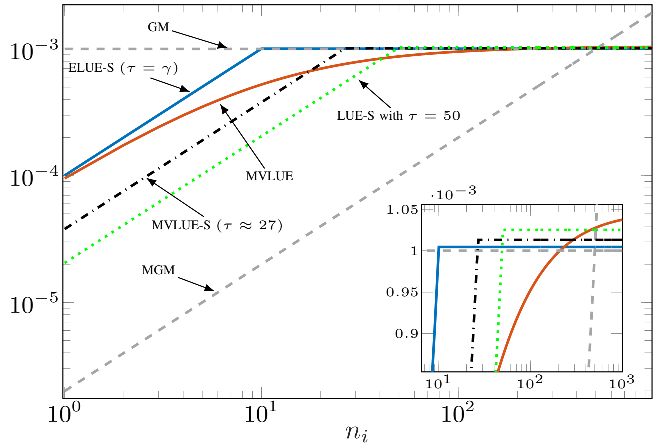


Fig. 1: Comparison between the different weight profiles for $\gamma = 10$ and $\boldsymbol{n}$ linearly spanning the interval $[1, 1000]$.

applicability too, irrespective of the fact that their design assumptions are actually met or not; they indeed represent intuitive ways to combine group means as either an overall average or a sample-size-weighted average of local averages, with no dependency on unknown parameters (namely variance values). One may expect that their performance is inferior to the MVLUE on heterogeneous data, as the MVLUE includes GM and MGM as limit cases and represents the stochastic counterpart (for random $\sigma_i^2$s) of the optimal BLUE (that knows the true variances). However, the drawback of the MVLUE, besides its higher conceptual and computational complexity (discussed in details in Sec. III-B), is the dependency on the parameter $\gamma$ linked to the moments $\mathcal{E}$ and $\sigma_\theta^2$: since the latter are typically unknown, the MVLUE cannot be implemented in exact form. Furthermore, mismatches to its design assumptions may be found in practice, which could harm its optimality even if the true $\gamma$ were used. This motivates the quest for a generally-applicable approach, able to cope with heterogeneous noise in the low-sample regime as well as uncertainty on the data model, indeed the aim of this work.

## III. Linear Unbiased Estimation with Saturated Sample-Size based Weights

Motivated by the discussion in Sec. II and inspired by weight trimming reviewed in Sec. I-B, we propose to consider saturated sample-size based weights, i.e., the LUE-S family, for the problem formulated above — in summary, distribution-agnostic linear unbiased estimation of a global deterministic parameter $\mu$ based on independent but non-identically distributed local means $\hat{\theta}_i$ (ref. eqs. (1)-(3)), with no information on variance values and considering possibly low-sampled $\boldsymbol{n}$ — which has different peculiarities compared to the classical statistical literature: in particular, it hampers sample estimation of local variances, implying that state-of-the-art approaches like ANOVA become very inaccurate. The use of saturated (trimmed) weights, despite quite popular in applications, has not received much theoretical attention. In the following, we provide a comprehensive analysis, with a number of theoretical contributions as well as numerical assessment via both simulations and real-world data, as summarized in Sec. I-A.

### A. Definition and principled derivation

We define the LUE-S family of estimators, denoted by $\hat{\mu}^{\text{LUE-S}}$, as (3) with weights linear in $n_i$ but saturating at $\tau \geq 0$, i.e.

$$w_i^{\text{LUE-S}} = \frac{\min(n_i, \tau)}{\sum_{j=1}^{N} \min(n_j, \tau)}. \tag{13}$$

Two limit cases of (13) are instances of the LS family: specifically, for $\tau \leq \min(\boldsymbol{n})$ one obtains constant weights $1/N$ hence the MGM estimator (5) is retrieved, while $\tau \geq \max(\boldsymbol{n})$ leads to the GM estimator (4) that uses $w_i \propto n_i$.

The structure in (13) has an interesting "water-filling" interpretation: by regarding $n_i$s as resource demands and $w_i$s as shares of allocated resource, such a scheme is *max-min fair*, i.e., small demands are firstly satisfied, then the remaining amount of resource is equally divided among the others [42], [43].[6] Such an appealing property in our case means that each $\hat{\theta}_i$ will be weighted according to its own $n_i$, which is reasonable since the higher the sample size, the better the accuracy, but at the same time no group will be weighted more than what is feasible in order not to underweight smaller-sized groups. Clearly, in this respect the value of $\tau$ plays a role.

Before discussing possible strategies for its setting, we prove the link between the proposed LUE-S family and the Maximum Entropy (ME) principle, which is related to the max-min fair interpretation discussed above.

The ME principle is a fundamental approach to deal with uncertainty. In our problem, uncertainty may come from unmet assumptions of the models, estimation error on the required parameters, imperfect knowledge of the values of $n_i$s (for instance, due to approximate counting, e.g., in stream processing, missing data, and/or quantization), just to name a few. To cope with such aspects, adhering to Occam's razor rationale — which underlies many ideas in signal processing, from model order selection to compressed sensing and sparse learning [33] — one should opt for the solution requiring the least number of hypotheses [45], [46], possibly incorporating only general known aspects while assuming nothing specific about what is unknown. Following statistical physics and information-theoretic arguments, a ME solution thus represents the least biased choice, since most feasible solutions have entropy close to the maximum (entropy concentration theorem [47]) and any other solution with lower entropy (less information) would inject into the model unwarranted assumptions or information that is not available [48], [49].[7]

In the considered problem, since weights are non-negative and sum up to one, they have the properties of a probability mass function (pmf); we can thus set up an optimization problem to find the discrete distribution with ME. If no constraints are considered besides the probability simplex ($\boldsymbol{w} \geq \boldsymbol{0}$

[6] As a result, a single share never exceeds the corresponding demand, and at the same time the scheme prevents that large demands would exhaust the resource (so starving small demands). The fact that the allocation is monotone in the whole range but flat in the higher range (hence the term "water-filling" [43]) is "fair" in the sense that there is no way to increase a share without decreasing already smaller shares [42], [44].

[7] A ME distribution agrees with everything that is known but avoids assuming anything that is unknown, thus "it is a transcription into mathematics of an ancient principle of wisdom" [50]. Equivalently, the ME approach to uncertainty is the best way to avoid unnecessary assumptions [51].

with $\sum_{i=1}^{N} w_i = 1$), the well-known solution is the uniform distribution, which coincides with the MGM $w_i^{\text{MGM}} = 1/N$ [52]. To obtain a non-trivial solution, further constraints need to be considered. Again, by the Occam's razor, we opt for the minimalistic choice of a mild constraint in the form of upper bound, the simplest option being proportional to the sample size. Indeed, this is a plain, linear relationship also found in other state-of-the-art estimators, namely the GM and BLUE. Moreover, for the ANOVA and MVLUE it is a simple matter to rewrite the weights as $w_i = \frac{1}{Z} \frac{n_i}{n_i + a_i}$ where $Z$ is the normalization constant (to satisfy $\sum_{i=1}^{N} w_i = 1$) and $a_i$ are positive parameters; thus, being $n_i \geq 1$, it follows that

$$w_i \leq \alpha n_i \tag{14}$$

with $\alpha = \frac{1}{Z}$. All such considerations motivate the choice of constraints of type (14) for the ME problem, i.e.

$$\boldsymbol{w}^{\star} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} w_i \log w_i \quad \text{s.t.} \begin{cases} \mathbf{0} \leq \boldsymbol{w} \leq \alpha \boldsymbol{n} \\ \sum_{i=1}^{N} w_i = 1 \end{cases} \tag{15}$$

where the objective function is the usual negative[8] Shannon's entropy. We have the following result.

**Proposition 1.** *For the problem of linear unbiased estimation of $\mu$ based on $\hat{\theta}_i$ in* (1)-(3)*, the solution of the ME problem* (15) *is given by the LUE-S* (13)*, with $\tau \propto 1/\alpha$ a reparametrization, i.e., a function of $\alpha \geq 1/\sum_{i=1}^{N} n_i$.*

*Proof.* The proof is a slight variant of the classical Lagrangian-based derivation of the uniform pmf (coinciding with $w_i^{\text{MGM}} = 1/N$) as ME distribution, with additional KKT conditions for the upper-bound constraints (14): this leads to $w_i$s that are either equal to a constant $\leq 1/N$ or trimmed to $\alpha n_i$, that is the LUE-S (13) after reparametrization. A self-consistent detailed proof is reported in the supplemental material. □

A further interesting result is that the LUE-S represents, in a sense, a kind of "attractor" for an entire class of objective functions, not only the ME. In particular, it is possible to replace the entropy term $w_i \log w_i$ with any other strictly convex (continuously differentiable) function $g(w_i)$, being the symmetry of the problem still preserved, as stated below.

**Proposition 2.** *The solution of the optimization problem*

$$\boldsymbol{w}^{\star} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} g(w_i) \quad \textit{s.t.} \begin{cases} \mathbf{0} \leq \boldsymbol{w} \leq \alpha \boldsymbol{n} \\ \sum_{i=1}^{N} w_i = 1 \end{cases} \tag{16}$$

*where $g : \mathbb{R} \mapsto \mathbb{R}$ is any strictly convex continuously differentiable function, is given by the LUE-S* (13)*, with $\tau \propto 1/\alpha$ a reparametrization, i.e., a function of $\alpha \geq 1/\sum_{i=1}^{N} n_i$.*

*Proof.* See supplemental material. □

It is worth noting that the infinite set of possible $g$ includes quadratic functions, leading in particular to the minimization of the $\ell_2$ (Euclidean) norm; this is tantamount to reducing the variance of the weights, as it is known that $\ell_2$-norm regularization introduces *shrinkage* for the sake of MSE reduction (e.g., ridge regression) [33]. This supports the practice of weight trimming (or winsorizing) found in applications, as recalled in Sec. I-B, which is also the rationale behind the LUE-S. In the following we provide a thorough analysis and demonstration of the merits of such a family of estimators.

### *B. Complexity analysis*

With the exception of MGM (constant weights), all state-of-the-art estimators require to explicitly compute all $N$ weights. The MVLUE, in particular, requires about $4N$ flops (counting additions and multiplications 1 flop each). By contrast, the LUE-S requires to compute only weights for $n_i < \tau$, the remaining ones being all equal. This amounts to about $(k_\tau - 1) + 3 + k_\tau + 1 + (N - 1) \approx N + 2k_\tau = (1 + 2\zeta)N$ flops, where $\zeta = k_\tau / N \in [0, 1]$ is the fraction of $n_i < \tau$ out of the total $N$. The complexity is thus i) worst-case: $3N$; ii) best-case: $N$; iii) average case $(1+2p)N$ where $p = \text{Prob}(n_i < \tau)$. Differently put, the MVLUE is computationally more expensive by a factor ranging from 133% to 400%, which is very significant especially for large-scale problems. Similar considerations apply to the other weight profiles.

### *C. Distributional Analysis*

In this section we derive and analyze the distribution of the weights $w_i^{\text{LUE-S}}$ in (13) with respect to the vector $\boldsymbol{n}$.

**Proposition 3.** *The pdf of the weights $w_i^{\textit{LUE-S}}$ in* (13)*, for i.i.d. random variables having common pdf $f_S$ characterizing the distribution of the sample sizes, and $p = \text{Prob}(n_i < \tau)$, is*

$$\begin{aligned} f_W(w) = {} & \frac{1}{w^2} \int_{\mathcal{I}_Z \cap \mathcal{I}_Y} z f_S(z)\, f_Y^{(c)}\!\left(z \frac{1-w}{w}\right) \mathrm{d}z \\ & + \tau(1-p)\, f_Y^{(c)}\!\left(\tau \frac{1-w}{w}\right) \mathbb{1}_{\{w > \frac{1}{N}\}} \\ & + \frac{\tau w^2 (N-1)(1-p)^{1/N}}{(1-w)^2} f_S(\tau) \mathbb{1}_{\{w < \frac{1}{N}\}} \\ & + (1-p)^N \delta\left(w - \frac{1}{N}\right) \end{aligned} \tag{17}$$

*where $f_Y^{(c)}$ denotes the continuous part of $f_Y$, the latter being the pdf of $y_i = \sum_{j \neq i}^{N} \min(n_j, \tau)$, and $\mathcal{I}_Z \cap \mathcal{I}_Y$ the intersection of the supports of $f_Z$ and $f_Y$, with $f_Z$ the pdf of $z_i = \min(n_i, \tau)$.*

*Proof.* See Appendix B, where particular expressions for Uniform and Pareto distribution of $\boldsymbol{n}$, i.e., $S \sim U(a, b)$ and $S \sim P(\alpha)$, are also reported. □

Fig. 2 shows the pdf of $\tilde{w}_i$ for the Uniform (left) and Pareto (right) sample size. The histogram, obtained via Monte Carlo simulation, aligns well with the theoretical expression. In both cases, a Dirac $\delta$ function is located at $1/N$, which is also the mean of the distribution. The $\delta$ impact on the histogram is significant in the Uniform case but not visible in the Pareto case, since its amplitude is $\sim 10^{-9}$. Remarkably, the weight distribution for the proposed LUE-S estimator adapts to the

[8] As customary, entropy maximization has been recast as minimization problem by taking the negative entropy as objective function.

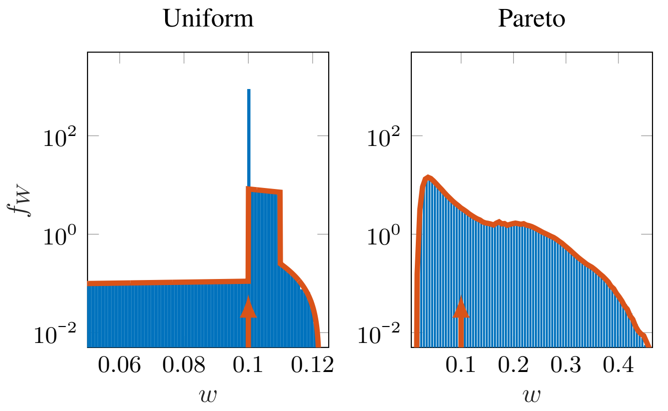


Fig. 2: Comparison between Monte Carlo histogram and analytical $f_W$ for Uniform and Pareto distributed sample size, $N = 10$ and $\tau = \gamma = 10$. Both distributions have a Dirac $\delta$ at $1/N = 0.1$ (arrow), which is clearly visible in the Uniform histogram, but not in Pareto one, since its amplitude is $\sim 10^{-9}$.

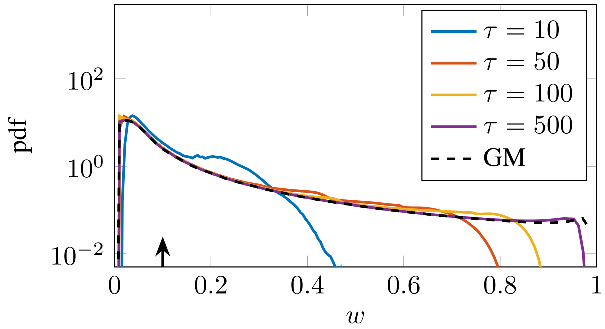


Fig. 3: Distribution of the weights $f_W$ for Pareto distributed sample size, number of groups $N = 10$ and varying $\tau$. All distributions have a Dirac $\delta$ located at $1/N$, indicated with an arrow.

sample size distribution, producing notable differences in the weight profile. In contrast, the MGM has constant weights, i.e., a $\delta$ with unitary amplitude located at $1/N$, regardless of the sample size distribution. The effect of the differing weight profile on the estimation error will be assessed in Sec. VI.

Fig. 3 shows how the weight distribution of the LUE-S changes with $\tau$, and compares it to that of the GM estimator. The distributions are obtained with respect to a vector $\boldsymbol{n}$ of $N = 10$ i.i.d. Pareto random variables. All distributions feature a Dirac $\delta$ function, indicated with an arrow, localized at $1/N$, which is also the mean of the distribution (see also Proposition 4 later). We can observe that varying $\tau$ significantly impact the shape of the weights distribution, reinforcing the interpretation of the LUE-S as a family of estimators. Notice also that such a distribution differs from that of the GM, MGM, and MVLUE; precisely, it collapses to the distribution of the GM only for $\tau \geq \max(\boldsymbol{n})$, and to that of the MGM for $\tau \leq \min(\boldsymbol{n})$. Lastly, it approximates the distribution of the MVLUE for $\tau = \gamma$: such a particular choice, referred to as Empirical LUE-S (ELUE-S) will be analyzed in details in Sec. IV-C.

## IV. Setting of $\tau$

In this section we discuss different strategies to set the cut-off parameter $\tau$ of $\hat{\mu}^{\text{LUE-S}}$. Beforehand, it is worth pointing out that although both MVLUE and LUE-S are parametric estimators, there is an important difference: $\tau$ in the LUE-S is a degree of freedom, while $\gamma$ in the MVLUE is the ratio of two statistical moments, typically unknown in practice. Thus, the MVLUE cannot be implemented in exact form. We will propose in Sec. V several estimators for $\gamma$, which will also enable a comparison with the ELUE-S (LUE-S with $\tau = \gamma$).

As to the setting of $\tau$, this is unfortunately related to the same issue found in robust statistics, i.e., seeking a general rule for choosing the trimming level, still an open problem [53]. In fixed cut-off estimators, the level that separates outliers from non-outliers is established prior to data collection. In fact, while theoretically the optimal cut-off could be derived if the distribution of the data were known, in practical applications the latter is unknown and historic data and expert judgement are the best tools for defining the most suitable fixed cut-off, to be reviewed periodically [54]. Indeed, while methods for automatic setting have been proposed in the survey sampling literature (e.g., based on the interquartile range compared to the median weight, or some empirical percentile), many real-world survey analysts continue to choose the cut-off in an ad-hoc manner [55]. In the following we discuss some possible approaches for setting $\tau$ in the context of this work.

### A. Minimum and Maximum $\ell_2$-norm weights

A first simple strategy is to look at the variance of the weights, as in fact the goal of trimming (saturate) weights is to reduce their variance to some extent (shrinkage), as discussed in Sec. I-B. We have the following result.

**Proposition 4.** *The weights of any linear unbiased estimator of $\mu$ have mean $1/N$ irrespective of $\boldsymbol{n}$, hence the minimum and maximum variance of $w_i^{\text{LUE-S}}$s correspond to the minimum and maximum $\ell_2$-norm solutions with respect to $\tau$; these are obtained for $\tau \leq \min(\boldsymbol{n})$ and $\tau \geq \max(\boldsymbol{n})$, and coincide with the MGM and GM estimators, respectively.*

*Proof.* See Appendix C. □

Notice that for the minimum-norm case the weight profile is constantly equal to $1/N$ (MGM) hence it achieves the minimum possible variance value of zero. The maximum-norm case can be instead interpreted as a degenerate saturation in one point $(\max(\boldsymbol{n}))$, yielding maximum variance within the LUE-S family. As a whole, such a result shows that looking at the weight variance does not lead to a convenient way to set $\tau$. A better approach is to look at the variance of the estimator instead of the variance of the weights, as shown below.

### B. Minimum estimation variance weights (MVLUE-S)

In the following we aim at determining the optimal value $\tau^*$ that minimizes the total estimation variance. In general, a minimum-variance linear unbiased estimator for $\mu$ is obtained by solving an optimization problem of the type (A.11), where the variance of the group means $\hat{\theta}_i$s depends on the assumed observation model. In particular, recalling Sec. II, we have that

$$\text{VAR}[\hat{\theta}_i] = \begin{cases} \frac{\bar{\sigma}_i^2}{n_i} & \text{BLUE} \\ \sigma_\theta^2 + \frac{\sigma_i^2}{n_i} & \text{random-effect} \\ \sigma_\theta^2 + \frac{\mathcal{E}}{n_i} & \text{MVLUE} \end{cases} . \tag{18}$$

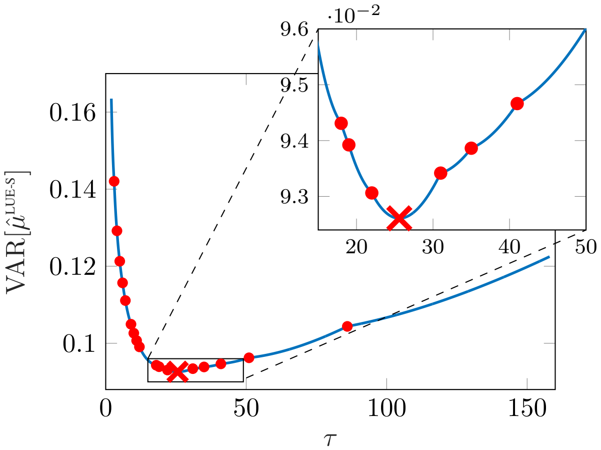


Fig. 4: Example of variance of the LUE-S as function of the threshold $\tau$ for $n_i$ generated as Pareto with $\alpha = 0.9$ (considering the MVLUE case, third line in eq. (18), for $\gamma = 40$). In this realization we have $\breve{N} = 19$ unique values $\breve{n}_\ell$ out of $N = 100$ generated $n_i$. Dots indicate the minimum values of the variance over each interval $[\breve{n}_\ell, \breve{n}_{\ell+1}]$, attained at $\tau^\star_\ell$, whereas the cross mark indicates the global minimum attained at $\tau^*$.

Clearly, if one knew which model exactly matches the data at hand, there would be no need of alternative estimators, as the optimal one (for that case, i.e., under the related assumptions) would be available. In this work we are instead considering that in real-world applications there is uncertainty in this respect, motivating the interested in the LUE-S (as discussed at the beginning of Sec. III and III-A). Therefore, it is reasonable to adopt a distribution-agnostic and very general type of estimator, that is the LUE-S family, but then use for its parameter setting one of the variances in (18) as a proxy for the unknown actual variance — the choice should be the one expected to be the closest, so moderately reintroducing in the estimator this partial (uncertain) prior knowledge. The general result is given by the following proposition.

**Proposition 5.** *Let us assume, without loss of generality, that the elements of $\boldsymbol{n}$ are sorted in ascending order and be $\breve{\boldsymbol{n}}$ the vector containing the $\breve{N} \leq N$ unique elements $\breve{n}_\ell$ of $\boldsymbol{n}$. The optimal $\tau^*$ can be obtained by selecting, among the points*

$$\tau^\star_\ell = \begin{cases} r_\ell & \textit{if } \breve{n}_\ell < r_\ell < \breve{n}_{\ell+1} \\ \breve{n}_\ell & \textit{if } r_\ell \leq \breve{n}_\ell \\ \breve{n}_{\ell+1} & \textit{if } r_\ell \geq \breve{n}_{\ell+1} \end{cases}, \quad \ell = 1, \ldots, \breve{N} - 1 \quad (19)$$

*the one that yields the minimum value of the variance of $\hat{\mu}^{\text{LUE-S}}$ over all intervals $[\breve{n}_\ell, \breve{n}_{\ell+1}]$ (eq. (D.3)), where*

$$r_\ell = \frac{(N - k_\tau) \sum_{i=1}^{k_\tau} n_i^2 \gamma_i}{\sum_{i=1}^{k_\tau} n_i \sum_{i=k_\tau+1}^{N} \gamma_i}, \quad (20)$$

*and $k_\tau \in \{1, \ldots, N\}$ counts the number of $n_i < \tau$.*

*Proof.* See Appendix D. □

Proposition 5 leads to an estimator we will refer to as Minimum Variance Unbiased Estimator with Saturated sample-size based weights (MVLUE-S), whose weights are therefore

$$w_i^{\text{MVLUE-S}} \overset{\text{def}}{=} \frac{\min(n_i, \tau^*)}{\sum_{j=1}^{N} \min(n_j, \tau^*)}. \quad (21)$$

As formally shown in Appendix D, the variance of $\hat{\mu}^{\text{LUE-S}}$ (eq. (D.4)) is a piecewise function of $\tau$, obtained as disjoint union of the variances over each interval $[\breve{n}_\ell, \breve{n}_{\ell+1}]$, the latter exhibiting at most one stationary point $r_\ell$; therefore, the global optimizer $\tau^*$ can be readily obtained by selecting the point $\tau^\star_\ell$, computed via eqs. (19)-(20), that yields the minimum variance. Fig. 4 shows an illustrative example, to facilitate a clarifying visualization of the described behavior.

### *C. Empirical Linear Unbiased Estimator with Saturated Sample-Size based Weights (ELUE-S)*

The optimization in Sec. IV-B may be undesirable for practical use, especially for large problem instances. Moreover, as discussed at the beginning of Sec. IV, optimal setting of a trimming level in the general case is an open problem, and may be even unsolvable in presence of uncertainty due to the difficulty of defining the most appropriate optimization criterion; therefore, a very popular approach remains a fixed manual setting. However, we discuss here an alternative formulation where $\tau$ is replaced by a simple value depending only on low-order statistics that can be estimated directly from the data through an explicit formula. Since this idea is loosely reminiscent of the Empirical Bayes approach, the resulting estimator will be referred to as Empirical Linear Unbiased Estimator with Saturated sample-size based weights (ELUE-S)[9]. We will show in particular that the setting $\tau = \gamma$ with $\gamma$ replaced by a suitable estimator $\hat{\gamma}$ is a viable option, which can be justified by drawing a connection with the MVLUE.

We start by noticing that the MVLUE weighs samples differently based on their sizes, but exhibits a kind of asymptotic saturation for large $n_i$. In fact, consider the function $w(x) = \frac{x}{x+\gamma}$ and let us study its limit behaviour. Straightforward investigation reveals that the derivative $\frac{\mathrm{d}w}{\mathrm{d}x} = \frac{\gamma}{(x+\gamma)^2}$ yields the first order approximation $w(x) \approx w(0) + \frac{\mathrm{d}w}{\mathrm{d}x}\big|_0 x = x/\gamma$ for small argument $x$, while for large argument $w(x) \approx 1$. Under these two regimes, $w(x)$ can be approximated by straight lines for $x \gg \gamma$ or $x \ll \gamma$, having intersection at $x = \gamma$, as also graphically apparent in Fig. 1.

The asymptotic behavior above is consistent with the observation that the terms in the summation of the total variance of the MVLUE (see eq. (A.14)) can be approximated as

$$w_i^2 \left( \frac{n_i + \gamma}{n_i} \right) \approx \begin{cases} w_i^2 \dfrac{\gamma}{n_i} & n_i \ll \gamma \\ w_i^2 & n_i \gg \gamma \end{cases} \quad (22)$$

and the resulting optimization problem returns as optimal solution exactly the ELUE-S (details are reported in Appendix E). We will analyze in Sec. VI the performance of this estimator, also in comparison with other choices of $\tau$.

## V. Hyperparameter Estimation

In this section we propose different estimators $\hat{\gamma}$, which will be used in the implementation of the ELUE-S (but allow

[9] Notice however that the proposed approach cannot be considered Bayesian, since the parameter $\mu$ to be estimated is deterministic and in the problem formulation at hand all probability distributions are unknown. Consequently, no statistical model is available for Bayesian inference, which requires the posterior distribution. The proposed approach is indeed distribution-agnostic.

for implementation of the MVLUE as well). Their relative performance, and the consequent impact on the estimation of $\mu$, will be then assessed in Sec. VI.

Recalling the definition (11), we observe that in order to estimate $\gamma$ we need to estimate the ratio between the two moments $\mathcal{E}$ and $\sigma_\theta^2$. To this aim, a possible approach is to obtain $\hat{\gamma}$ as the ratio between any estimators of $\mathcal{E}$ and $\sigma_\theta^2$, i.e., $\hat{\gamma} = \hat{\mathcal{E}}/\hat{\sigma}_\theta^2$. Although in general the ratio estimator is not the ratio of the estimators, the latter is an unbiased estimator of the ratio (to first order, up to higher-order terms in Taylor expansion) if the two estimators (numerator and denominator) are unbiased [56]. The simplest choice is to use the sample mean and sample variance of $\hat{\theta}_i$'s, respectively, that is

$$\hat{\mathcal{E}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}_i^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i-1}\sum_{m=1}^{n_i}(x_{i,m}-\hat{\theta}_i)^2, \tag{23}$$

$$\begin{aligned}\hat{\sigma}_\theta^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\theta}_i - \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i\right)^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\theta}_i - \mu^{\text{MGM}}\right)^2.\end{aligned} \tag{24}$$

It is easy to prove that $\hat{\mathcal{E}}$ is an unbiased estimator of $\mathcal{E}$ ($\mathrm{E}[\hat{\mathcal{E}}] = \mathcal{E}$), while for $\hat{\sigma}_\theta^2$ it can be shown that

$$\mathrm{E}[\hat{\sigma}_\theta^2] = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\mathcal{E} + \sigma_\theta^2. \tag{25}$$

Eq. (25) reveals that the variance estimator (24) is biased, but suggests a natural way to reduce the bias. In particular, it is possible to consider the following adjusted estimator

$$\hat{\sigma}_{\theta,\text{USS}}^2 = \hat{\sigma}_\theta^2 - \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\hat{\mathcal{E}} \tag{26}$$

which tries to compensate the bias by leveraging the unbiased estimator $\hat{\mathcal{E}}$ in (23). We use the label Unweighted Sum of Squares (USS) for this proposed estimator since (26) can be seen as the generalization to random $\sigma_i^2$s of the USS estimator available in the literature for deterministic $\sigma_i^2$'s [21].

A more sophisticated approach is to consider a weighted version of the sample variance estimator (24), with the aim of mitigating the impact of wrong sample variance estimation. This may be particularly suitable when the sample size varies wildly across the groups, and/or it is not sufficiently large to guarantee an accurate estimation. Let us thus consider the generic weighted estimator

$$\hat{\sigma}_{\theta,\text{weighted}}^2 = \frac{V_1}{V_1^2 - V_2}\sum_{i=1}^{N} v_i\left(\hat{\theta}_i - \frac{1}{V_1}\sum_{j=1}^{N}v_j\hat{\theta}_j\right)^2 \tag{27}$$

where $V_1 = \sum_{i=1}^{N} v_i$ and $V_2 = \sum_{i=1}^{N} v_i^2$. It is easy to prove that the bias of $\hat{\sigma}_{\theta,\text{weighted}}^2$ is

$$\mathrm{E}[\hat{\sigma}_{\theta,\text{weighted}}^2] = \frac{1}{\bar{V}}\sum_{i=1}^{N}\frac{1}{n_i}\left(v_i - \frac{v_i^2}{V_1}\right)\mathcal{E} + \sigma_\theta^2 \tag{28}$$

where we defined for convenience $\bar{V} = (V_1 - V_2/V_1)$. Clearly, in the weighted case the expression for the bias is more involved, but nevertheless it can still be compensated by using the unbiased estimator of $\mathcal{E}$, as done in the USS approach.

| | Parameter | Value |
|---|---|---|
| Deterministic | $N$ | 50 |
| | $\alpha$ | 0.9 |
| | $\mu$ | 50 |
| | $\sigma_\theta^2$ | 0.1 |
| | $\mathcal{E}$ | 2 |
| Random | $n_i$ | $P(\alpha)$ |
| | $\theta_i$ | $\mathcal{N}(\mu,\sigma_\theta^2)$ |
| | $\sigma_i^2$ | Rayleigh($\mathcal{E}\sqrt{2/\pi}$) |
| | $x_{i,m}$ | Gumbel $(\mu_{\text{Gumbel}}, \beta_{\text{Gumbel}})$ |

TABLE I: Simulation parameters

A reasonable choice for the weights is $v_i = n_i$, which, when corrected for the bias using the unbiased estimator $\hat{\mathcal{E}}$, gives

$$\hat{\sigma}_{\theta,\text{ANOVA}}^2 = \frac{1}{\bar{V}}\left(\sum_{i=1}^{N} n_i\left(\hat{\theta}_i - \hat{\mu}^{\text{GM}}\right)^2 - (N-1)\hat{\mathcal{E}}\right) \tag{29}$$

where we have highlighted how this choice of the weights implies the use of $\hat{\mu}^{\text{GM}}$ as sample estimator for the mean. We label the proposed estimator (29) as ANOVA since, again, it can be seen as the generalization to random $\sigma_i^2$s of the well-known variance estimator for deterministic $\sigma_i^2$s used in standard ANOVA [21].

With these definitions of the $\sigma_\theta^2$ estimators, the proposed unbiased estimators of $\gamma$ are consequently defined as

$$\hat{\gamma}_{\text{USS}} = \hat{\mathcal{E}}/\hat{\sigma}_{\theta,\text{USS}}^2 \tag{30}$$

and

$$\hat{\gamma}_{\text{ANOVA}} = \hat{\mathcal{E}}/\hat{\sigma}_{\theta,\text{ANOVA}}^2. \tag{31}$$

## VI. Numerical results

We analyze the ELUE-S by comparing its performance against different approaches. The analysis is first performed via Monte Carlo simulations, then an illustrative application to the COVID-19 dataset provided is discussed.

### A. Simulations

Synthetic data are obtained by simulating $N = 50$ data sources or groups, having cardinality distributed as Pareto with $\alpha = 0.9$. This gives sample sizes $n_i$s with very uneven distribution (theoretically, infinite mean and variance). The observations (raw data) $x_{i,m}$ are generated from a Gumbel distribution with location and scale parameters set to have local mean $\theta_i$ and local variance $\sigma_i^2$. Moreover, the local means $\theta_i$s are Gaussian distributed with mean $\mu = 50$ (the global parameter of interest) and variances $\sigma_\theta^2 = 0.1$, while the local variances $\sigma_i^2$s are Rayleigh distributed with mean $\mathcal{E} = 2$. Such settings[10], summmarized in Table I, represent a non-trivial scenario where many low samples, but also a few large ones,

[10] For reproducibility, we remind that, to obtain a mean $\mathcal{E}$, the parameter of the Rayleigh distribution must be set to $\mathcal{E}\sqrt{2/\pi}$. Likewise, to obtain mean $\theta_i$ and local variance $\sigma_i^2$, the scale and location parameters of the Gumbel distribution must be set to $\beta_{\text{Gumbel}} = \sqrt{(6\sigma_i^2/\pi^2)}$ and $\mu_{\text{Gumbel}} = \theta_i - \gamma_{\text{em}}\beta_{\text{Gumbel}}$, respectively, where $\gamma_{\text{em}}$ is the Euler–Mascheroni constant.

are present, hence local sample estimators of the variance may be inaccurate. Moreover, both means and variances locally vary (due to the random generation), so making the whole dataset heterogeneous, though bringing information about the same underlying common (global) parameter $\mu$.

We will first focus on the estimation of the statistics needed to estimate the hyperparameter $\gamma$, which will in turn be used to provide an estimate of $\mu$ for the ELUE-S.

The three plots in Fig. 5 show the pdf estimates of the different quantities involved in the estimation process obtained with $M = 10^6$ Monte Carlo runs. In the plots, the true values are illustrated as vertical loosely dashed lines. Specifically, Figs. 5a and 5b report the performance of the estimators of $\mathcal{E}$ and $\sigma_\theta^2$, respectively. The pdf estimate of $\hat{\mathcal{E}}$ in Fig. 5a is centered around the true value $\mathcal{E}$, which is not surprising, as the estimator (23) is unbiased. More revealing, Fig. 5b shows the performance of the different estimators of $\sigma_\theta^2$ derived in Sec. V, i.e., $\hat{\sigma}_\theta^2$, $\hat{\sigma}_{\theta,\text{USS}}^2$, and $\hat{\sigma}_{\theta,\text{ANOVA}}^2$. As expected, the bias of $\hat{\sigma}_\theta^2$ greatly affects its performance, as its pdf estimate is centered around a completely wrong value. The USS is able to adjust for the bias, however we can still see how the estimates are highly scattered (large variance). Clearly, a thoughtful choice of the weights, as in the case of the ANOVA estimator, leads to better performance, which in turns yields a more accurate $\hat{\gamma}$, as illustrated in Fig. 5c.

We now analyze the bias and MSE/variance of the estimators of $\gamma$. Fig. 6 reports the MSE as function of $N$, together with a bias-variance analysis. Considering the values of the parameters as above, the ground truth is $\gamma = 20$. Fig. 6a shows the superiority of $\hat{\gamma}_{\text{ANOVA}}$ with respect to the biased estimator $\hat{\gamma}$ and the $\hat{\gamma}_{\text{USS}}$, in particular when $N$ is very large. The bias of $\hat{\gamma}$ does not vanish as $N$ grows, therefore its performance does not improve with the number of groups; on the contrary, it worsens as $N$ gets larger than 10, as it becomes more probable to draw wild $n_i$ from the Pareto distribution. Figs. 6b-6c show the bias-variance trade-off for $\hat{\gamma}_{\text{ANOVA}}$ and $\hat{\gamma}_{\text{USS}}$, respectively. The bias correction implemented in $\hat{\gamma}_{\text{USS}}$ guarantees that, as $N$ increases, the estimator performance improves; however, it still suffers from the heterogeneity of the sample sizes, which $\hat{\gamma}_{\text{ANOVA}}$ tries instead to mitigate, outperforming its competitors.

We now assess how the performance of such estimators impact onto the final estimation accuracy on $\mu$. Figs. 7 and 8 show the estimates produced by the different versions of the ELUE-S, namely the ELUE-S, ELUE-S$_{\text{USS}}$, and ELUE-S$_{\text{ANOVA}}$ (which use the corresponding estimators of $\gamma$), compared to the GM/MGM estimators which are the natural competitors. The performance of the estimators are also compared against the (non-implementable) oracle estimator LUE-S$_{\text{ORACLE}}$ that uses the true value of $\gamma$, which represents a theoretical benchmark. As the LUE-S is an unbiased estimator irrespective of the choice of weights, the pdfs in Fig. 7 are all centered around the true value of the parameter $\mu$ to be ultimately estimated. Different weighting strategies however strongly affect the estimator variance. For the case at hand, the best performance, attained obviously by the oracle estimator, are almost matched by $\hat{\mu}_{\text{ANOVA}}^{\text{ELUE-S}}$, as a consequence of the better estimate of the hyperparameter $\gamma$ as from Fig. 5c. Slightly worse performance are obtained by $\hat{\mu}_{\text{USS}}^{\text{ELUE-S}}$, while $\hat{\mu}^{\text{ELUE-S}}$ is the worst among the three solutions. A large performance gap is instead evident with the competitors, especially the $\hat{\mu}^{\text{GM}}$ which performs significantly worse.

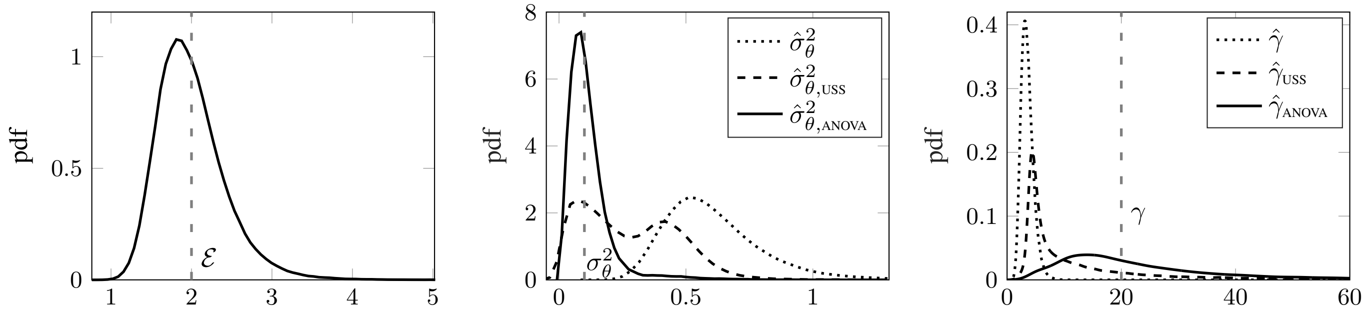
To better highlight the performance differences among the estimators, in particular with reference to larger errors, the complementary cumulative distribution function (CCDF) of the squared estimation error $(\hat{\mu} - \mu)^2$ is plotted in Fig. 8, in doubly-logarithmic scale. It is evident the excellent performance of $\hat{\mu}_{\text{ANOVA}}^{\text{ELUE-S}}$, very close to the theoretical ideal benchmark (oracle), while a certain fraction of outlying error values affect the tail behavior of the other estimators. The competitors $\hat{\mu}^{\text{GM}}$ and $\hat{\mu}^{\text{MGM}}$ lead to systematically larger errors in the whole range, with very significant differences on higher MSE values.

Fig. 9 shows the asymptotic performance of the different estimators of $\mu$ in terms of MSE with respect to $N$. We can see how the behaviour of the estimators is consistent in the asymptotic regime, with $\hat{\mu}_{\text{ANOVA}}^{\text{ELUE-S}}$ achieving the same performance of the Oracle estimator (which knows the true $\gamma$). It is also interesting to notice that the GM estimator, when $N$ is very small ($< 10$), performs similarly with respect to its competitors. This is probably due to the fact that the other more complex estimators rely on the possibility to obtain good estimates of the statistics of the problem ($\hat{\mathcal{E}}$, $\hat{\sigma}_\theta^2$). This is particularly difficult when $N$ is very small, therefore the performance of the estimators are more or less all the same. As $N$ grows, and the $n_i$s become more and more "wild", the GM fails in copying with the heterogeneity of the data, and its performance degrades. This interpretation is also supported by the fact that the $\hat{\mu}_{\text{ANOVA}}^{\text{ELUE-S}}$ error curve does not achieves the same error of the Oracle estimator.

Fig. 10 shows the CCDF of the estimation error attained by the MVLUE with $\gamma$ obtained using $\hat{\gamma}_{\text{ANOVA}}$ and the estimation error of the LUE-S for $\tau = 1, \ldots, \max(\boldsymbol{n})$. We can clearly see that, for several choices of $\tau$, the LUE-S outperforms the MVLUE in terms of MSE (colored curves). This is because the MVLUE is optimal only if its assumptions are exactly met, while the LUE-S has the flexibility to possibly approach the optimal estimator for known variances (BLUE). However, how to obtain the optimal $\tau$ in the most general case is still an open problem, as discussed in Sec. IV. Therefore, the choice of the ELUE-S ($\tau = \gamma$) is a convenient one, as it remains close to the MVLUE but is more flexible; moreover, it depends on a single parameter for which we have provided estimators (Sec. V) to adapt to the data; finally, it has low conceptual as well computational complexity (as discussed in Sec. III-B).

### B. Application to a real-world dataset: COVID-19

Estimation of epidemiological quantities such as reproduction number $R_0$ or death rate is of paramount importance to monitor, predict, and counteract the spread of viruses. However, this is a particularly difficult task, as such indicators try to summarize the enormous complexity of the epidemic into single values. Moreover, the extreme heterogeneity of the data at hand significantly challenges the performance of estimators. The recent COVID-19 pandemic has caused a surge in the scientific literature around these topics [57]–[59]. In practice, however, simple indicators are typically used by

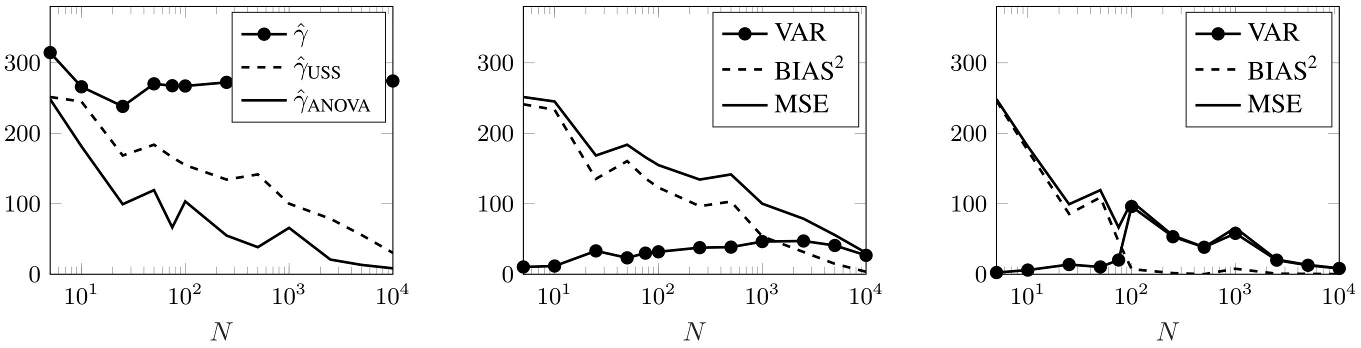


(a) Pdf of the $\hat{\mathcal{E}}$ estimator

(b) Pdf of the $\hat{\sigma}^2_\theta$, $\hat{\sigma}^2_{\theta,\text{USS}}$, and $\hat{\sigma}^2_{\theta,\text{ANOVA}}$ estimators

(c) Pdf of the $\hat{\gamma}$, $\hat{\gamma}_{\text{USS}}$, and $\hat{\gamma}_{\text{ANOVA}}$ estimators

Fig. 5: Comparison of different estimators for hyperparameter estimation.



(a) Comparison of MSE[$\hat{\gamma}$] = $|\gamma - \hat{\gamma}|^2$

(b) USS Bias-Variance Trade-Off

(c) ANOVA Bias-Variance Trade-Off

Fig. 6: Mean Square Error analysis of the $\gamma$ estimators as function of the number of groups $N$.

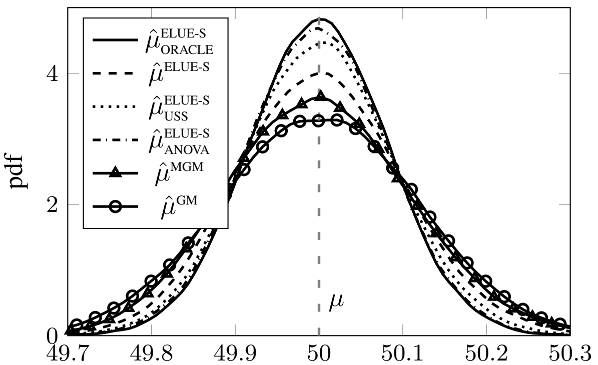


Fig. 7: Pdf of $\hat{\mu}$ obtained with Monte Carlo simulation

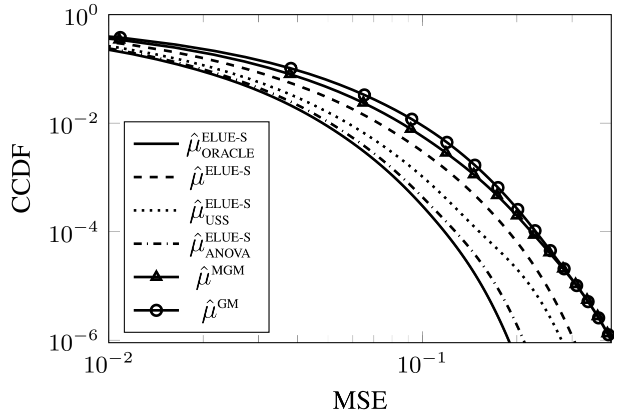


Fig. 8: CCDF of the squared estimation error $(\hat{\mu} - \mu)^2$.

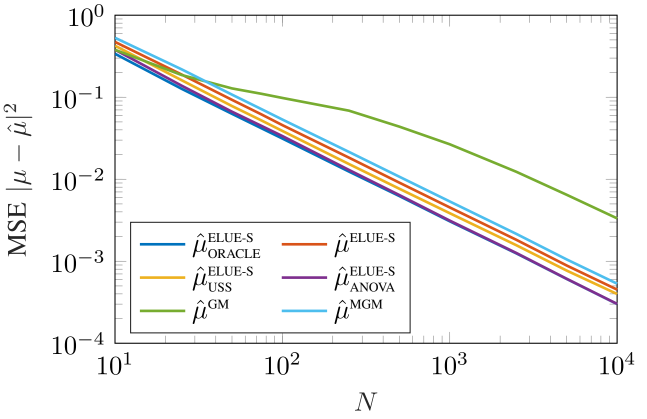


Fig. 9: Mean Square Error analysis of the $\mu$ estimators as function of the number of groups $N$.

decision makers to take informed actions, which makes very important their accurate estimation. Hereafter we will focus on the problem of estimating the COVID-19 death rate by using the data available from The New York Times official repository [60]. The time window goes from the pandemic surge on March 2020 up to January 2021, before the beginning of the vaccination campaign. The dataset consists of the daily cumulative number of cases and cumulative deaths for each county in the US. We denote $n_i(t)$ and $x_i(t)$ the number of cases and deaths, respectively, on day $t$ and for the $i$-th county.

Assuming the number of deaths to be the sum of binary variables indicating whether the $m$-th individual survived the

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2023.3293908

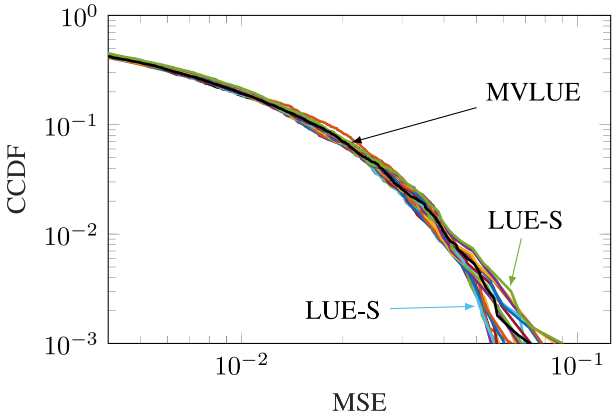


Fig. 10: CCDF of the MVLUE estimation error with $\gamma$ (black solid line) and of the LUE-S estimation error for $\tau = 1, \ldots, \max(\boldsymbol{n})$ (colored lines). For several choices of $\tau$, the LUE-S outperforms the MVLUE in terms of MSE.

infection, i.e., $x_i(t) = \sum_{m=1}^{n_i(t)} x_{i,m}(t)$ and $x_{i,m}(t) = [0, 1]$, our goal is to estimate the mortality of the infection $\mu$, which is also the probability of $x_{i,m}$ to be equal 1.

Figs. 11-14 illustrate the characteristic of the dataset along the considered time span. In Fig. 11 we can see the evolution of $N$, which is the number of US counties that at time $t$ have $n_i \geq 1$ cases (dates are in MM/DD format). As one may expect, as the pandemic spreads, the number of counties which started testing and reporting cases of infected individuals increased rapidly by the end of March 2020 and reached its maximum by the end of August 2020. The distribution of $n_i$ changed also dramatically during the pandemic, as illustrated in Fig. 12 where we show the empirical CCDF of $\boldsymbol{n}(t)$ in doubly-logarithmic scale. The CCDF helps visualizing the long tail of the distribution, whose upper bound of the support increases by two orders of magnitude during the considered time span. The changes in the shape of distribution are clearly visible in Fig. 13, where the empirical pdf of $\log(\boldsymbol{n}(t))$ is shown for five different days (dates are in YY/MM/DD format). The logarithmic transformation is applied to facilitate the comparison and highlight the main features of the distribution, which otherwise would be hindered by the heavy tail. Finally, we are also interested in the distribution of $\hat{\boldsymbol{\theta}}(t)$, shown in Fig. 14. Such a distribution is an indicator of the heterogeneity of the local measures: its "wildness", together with an uneven distribution of $\boldsymbol{n}$, directly affects the estimation problem (3). Here it is evident that for the first months of the pandemic, the local means distribution is significantly wide, with an heavy tail that slowly disappears as we move forward in time.

Fig. 15 shows the result of the estimation of the death rate using the different proposed estimators and the natural competitors. In particular, we compare the GM and MGM with the ELUE-S based on the different $\gamma$ estimators discussed in Sec. V. We can observe several interesting features in the plot. During the first days of the pandemic, estimates vary abruptly: due to the scarcity of data (many sample sizes are $n_i = 1$ or 2), we are in a regime of low sample and high heterogeneity, and the estimate are consequently unreliable. After some time, as the sample size increases, the estimate becomes apparently more reliable and, also, the difference between the estimators can be clearly appreciated. Since no ground

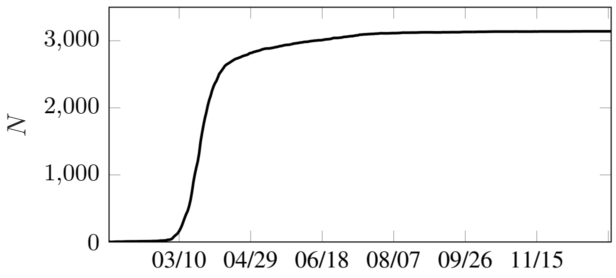


Fig. 11: Number of counties with $n_i(t) > 0$.

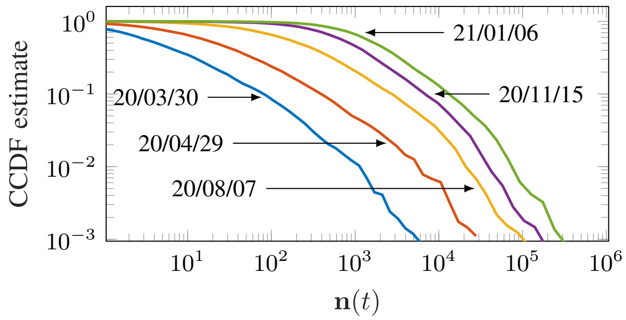


Fig. 12: CCDF estimates of $\boldsymbol{n}$ on five different dates.

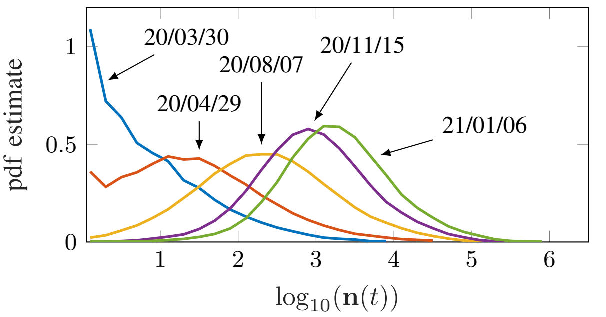


Fig. 13: Pdf estimates of $\log(\boldsymbol{n})$ on five different dates.

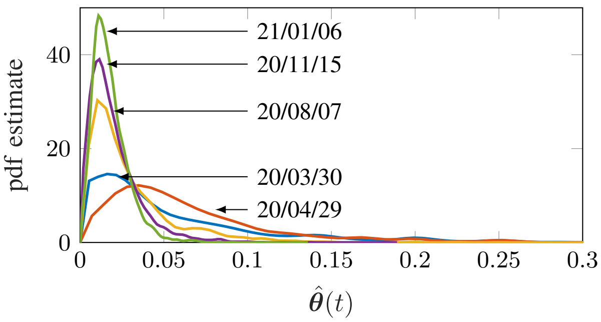


Fig. 14: Pdf estimates of $\hat{\boldsymbol{\theta}}(t)$ on five different dates.

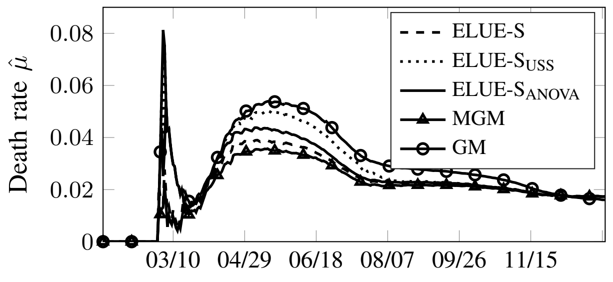


Fig. 15: Comparison of the Covid-19 death rate estimate obtained with different estimators.

truth is available, no absolute error comparison is possible. Nevertheless, we can observe that the estimate provided by the ELUE-S$_{\text{ANOVA}}$ lies almost at the center between the other approaches. This, together with the observation that the GM and MGM can be interpreted as limit cases of the ELUE-S and given the analysis in Sec. VI-A, suggests that they are likely under- and over-estimating, respectively, the parameter of interest. The proposed ELUE-S instead, with its particular weighting strategy, is capable of counteracting effectively both heterogeneity and presence of outliers in the data, providing more trustworthy results. As discussed in Sec. IV-B, when most of the sample sizes are large enough to provide reliable local estimates (last two curves in Fig. 13), the weight profile becomes less crucial and all curves converge to a similar value.

## VII. Conclusion

The paper addressed the challenging problem of distribution-agnostic linear unbiased estimation of a global deterministic parameter from non-homogeneous data. This setup is found in many practical cases where joint processing of different groups of data is performed, or data contamination from outliers and/or structural heavy-tailedness of the underlying random process exist. Special consideration has been given to scenarios in which the applicability of sample estimators of the local variance is limited due to low-sample at each source or group. This means in particular that standard state-of-the-art approaches, including ANOVA techniques, yield poor results.

To tackle such a problem under uncertain conditions on the data model, an estimation approach has been investigated where the idea of winsorizing is applied to the sample size values determining the weight profile, so obtaining a family of estimators with trimmed weights, proportional to the sample size but with a proper saturation. A comprehensive theoretical analysis has been performed. Finally, proper estimators for the hyperparameters have been proposed and analyzed.

Numerical simulations results have shown that the proposed approach outperforms state-of-the-art alternative approaches based on least squares. Moreover, illustrative application to COVID-19 data analysis has been presented, which highlighted the challenges of dealing with unbalanced, heterogeneous data, and confirmed the merits of the proposed approach.

## References

[1] S. M. Kay, *Fundamentals of statistical signal processing*. Prentice Hall PTR, 1993.

[2] Y.-R. Tsai and C.-J. Chang, "Cooperative information aggregation for distributed estimation in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 59, no. 8, pp. 3876–3888, 2011.

[3] I. Nevat, G. W. Peters, and I. B. Collings, "Random field reconstruction with quantization in wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 61, no. 23, pp. 6020–6033, 2013.

[4] K. Radnosrati, G. Hendeby, and F. Gustafsson, "Exploring positive noise in estimation theory," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3590–3602, 2020.

[5] J.-J. Xiao and Z.-Q. Luo, "Decentralized estimation in an inhomogeneous sensing environment," *IEEE Transactions on Information Theory*, vol. 51, no. 10, pp. 3564–3575, 2005.

[6] J. Li and G. AlRegib, "Distributed estimation in energy-constrained wireless sensor networks," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3746–3758, 2009.

[7] M. Pesavento and A. Gershman, "Maximum-likelihood direction-of-arrival estimation in the presence of unknown nonuniform noise," *IEEE Transactions on Signal Processing*, vol. 49, no. 7, pp. 1310–1324, 2001.

[8] B. Liao, S.-C. Chan, L. Huang, and C. Guo, "Iterative methods for subspace and doa estimation in nonuniform noise," *IEEE Transactions on Signal Processing*, vol. 64, no. 12, pp. 3008–3020, 2016.

[9] J. Font-Segura and X. Wang, "Glrt-based spectrum sensing for cognitive radio with prior information," *IEEE Transactions on Communications*, vol. 58, no. 7, pp. 2137–2146, 2010.

[10] K. Panwar, P. Babu, and P. Stoica, "Maximum likelihood algorithm for time-delay based multistatic target localization," *IEEE Signal Processing Letters*, vol. 29, pp. 847–851, 2022.

[11] I. Fijalkow, E. Heiman, and H. Messer, "Parameter estimation from heterogeneous/multimodal data sets," *IEEE Signal Processing Letters*, vol. 23, no. 3, pp. 390–393, 2016.

[12] Y. Chen, S. Kar, and J. M. Moura, "Resilient distributed parameter estimation with heterogeneous data," *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 4918–4933, 2019.

[13] G. Papageorgiou, P. Bouboulis, and S. Theodoridis, "Robust linear regression analysis—a greedy approach," *IEEE Transactions on Signal Processing*, vol. 63, no. 15, pp. 3872–3887, 2015.

[14] A. M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust estimation in signal processing: A tutorial-style treatment of fundamental concepts," *IEEE Signal Processing Magazine*, vol. 29, no. 4, 2012.

[15] A. Aubry, A. D. Maio, S. Marano, and M. Rosamilia, "Single-pulse simultaneous target detection and angle estimation in a multichannel phased array radar," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6649–6664, 2020.

[16] A. Aubry, A. De Maio, S. Marano, and M. Rosamilia, "Structured covariance matrix estimation with missing-(complex) data for radar applications via expectation-maximization," *IEEE Transactions on Signal Processing*, vol. 69, pp. 5920–5934, 2021.

[17] A. Aubry, A. De Maio, L. Lan, and M. Rosamilia, "Adaptive radar detection and bearing estimation in the presence of unknown mutual coupling," *IEEE Transactions on Signal Processing*, vol. 71, pp. 1248–1262, 2023.

[18] A. Coluccia, F. Ricciato, and P. Romirer-Maierhofer, "Robust estimation of mean failure probability in access networks," *Computer Networks*, vol. 73, pp. 282–301, 2014.

[19] A. Coluccia, "Robust estimation of the mean probability of binary events: A low-complexity minimax approach," in *2013 18th International Conference on Digital Signal Processing (DSP)*, 2013.

[20] A. Coluccia and G. Notarstefano, "Distributed bayesian estimation of arrival rates in asynchronous monitoring networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 5050–5054.

[21] P. S. Rao, J. Kaplan, and W. G. Cochran, "Estimators for the one-way random effects model with unequal error variances," *Journal of the American Statistical Association*, vol. 76, no. 373, pp. 89–97, 1981.

[22] A. Coluccia, "Robust opportunistic inference from non-homogeneous distribution-free measurements," *IEEE Transactions on Signal Processing*, vol. 64, no. 15, pp. 3945–3954, 2016.

[23] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *Journal of the American Statistical Association*, no. 47, p. 663–685, 1952.

[24] J. Robins, A. Rotnitzky, and L. Zhao, "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, no. 89, p. 846–866, 1994.

[25] M. R. Elliott, "Model averaging methods for weight trimming," *Journal of Official Statistics*, no. 24, p. 517–540, 2008.

[26] D. Basu, "An essay on the logical foundations of survey sampling, part 1," in *Foundations of Statistical Inference*, V. P. Godambe and D. A. Sprott, Eds. Toronto: Holt, Rinehart and Winston, 1971, p. 203–233.

[27] J. Rao, "Alternative estimators in pps sampling for multiple characteristics," *Sankhya A*, no. 28, p. 47–60, 1966.

[28] D. Haziza and J.-F. Beaumont, "Construction of weights in surveys: A review," *Statistical Science*, no. 32, p. 206–226, 2017.

[29] J.-F. Beaumont and A. Alavi, "Robust generalized regression estimation," *Survey Methodol.*, no. 30, p. 195–208, 2004.

[30] F. Potter, "A study of procedures to identify and trim extreme sampling weights," in *Proceedings of the Survey Research Methods Section, American Statistical Association*. Alexandria, VA: American Statistical Association, 1990, p. 225–230.

[31] K. Henry and R. Valliant, "Methods for adjusting survey weights when estimating a total," in *Proceedings of the 2012 Federal Committee on Statistical Methodology Research Conference*, 2012.

[32] J.-F. Beaumont, "A new approach to weighting and inference in sample surveys," *Biometrika*, vol. 95, no. 3, pp. 539–553, 2008.

[33] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective, 2nd ed.* Academic Press, 2020.

[34] F. Potter, "The effect of weight trimming on nonlinear survey estimates," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, 1993.

[35] D. Scharfstein, A. Rotnitzky, and J. Robins, "Adjusting for non-ignorable drop-out using semiparametric non-response models," *Journal of the American Statistical Association*, no. 94, p. 1096–1120, 1999.

[36] B. Lee, J. Lessler, and E. Stuart, "Weight trimming and propensity score weighting," *PLoS ONE*, no. 6, p. e18174, 2011.

[37] P. J. Huber and E. M. Ronchetti, *Robust Statistics, second ed.* Wiley, 2009.

[38] P. J. Huber, *Robust statistical procedures*. SIAM, 1996.

[39] A. Guruswamy, R. S. Blum, S. Kishore, and M. Bordogna, "On the optimum design of l-estimators for phase offset estimation in ieee 1588," *IEEE Transactions on Communications*, vol. 63, no. 12, 2015.

[40] D. Pastor and F.-X. Socheleau, "Robust estimation of noise standard deviation in presence of signals with unknown distributions and occurrences," *IEEE transactions on Signal Processing*, vol. 60, no. 4, pp. 1545–1555, 2012.

[41] W. J. Dixon, "Simplified estimation from censored normal samples," *The Annals of Mathematical Statistics*, no. 31, p. 385–391, 1960.

[42] D. P. Bertsekas, R. G. Gallager, and P. Humblet, *Data networks*. Prentice-Hall International New Jersey, 1992, vol. 2.

[43] A. Coluccia, A. D'Alconzo, and F. Ricciato, "On the optimality of max–min fairness in resource allocation," *annals of telecommunications-annales des télécommunications*, vol. 67, no. 1-2, pp. 15–26, 2012.

[44] S. Keshav and S. Kesahv, *An engineering approach to computer networking: ATM networks, the Internet, and the telephone network*. Addison-Wesley Reading, 1997, vol. 116.

[45] V. Kreinovich, S. Longpre, and L. Ginzburg, "Why is selecting the simplest hypothesis (consistent with data) a good idea? a simple explanation," *Bulletin of the European Association for Theoretical Computer Science (EATCS)*, no. 77, 2002.

[46] W. Kirchherr, M. Li, and P. Vitanyi, "The miraculous universal distribution," *The Mathematical Intelligencer*, no. 19, 1997.

[47] E. Jaynes, "On the rationale of maximum-entropy methods," *Proceedings of the IEEE*, no. 70, 1982.

[48] S. Guiasu and A. Shenitzer, "The principle of maximum entropy," *The Mathematical Intelligencer*, no. 7, 1985.

[49] E. Jaynes, "Information theory and statistical mechanics i/ii," *Physical Review*, no. 106/108, 1957.

[50] ——, *Notes on Present Status and Future Prospects*. Springer Netherlands, 1991, pp. 1–13.

[51] L. Roberts, "A discipline for the avoidance of unnecessary assumptions," *ASTIN Bulletin*, 1971.

[52] T. M. Cover and J. A. Thomas, *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006.

[53] S. Viviani and A. Farcomeni, "Trimmed cox regression for robust estimation in survival studies," *https://www.dss.uniroma1.it/sites/default/files/vecchie-pubblicazioni/RT_7_2010_Viviani.pdf*, 2010.

[54] S. Hicks, M. Fetter, and S. Cowles, "An evaluation of truncation estimators for improving state estimates of total hogs," *National Agricultural Statistics Service*, no. SRB research report 95-02, 1995.

[55] C.-Y. Lin, E. Kaizar, D. Faries, and J. Johnston, "A comparison of reweighting estimators of average treatment effects in real world populations," *Pharmaceutical Statistics*, no. 20, p. 765–782, 2021.

[56] R. C. Elandt-Johnson and N. L. Johnson, *Survival models and data analysis*. John Wiley & Sons, 1980, vol. 110.

[57] Y.-C. Chen, P.-E. Lu, C.-S. Chang, and T.-H. Liu, "A time-dependent sir model for covid-19 with undetectable infected persons," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 4, pp. 3279–3294, 2020.

[58] V. La Gatta, V. Moscato, M. Postiglione, and G. Sperli, "An epidemiological neural network exploiting dynamic graph structured data applied to the covid-19 outbreak," *IEEE Transactions on Big Data*, 2020.

[59] X. Chen, G. Zhu, L. Zhang, Y. Fang, L. Guo, and X. Chen, "Age-stratified covid-19 spread analysis and vaccination: A multitype random network approach," *IEEE Transactions on Network Science and Engineering*, 2021.

[60] "The new york times. (2021). coronavirus (covid-19) data in the united states," https://github.com/nytimes/covid-19-data, accessed: 2021-01-23.

[61] F. Grassi and A. Coluccia, "On the sum of random samples with bounded pareto distribution," *Signal Processing*, vol. 192, p. 108389, 2022.

[62] D. H. Bailey and P. N. Swarztrauber, "A fast method for the numerical evaluation of continuous fourier and laplace transforms," *SIAM Journal on Scientific Computing*, vol. 15, no. 5, pp. 1105–1110, 1994.

[63] D. Levin, "Fast integration of rapidly oscillatory functions," *Journal of Computational and Applied Mathematics*, vol. 67, no. 1, pp. 95–101, 1996.

[64] T. Ooura and M. Mori, "A robust double exponential formula for fourier-type integrals," *Journal of computational and applied mathematics*, vol. 112, no. 1-2, pp. 229–241, 1999.

[65] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[66] A. Beck, *Introduction to nonlinear optimization: Theory, algorithms, and applications with MATLAB*. SIAM, 2014.

[67] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

## Appendix A
## State-of-the-art estimators

### A. Grand Mean (GM)

Assuming for all measurements a white model, i.e., (2) with $\mathrm{E}[e_{i,m}] = 0$ and $\mathrm{VAR}[e_{i,m}] = \sigma^2$, the optimal estimator is the Grand Mean (GM), solution of the LS optimization problem

$$\hat{\mu}^{\text{GM}} = \arg\min_{\tilde{\mu}} \sum_{i=1}^{N} \sum_{m=1}^{n_i} (x_{i,m} - \tilde{\mu})^2. \tag{A.1}$$

Simple derivative computation yields $w_i^{\text{GM}} \propto n_i$, i.e., eq. (4). Notice that if $\sigma^2$ were known, the weighted LS approach would standardize the errors by $\sigma$, i.e., $\frac{x_{i,m}-\tilde{\mu}}{\sigma}$ would replace $x_{i,m} - \tilde{\mu}$ in (A.1); this trivially leads again to the GM.

### B. Mean of Group Means (MGM)

Under the assumption of homogeneity on local averages, i.e., $x_{i,m}$ might have unequal variances but the latter become (possibly approximately) the same after the local averaging operation, the optimal LS estimator is the so-called Mean of Group Means (MGM) with constant weights $w_i^{\text{MGM}} \overset{\text{def}}{=} 1/N$: it solves the problem

$$\hat{\mu}^{\text{MGM}} = \arg\min_{\tilde{\mu}} \sum_{i=1}^{N} (\hat{\theta}_i - \tilde{\mu})^2 \tag{A.2}$$

which returns eq. (5).

### C. Best Linear Unbiased Estimator (BLUE)

To cope with the more general non-homogeneous case of unequal variances

$$\mathrm{VAR}[e_{i,m}] = \bar{\sigma}_i^2 \tag{A.3}$$

the (weighted) LS approach standardizes each error by the corresponding standard deviation. The resulting estimator, often called best linear unbiased estimator (BLUE) [1], [5], [6], is thus the solution of the problem

$$\hat{\mu}^{\text{BLUE}} = \arg\min_{\tilde{\mu}} \sum_{i=1}^{N} \sum_{m=1}^{n_i} \left( \frac{x_{i,m} - \tilde{\mu}}{\bar{\sigma}_i} \right)^2 \tag{A.4}$$

which by a simple calculation of the derivative yields eq. (6).

### *D. Analysis-of-Variance (ANOVA)*

The non-homogeneous model can be further generalized as the so-called *random-effect model*. Indeed, to accommodate the different local conditions resulting from inter-group variability, a hierarchical model is considered in which measurements $x_{i,m}$, $m = 1, \dots, n_i$ have unknown pdf depending on a local random variable $\theta_i$, with variance $\sigma_\theta^2$ in addition to the group variances $\sigma_i^2$, regarded instead as deterministic parameters. More specifically, heterogeneous noise terms $e_{i,m}$ are modeled as the sum of two random components, i.e.

$$e_{i,m} = \beta_i + \tilde{e}_{i,m} \tag{A.5}$$

which the observation model (2) can be rewritten as

$$\begin{cases} x_{i,m} = \theta_i + \tilde{e}_{i,m} \\ \theta_i = \mu + \beta_i \end{cases} \tag{A.6}$$

under the assumptions that $\beta_i$ and $\tilde{e}_{i,m}$ are (independent) zero-mean random variables with[11]

$$\mathrm{VAR}[\beta_i] = \sigma_\theta^2, \text{ hence } \mathrm{E}[\theta_i] = \mu \tag{A.7}$$

$$\mathrm{VAR}[\tilde{e}_{i,m}] = \sigma_i^2, \text{ hence } \mathrm{E}[x_{i,m}|\theta_i] = \theta_i \tag{A.8}$$

which imply that $\mathrm{E}[\mathrm{E}[x_{i,m}|\theta_i]] = \mu$. Denoting by $f_\Theta$ the unknown common pdf of $\theta_i$s, and assuming that the first two moments of $f_\Theta$ exist, the law of total variance can be used to compute the variances of $x_{i,m}$ required by the LS approach. It can be then straightforwardly shown for eq. (1) that $\mathrm{VAR}[\hat{\theta}_i] = \sigma_\theta^2 + \sigma_i^2/n_i$, to be used as inverse weights for $\hat{\theta}_i$s; then, replacing the true (unknown) variances with sample estimators leads to the ANOVA approach in eq. (8) [21].

### *E. Minimum-Variance Linear Unbiased Estimator (MVLUE)*

A limitation of the ANOVA estimator is that it requires sufficiently large values of the sample sizes $n_i$s to properly estimate the unknown variances. To cope with the challenging scenario where $n_i$ can be very small and with wide variability (heavy-tailed distribution), for which the ANOVA approach would be very inaccurate, in [22] not only the $\theta_i$s but also the variances $\sigma_i^2$s are modeled as random variables. Denoting by $f_\Sigma$ the unknown pdf of the latter and again assuming the first two moments exist, we have that

$$\sigma_i^2 \overset{\text{def}}{=} \mathrm{VAR}[x_{i,m}|\theta_i, \sigma_i^2] \in \mathbb{R}_+ \tag{A.9}$$

is the (conditional) variance of $x_{i,m}$ and

$$\sigma_\theta^2 \overset{\text{def}}{=} \mathrm{VAR}[\theta_i] \in \mathbb{R}_+ \tag{A.10}$$

the (unconditional) variance of $\theta_i$; they both account for the heteroscedasticity of the random effects.

It is worth remarking two main aspects that differentiate this formulation from standard random-effect models. First, since the pdfs of the different random variables at play are unknown, optimal estimation approaches that assume specific statistical models (ML or MAP/MMSE) cannot be adopted. Second, in standard random-effect models such as ANOVA and its variants, the conditional variances $\sigma_i^2$s are deterministic parameters, to be estimated together with $\sigma_\theta^2$ (and $\mu$) [21], often assuming exactly or approximately equal sample sizes (balanced case). However, local variance estimators provide inaccurate results for small sample size, and additional issues are found in unbalanced cases characterized by significant sample size variations across groups. As said, to overcome such limitations in [22] $\sigma_i^2$s are regarded as random variables instead of deterministic parameters with mean $\mathcal{E}$ in eq. (9) and the minimum-variance linear unbiased estimator (MVLUE) is derived accordingly. Such an estimator $\hat{\mu}^{\text{MVLUE}} = \sum_{i=1}^N w_i^{\text{MVLUE}} \hat{\theta}_i$ is obtained by solving the optimization problem

$$\boldsymbol{w}^{\text{MVLUE}} = \underset{\boldsymbol{w} \geq \boldsymbol{0}, \sum_{i=1}^N w_i = 1}{\arg\min} \mathrm{VAR}[\hat{\mu}(\boldsymbol{w})] \tag{A.11}$$

where

$$\mathrm{VAR}[\hat{\mu}(\boldsymbol{w})] = \sum_{i=1}^n w_i^2 \, \mathrm{VAR}[\hat{\theta}_i] \tag{A.12}$$

and the constraint $\sum_{i=1}^N w_i = 1$ serves to obtain an unbiased estimator, being $\mathrm{E}[\hat{\mu}] = \sum_{i=1}^N w_i \mathrm{E}[\mathrm{E}[\hat{\theta}_i]] = \mu \sum_{i=1}^N w_i$. From the law of total variance it follows that [22]

$$\mathrm{VAR}\left[\hat{\theta}_i\right] = \sigma_\theta^2 + \frac{1}{n_i}\mathcal{E} \tag{A.13}$$

hence problem (A.11) is equivalent to

$$\boldsymbol{w}^{\text{MVLUE}} = \underset{\boldsymbol{w} \geq \boldsymbol{0}, \sum_{i=1}^N w_i = 1}{\arg\min} \sum_{i=1}^N w_i^2 \left(1 + \frac{\gamma}{n_i}\right) \tag{A.14}$$

where $\gamma$ in eq. (11) is the ratio of the statistics $\mathcal{E}$ and $\sigma_\theta^2$ describing local and global variability, respectively. Indeed, (A.13)-(A.14) indicate that variability in each group has both global ($\sigma_\theta^2$) and local ($\mathcal{E}$) causes, and a larger sample size can mitigate only the latter. Fig. 16 summarizes the scenario.

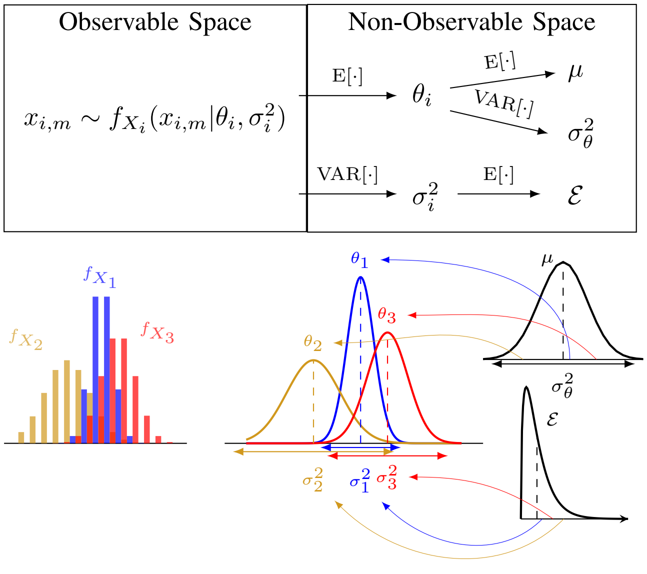


Fig. 16: Schematic representation of the hierarchical model with relevant distributions and parameters. From left to right: observable empirical data (represented as histograms), corresponding unknown theoretical pdfs with relevant parameters (Gaussian pdfs are used only for the sake of visualization), and underlying process-related unknown pdfs for the hyperparameters.

[11] It is actually sufficient to assume that the conditional mean of the measurements is proportional to $\theta_i$, i.e., $\mathrm{E}[x_{i,m}|\theta_i] \propto \theta_i$, since it is always possible to renormalize the data to get rid of the proportionality factor.

The optimal solution of (A.14) is eq. (10) [22]. A low $\gamma$ corresponds to limited local variability, where the intra-group fluctuation of the $x_{i,m}$s around their mean value $\theta_i$ is significantly smaller than the inter-group dispersion of the $\theta_i$'s. In that case, the problem setup approximately meets the homogeneous assumption of the MGM. Conversely, from (A.14) it follows that a large value of $\gamma$ yields a proportional increase of the total estimation variance, and the actual possibility to end up with a satisfactory estimate depends on the availability of sufficiently large samples. In that case, the strategy of the GM is close to optimal, since local estimates are weighted proportionally to their sample size, so limiting the contribution of small-size groups to the total variance.

## Appendix B
## Proof of Proposition 3

To start with, let us recall that in the general case the pdf $f_{\tilde{W}}$ characterizing the normalized weights $\tilde{w}_i = \frac{\tilde{z}_i}{\tilde{z}_i + \tilde{y}_i}$, where $\tilde{z}_i = g(n_i)$ and $\tilde{y}_i = \sum_{j \neq i}^{N} \tilde{z}_j$, can be written as [22]

$$\begin{aligned} f_{\tilde{W}}(\tilde{w}) &= \frac{1}{\tilde{w}^2} \int \tilde{z} f_{\tilde{Z}}(\tilde{z}) f_{\tilde{Y}}\left(\tilde{z}\left(\frac{1}{\tilde{w}} - 1\right)\right) \mathrm{d}\tilde{z} \\ &= \frac{1}{\tilde{w}^2} \int \tilde{z} f_{\tilde{Z}}(\tilde{z}) f_{\tilde{Y}}\left(\tilde{z}\tilde{w}'\right) \mathrm{d}\tilde{z} \end{aligned} \tag{B.1}$$

where $f_{\tilde{Z}}$ and $f_{\tilde{Y}}$ are the pdfs of $\tilde{z}_i$ and $\tilde{y}_i$, respectively, and we introduced the shorthand $\tilde{w}' \overset{\text{def}}{=} 1/\tilde{w} - 1$. We are interested in developing further (B.1), which holds true for any choice of $\tilde{w}_i$, for the case at hand. To this aim, we denote $z_i \overset{\text{def}}{=} \min(n_i, \tau)$, whereas $y_i$, $w_i$, and $w_i'$ follow immediately from the definitions above. Then, we rewrite

$$w_i = \frac{\chi_i n_i + (1 - \chi_i)\tau}{\sum_{j=1}^{N} [\chi_j n_j + (1 - \chi_j)\tau]} = \frac{\tau + \chi_i(n_i - \tau)}{\tau N + \sum_{j=1}^{N} \chi_j(n_j - \tau)}$$

where $\chi_i$s are i.i.d. Bernoulli random variables taking on value 1 with probability $p \overset{\text{def}}{=} \Pr\{n_i < \tau\} = F_S(\tau)$. Clearly, $Z$ is a mixed random variable and its probability function can be written as $f_Z(z) = (1-p)\delta(z - \tau) + f_S(z)\mathbb{1}_{\{z<\tau\}}$. The common pdf of the $y_i$s can be obtained by means of the law of total probability as

$$\begin{aligned} f_Y(y) &= (1-p)^{N-1} \delta\left(y - \tau(N-1)\right) \\ &\quad + \sum_{k=1}^{N-1} \binom{N-1}{k} p^k (1-p)^{N-1-k} f_{T_k}(y) \\ &\overset{\text{def}}{=} f_Y^{(d)}(y) + f_Y^{(c)}(y) \end{aligned} \tag{B.2}$$

where $f_Y^{(d)}$ and $f_Y^{(c)}$ denote the discrete and continuous part of $f_Y$, and $f_{T_k}$ is the pdf of the sum of $k$ i.i.d. random variables distributed as $S$, but conditioned on $S < \tau$, i.e., the sum of $k$ truncated random variables. Hence, (B.1) can be rewritten as the sum of four contributions, that is eq. (17) once recalling that $w' = \frac{1-w}{w}$ and denoting by $\mathcal{I}_Z \cap \mathcal{I}_Y$ the intersection of the supports of $f_Z$ and $f_Y$.

For the uniform distribution, assuming $a < \tau < b$, the pdf of $Z$ becomes $f_Z(z) = (1-p)\delta(z-\tau) + \frac{1}{b-a}\mathbb{1}_{\{a \le z < \tau\}}$, where $p = \frac{\tau - a}{b - a}$. The pdf of $y_i$ can be obtained by noticing that $f_{T_k}$ appearing in (B.2) is the pdf of the sum of $k$ random variables uniformly distributed in $(a, \tau)$, called Irwin-Hall distribution:

$$f_{T_k}(t) = \frac{(\tau - a)^{-1}}{2(k-1)!} \sum_{\ell=0}^{k} (-1)^\ell \binom{k}{\ell} \frac{(\varphi_k(t) - \ell)^k}{|\varphi_k(t) - \ell|} \tag{B.3}$$

where the function $\varphi_k(t) = (t - ka - \tau(N-k))/(\tau - a)$ shifts and rescales each term by a quantity depending on the number of $n_i$s below and above $\tau$.

Let us now consider the pdf of the normalized weights $f_W$. Firstly, we observe that the integration domain can be written as $\mathcal{I}_Z \cap \mathcal{I}_Y = [a \max(1, \frac{N-1}{w}), \tau \min(1, \frac{N-1}{w})] = [A, B]$. By isolating one factor of the $k$-th power appearing in (B.3), the sign function is obtained. Then, substituting (B.3) in (17), we observe that the order of the two summations and the integral can be exchanged, and the integration domain splits in two terms with opposite sign. The splitting value is obtained as

$$\frac{\varphi_k(zw') - \ell}{|\varphi_k(zw') - \ell|} = \begin{cases} 1 & z > \frac{(\tau-a)(\ell-k)+\tau(N-1)}{w'} = \rho \\ -1 & \text{otherwise} \end{cases} . \tag{B.4}$$

Finally, the integral in (17) can be rewritten as

$$\begin{aligned} \Phi_k(zw') &= \int z \left(\varphi_k\left(zw'\right) - \ell\right)^{k-1} \mathrm{d}z \\ &= -\left(\frac{\tau - a}{w'}\right)^2 \frac{(\varphi_k(-kzw') - \ell)(\varphi_k(zw') - \ell)^k}{k(k+1)} \end{aligned} \tag{B.5}$$

and evaluated as

$$\begin{cases} \Phi_k(B) - \Phi_k(A) & \rho < A \\ \Phi_k(B) - \Phi_k(A) - 2\Phi_k(\rho) & A < \rho < B \; . \\ \Phi_k(A) - \Phi_k(B) & \rho > B \end{cases} \tag{B.6}$$

When the sample size is Pareto-distributed, the common pdf of the $z_i$s is $f_Z(z) = (1-p)\delta(z - \tau) + \frac{\alpha}{z^{\alpha+1}}\mathbb{1}_{\{1 \le z < \tau\}}$, with $p = (1 - \tau^{-\alpha})\mathbb{1}_{\{\tau \ge 1\}}$. In the Pareto case the distribution of $T_k$ is available in terms of the integral function

$$\begin{aligned} f_{T_k}(y) = \frac{1}{\pi} \Bigg[ \int_0^\infty & \mathrm{e}^{-\mathrm{i}\left(\left(y + k(\beta - \mu)\right)\xi + \frac{\alpha k \pi}{2}\right)} \Bigg( \frac{\alpha}{p} (\beta\xi)^\alpha \\ & \left(\Gamma(-\alpha, -\mathrm{i}\beta\xi) - \Gamma(-\alpha, -\mathrm{i}(\beta + \tau - \mu)\xi)\right) \Bigg)^k \mathrm{d}\xi \Bigg] \end{aligned} \tag{B.7}$$

where $\Gamma(\cdot, \cdot)$ is the upper incomplete Gamma function. As discussed in [61], the integral in (B.7) is essentially a shifted version of the Fourier transform and it can be efficiently computed by leveraging numerical methods for integration of functions expressed as product between an oscillatory part and decaying part [62]–[64].

## Appendix C
## Proof of Proposition 4

Since $w_i$ is a function of $n_i$ only, and $n_i$s are i.i.d. (when regarded as random variables), all $w_i$ have the same mean. Thus, computing the expected value at both sides of $\sum_i w_i = 1$, one obtaines $N\mathrm{E}[w_i] = 1$ from which it follows that the mean of any linear unbiased estimator of $\mu$ is equal to $1/N$ irrespective of the distribution of the $n_i$s.

Moreover, since $z_i = \min(n_i, \tau)$ is an increasing function of $\tau$, and $w_i = z_i / \sum_i z_i$, by sorting $n_i$s in ascending order it trivially follows that $z_i \leq z_i' \overset{\text{def}}{=} \min(n_i, \tau')$ for $\tau' \geq \tau$. Likewise, $\sum_i z_i' \geq \sum_i z_i$. Thus, $w_i' \overset{\text{def}}{=} z_i' / \sum_i z_i'$ will have $z_i' = z_i$ for $n_i \leq \tau$ and $z_i' > z_i$ for larger values of $n_i$, and all $w_i'$ will be divided by a normalization factor $\sum_i z_i'$ larger than that appearing in $w_i$. The result is that $w_i < w_i'$ for $n_i < \tau$, and vice versa for the upper range where the saturation to the value $\tau'$ exceeds that to the value $\tau$, so increasing the dispersion of weights with respect to their mean value $1/N$. From these observations it follows that in general the variance of $w_i$s is always increasing in $\tau$, hence $\tau = \min(\boldsymbol{n})$ (or smaller) yields the minimum variance for the weights, while $\tau = \max(\boldsymbol{n})$ (or larger) yields maximum variance. As $\sum_i (w_i - 1/N)^2 = \|\boldsymbol{w}\|_2^2 - 1/N^2$, the thesis follows.

## Appendix D
## Proof of Proposition 5

Being the elements of $\boldsymbol{n}$ sorted in ascending order (without loss of generality), eq. (13) can be conveniently rewritten as

$$w_i^{\text{LUE-S}} = \frac{(\tau - n_i)\mathbb{1}_{\{\tau < n_i\}} + n_i}{\sum_{j=1}^{N} (\tau - n_j)\mathbb{1}_{\{\tau < n_j\}} + n_j}. \tag{D.1}$$

Substituting back in (A.12) gives the following expression

$$\text{VAR}[\hat{\mu}^{\text{LUE-S}}] \propto \frac{\sum_{i=1}^{N} \gamma_i \left[(\tau - n_i)\mathbb{1}_{\{\tau < n_i\}} + n_i\right]^2}{\left(\sum_{j=1}^{N} \left[(\tau - n_j)\mathbb{1}_{\{\tau < n_j\}} + n_j\right]\right)^2} \tag{D.2}$$

where we used the shorthand $\gamma_i = \text{VAR}[\hat{\theta}_i]$ (ref. eq. (18)).

Let us consider the vector $\breve{\boldsymbol{n}}$ with element $\breve{n}_\ell$ and length $\breve{N} \leq N$ containing the unique elements of $\boldsymbol{n}$. For $\tau \in [\breve{n}_\ell, \breve{n}_{\ell+1}]$, the sums in (D.2) split in two parts, for $n_i \leq \tau$ and $n_i > \tau$, or equivalently for $i \leq k_\tau$ and $i > k_\tau$, respectively, where $k_\tau \in \{1, \ldots, N\}$ counts the number of $n_i < \tau$. By introducing the map $\text{VAR}_\ell[\hat{\mu}^{\text{LUE-S}}](\tau) \colon [\breve{n}_\ell, \breve{n}_{\ell+1}] \mapsto \mathbb{R}_+$

$$\text{VAR}_\ell[\hat{\mu}^{\text{LUE-S}}] = \frac{\sum_{i=1}^{k_\tau} n_i^2 \gamma_i + \tau^2 \sum_{i=k_\tau+1}^{N} \gamma_i}{\left(\sum_{j=1}^{k_\tau} n_j + (N - k_\tau)\tau\right)^2}, \quad \tau \in [\breve{n}_\ell, \breve{n}_{\ell+1}] \tag{D.3}$$

where we have omitted the explicit dependency on $\tau$ to simplify the notation, the total variance (D.2) can be rewritten as union of functions with disjoint supports, i.e.

$$\text{VAR}[\hat{\mu}^{\text{LUE-S}}] = \bigsqcup_{\ell=1}^{\breve{N}-1} \Big\{ (\tau, \text{VAR}_\ell[\hat{\mu}^{\text{LUE-S}}]) \bigm| \tau \in [\breve{n}_\ell, \breve{n}_{\ell+1}] \Big\}. \tag{D.4}$$

Consequently, (D.4) is a piecewise function at least continuous over its domain, hence by the Weierstrass theorem it admits a global minimum. The global minimizer, i.e., the optimal threshold $\tau^*$, thus belongs to the set of all the minimizers $\tau_\ell^*$ of $\text{VAR}_\ell[\hat{\mu}^{\text{LUE-S}}]$ for $\ell = 1, \ldots, \breve{N} - 1$. Moreover, (D.3) is the ratio of two quadratic functions in $\tau$, hence its derivative is easily computed and shows that there may exist at most one stationary point over each interval. Since the function is decreasing/increasing before/after such a point, the latter qualifies as a possible local minimum only if it falls in the interior of the interval; otherwise, the minimum over the considered interval is attained at one of the two boundaries (left or right, see inset in Fig. 4 which provides an illustrative example). It is easy to show that the root of the derivative of (D.3) (with respect to $\tau$) is given by eq. (20), hence the expression for the set of points $\tau_\ell^*$ reported in eq. (19) follows. The global minimizer $\tau^*$ is readily obtained by evaluating (D.3) at all $\tau_\ell^*$ for $\ell = 1, \ldots, \breve{N} - 1$ and selecting the one that yields the smallest value.

## Appendix E
## Derivation of the ELUE-S as approximate MVLUE

Under these approximations, the optimization problem (A.14) can be conveniently rewritten with a cost function split in two parts, with associated Lagrangian function

$$\sum_{i=1}^{k_\gamma} w_i^2 \frac{\gamma}{n_i} + \sum_{i=k_\gamma+1}^{N} w_i^2 + \lambda(\sum_{i=1}^{N} w_i - 1)$$

where $k_\gamma \in \{1, \ldots, N\}$ denotes the maximum integer such that $n_{k_\gamma} \leq \gamma$. By taking the derivative with respect to $w_i$ and posing equal to zero, we obtain

$$\begin{cases} 2w_i \dfrac{\gamma}{n_i} + \lambda = 0 & i = 1, \ldots, k_\gamma \\ 2w_i + \lambda = 0 & i = k_\gamma + 1, \ldots, N \end{cases}$$

from which the value of $w_i$ as function of the Lagrange multiplier $\lambda$ easily follows. By exploiting the constraint $\sum_{i=1}^{N} w_i = 1$ we obtain $\lambda = -\frac{2}{\frac{1}{\gamma}\sum_{i=1}^{k_\gamma} n_i + (N - k_\gamma)}$ and, finally,

$$w_i = \begin{cases} \dfrac{n_i}{\sum_{i=1}^{k_\gamma} n_i + \gamma(N - k_\gamma)} & n_i \leq \gamma \\ \frac{\gamma}{\sum_{i=1}^{k_\gamma} n_i + \gamma(N - k_\gamma)} & n_i > \gamma \end{cases} \tag{E.1}$$

which is equivalent to the more convenient expression

$$w_i = \frac{\min(n_i, \gamma)}{\sum_{j=1}^{N} \min(n_j, \gamma)}. \tag{E.2}$$

From (E.2) it is immediate to realize that such weights are formally identical to those of the LUE-S (13) for $\tau = \gamma$.