Bernoulli **25**(4B), 2019, 3459–3478
https://doi.org/10.3150/18-BEJ1096

Nonparametric empirical Bayes improvement of shrinkage estimators with applications to time series

EITAN GREENSHTEIN[1], ARIEL MANTZURA[2] and YA'ACOV RITOV[3]

[1]*Israel Central Bureau of Statistics, Kanfei Nesharim 66, 9546456 Jerusalem, Israel.*
E-mail: eitan.greenshtein@gmail.com
[2]*Bank of Israel; Eliezer Kaplan 3, P.O. Box 780, 91007 Jerusalem, Israel.*
E-mail: ariel.mansura@boi.org.il
[3]*University of Michigan and the Hebrew University of Jerusalem; Department of Statistics, University of Michigan, 1085 South University, Ann Arbor, MI 48109-1107, USA. E-mail: yritov@umich.edu*

We consider the problem of estimating a vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ under a squared loss, based on independent observations $Y_i \sim N(\mu_i, 1)$, $i = 1, \ldots, n$, and possibly extra structural assumptions. We argue that many estimators are asymptotically equal to $\hat{\mu}_i = \alpha\tilde{\mu}_i + (1-\alpha)Y_i + \xi_i = \tilde{\mu}_i + (1-\alpha)(Y_i - \tilde{\mu}_i) + \xi_i$, where $\alpha \in [0, 1]$ and $\tilde{\mu}_i$ may depend on the data, but is not a function of Y_i, and $\sum \xi_i^2 = o_p(n)$.

We consider the optimal estimator of the form $\tilde{\mu}_i + g(Y_i - \tilde{\mu}_i)$ for a general, possibly random, function g, and approximate it using nonparametric empirical Bayes ideas and techniques. We consider both the retrospective and the sequential estimation problems. We elaborate and demonstrate our results on the case where $\hat{\mu}_i$ are Kalman filter estimators. Simulations and a real data analysis are also provided.

Keywords: empirical Bayes; exchangeable; Kalman filter; shrinkage estimators

1. Introduction

Consider the problem of estimating the values of $\mu_1, \ldots, \mu_n$ based on the observations $Y_1, \ldots, Y_n$, where μ_i may be deterministic or random, and conditional on the μs, $Y_1, \ldots, Y_n$ are independent, and $Y_i \sim N(\mu_i, 1)$. We write $Y_i = \mu_i + \varepsilon_i$. We use bold to denote a vector representation of sample values, e.g., $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$.

The performance of an estimator $\boldsymbol{\delta} = \boldsymbol{\delta}(\boldsymbol{Y})$ for $\boldsymbol{\mu}$ is evaluated according to its squared error risk: $E\|\boldsymbol{\delta} - \boldsymbol{\mu}\|^2$. The estimator $\boldsymbol{\delta}_1$ *strictly* asymptotically improves upon $\boldsymbol{\delta}_2$ if $E\|\boldsymbol{\delta}_1 - \boldsymbol{\mu}\|^2 \le cE\|\boldsymbol{\delta}_2 - \boldsymbol{\mu}\|^2 + o(n)$ for some $0 \le c < 1$. If the last inequality holds with $c = 1$, we say that $\boldsymbol{\delta}_1$ asymptotically improves upon $\boldsymbol{\delta}_2$.

In situations where $Y_1, \ldots, Y_n$ are perceived as "exchangeable", the nonparametric empirical Bayes approach is appealing. In this paper, "exchangeable" is not meant in the usual sense in probability, in particular, for deterministic μs, the values $\mu_1, \ldots, \mu_n$ are not assumed to be all equal. By "perceived exchangeable" we mean that the statistician does not distinguish between $Y_1, \ldots, Y_n$ a priory, that is, the index does not carry any information. Therefore, an appealing approach is to apply the same estimating function $\delta(Y_i)$, when estimating each μ_i, $i = 1, \ldots, n$, that is, confined to estimators of the type $\boldsymbol{\delta}(\boldsymbol{Y}) = (\delta(Y_1), \ldots, \delta(Y_n))$. Such estimation functions are called coordinate-wise.

A central task in empirical Bayes theory is to approximate the optimal coordinate-wise $\boldsymbol{\delta}$ in a 'perceived exchangeable' setup. However, we are interested in situations where $Y_1, \ldots, Y_n$ should *not* be treated as exchangeable, due to structural conditions, covariates, and assumptions. For example, suppose $Y_1, Y_2, \ldots$ is a time series modeled by a state-space. In such situations we strive to transform the setup into one that may be naturally treated as exchangeable and thus classical empirical Bayes approaches may be naturally applied.

Our suggested method yields a way to improve the performance of some classical estimators, which are described in the following through their general 'canonical' presentation and through examples.

Canonical estimators $\hat{\boldsymbol{\mu}}$. As the examples below show, many estimators of μ_i are built out of two components. One is the the value of Y_i itself, and the other is an estimator $\tilde{\mu}_i$ based on the rest of the observations. Typically, the final estimate is approximately a linear combination of the two:

$$\begin{aligned}\hat{\mu}_i &= \alpha\tilde{\mu}_i + (1-\alpha)Y_i + \xi_i \\ &= \tilde{\mu}_i + (1-\alpha)(Y_i - \tilde{\mu}_i) + \xi_i, \qquad i = 1, \ldots, n,\end{aligned}$$

where the predictor $\tilde{\mu}_i$ is a parametric function of the available data excluding Y_i, and the ξ_i are some error terms, satisfying $E\|\boldsymbol{\xi}\|^2 = o(n)$, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$. In vector notations:

$$\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}} + (1-\alpha)(\boldsymbol{Y} - \tilde{\boldsymbol{\mu}}) + \boldsymbol{\xi}. \tag{1}$$

1.1. Examples

1.1.1. *Stein estimator*

The Stein estimator is of special importance (see, e.g., Lehmann and Casella [15]). The standard form of the James–Stein estimator is that of (1) with a non-random $\tilde{\boldsymbol{\mu}}$, typically $\tilde{\mu}_i \equiv 0$:

$$\hat{\mu}_i = \tilde{\mu}_i + \left(1 - \frac{n-2}{\|\boldsymbol{Y} - \tilde{\boldsymbol{\mu}}\|^2}\right)(Y_i - \tilde{\mu}_i).$$

The above class of estimators may be viewed as shrinking towards $\tilde{\boldsymbol{\mu}}$. In the classical James–Stein estimator, $\tilde{\boldsymbol{\mu}}$ does not depend on the data while α depends on $\tilde{\boldsymbol{\mu}}$ and the data:

$$\alpha = \frac{1}{1 + \|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^2/n} = \frac{n-2}{\|\boldsymbol{Y} - \tilde{\boldsymbol{\mu}}\|^2} + o_p(1).$$

Moreover, the smaller $\|\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}\|^2$ is, the smaller is the risk of the corresponding Stein estimator. In particular, if we shrink towards the true mean, i.e., $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}$, the risk of the corresponding Stein estimator is $o(n)$.

Asymptotically, the James–Stein estimator converges to the best linear correction of the vector of *a priori* guesses, and to the Bayes estimator if $v_i = \mu_i - \tilde{\mu}_i$, $i = 1, \ldots, n$, are i.i.d. normal with mean 0 (Efron and Morris [11]). This motivates shrinking towards a data dependent $\tilde{\boldsymbol{\mu}}$ which is a good initial guess for $\boldsymbol{\mu}$, see, for example, $\tilde{\mu}_i \equiv \bar{Y}$.

1.1.2. *Fay–Herriot and small area estimation*

Fay–Herriot [12] generalized the above line of thought to the case where there are explanatory variables. Suppose that each Y_i is accompanied by a vector X_i of explanatory variables, independent of ε_i. Let $\hat{\beta}$ be the usual least squares, $\hat{\beta} = \arg\min_b \|\boldsymbol{Y} - Xb\|^2$, where X is a design matrix. Then, under mild conditions, we expect $\hat{\beta} \xrightarrow{\text{p}} \beta$, which is independent of $\varepsilon_1, \ldots, \varepsilon_n$. We now apply the previous arguments to the estimator

$$\hat{\boldsymbol{\mu}} = X\hat{\beta} + \left(1 - \frac{n}{\|\boldsymbol{Y} - X\hat{\beta}\|^2}\right)(\boldsymbol{Y} - X\hat{\beta}).$$

Letting $X\beta$ play the role of $\tilde{\boldsymbol{\mu}}$, the last equation has the form (1).

The Fay–Herriot type estimators play a major role in *small area estimation* (Rao [16]). A treatment of such estimators in this context, in the spirit of the current paper, is given in, Cohen et al. [8]. More generally, estimators of the type (1) are used in small area estimation and termed 'composite estimators'. The $\tilde{\boldsymbol{\mu}}$ term is often referred to as the 'synthetic' part of $\hat{\boldsymbol{\mu}}$. The synthetic part is based on model assumptions and information 'borrowed' from neighboring areas.

1.1.3. *Kalman filter*

Consider a state-space model, for example, μ_i is a Gaussian ARIMA process and $Y_i = \mu_i + \varepsilon_i$, $\varepsilon_i \sim N(0, 1)$, ε_i are independent and independent of μ_i, $i = \ldots, -1, 0, 1, 2, \ldots$.

Let $\mathcal{D}_i$ be the index set of the available observations at time i and excluding the ith observation. In a retrospective estimation $\mathcal{D}_i = (\ldots, i-2, i-1, i+1, i+2, \ldots)$ and in sequential estimation $\mathcal{D}_i = (\ldots, i-3, i-2, i-1)$.

Suppose the random $\ldots \mu_{-1}, \mu_0, \mu_1, \ldots$ is a Gaussian process. Let $\tilde{\mu}_i = E(\mu_i | Y_j, j \in \mathcal{D}_i)$. Then, under squared loss, the optimal Kalman-filter estimator for μ_i is:

$$\begin{aligned}
\hat{\mu}_i &\equiv E\big(\mu_i | Y_j, j \in \mathcal{D}_i \cup \{i\}\big) \\
&= \alpha_i E(\mu_i | Y_j, j \in \mathcal{D}_i) + (1 - \alpha_i) Y_i \\
&= \alpha_i \tilde{\mu}_i + (1 - \alpha_i) Y_i \\
&= \tilde{\mu}_i + (1 - \alpha_i)(Y_i - \tilde{\mu}_i).
\end{aligned}$$

In the above $\alpha_i = 1/(\tau_i^2 + 1)$, where τ_i^2 is the variance of μ_i given $\{Y_j, j \in \mathcal{D}_i\}$. The second equality above is obtained due to the Gaussianity of the μ-process, regardless of special state-space structure. This is a shrinkage estimator that shrinks towards a random $\tilde{\mu}_i$. When the μ-process is stationary $\alpha_i = \alpha + o(1)$. In a stationary state space $\hat{\mu}$ may be presented as a linear combination of the observed Ys, see, for example, Brockwell and Davis [2], it may be seen that $(1 - \alpha_i)$ equals to the coefficient of Y_i under the linear representation of $\hat{\mu}_i$. When the parameters of the stationary state space are known, the canonical representation is obvious given the coefficients of the observed Ys and also by the above derivation; this is also true in typical scenarios where the coefficients are consistently estimated in appropriate rates under stationarity, see Brockwell and Davis [2].

When the μ-process is indeed Gaussian the above estimators cannot be asymptotically improved. However, consider for example a situation where the μ-process is a *non-Gaussian ARMA*. We will argue in the sequel that the Kalman-filter estimator, although optimal among linear filters, is not globally optimal and may be asymptotically strictly improved.

Summary: As demonstrated above, under a stationary state space model $\hat{\mu}_i$ has the canonical form with an appropriate α. The estimator $\hat{\mu}_i$ is optimal when the μ process is Gaussian, it is the best linear estimator under a general state-space model regardless of whether the μ-process is Gaussian. The canonical form is obtained also when the parameters of the assumed model are estimated.

When the true model is in fact different than the assumed model whose parameters are estimated, then, under mild conditions, the estimator $\hat{\mu}_i$ still has the canonical form if the true model is stationary. This follows since due to stationarity the estimated parameters (under the wrong model) converge.

1.2. The main ideas

The James–Stein estimator with $\tilde{\mu}_i \equiv 0$ and more general compound decision and empirical Bayes analyses for estimating $\boldsymbol{\mu}$ are appealing when $Y_1, \ldots, Y_n$ are perceived as exchangeable, and μ_i are estimated in a coordinate-wise manner. This is not the situation we study. For instance, consider the case where the μ-process is an AR(1). In a setup where $Y_1, Y_2, \ldots$ are not perceived as exchangeable, we should not use such approaches directly. We strive to transform the original problem to one that the statistician may perceive as (approximately) exchangeable.

The way to achieve exchangeability in our more general setup is by subtracting the predictor $\tilde{\mu}_i$ from Y_i. Let

$$\boldsymbol{Z} = \boldsymbol{Y} - \tilde{\boldsymbol{\mu}},$$

denote

$$\boldsymbol{\nu} = \boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}.$$

Then $Z_i = \nu_i + \varepsilon_i$. Note, ε_i is independent of μ_i, and $\tilde{\mu}_i$ is not a function of Y_i, thus ε_i is independent of ν_i. Therefore:

$$\mathcal{L}(Z_i | \boldsymbol{\mu}, \nu_i) = N(\nu_i, 1), \qquad i = 1, \ldots, n. \tag{2}$$

The sequence $Z_1, \ldots, Z_n$ may be considered as exchangeable, since prior to observing Z_i, the best guess of the statistician for the value of $E(Z_i) = \nu_i$ is zero. Therefore, estimating ν_i in a coordinate-wise manner by $\delta(Z_i)$ with the same δ for each i, $i = 1, 2, \ldots$ is natural.

Since $\tilde{\mu}_i$, $i = 1, \ldots, n$ are observed, estimating ν_i is equivalent to estimating μ_i, $i = 1, \ldots, n$. We can, therefore, consider the pair ν_i and its measurement Z_i instead of the pair μ_i and its measurement Y_i. However, the situation cannot be trivially reduced to a normal compound decision problem by conditioning on $\boldsymbol{\nu}$. The reason is that although the distribution of Z_i conditional on ν_i and $\boldsymbol{\mu}$ is $N(\nu_i, 1)$, the dependence of Z_i on $\boldsymbol{\nu}$ often implies that its distribution conditional on $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ is very different from $N(\nu_i, 1)$. This may be seen in the following trivial example.

Example 1.1. Let $Y_i \sim N(\mu_i, 1)$ be independent. For simplicity, the μs are non-random, and suppose $\mu_1 = \cdots = \mu_n = 0$. Let $\tilde{\mu}_i = Y_{i-1}$, $i > 1$, and let $\tilde{\mu}_1 = 0$. Then $\nu_i = -Y_{i-1}$, $i > 1$, and $\mathcal{L}(Z_i|\boldsymbol{\mu}, \nu_i) = N(\nu_i, 1)$, $i = 2, \ldots, n$. However, obviously $\mathcal{L}(Z_i|\boldsymbol{\mu}, \boldsymbol{\nu})$ is degenerate, $i = 1, \ldots, (n-1)$.

In this example, the considerations and model that lead the statistician to $\tilde{\mu}_i$ are not helpful. In fact, applying a standard nonparametric empirical Bayes procedure on the initial sequence $Y_1, \ldots, Y_n$, would yield a much better estimator for $\mu_1, \ldots, \mu_n$. The purpose of this paper is to show a way of improving canonical estimators that are based on $\tilde{\mu}_i$ and Y_i, *regardless* of whether the considerations/model that lead to $\tilde{\mu}_i$ are 'helpful'/'true'. Indeed, the assumptions stated in the next section are satisfied in this example, and our main improvement results are implied.

A key role in the following analysis is played by the distribution

$$G = G_{\boldsymbol{\mu}} \triangleq \sum_{i=1}^{n} \frac{1}{n} \mathcal{L}(\nu_i|\boldsymbol{\mu}). \tag{3}$$

It is the marginal distribution of ν_I given $\boldsymbol{\mu}$, for a randomly selected index I, uniformly distributed on $\{1, \ldots, n\}$.

Example 1.1 (continued). In light of our example, as $n \to \infty$ the distribution $G_{\boldsymbol{\mu}} = G^n_{\boldsymbol{\mu}}$ of ν_I, converges to $N(0, 1)$. Here, we need $n \to \infty$ because of the point mass at zero, which is implied since since $\nu_1 = 0$, a.s.

Our plan is the following. We condition on $\boldsymbol{\mu}$ and approximate the optimal improvement function $\tilde{\delta}$. The latter satisfies:

$$\begin{aligned} \tilde{\delta} = \tilde{\delta}_{\boldsymbol{\mu}} &\triangleq \arg\min_g E\Big(\sum \big(\mu_i - \tilde{\mu}_i - g(Y_i - \tilde{\mu}_i)\big)^2 | \boldsymbol{\mu}\Big) \\ &= \arg\min_g E\bigg(\sum_i \big(\nu_i - g(Z_i)\big)^2 | \boldsymbol{\mu}\bigg). \end{aligned} \tag{4}$$

By convexity, $\tilde{\delta}_{\boldsymbol{\mu}}$ is unique. In Lemma 1, we prove that $\tilde{\delta}$ has the representation:

$$\delta = \delta_{\boldsymbol{\mu}} = \arg\min_g \int E\big(\nu - g(Z)\big)^2 dG_{\boldsymbol{\mu}}(\nu), \tag{5}$$

that is, $\tilde{\delta}_{\boldsymbol{\mu}} = \delta_{\boldsymbol{\mu}}$.

The last minimization problem is standard and the minimizer is the standard Bayes estimator:

$$\delta_{\boldsymbol{\mu}}(z) = E(\nu|Z = z),$$

where the conditional expectation is under the model where $\nu \sim G_{\boldsymbol{\mu}}$ and $\mathcal{L}(Z|\nu) = N(\nu, 1)$. Denote:

$$\boldsymbol{\delta} = \big(\delta_{\boldsymbol{\mu}}(Z_1), \ldots, \delta_{\boldsymbol{\mu}}(Z_n)\big).$$

Similarly denote $\tilde{\boldsymbol{\delta}}$.

The task of approximating $\delta_{\boldsymbol{\mu}}$ based on $Y_1, \ldots, Y_n$, without knowing $G \equiv G_{\boldsymbol{\mu}}$ will be performed along the lines of the method suggested by Brown and Greenshtein [5] – see the following subsection. The approximation is denoted $\hat{\delta} \equiv \hat{\delta}_{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\delta}}$. We will show:

$$\hat{\delta}_{\boldsymbol{\mu}} \approx \delta_{\boldsymbol{\mu}}.$$

First, consider the following improvement of $\hat{\boldsymbol{\mu}}$:

$$\boldsymbol{\mu}^I = \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\delta}} = \tilde{\boldsymbol{\mu}} + \boldsymbol{\delta}.$$

Obviously, $\boldsymbol{\mu}^I$ improves upon the estimator $\hat{\boldsymbol{\mu}}$. This follows since $\hat{\boldsymbol{\mu}}$ is restricted to functions that are asymptotically equivalent to $\alpha\tilde{\boldsymbol{\mu}} + (1-\alpha)\boldsymbol{Y} = \tilde{\boldsymbol{\mu}} + (1-\alpha)\boldsymbol{Z}$, that is, where the term that corrects $\tilde{\boldsymbol{\mu}}$ is restricted to be proportional to $\boldsymbol{Z}$. However, $\boldsymbol{\mu}^I$ is not a function of the observed data $Y_1, \ldots, Y_n$, since $\boldsymbol{\delta}$ depends on the unknown $\boldsymbol{\mu}$. The proper estimator is based on the estimator $\hat{\boldsymbol{\delta}}$, specifically:

$$\hat{\boldsymbol{\mu}}^I = \tilde{\boldsymbol{\mu}} + \hat{\boldsymbol{\delta}}.$$

We will show that $\hat{\boldsymbol{\mu}}^I$ and $\boldsymbol{\mu}^I$ are asymptotically equivalent, whereas $\hat{\boldsymbol{\mu}}^I$ asymptotically improves upon $\hat{\boldsymbol{\mu}}$. As will be shown, the asymptotical improvement is often strict.

More on creating exchangeability. Exchangeability is commonly lost, under a heterogenous known variance setup, i.e., $\mathrm{Var}(Y_i) = \sigma_i^2$, such an heterogeneity may occur, for example, due to different sample sizes in different unit groups. An immediate way to restore homogeneity and exchangeability is to normalize through dividing Y_i by its known σ_i and estimate μ_i/σ_i, or apply some other variance stabilizing transformation. This changes the target parameter, but more importantly it might mask valuable information that could be provided by σ_i as a covariate. A possible way to handle it is to split the data into subgroups with members that have similar values of σ_i, and treat each subgroup as exchangeable. This may also be done when other covariates are present. In addition, one may apply classes of estimation functions that do not work in a coordinate-wise manner but also depend on covariates. Choosing a member from a class of non coordinate-wise estimation functions would not involve classical considerations of empirical Bayes. Choosing a concrete member from a class typically involves deriving good (unbiased) estimators of their risks. See, for example, Xie et al. [22], Brown et al. [7], Weinstein et al. [21], Banarejee et al. [1].

Our approach under heterogeneity of σ_i (and presence of other covariates) could add to the above by letting $\tilde{\mu}_i$ also depend on σ_i for reasonable functional dependence. We stress, this functional dependence is not assumed "true" in any sense, as may be seen in the sequel. This way we utilize the information provided by σ_i. After utilizing this information, normalizing through dividing by σ_i might not be too harmful. The advantage of this approach is that after transforming the problem via subtracting $\tilde{\mu}$, we may naturally use pure empirical Bayes techniques. This approach was used in a Fay–Herriot model with heterogeneous variances, in Cohen et al. [8].

1.3. Essentials of normal empirical Bayes

In this subsection we recall some facts and ideas of the normal nonparametric empirical Bayes approach. The ideas of Compound-Decision and empirical Bayes were developed by Robbins [17–19]. Copas [9], Zhang [23], and Efron [10] provide good introductions of those ideas and their various applications.

Suppose that ν is distributed G and $(Z|\nu) \sim N(\nu, 1)$. It is desired to approximate δ – the Bayes decision under squared risk, $\delta(z) = E(\nu|Z = z)$, based on the observed $Z_1, \ldots, Z_n$, $Z_i \sim N(\nu_1, 1)$ when G is completely unknown. In the following, we will give a useful representation of δ based on the unknown G.

Let

$$f(z) = \int \varphi(z - \nu)\, dG(\nu), \tag{6}$$

where φ is the standard normal probability density function. In the case where $G = G_{\boldsymbol{\mu}}$, we might want to emphasize it by writing $f = f_{\boldsymbol{\mu}}$.

The following equation (7) is known as Tweedie's formula, it may also be found in Brown [4],

$$\delta(z) = z + \frac{f'(z)}{f(z)}, \tag{7}$$

where f' is the derivative of f.

The last presentation is useful since it shows that the Bayes procedure is a function of the hard-to-estimate G only through its corresponding easy-to-estimate f and f'. We define our estimator $\hat{\delta}$ for the Bayes decision under G, through kernel estimation of f and f'. Consider, for simplicity, the two kernel estimators

$$\hat{f}(z) = n^{-1} \sum_j K_\sigma(Z_j - z),$$

$$\hat{f}'(z) = n^{-1} \sum_j K'_\sigma(Z_j - z),$$

where $K_\sigma(z) = \sigma^{-1} K(z/\sigma)$, $\sigma = \sigma_n$ and K' is the derivative of K. The specific kernel we use is the normal kernel. We define

$$\hat{\delta}^*(z) = z + \frac{\hat{f}'(z)}{\hat{f}(z)}.$$

For technical reasons, we truncate the above as follows:

$$\hat{\delta}(Z) = Z + \big(\hat{\delta}^*(Z) - Z\big) \times I\big(|\hat{\delta}^*(Z) - Z\big)| < M_n), \tag{8}$$

where $M_n \to \infty$. A convenient (though not essential) choice for us is $M_n = (\log n)^{0.15}$ (see the Appendix). In all of our simulations and real data analyses, we took $M_n \equiv 3$. The smoothing

parameter σ_n should converge to 0, but the convergence can be very slow. For simplicity, we adopt the recommendation of Brown and Greenshtein [5] for

$$\sigma_n = (\log n)^{-1/2}.$$

GMLE. Consistency results in the estimation of δ, similar to those that we derive, could be obtained by estimating the appropriate mixing distribution G via GMLE, as in Jiang and Zhang [13] or Koenker and Mizera [14]. The advantage of one approach relative to the other in terms of rates is beyond the scope of this paper.

1.4. The rest of the paper

Consider a pair of estimators $\tilde{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\mu}}$, such that:

$$\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}} + (1-\alpha)(\boldsymbol{Y} - \tilde{\boldsymbol{\mu}}) + \boldsymbol{\xi}, \qquad E\|\boldsymbol{\xi}\|^2 = o(n).$$

Assume further that $\tilde{\mu}_i \in \mathcal{F}_i \equiv \mathcal{F}(Y_j, j \in \mathcal{D}_i)$, the smallest σ-algebra that is generated by Y_j, $j \in \mathcal{D}_i$, where $\mathcal{D}_i \cup \{i\}$ is the index set of the available observations at time i.

Our main results, Theorem 3 (retrospective filtering) and Theorem 6 (sequential filtering), show that under mild conditions and for a suitably defined random function $\hat{\boldsymbol{\delta}}(\cdot)$, the estimator $\hat{\boldsymbol{\mu}}^I = \tilde{\boldsymbol{\mu}} + \hat{\boldsymbol{\delta}}(\boldsymbol{Y} - \tilde{\boldsymbol{\mu}})$ asymptotically improves upon $\hat{\boldsymbol{\mu}}$. Typically it is strictly so. Theorem 2 shows that $\hat{\boldsymbol{\delta}}$ and $\boldsymbol{\delta}$ are asymptotically equivalent.

In Section 5, we present simulation results comparing our $\hat{\boldsymbol{\mu}}^I$ with a Kalman-filter $\hat{\boldsymbol{\mu}}$, when the μ-process is a non-Gaussian AR(1). In Section 6, we analyze a real data example where various ARIMA models and their corresponding Kalman filters $\hat{\boldsymbol{\mu}}$ are applied, and their performance is compared with their improvement counterparts $\hat{\boldsymbol{\mu}}^I$.

2. Assumptions

In this section, we state all of our assumptions. No attempt was made to give the weakest possible conditions, for example, in terms of the various assumed powers of $\log n$. Our considerations were mainly to ease the presentation.

It should be noted that our results do not depend on whether the various model assumptions that lead to the $\tilde{\mu}_i$ estimators, $i = 1, 2, \ldots$ are indeed true. Our modest requirement is just that those estimators will not be too crude in the sense that ν_i, the mean of $Z_i = Y_i - \tilde{\mu}_i$, will not be too large. This is formulated in Assumption 1. That assumption implies that $\nu_1, \ldots, \nu_n$ are not too spread, which implies that the magnitude of the random variable $f(Z) \equiv f_{\boldsymbol{\mu}}(Z)$ is "large enough" with "high" probability (see the Appendix).

Recall that $G \equiv G_{\boldsymbol{\mu}}$ is the conditional distribution of the estimation error ν_i, of the initial estimator $\tilde{\mu}_i$. Our formulation is for a triangular array, where $G_{\boldsymbol{\mu}} \equiv G^n_{\boldsymbol{\mu}}$, which is expected to be tight.

Assumption 1. Let $\boldsymbol{C}_n = (-D_n, D_n)$, where $D_n = \kappa\sqrt{\log n}$. For large enough κ

$$G_{\boldsymbol{\mu}}^n(\boldsymbol{C}_n) > 1 - \frac{1}{\log n},$$

uniformly in $\boldsymbol{\mu}$.

Consider Assumption 1, in the simple context where the μ-process is a random walk with bounded steps. Consider $\tilde{\mu}_i$ that satisfy $\tilde{\mu}_i = Y_{i-1} + \mathcal{O}_p(\sqrt{\log(n)})$. Then Assumption 1 is satisfied, although the μ-process itself is non-stationary.

The following assumption is needed in order to control the variance of our kernel estimates $\hat{f}$ and $\hat{f}'$, when invoking the normal kernels $K_{\sigma_n}(Z - z)$ with bandwidth $\sigma_n = (\log n)^{-1/2}$.

Assumption 2. For $\psi = f_{\boldsymbol{\mu}}$, $f'_{\boldsymbol{\mu}}$ and for $\sigma_n = (\log n)^{-1/2}$ the following holds uniformly for $\boldsymbol{\mu}$ and z:

$$\operatorname{Var}\big(\hat{\psi}(z)|Z_i = z, \boldsymbol{\mu}\big) = \mathcal{O}\big((\log n)^{-4}\big), \qquad i = 1, \ldots, n, \tag{9}$$

$$E\big(\hat{\psi}(z)|Z_i = z, \boldsymbol{\mu}\big) = \psi(z) + \mathcal{O}(1/\log n). \tag{10}$$

The last assumption is very mild. It implies that the conditional variances of the kernel estimators approach zero in a $1/(\log n)^4$ rate, rather than the $1/n\sigma_n^2$ rate for independent observations. Our sequence is not of i.i.d. observations, but typically will be strongly mixing. Similarly, the assumption about the difference between $E(\hat{\psi}(z)|Z_i = z, \boldsymbol{\mu})$ and $\psi(z)$ is mild. The difference is smaller than $|E(\hat{\psi}(z)|Z_i = z, \boldsymbol{\mu}) - E(\hat{\psi}(z)|\boldsymbol{\mu})| + |E(\hat{\psi}(z)|\boldsymbol{\mu}) - \psi(z)|$. The second term in the last expression is the bias of the kernel estimator with a symmetric kernel and bandwidth σ_n, which is $\mathcal{O}(\sigma_n^2) = \mathcal{O}(1/\log n)$. The first term controls the affect of a single Z_i on the conditional expectation of $\hat{\psi}$, which is expected to be $\mathcal{O}(1/n\sigma_n)$ under mixing conditions.

Our last assumption is about the structure of the estimator $\hat{\boldsymbol{\mu}}$ which we improve upon. Those are the canonical estimators described in the Introduction, see also the examples.

Assumption 3. Suppose that there exists a fixed $\alpha \in [0, 1]$, $\alpha \equiv \alpha_{\boldsymbol{\mu}}$ may depend on $\boldsymbol{\mu}$, so that

$$\hat{\boldsymbol{\mu}} = \alpha\tilde{\boldsymbol{\mu}} + (1 - \alpha)\boldsymbol{Y} + \boldsymbol{\xi}, \qquad \text{where } E\big(\|\boldsymbol{\xi}\|^2|\boldsymbol{\mu}\big) = o(n) \text{ uniformly in } \boldsymbol{\mu}; \tag{11}$$

3. Retrospective empirical Bayes estimation

In this section, we consider the retrospective estimation of μ_i, where the entire data set is given. The sequential case is considered in the next section. Our goal is to approximate the ideal $\boldsymbol{\mu}^I$, by an asymptotically equivalent estimator of $\boldsymbol{\mu}$ which depends only on the data. The performance of any estimator of the form $\tilde{\boldsymbol{\mu}} + \boldsymbol{g}$ as an estimator for $\boldsymbol{\mu}$, may be evaluated through the performance of $\boldsymbol{g}$ as an estimator of $\boldsymbol{\nu}$, since: $E\|\tilde{\boldsymbol{\mu}} + \boldsymbol{g} - \boldsymbol{\mu}\|^2 = E\|\boldsymbol{g} - \boldsymbol{\nu}\|^2$.

We start by proving that the two functions defined in (4) and (5) are the same.

Lemma 1.

$$E\|\boldsymbol{\delta}-\boldsymbol{\nu}\|^2 = E\|\tilde{\boldsymbol{\delta}}-\boldsymbol{\nu}\|^2.$$

Thus, $\boldsymbol{\delta} \equiv \tilde{\boldsymbol{\delta}}$.

Proof. Since $\tilde{\delta}$ was defined as the minimizer of the LHS of (4), it is enough to show that conditional on every $\boldsymbol{\mu}$,

$$E\Big(\sum\big(\tilde{\delta}_{\boldsymbol{\mu}}(Z_i)-\nu_i\big)^2\big|\boldsymbol{\mu}\Big) \geq E\Big(\sum\big(\delta_{\boldsymbol{\mu}}(Z_i)-\nu_i\big)^2\big|\boldsymbol{\mu}\Big).$$

Denote $R(\nu, g) = E(g(Z)-\nu)^2$, for $Z \sim N(\nu, 1)$. Recall that the conditional distribution of Z_i conditional on ν_i and $\boldsymbol{\mu}$ is $N(\nu_i, 1)$ – see (2). For a random permutation π,

$$\begin{aligned}
E\Big(\sum\big(\tilde{\delta}_{\boldsymbol{\mu}}(Z_i)-\nu_i\big)^2\big|\boldsymbol{\mu}\Big) &= nE\big(\tilde{\delta}_{\boldsymbol{\mu}}(Z_{\pi(1)})-\nu_{\pi(1)}\big)^2\big|\boldsymbol{\mu}\big) \\
&= nE\Big(E\Big(\big(\tilde{\delta}_{\boldsymbol{\mu}}(Z_{\pi(1)})-\nu_{\pi(1)}\big)^2\big|\boldsymbol{\mu},\nu_{\pi(1)}\Big)\big|\boldsymbol{\mu}\Big) \\
&= n\int R(\nu,\tilde{\delta}_{\boldsymbol{\mu}})\,dG_{\boldsymbol{\mu}}(\nu) \\
&\geq n\int R(\nu,\delta_{\boldsymbol{\mu}})\,dG_{\boldsymbol{\mu}}(\nu) \\
&= E\Big(\sum\big(\delta_{\boldsymbol{\mu}}(Z_i)-\nu_i\big)^2\big|\boldsymbol{\mu}\Big). \qquad (12)
\end{aligned}$$

The lemma is concluded by the uniqueness of the minimizer of a strictly convex function. □

However, $\delta \equiv \delta_{\boldsymbol{\mu}}$ is unknown since $\boldsymbol{\mu}$ is unknown. The following theorem shows that it can be well approximated using the observations.

Theorem 2. *If $\hat{\delta}$ is defined as in (8) with $M_n = \mathcal{O}((\log n)^{0.15})$, then*

$$E\|\hat{\boldsymbol{\delta}}-\boldsymbol{\nu}\|^2 \leq E\|\boldsymbol{\delta}-\boldsymbol{\nu}\|^2 + o(n).$$

The proof is given in the Appendix.

By definition, $\boldsymbol{\mu}^I$ improves over $\hat{\boldsymbol{\mu}}$. Lemma 1 and Theorem 2 implies that $\boldsymbol{\mu}^I \approx \hat{\boldsymbol{\mu}}^I$ and hence $\hat{\boldsymbol{\mu}}^I$ improves over $\hat{\boldsymbol{\mu}}$, at least asymptotically. We aim to formulate a more general result, where in addition, it is stated and shown when $\hat{\boldsymbol{\mu}}^I$ is asymptotically strictly better than $\hat{\boldsymbol{\mu}}$. For this purpose, we present a few more considerations.

Only when $\boldsymbol{\delta}$ is asymptotically linear, $\boldsymbol{\mu}^I$ may be equivalent to $\hat{\boldsymbol{\mu}}$, and consequently $\hat{\boldsymbol{\mu}}^I$ would not strictly improve over $\hat{\boldsymbol{\mu}}$. It follows from (7) that this happens when $f'(z)/f(z) = (\log f(z))'$ is approximately proportional to z. Consider a sequence $\boldsymbol{\mu}$ and the corresponding $G_{\boldsymbol{\mu}}$. Suppose $G_{\boldsymbol{\mu}}$ converges weakly to G_0. Since the corresponding sequence $f(z)$ is a convolution of a Gaussian kernel with the 'prior' $G_{\boldsymbol{\mu}}$, $(\log f(z))'$ is linear only if G_0 is Gaussian.

Recall that in a normal Bayesian problem under squared loss, a prior $G_0 = N(0, \tau^2)$ and observation $Z \sim N(\nu, 1)$, the Bayes decision is $\hat{\nu} = Z\tau^2/(\tau^2 + 1)$. In particular, setting $\tau^2 = (1-\alpha)/\alpha$, the Bayes decision is $(1-\alpha)Z$. Together with Lemma 1 and Theorem 2 we established the following.

Theorem 3. *If Assumptions 1–3 hold*:

(i)
$$E\left\|\hat{\boldsymbol{\mu}}^I - \boldsymbol{\mu}\right\|^2 \le E\left\|\boldsymbol{\mu}^I - \boldsymbol{\mu}\right\|^2 + o(n) \le E\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 + o(n).$$

(ii) *For $\boldsymbol{\mu} \equiv \boldsymbol{\mu}_n$, suppose that $\alpha_n \xrightarrow{\mathrm{p}} \alpha$ and $G_{\boldsymbol{\mu}} \rightsquigarrow G_0 \ne N(0, (1-\alpha)/\alpha)$. Then there exists $c < 1$ such that for large enough n*:
$$E\left(\left\|\hat{\boldsymbol{\mu}}^I - \boldsymbol{\mu}\right\|^2\right) < cE\left(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2\right).$$

Part (i) of Theorem 3 assures us that asymptotically the improved estimator $\hat{\boldsymbol{\mu}}^I$, does as good as $\hat{\boldsymbol{\mu}}$. Part (ii) implies that typically the improved estimator is asymptotically strictly better. Obviously the asymptotic improvement is not always strict since, for example, the Kalman filter is optimal under a Gaussian state-space model.

Heuristically, the value of c is expected to get smaller as the distribution G_0 'resembles' a normal distribution less. The canonical $\hat{\mu}$ is implicitly motivated under Normal G_0. In particular in cases of a sparse G_0, which is expected in situations where $\tilde{\mu}_i$ does a good job in most cases, thus the corresponding $\nu_i \approx 0$, but in some exceptional cases (or, indices i) the considerations leading to $\tilde{\mu}_i$ are very wrong and $\nu_i = \mu_i - \tilde{\mu}_i$ are large.

Example 1.1 (concluded). In our example, $G_{\boldsymbol{\mu}}$ converges to $N(0, \tau^2)$, $\tau^2 = 1$. In light of the above $\hat{\boldsymbol{\mu}}^I \approx \hat{\boldsymbol{\mu}}$ asymptotically. By the above:
$$\hat{\mu}_i \approx \tilde{\mu}_i + \frac{\tau^2}{\tau^2+1}Z_i = Y_{i-1} + \frac{1}{2}(Y_i - Y_{i-1}) = \frac{Y_{i-1} + Y_i}{2}.$$

The above obviously does not converge to the optimal coordinate-wise decision function, which is determined by $\delta^{\mathrm{opt}}(Y_i) \equiv 0$. It only asymptotically improves upon the optimal among canonical estimator based on $\tilde{\mu}_i$ and Y_i. The improvement is not strict, since $\hat{\boldsymbol{\mu}}^I \approx \hat{\boldsymbol{\mu}}$.

4. Sequential estimation

We now consider the sequential case, where $\mathcal{D}_i = \{1, \ldots, i-1\}$ and $\tilde{\mu}_i$ is $\mathcal{F}_{i-1} = \sigma(Y_1, \ldots, Y_{i-1})$ measurable. The definition of the different estimators is the same as in the previous section with the necessary adaption to the current sets $\mathcal{D}_i$, so $\hat{\mu}_i$ and $\tilde{\mu}_i$ of this section are sequential. Our aim is to find a sequential estimator, denoted $\hat{\boldsymbol{\mu}}^{\mathrm{IS}}$, that satisfies $E\|\hat{\boldsymbol{\mu}}^{\mathrm{IS}} - \boldsymbol{\mu}\|^2 + o(n) \le E\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$. Here, $\hat{\boldsymbol{\mu}} \equiv \hat{\boldsymbol{\mu}}^S$ is sequential, it has the form $\hat{\mu}_i = \tilde{\mu}_i + \alpha Z_i$ where $\tilde{\mu}_i \equiv \tilde{\mu}_i^S \in \mathcal{F}_{i-1}$; we omit the superscript S. In general, by a sequential estimator $\hat{\boldsymbol{\mu}}^{\mathrm{IS}} = (\hat{\mu}_1^{\mathrm{IS}}, \ldots, \hat{\mu}_n^{\mathrm{IS}})$ we mean that

$\hat{\mu}_i^{\mathrm{IS}} \in \mathcal{F}_i$, $i = 1, 2, \ldots$. A natural approach, which indeed works, is to let $\hat{\mu}_i^{\mathrm{IS}} = \tilde{\mu}_i + \hat{\delta}^i$, where $\hat{\delta}^i$ is defined as in (8), but with $\hat{f} = \hat{f}^i$ restricted to the available data $Z_1, \ldots, Z_i$, $i = 1, \ldots, n$. Let

$$\hat{\boldsymbol{\delta}}^S = \left(\hat{\delta}^1, \ldots, \hat{\delta}^n\right).$$

Define

$$\hat{\boldsymbol{\mu}}^{\mathrm{IS}} = \tilde{\boldsymbol{\mu}} + \hat{\boldsymbol{\delta}}^S.$$

The following lemma is an abstract version of a result in Samuel [20] and is the key to the main result of this section.

Lemma 4. *Let Δ be some set and $f_j : \Delta \to \mathbb{R}$, $j = 1, 2, \ldots, n$. Let $\delta_i \in \Delta$ satisfies $\sum_{j=1}^{i} f_j(\delta_i) \le \inf_{\delta\in\Delta} \sum_{j=1}^{i} f_j(\delta) + \zeta_i$, $i = 1, \ldots, n$. Then*

$$\sum_{i=1}^{n} f_i(\delta_i) \le \inf_{\delta\in\Delta} \sum_{i=1}^{n} f_i(\delta) + \sum_{i=1}^{n} \zeta_i.$$

Proof. By a trivial telescoping argument:

$$\begin{aligned}
\sum_{i=1}^{n} f_i(\delta_i) &= \sum_{i=1}^{n} \left(\sum_{j=1}^{i} f_j(\delta_i) - \sum_{j=1}^{i-1} f_j(\delta_i) \right) \\
&= \sum_{j=1}^{n} f_j(\delta_n) - \sum_{i=2}^{n} \left(\sum_{j=1}^{i-1} f_j(\delta_i) - \sum_{j=1}^{i-1} f_j(\delta_{i-1}) \right) \\
&\le \sum_{j=1}^{n} f_j(\delta_n) - \sum_{i=2}^{n} \left(\sum_{j=1}^{i-1} f_j(\delta_i) - \inf_{\delta\in\Delta} \sum_{j=1}^{i-1} f_j(\delta) - \zeta_{i-1} \right) \\
&\le \sum_{j=1}^{n} f_j(\delta_n) + \sum_{i=1}^{n-1} \zeta_i.
\end{aligned}$$

□

We consider the lemma with

$$f_j(\delta) = \iint \left(\delta(z) - \nu\right)^2 \varphi(z - \nu)\, dz\, dG_j(\nu), \tag{13}$$

where δ is a decision function, φ is the standard normal density, and $G_i = \mathcal{L}(\nu_i | \mu_1, \ldots, \mu_n)$. Note that $G_i = \mathcal{L}(\nu_i | \mu_1, \ldots, \mu_i)$ since $\nu_i = \mu_i - \tilde{\mu}_i$ and $\tilde{\mu}_i$ is estimated sequentially.

Corollary 5. *Let* $\delta^i = \arg\min_\delta \sum_{j=1}^i \iint (\delta(z)-\nu)^2 \varphi(z-\nu)\,dz\,dG_j(\nu)$, $i=1,\dots,n$. *Then*

$$\sum_{i=1}^n \int \big(\delta^i(z)-\nu\big)^2 \varphi(z-\nu)\,dz\,dG_i(\nu) \le \sum_{i=1}^n \int \big(\delta^n(z)-\nu\big)^2 \varphi(z-\nu)\,dz\,dG_i(\nu).$$

Let $\boldsymbol{\mu}^i = \boldsymbol{\mu}_n^i = (\mu_1,\dots,\mu_i)$. Define other partial vectors similarly and define $G_{\boldsymbol{\mu}^i}$ as in (3) to be the distribution conditional on $\boldsymbol{\mu}^i$ and restricted to $\boldsymbol{\nu}^i$. Clearly:

$$\begin{aligned}\delta^i &= \arg\min \sum_{j=1}^i \iint \big(\delta(z)-\nu\big)^2 \varphi(z-\nu)\,dz\,dG_j(\nu)\\ &= \arg\min n \iint \big(\delta(z)-\nu\big)^2 \varphi(z-\nu)\,dz\,dG_{\boldsymbol{\mu}^i}(\nu).\end{aligned}$$

Consider a sequential procedure $\hat{\boldsymbol{\mu}}$, satisfying Assumption 3, $\hat{\boldsymbol{\mu}} = \tilde{\boldsymbol{\mu}} + \alpha \boldsymbol{Z} + \boldsymbol{\xi}$, where $\alpha = \alpha_{\boldsymbol{\mu}}$ and $E\boldsymbol{\xi}^2 = o(n)$. Let $\boldsymbol{\delta}^S = (\delta^1,\dots,\delta^n)$, $\boldsymbol{\mu}^{\mathrm{IS}} = \tilde{\boldsymbol{\mu}} + \boldsymbol{\delta}^S$. By definitions and by Corollary 5, we have:

$$\begin{aligned}E\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 &= E\|\alpha \boldsymbol{Z} - \boldsymbol{\nu}\|^2 + o(n)\\ &\ge E\sum \big(\delta^n(Z_i) - \nu_i\big)^2 + o(n)\\ &\ge E\sum \big(\delta^i(Z_i) - \nu_i\big)^2 + o(n)\\ &= E\big\|\boldsymbol{\mu}^{\mathrm{IS}} - \boldsymbol{\nu}\big\|^2 + o(n).\end{aligned}$$

Moreover, as explained in the previous section, unless the μ process itself is Gaussian, δ^i is asymptotically strictly better than any linear stationary estimator, in particular $(1-\alpha)z$. Finally, as proved in the previous section for the retrospective case, for large i, $E(\hat{\delta}^i - \delta^i)^2 = o(1)$.

Thus, we can obtain our main result of this section:

Theorem 6.

(i) *Under Assumptions 1–3*:

$$E\big\|\hat{\boldsymbol{\mu}}^{\mathrm{IS}} - \boldsymbol{\mu}\big\|^2 \le E\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 + o(n).$$

(ii) *For* $\boldsymbol{\mu} \equiv \boldsymbol{\mu}_n$, *suppose that* $\alpha_{\boldsymbol{\mu}_n} \xrightarrow{\mathrm{p}} \alpha_0$ *and* $G_{\boldsymbol{\mu}} \rightsquigarrow G_0 \ne N(0,(1-\alpha_0)/\alpha_0)$. *Then there exists* $c<1$ *such that for large enough* n:

$$E\big(\big\|\hat{\boldsymbol{\mu}}^{\mathrm{IS}} - \boldsymbol{\mu}\big\|^2\big) < cE\big(\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2\big).$$

5. Simulations

We now present simulation results for the following state-space model. Further numerical studies and analysis of real data of our improvement method, in the context of Fay Herriot estimators, may be found in Cohen et al. [8].

$$\begin{aligned} Y_i &= \mu_i + \varepsilon_i \\ \mu_i &= \phi\mu_{i-1} + U_i, \qquad i = 1, \ldots, n, \end{aligned} \tag{14}$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $N(0, 1)$ and independent of $U_1, \ldots, U_n$. The variables U_i, $i = 1, \ldots, n$ are independent, $U_i = X_i I_i$ where $X_i \sim N(0, v^2)$ are independent, while $I_1, \ldots, I_n$ are i.i.d. Bernoulli with mean of 0.1, independent of each other and of $X_1, \ldots, X_n$. We study the twelve cases that are determined by $\phi = 0.25, 0.75$ and $v = 0, 1, \ldots, 5$. In each case, we investigate both the sequential and the retrospective setups. In the case $\phi = 0.25$ the values of $(1 - \alpha)$ that correspond to $v = 1, 2, 3, 4, 5$, in a sequential setup are 0.096, 0.294, 0.48, 0.62, 0.72. In the retrospect setup, the corresponding values are very slightly smaller.

If $U_1, \ldots, U_n$ were i.i.d. normal, the data would follow a Gaussian state-space model, and the corresponding Kalman filter estimator $\hat{\boldsymbol{\mu}}^K \equiv \hat{\boldsymbol{\mu}}$ would be optimal. Since the U_i's are not normal, the corresponding AR(1) Kalman filter estimator is not optimal (except in the degenerate case, $v = 0$), though it is optimal among linear filters. This is reflected in our simulation results where for the cases $v = 0$ and $v = 1$ our "improved" method $\hat{\boldsymbol{\mu}}^I$ performs slightly worse than the Kalman filter estimator $\hat{\boldsymbol{\mu}}^K$. It improves in all the rest. The above is stated and proved formally in the following proposition. It could also be shown indirectly by applying part (ii) of Theorem 3.

Proposition 7. *Consider the state-space model, as defined by* (14). *If U_i are not normally distributed then*

$$E\left\|\hat{\boldsymbol{\mu}}^I - \boldsymbol{\mu}\right\|^2 \le cE\left\|\hat{\boldsymbol{\mu}}^K - \boldsymbol{\mu}\right\|^2,$$

for a constant $c \in (0, 1)$ and large enough n.

Proof. We sketch the proof. Given the estimators $\tilde{\mu}_i^K$ and $\hat{\mu}_i^K$, $i = 1, \ldots, n$, let $Z_i = Y_i - \tilde{\mu}_i^K$. Then $Z_i = \mu_{i-1} + U_i - \tilde{\mu}_i^K + \varepsilon_i = \nu_i + \varepsilon_i$. The distribution $\tilde{G}^i$ of ν_i may be normal only if U_i is normal, since U_i is independent of μ_{i-1} and $\tilde{\mu}_i^K$. The distributions $\tilde{G}^i$ converge to a distribution G as i and $n - i$ approach infinity. As before, G is normal only if U_i are normal. Now, the asymptotically optimal estimator for ν_i under squared loss given the observation Z_i is $\hat{\nu}_i$, where $\hat{\nu}_i$ is the Bayes estimator under a prior G on ν_i, based on an observation $Z_i \sim N(\nu_i, 1)$. This Bayes estimator is linear and coincides with the KF estimator $\hat{\mu}_i^K$, only if G is normal. □

Analogous discussion and situation are also valid in the sequential case.

In our simulations, the parameters ϕ and $\mathrm{VAR}(U_i)$ are treated as known. Alternatively, the maximum likelihood estimation, assuming (wrongly) normal innovations U_i, yields results similar to those reported in Table 1.

The simulation results in Table 1 are for the case $n = 500$. In order to speed the asymptotics, we allowed a 'warm up' of 100 observations prior to the $n = 500$ in the sequential case, we also

Table 1. Simulation: Mean Squared error of the two estimators

ϕ	0.25						0.75					
v	0	1	2	3	4	5	0	1	2	3	4	5
						Retrospective filter:						
$\widehat{\mu}^{\dagger}$	0	71	156	226	290	333	0	49	147	235	301	350
$\hat{\mu}^{I\ddagger}$	23	66	125	148	160	177	24	91	166	215	253	271
						Sequential filter:						
$\widehat{\mu}^{\dagger}$	0	47	145	234	309	355	0	83	187	264	325	372
$\hat{\mu}^{\mathrm{IS}\ddagger}$	39	81	129	147	159	158	34	112	184	216	239	253

† Kalman filter, ‡ Improved.

allowed a 'warm up' of 50 in both sides of the $n = 500$ observations in the retrospective case. Each entry in the table is based on 100 simulations. In each realization, we recorded $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2$ and $\|\hat{\boldsymbol{\mu}}^I - \boldsymbol{\mu}\|^2$, and each entry is based on the corresponding average. The same is true in the sequential case with $\|\hat{\boldsymbol{\mu}}^{\mathrm{IS}} - \boldsymbol{\mu}\|^2$.

It may be seen that when the best linear filter is optimal or nearly optimal (when $v = 0$ or approximately so), our improved method is slightly worse than the Kalman filter estimator, however as v increases, the advantage of the improved method may become significant.

6. Real data example

In this section, we demonstrate the performance of our method on real data taken from the FX (foreign exchange) dollar-shekel market in Israel. The data consists of the daily number of swaps – purchase of one currency for another with a given value date while simultaneously selling back of the same amount with a different value date. We consider only purchases of between 5 million and 20 million dollars. The time period is January 2nd, 2009 to December 31st, 2013, a total of $n = 989$ business days. The average number of daily purchases is 24, with the daily number ranging from 2 to 71. In our analysis, we used the first 100 observations as a 'warm up', similar to the way it was done in our simulations section.

We denote by X_i, $i = 1, \ldots, n$, the number of purchases on day i and assume that they have a Poisson distribution: $X_i \sim \mathrm{Po}(\lambda_i)$. The data were transformed to $Y_i = 2\sqrt{X_i + 0.25}$ as in Brown et al. [3] and Brown, Greenshtein and Ritov [6] in order to get an (approximately) normal variable with a variance of $\sigma^2 = 1$.

The assumed model in this section is the following state space system of equations:

$$Y_i = \mu_i + \varepsilon_i$$
$$\mu_i \sim \mathrm{ARIMA}(p, d, q), \qquad i = 1, \ldots, n,$$

where $\mu_i = 2\sqrt{\lambda_i}$ and $\varepsilon_i \sim N(0,1)$ are independent of each other and of the ARIMA(p,d,q) process. We consider the following three special cases of ARIMA(p,d,q): AR(1), AR(2), and ARIMA(1, 1, 0).

Under each model there are induced Kalman filter estimators $\tilde{\boldsymbol{\mu}} \equiv \tilde{\boldsymbol{\mu}}^K$, and $\hat{\boldsymbol{\mu}} \equiv \hat{\boldsymbol{\mu}}^K$. Similarly, the improved estimator $\hat{\boldsymbol{\mu}}^I$ is defined. We denote the sequential and retrospective estimators similarly with no danger of confusion.

After estimating μ_i, we transform the result back to get the estimator $\hat{\lambda}_i^J$ for λ_i, $\hat{\lambda}_i^J = 0.25\hat{\mu}_i^{J2}$, $i = 1,\ldots,n$, $J \in \{\text{'}I\text{'}, \text{'}K\text{'}\}$ where, $\hat{\mu}_i^J$ is the estimator of μ_i by method J. We evaluate the performances of both estimation methods by the following non-standard cross-validation method as described in Brown et al. [6]. It is briefly explained in the following.

Let $p \in (0,1)$, $p \approx 1$, and let $U_1,\ldots,U_n$ be independent given $X_1,\ldots,X_n$, where $U_i \sim B(X_i, p)$ are binomial variables. It is known that $U_i \sim \text{Po}(p\lambda_i)$, $V_i = X_i - U_i \sim \text{Po}((1-p)\lambda_i)$, and they are independent given $\lambda_1,\ldots,\lambda_n$. We will use the 'main' sub-sample $U_1,\ldots,U_n$ for the construction of both estimators (Kalman filter and Improved) while the 'auxiliary' sub-sample $V_1,\ldots,V_n$ is used for validation. Consider the following function,

$$\begin{aligned}
\rho(J;\mathbf{U},\mathbf{V}) &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{\lambda}_i^J}{p} - \frac{V_i}{(1-p)}\right)^2 \\
&= \frac{1}{np^2}\sum_{i=1}^{n}\left(\hat{\lambda}_i^J - p\lambda_i\right)^2 + \frac{1}{n(1-p)^2}\sum_{i=1}^{n}\left(V_i - (1-p)\lambda_i\right)^2 \\
&\quad - \frac{2}{n}\sum_{i=1}^{n}\left(\frac{\hat{\lambda}_i^J}{p} - \lambda_i\right)\left(\frac{V_i}{(1-p)} - \lambda_i\right) \\
&= \frac{1}{np^2}\sum_{i=1}^{n}\left(\hat{\lambda}_i^J - p\lambda_i\right)^2 + A_n + R_n(J), \qquad J \in \left\{\text{'}K\text{'}, \text{'}I\text{'}\right\}.
\end{aligned}$$

The term $R_n(J) = \mathcal{O}_p(n^{-1/2})$ and will be ignored. We estimate A_n by the method of moments:

$$\hat{A}_n = \frac{1}{n(1-p)^2}\sum_{i=1}^{n} V_i.$$

We repeat the cross-validation process 500 times and average the computed values of $\rho(J;\mathbf{U},\mathbf{V}) - \hat{A}_n$. When p is close to 1, the average obtained is a plausible approximation of the average squared risk in estimating λ_i, $i = 101,\ldots,989$. By the above method we also approximated the average risk of the naive estimator $\hat{\lambda}_i^N = X_i$, $i = 101,\ldots,989$. The approximations for the retrospective and sequential cases are displayed in Tables 2 and A.4, respectively. The estimated ARIMA coefficients for the various models are given in Table 3.

From Table 2, we may observe that in the retrospective case the improved method does uniformly better than the naive estimator and the Kalman filter. In fact, in all situations, except for

Table 2. The retrospective case: Cross-validation estimation of the average squared risk

$p = 0.95$	AR(1)	AR(2)	ARIMA(1, 1, 0)
Kalman filter – $\hat{\lambda}_i^K$	27.1	19.4	20.4
Improved method – $\hat{\lambda}_i^I$	18.7	18.5	17.4
Naive method – $\hat{\lambda}_i^N$	26.4	26.4	26.4

a small deterioration under the ARIMA(1, 1, 0) with sequential filtering, the performance of the improved method is quite uniform, showing its robustness against model misspecification.

It is somewhat surprising that the Kalman filter under AR(1) with retrospective estimation does not do better than the naive filter, but does considerably better in the sequential case. The reason is that the AR(1) model does not fit the data well. When it is enforced on the data, the Kalman filter gives too much weight to the surrounding data, and too little to the "model free" naive estimator. This result shows the robustness of our estimator.

In fact, we did a small simulation, where the process was AR(2), with the parameters as estimated for the data. When an AR(1) was fitted to the data, the retrospective Kalman filter was strictly inferior to the sequential one.

In the sequential case, Table A.4, the improved method does better than the naive method, but contrary to the non-sequential case, it improves slightly upon the Kalman filter only in the AR(1) and AR(2) models, while in the ARIMA(1, 1, 0) model the Kalman filter does slightly better.

Appendix: Proof of Theorem 2

We assume wlog that all the decision functions $g = g(Z)$ involved and their estimates $\hat{g} = \hat{g}(Z)$, are within a $M_n = (\log n)^{0.15}$ distance from the observed Z. This may be assumed wlog by truncating and obtaining an asymptotically equivalent procedure g for any procedure g^* (for every choice $M_n \to \infty$) as follows:

$$g(Z) = Z + \big(g^*(Z) - Z\big) \times I\big(|g^*(Z) - Z\big)| < M_n). \tag{15}$$

The specific $M_n = \log(n)^{0.15}$ truncation is convenient for the following, but not essential.

Table 3. The retrospective case: Parameter estimation

$p = 0.95$	AR(1)	AR(2)	ARIMA(1, 1, 0)
α	12.38	14.218	0.01
ϕ_1	−0.28	−0.341	−0.6
ϕ_2		−0.124	
σ^2	3.4	3.4	4.7

Table A.4. The sequential case: Cross-validation approximation of the average squared risk

$p = 0.95$	AR(1)	AR(2)	ARIMA(1, 1, 0)
Kalman filter – $\hat{\lambda}_i^K$	19.2	19.2	21.2
Improved method – $\hat{\lambda}_i^I$	19.0	19.2	22.6
Naive method – $\hat{\lambda}_i^N$	26.4	26.4	26.4

In the following all the expectations are conditional on $\boldsymbol{\mu}$, however in order to simplify notations we will completely suppress $\boldsymbol{\mu}$ in the notations.

It is enough to show that $E\|\boldsymbol{\delta} - \hat{\boldsymbol{\delta}}\|^2 = o(n)$.

Let π be a random permutation, denote $\pi(1) \equiv I$, then the random index I is distributed uniformly on $\{1, \ldots, n\}$. Denote $W = Z_{\pi(1)} \equiv Z_I$. Then, the density of W, is:

$$f(w) = \int \varphi(w - v)\, dG(v).$$

Denote $\mathcal{Z} = (Z_{(1)}, \ldots, Z_{(n)})$, the order statistic of $Z_1, \ldots, Z_n$, then conditional on $\mathcal{Z}$ the random function $\hat{\delta}$ is fixed, denoted $\hat{\delta}^{\mathcal{Z}}$. Conditional on $\mathcal{Z}$, W is uniform on $\{Z_1, \ldots, Z_n\}$; hence, conditional on $\mathcal{Z}$, $\hat{\delta}(W)$ is uniform on $\{\hat{\delta}^{\mathcal{Z}}(Z_1), \ldots, \hat{\delta}^{\mathcal{Z}}(Z_n)\}$, and $E((\hat{\delta}(W) - \delta(W))^2|\mathcal{Z}) = \frac{1}{n}\sum(\hat{\delta}^{\mathcal{Z}}(Z_i) - \delta(Z_i))^2$. Hence,

$$\begin{aligned} E\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|^2 &= E\sum_i \big(\hat{\delta}(Z_i) - \delta(Z_i)\big)^2 \\ &= E\left(E\sum_i \big(\hat{\delta}(Z_i) - \delta(Z_i)\big)^2|\mathcal{Z}\right) \\ &= EnE\big(\hat{\delta}(W) - \delta(W)\big)^2|\mathcal{Z}) \\ &= nE\big(\hat{\delta}(W) - \delta(W)\big)^2. \end{aligned}$$

Thus, we may write:

$$E\|\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}\|^2 = nEE\big(\big(\hat{\delta}(W) - \delta(W)\big)^2|W\big) = n\int E\big(\big(\hat{\delta}(W) - \delta(W)\big)^2|W = w\big) f(w)\, dw. \quad (16)$$

In the following, we bound the expected squared difference, between (the truncated versions of) $\hat{\delta}(W)$ and $\delta(W) = W + \frac{f'(W)}{f(W)}$.

Let

$$A = \big\{w|f(w) > \log(n)^{-0.9}\big\}.$$

By Assumption 1, $P(A^c) = \mathcal{O}(D_n/\log(n)^{0.9})$,

$$\begin{aligned}
&\int E\big(\big(\hat{\delta}(W) - \delta(W)\big)^2 | W = w\big) f(w)\, dw \\
&= \int_A E\big(\big(\hat{\delta}(W) - \delta(W)\big)^2 | W = w\big) f(w)\, dw + \mathcal{O}\left(\frac{4M_n^2 D_n}{\log(n)^{0.9}}\right) \\
&= \int_A E\big(\big(\hat{\delta}(W) - \delta(W)\big)^2 | W = w\big) f(w)\, dw + o(1).
\end{aligned}$$

For the last equality, recall that by Assumption 1, $D_n = \mathcal{O}(\sqrt{\log(n)})$.

In order to show that the above integral is of order $o(1)$, it is enough to show that for each $w \in A$

$$E\big(\big(\hat{\delta}(W) - \delta(W)\big)^2 | W = w\big) = o(1).$$

For every $w \in A$, define the event:

$$B_w = \left[\big|\hat{f}'(w) - f'(w)\big| < \frac{1}{\log(n)} \cap \big|\hat{f}(w) - f(w)\big| < \frac{1}{\log(n)}\right].$$

By Assumption 2 and Chebyshev's inequality, the probability of B_w^c-the complementary to the event B_w, satisfy

$$P\big(B_w^c\big) = o\big(M_n^{-2}\big)$$

uniformly in $w \in A$. Hence, it is enough to show that for $w \in A$

$$E\big(\big(\hat{\delta}(w) - \delta(w)\big)^2 | W = w\big)\mathcal{I}(B_w) = o(1), \tag{17}$$

here $\mathcal{I}$ is an indicator of the corresponding event. Recall that, δ and $\hat{\delta}$ are M_n truncations of $\delta(w) = w + \frac{f'(w)}{f(w)}$ and $\hat{\delta}(w) = w + \frac{\hat{f}'(w)}{\hat{f}(w)}$. It may be checked now that (17) follows.

Acknowledgements

This research was partially supported by ISF grant 1770/15 and by NSF grant 1712962.

References

[1] Banarejee, T., Mukherjee, G. and Sun, W. (2018). Adaptive sparse estimation with side information.
[2] Brockwell, P.J. and Davis, R.A. (1991). *Time Series*: *Theory and Methods*, 2nd ed. *Springer Series in Statistics*. New York: Springer. MR1093459
[3] Brown, L., Cai, T., Zhang, R., Zhao, L. and Zhou, H. (2010). The root-unroot algorithm for density estimation as implemented via wavelet block thresholding. *Probab. Theory Related Fields* **146** 401–433. MR2574733

[4] Brown, L.D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann. Math. Stat.* **42** 855–903. MR0286209
[5] Brown, L.D. and Greenshtein, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *Ann. Statist.* **37** 1685–1704. MR2533468
[6] Brown, L.D., Greenshtein, E. and Ritov, Y. (2013). The Poisson compound decision problem revisited. *J. Amer. Statist. Assoc.* **108** 741–749. MR3174656
[7] Brown, L.D., Mukherjee, G. and Weinstein, A. (2018). Empirical Bayes estimates for a two-way cross-classified model. *Ann. Statist.* **46** 1693–1720. MR3819114
[8] Cohen, N., Greenshtein, E. and Ritov, Y. (2013). Empirical Bayes in the presence of explanatory variables. *Statist. Sinica* **23** 333–357. MR3076170
[9] Copas, J.B. (1969). Compound decisions and empirical Bayes. (With discussion.). *J. Roy. Statist. Soc. Ser. B* **31** 397–425. MR0269013
[10] Efron, B. (2010). *Large-Scale Inference. Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge: Cambridge Univ. Press. MR2724758
[11] Efron, B. and Morris, C. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597
[12] Fay, R.E. III and Herriot, R.A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277. MR0548019
[13] Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. *Ann. Statist.* **37** 1647–1684. MR2533467
[14] Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* **109** 674–685. MR3223742
[15] Lehmann, E.L. and Casella, G. (2003). *Theory of Point Estimation*, *Springer Texts in Statistics*. New York: Springer.
[16] Rao, J.N.K. (2003). *Small Area Estimation. Wiley Series in Survey Methodology*. Hoboken, NJ: Wiley Interscience. MR1953089
[17] Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 1950 131–148. Berkeley and Los Angeles: Univ. California Press. MR0044803
[18] Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 157–163. Berkeley and Los Angeles: Univ. California Press. MR0084919
[19] Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Stat.* **35** 1–20. MR0163407
[20] Samuel, E. (1965). Sequential compound estimators. *Ann. Math. Stat.* **36** 879–889. MR0183055
[21] Weinstein, A., Ma, Z., Brown, L.D. and Zhang, C.-H. (2018). Group-linear empirical Bayes estimates for a heteroscedastic normal mean. *J. Amer. Statist. Assoc.* **113** 698–710. MR3832220
[22] Xie, X., Kou, S.C. and Brown, L.D. (2012). SURE estimates for a heteroscedastic hierarchical model. *J. Amer. Statist. Assoc.* **107** 1465–1479. MR3036408
[23] Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods. *Ann. Statist.* **31** 379–390. MR1983534

Received October 2017 and revised October 2018