Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 138 (2008) 3069–3085
journal of statistical planning and inference
www.elsevier.com/locate/jspi

A pre-test like estimator dominating the least-squares method☆

Yonina C. Eldar*, Jacob Slava Chernoi

Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel

Received 7 March 2007; received in revised form 29 September 2007; accepted 4 December 2007
Available online 23 December 2007

Abstract

We develop a pre-test type estimator of a deterministic parameter vector $\boldsymbol{\beta}$ in a linear Gaussian regression model. In contrast to conventional pre-test strategies, that do not dominate the least-squares (LS) method in terms of mean-squared error (MSE), our technique is shown to dominate LS when the effective dimension is greater than or equal to 4. Our estimator is based on a simple and intuitive approach in which we first determine the linear minimum MSE (MMSE) estimate that minimizes the MSE. Since the unknown vector $\boldsymbol{\beta}$ is deterministic, the MSE, and consequently the MMSE solution, will depend in general on $\boldsymbol{\beta}$ and therefore cannot be implemented. Instead, we propose applying the linear MMSE strategy with the LS substituted for the true value of $\boldsymbol{\beta}$ to obtain a new estimate. We then use the current estimate in conjunction with the linear MMSE solution to generate another estimate and continue iterating until convergence. As we show, the limit is a pre-test type method which is zero when the norm of the data is small, and is otherwise a non-linear shrinkage of LS.

Keywords: Pre-test estimators; Dominating estimators; Regression analysis; Biased estimation; Mean-squared error criterion

1. Introduction

A vast variety of estimation problems in a broad range of application areas can be written in the form of a linear Gaussian regression model. In this class of problems, the goal is to estimate a deterministic parameter vector $\boldsymbol{\beta}$ from noisy observations $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where $\mathbf{X}$ is a known model matrix and $\mathbf{e}$ is a Gaussian noise vector. The maximum-likelihood estimate of $\boldsymbol{\beta}$ in this model is the well-known least-squares (LS) method, which chooses the vector $\hat{\boldsymbol{\beta}}$ that minimizes the data error $\|\mathbf{y} - \hat{\mathbf{y}}\|^2$ where $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ is the estimated data vector. In addition to being maximum-likelihood optimal, the LS approach is minimax over all choices of $\boldsymbol{\beta}$, and minimizes the variance among all unbiased strategies. However, even though unbiasedness may be appealing intuitively, it does not necessarily lead to a small estimation error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ (Efron, 1975). Therefore, in the past 50 years a large body of research has been devoted to the design of estimates that outperform LS in an estimation error sense (Judge and Bock, 1983, 1978; Hoerl and Kennard, 1970; Marquardt, 1970; Tikhonov and Arsenin, 1977; Mayer and Willke, 1973; Eldar and Oppenheim, 2003; Eldar et al., 2004, 2005; Ben-Haim and Eldar, 2007; Eldar, 2006).

A direct measure of estimation error is the mean-squared error (MSE). Since $\boldsymbol{\beta}$ is deterministic, the MSE depends on $\boldsymbol{\beta}$, and therefore cannot be minimized for all parameter values. Consequently, one method may be better than another for some values of $\boldsymbol{\beta}$, but worse for other parameter values. For example, the trivial estimator $\hat{\boldsymbol{\beta}} = \mathbf{0}$ will have low MSE

☆ This work was supported in part by the Israel Science Foundation under Grant no. 1081/07.
* Corresponding author. Tel.: +972 4 8293256; fax: +972 4 8295757.
E-mail addresses: yonina@ee.technion.ac.il (Y.C. Eldar), jcher@techunix.technion.ac.il (J.S. Chernoi).

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.12.002

when the true norm of $\boldsymbol{\beta}$ is small, but will have large MSE otherwise. Nonetheless, a partial order between estimators can be imposed by using the concepts of domination and admissability (Lehmann and Casella, 1999). Specifically, an estimate $\hat{\boldsymbol{\beta}}_1$ is said to dominate a different estimate $\hat{\boldsymbol{\beta}}_2$ if the MSE of $\hat{\boldsymbol{\beta}}_1$ is never larger than that of $\hat{\boldsymbol{\beta}}_2$ for all choices of $\boldsymbol{\beta}$, and is strictly lower for at least one parameter value. An estimator is admissible if it is not dominated by any other approach.

In his seminal paper, Stein showed that for $\mathbf{X}=\mathbf{I}$ and white noise, the LS strategy is inadmissible when the parameter dimension is larger than 2 (Stein, 1956). Several years later, James and Stein developed a non-linear shrinkage of the conventional LS and proved that it dominates the LS solution (James and Stein, 1961). Various modifications of the James–Stein approach have since been developed that are applicable to the general linear model considered here (Strawderman, 1971; Alam, 1973; Berger, 1976); of these methods, Bock's (1975) seems to be most quoted. Recently, a blind minimax framework has been developed which results in LS dominating estimators that are more general than simple shrinkage techniques (Ben-Haim and Eldar, 2005, 2007).

Another common LS alternative aimed at reducing the MSE is the pre-test approach (Bancroft, 1944; Scolve et al., 1972; Bock et al., 1973; Judge and Bock, 1978, 1983; Giles and Giles, 1993). This estimation rule depends on a preliminary test performed on the data. The idea is to first check whether certain restrictions on the parameter are true; if the hypothesis that the restrictions are true is accepted, then an estimator incorporating these constraints is used. Otherwise, the LS method is employed. A special case is when the test is for $\boldsymbol{\beta}=\mathbf{0}$. In this setting, the test statistic is the weighted norm of the LS solution, and if the hypothesis is accepted then the estimate is $\hat{\boldsymbol{\beta}}=\mathbf{0}$. Throughout the paper, we will focus on this class of pre-test methods. Unfortunately, it can be shown that for many choices of $\boldsymbol{\beta}$ the pre-test strategy has larger MSE than the LS approach (Bock et al., 1973; Scolve et al., 1972). Nonetheless, it is the estimator of choice in many applied areas particularly in economics (Giles and Giles, 1993; Aigner and Judge, 1977).

Two of the main features of the pre-test technique is that it is $\mathbf{0}$ when the data norm is small, and is discontinuous, thus taking on the form of a test. An alternative strategy with the first property is the positive-part James–Stein estimate (Baranchik, 1964), which is a modification of the James–Stein approach in which the shrinkage factor is replaced by 0 whenever it becomes negative. The advantage of this technique is that it dominates the LS strategy for large enough effective dimension. However, in contrast to the pre-test method, by its definition the shrinkage is continuous. Furthermore, there is no real test performed on the data but rather the outcome $\hat{\boldsymbol{\beta}}=\mathbf{0}$ is an artifact of the original James–Stein method which allows for negative shrinkage.

To preserve the main features of the pre-test strategy it would be desirable to develop an estimator that dominates the LS method, and is based on preliminary testing of the data. In this paper we suggest an estimator with these properties, using a simple and intuitive approach. The idea behind our strategy is to first find the linear estimator that minimizes the MSE, which we refer to as the minimum MSE (MMSE) estimate. Unfortunately, since $\boldsymbol{\beta}$ is deterministic, the MMSE solution depends explicitly on $\boldsymbol{\beta}$ which is unknown. To circumvent the dependency on $\boldsymbol{\beta}$, we suggest substituting the LS value for $\boldsymbol{\beta}$ in the MMSE method, thus obtaining an approximate MMSE estimator. We then use this new value instead of $\boldsymbol{\beta}$ and re-apply the MMSE solution to the data. This process is continued iteratively until convergence. In Section 3 we discuss this strategy in detail and show that the iterations converge. Our final estimate is chosen as the resulting limit, which we refer to as the iterated MMSE (IMMSE). The IMMSE has a pre-test type form: the estimate is $\mathbf{0}$ when the data norm is small and is otherwise a specific non-linear shrinkage of LS. This shrinkage is different than that obtained from the James–Stein approach and does not have a positive-part form, i.e. it is discontinuous in the data. In Section 4 we prove that our technique dominates LS for all values of $\boldsymbol{\beta}$ as long as the effective dimension is greater than or equal to 4. We then demonstrate the performance of our method and compare it to the pre-test and positive-part Bock (which is a generalization of James–Stein to the general, dependent case) techniques in Section 5.

Throughout the paper, we use boldface lowercase letters for vectors in $\mathbb{R}^m$ and boldface uppercase letters for matrices in $\mathbb{R}^{n\times m}$. The set of non-negative integers is denoted by $\mathbb{N}$. The norm of a vector $\boldsymbol{\beta}$ is defined as $\|\boldsymbol{\beta}\|^2=\boldsymbol{\beta}'\boldsymbol{\beta}$, and the ith element of a vector $\mathbf{y}$ is represented by $\mathbf{y}_i$. The $m\times m$ identity matrix is written as $\mathbf{I}_m$, $(\cdot)'$ and $\mathrm{tr}(\cdot)$ are the transpose and the trace of the corresponding matrix, respectively, and $\hat{(\cdot)}$ is an estimated vector or matrix. For a scalar a, $[a]_+$ is equal to 0 for $a\leqslant 0$ and a otherwise. The indicator function $I_{[a,b)}(x)$ is equal to 1 for $a\leqslant x<b$ and 0 otherwise. Similar definitions hold for $I_{(a,b)}(x)$, $I_{[a,b]}(x)$ where a square bracket means that the corresponding end point is included and a regular bracket means that it is excluded. The statistical expectation is written as $E\{\cdot\}$, the non-central chi-square distribution with m degrees of freedom and non-centrality parameter λ is denoted by $\chi^2_{m,\lambda}$, and the normal distribution with mean $\boldsymbol{\beta}$ and covariance $\mathbf{I}_m$ is denoted by $\mathcal{N}(\boldsymbol{\beta},\mathbf{I}_m)$.

2. The pre-test and positive-part estimators

We treat the problem of estimating a deterministic parameter vector $\boldsymbol{\beta} \in \mathbb{R}^m$ from observations $\mathbf{y} \in \mathbb{R}^n$ which are related through the linear Gaussian regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \tag{1}$$

Here $\mathbf{X}$ is a known $n \times m$ model matrix with full rank m, and $\mathbf{e}$ is a zero-mean Gaussian random vector with known positive definite covariance $\mathbf{C} = E\{\mathbf{e}\mathbf{e}'\}$.

Our goal is to design an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ that is close to $\boldsymbol{\beta}$ in an MSE sense, where the MSE between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ is defined by

$$E\{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\} = \|E\{\hat{\boldsymbol{\beta}}\} - \boldsymbol{\beta}\|^2 + E\{\|\hat{\boldsymbol{\beta}} - E\{\hat{\boldsymbol{\beta}}\}\|^2\}. \tag{2}$$

Since the MSE in general depends on $\boldsymbol{\beta}$, it cannot be minimized for all $\boldsymbol{\beta}$. A popular approach to overcome this difficulty is to restrict attention to linear unbiased estimates of the form $\hat{\boldsymbol{\beta}} = \mathbf{G}\mathbf{y}$ for a matrix $\mathbf{G}$ such that $\mathbf{G}\mathbf{X} = \mathbf{I}$. We can then choose $\mathbf{G}$ to minimize the resulting MSE, which leads to the celebrated LS estimate:

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{y}. \tag{3}$$

However, this does not mean that the residual MSE is small. In fact, it is well known that the MSE of the LS method can be large in many estimation problems.

One of the popular approaches in practice to try and improve the LS performance is to use a pre-test estimator. The reasoning behind this strategy is that if the vector $\boldsymbol{\beta}$ to be estimated has small norm, then intuitively we should be able to reduce the MSE by choosing $\hat{\boldsymbol{\beta}} = \mathbf{0}$. In practice, however, we do not know the norm of $\boldsymbol{\beta}$. Therefore, the pre-test technique consists of first checking whether $\|\boldsymbol{\beta}\|$ is small based on the given data $\mathbf{y}$, and then using $\mathbf{0}$ as an estimate if this hypothesis is accepted, and the conventional LS method otherwise. More specifically, let

$$u = \hat{\boldsymbol{\beta}}_{\mathrm{LS}}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}} \tag{4}$$

be the test statistic, which is distributed as a non-central chi-square random variable with m degrees of freedom. Then the pre-test estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\mathrm{PT}} = \begin{cases} \mathbf{0}, & u \leqslant c, \\ \hat{\boldsymbol{\beta}}_{\mathrm{LS}}, & u > c. \end{cases} \tag{5}$$

Here c is determined by the required significance level α of the test so that $P(u > c) = \alpha$ where the probability is with respect to the null hypothesis so that $u \sim \chi^2_{m,0}$ (see Wan and Zou, 2003; Kibria and Saleh, 2006 and references therein on methods for choosing c).

To gain more insight into the statistic u of (4) note that $\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = \boldsymbol{\beta} + \mathbf{e}$ where $\mathbf{e}$ is a Gaussian noise vector with covariance $\mathbf{Q} = (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}$. Since the noise is colored, we can whiten it using

$$\mathbf{Q}^{-1/2}\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = \mathbf{Q}^{-1/2}\boldsymbol{\beta} + \tilde{\mathbf{e}}, \tag{6}$$

where $\tilde{\mathbf{e}}$ is a white Gaussian noise vector with covariance $\mathbf{I}$. Since $\tilde{\mathbf{e}}$ is white, a reasonable estimate of $\boldsymbol{\beta}'\mathbf{Q}^{-1}\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}\mathbf{Q}^{-1}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, which is equal to u.

Unfortunately, it can be shown that in most of the parameter range, the pre-test estimator is inferior to the LS approach and results in a larger MSE. In Fig. 1 we plot the MSE as a function of $\|\boldsymbol{\beta}\|^2/n$ using $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{PT}}$ assuming $\mathbf{X} = \mathbf{C} = \mathbf{I}$, $n = 15$ and a test with significance level $\alpha = 0.05$. For each value of $\|\boldsymbol{\beta}\|^2/n$, a random vector $\boldsymbol{\beta}$ was generated and normalized to the required norm. The MSE was computed by averaging over 20,000 noise realizations. As can be seen, when $\|\boldsymbol{\beta}\|$ is small the pre-test estimator performs well, however as the norm increases, the performance deteriorates.

To overcome this shortcoming of $\hat{\boldsymbol{\beta}}_{\mathrm{PT}}$, in Scolve et al. (1972) the authors propose an estimation strategy that resembles the pre-test concept by relying on the James–Stein approach. We refer to this method as the James–Stein pre-test

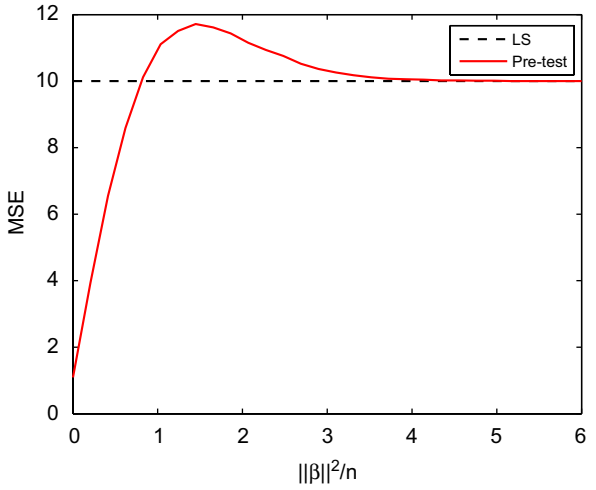

Fig. 1. MSE of the LS and pre-test estimators as a function of the norm of $\boldsymbol{\beta}$.

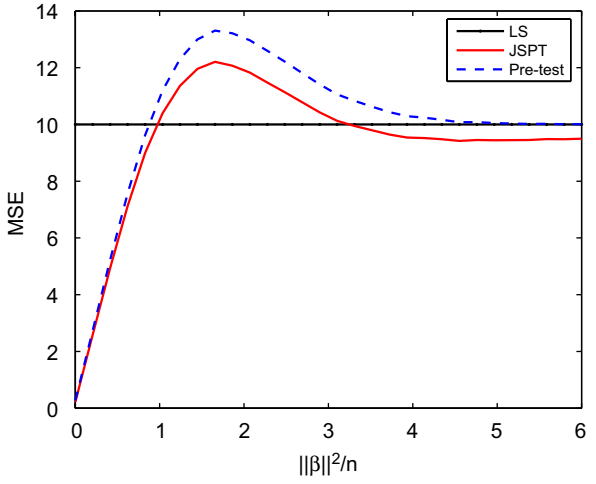

Fig. 2. MSE of the LS, pre-test and James–Stein pre-test estimators as a function of the norm of $\boldsymbol{\beta}$.

estimator. They first show that if $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ in (5) for $u > c$ is replaced by the positive-rule James–Stein estimate (Baranchik, 1964) given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{JS}} = \left[1 - \frac{r}{\|\hat{\boldsymbol{\beta}}_{\mathrm{LS}}\|^2}\right]_+ \hat{\boldsymbol{\beta}}_{\mathrm{LS}}, \tag{7}$$

with $0 \leqslant r \leqslant 2(m-2)$, then the resulting estimator has smaller MSE than $\hat{\boldsymbol{\beta}}_{\mathrm{PT}}$ for all values of $\boldsymbol{\beta}$. However, as can be seen in Fig. 2, when $c > r$ this approach still does not dominate $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$. In the figure, the MSE of the LS, pre-test and the James–Stein pre-test estimators are plotted as a function of $\|\boldsymbol{\beta}\|^2/n$ with $\alpha = 0.01$, $n = 10$, and r of (7) chosen as $r = 2.5$. The remaining conditions are the same as those in Fig. 1.

When $c \leqslant r$, the James–Stein pre-test estimator reduces to the conventional positive-part James–Stein method of (7) which is known to dominate $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$. Note, however, that this strategy is not truly based on a preliminary test as it is a continuous function of the observations. The thresholding effect resulting from the positive-part operation is more an artifact of the negative shrinkage in the conventional James–Stein method, than based on pre-test reasoning.

In the next section we develop an estimation approach similar in spirit to the pre-test strategy that results in a zero estimate when the norm is small, but dominates the LS method over the entire parameter space for large enough effective dimension.

3. The iterative MSE estimator

To develop our approach, suppose that we restrict attention to linear estimates of the form $\hat{\boldsymbol{\beta}} = \mathbf{G}\mathbf{y}$ for some $m \times n$ matrix $\mathbf{G}$. We would like to choose $\mathbf{G}$ to minimize the MSE. Using the fact that $E\{\mathbf{e}\} = 0$ and $E\{\mathbf{e}\mathbf{e}'\} = \mathbf{C}$, the MSE follows from (2) as

$$E\{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2\} = \boldsymbol{\beta}'(\mathbf{G}\mathbf{X} - \mathbf{I})'(\mathbf{G}\mathbf{X} - \mathbf{I})\boldsymbol{\beta} + \mathrm{tr}(\mathbf{G}\mathbf{C}\mathbf{G}'). \tag{8}$$

Noting that the MSE is convex in $\mathbf{G}$, the optimal $\mathbf{G}$ can be found by setting the derivative to 0 which results in[1]

$$\mathbf{G}\mathbf{C} + (\mathbf{G}\mathbf{X} - \mathbf{I})\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}' = 0. \tag{9}$$

Since $\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}' + \mathbf{C}$ is positive definite,

$$\mathbf{G} = \boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}' + \mathbf{C})^{-1}, \tag{10}$$

which, using the matrix inversion lemma, can be written as

$$\mathbf{G} = \left(1 - \frac{\boldsymbol{\beta}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\boldsymbol{\beta}}{1 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\boldsymbol{\beta}}\right)\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}'\mathbf{C}^{-1} = \frac{1}{1 + \boldsymbol{\beta}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\boldsymbol{\beta}}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}'\mathbf{C}^{-1}. \tag{11}$$

Evidently, the MMSE estimate depends on $\boldsymbol{\beta}$ and therefore cannot be implemented. Instead, we suggest iterating the MMSE solution (11). Specifically, given an initial estimate $\hat{\boldsymbol{\beta}}_0$ of $\boldsymbol{\beta}$, we obtain a new estimate by choosing $\hat{\boldsymbol{\beta}}_1 = \mathbf{G}(\hat{\boldsymbol{\beta}}_0)\mathbf{y}$ where $\mathbf{G}(\hat{\boldsymbol{\beta}}_0)$ is the matrix given by (11) with $\boldsymbol{\beta}$ replaced by $\hat{\boldsymbol{\beta}}_0$. We can then use $\hat{\boldsymbol{\beta}}_1$ in conjunction with $\mathbf{G}$ to further improve the solution. Continuing these iterations we obtain

$$\hat{\boldsymbol{\beta}}_k = \mathbf{G}(\hat{\boldsymbol{\beta}}_{k-1})\mathbf{y} = \frac{\hat{\boldsymbol{\beta}}_{k-1}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{y}}{1 + \hat{\boldsymbol{\beta}}_{k-1}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}_{k-1}}\hat{\boldsymbol{\beta}}_{k-1}. \tag{12}$$

In the remainder of this section we show that with $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ the iterations defined by (12) converge for large k to the IMMSE estimate $\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}$, which is a pre-test type method. In Section 4 we prove that $\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}$ dominates $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ when the effective dimension is greater than or equal to 4.

If we choose $\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, then $\hat{\boldsymbol{\beta}}_k$ of (12) can be expressed as

$$\begin{aligned}
&\hat{\boldsymbol{\beta}}_k = \alpha_k\hat{\boldsymbol{\beta}}_0, \quad k \geqslant 1,\\
&\alpha_k = \frac{\alpha_{k-1}^2 a(\mathbf{y})}{1 + \alpha_{k-1}^2 a(\mathbf{y})}, \quad k \geqslant 1,
\end{aligned} \tag{13}$$

with initial conditions

$$\hat{\boldsymbol{\beta}}_0 = \hat{\boldsymbol{\beta}}_{\mathrm{LS}}, \quad \alpha_0 = 1 \tag{14}$$

and

$$a(\mathbf{y}) = \hat{\boldsymbol{\beta}}_{\mathrm{LS}}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{y} = \hat{\boldsymbol{\beta}}_{\mathrm{LS}}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}. \tag{15}$$

Note that $a(\mathbf{y})$ is equal to the statistic u of (4) used in the pre-test estimator.

In Theorem 1 below we show that $\hat{\boldsymbol{\beta}}_k$ converges to a limit that takes on the form of a pre-test estimator.

---

[1] We use the following derivatives: For any symmetric $\mathbf{A}$,

$$\frac{\partial \mathrm{tr}(\mathbf{B}\mathbf{A}\mathbf{B}')}{\partial \mathbf{B}} = 2\mathbf{B}\mathbf{A}, \quad \frac{\partial \boldsymbol{\beta}'\mathbf{B}'\mathbf{B}\boldsymbol{\beta}}{\partial \mathbf{B}} = 2\mathbf{B}\boldsymbol{\beta}\boldsymbol{\beta}'.$$

Theorem 1. *Let* $\hat{\boldsymbol{\beta}}_k$ *be defined by the recursion* (13) *with initial conditions given by* (14). *Then* $\hat{\boldsymbol{\beta}}_{k+1} \to_{k\to\infty} \hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}$ *where*

$$\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}} = \begin{cases} 0, & a(\mathbf{y}) \leqslant 4, \\ \frac{1}{2}\left(1 + \sqrt{1 - \frac{4}{a(\mathbf{y})}}\right)\hat{\boldsymbol{\beta}}_{\mathrm{LS}}, & a(\mathbf{y}) > 4. \end{cases} \tag{16}$$

Proof. To study the convergence of $\hat{\boldsymbol{\beta}}_k$ it is sufficient to study the convergence of the sequence α_k. Now, suppose that α_k converges to a finite fixed point α. Then substituting $\alpha_k = \alpha_{k-1} = \alpha$ into (13), α must satisfy

$$\alpha(1 + \alpha^2 a(\mathbf{y})) = \alpha^2 a(\mathbf{y}), \tag{17}$$

from which we conclude that there are three possible fixed points:

$$\alpha^0 = 0, \quad \alpha^{\pm} = \frac{1}{2}\left(1 \pm \sqrt{1 - \frac{4}{a(\mathbf{y})}}\right), \tag{18}$$

where $\alpha^{\pm}$ exist only if $a(\mathbf{y}) \geqslant 4$.

To investigate which of the solutions are asymptotically stable, we rely on the fact that α is an asymptotically stable fixed point of the recursion $\alpha_k = f(\alpha_{k-1})$ if $|f'(\alpha)| < 1$ (Mickens, 1990, p. 71), where f' denotes the derivative of the function f''. In our case, $f(x) = x^2 a/(1 + x^2 a)$ where for brevity we denoted $a = a(\mathbf{y})$, and $f'(x) = 2xa/(1 + x^2 a)^2$. Since $f'(\alpha^0) = 0$ we conclude that when $a(\mathbf{y}) < 4$, $\alpha_0 = 0$ is asymptotically stable, and $\hat{\boldsymbol{\beta}}_{k+1} \to 0$. Next, suppose that $a(y) = 4$. In this case, $f'(\alpha^{\pm}) = 1$. Since $f''(\alpha^{\pm}) \neq 0$, we conclude that $\alpha^{\pm}$ are unstable and therefore $\hat{\boldsymbol{\beta}}_{k+1} \to 0$.

We now consider the case where $a(\mathbf{y}) > 4$. It is easy to see that $f'(\alpha^-) > 1$ so that this point is unstable, while $f'(\alpha^+) < 1$ rendering α^+ a stable fixed point. Thus, in principle, for $a(\mathbf{y}) > 4$ the iterations can converge to 0 or α^+. It remains to show that for $\alpha_0 = 1$, $\hat{\boldsymbol{\beta}}_{k+1} \to \alpha^+$. This can be verified by noting that $f'(x) > 0$ for $x > 0$ so that $f(x)$ is monotonically increasing, and $f'(x) < 1$ for all $\alpha_+ \leqslant x \leqslant 1$. □

To gain some insight into the IMMSE estimate, note that we can express $\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}$ for $a(\mathbf{y}) > 4$ as $\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}} = \zeta(a)\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ where

$$\zeta(a) = \frac{1}{2}\left(1 + \sqrt{1 - \frac{4}{a}}\right) \tag{19}$$

and for brevity we denoted $a = a(\mathbf{y})$. It is easy to see that $\zeta(4) = \frac{1}{2}$, $\zeta(a)$ is monotonically increasing for $a > 4$, and approaches 1 when $a \to \infty$. Therefore, for large values of a, $\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}} = \hat{\boldsymbol{\beta}}_{\mathrm{LS}}$. This is intuitively reasonable since large values of a imply that the noise is small compared to the norm of $\boldsymbol{\beta}$, in which case the LS is close to the true parameter value. Indeed, when the noise is exactly 0, $\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = \boldsymbol{\beta}$.

4. LS domination of the iterative MSE estimator

In this section we prove that $\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}$ dominates the LS method when the effective dimension is large enough.

We begin by considering the case in which $\mathbf{X} = \mathbf{C} = \mathbf{I}$. Under this assumption, the LS estimator reduces to $\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = \mathbf{y}$ and $a(\mathbf{y}) = \mathbf{y}'\mathbf{y} \sim \chi^2_{m,\lambda}$ with $\lambda = \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\beta}$.

Proposition 1. *Consider the model* (1) *with* $\mathbf{X} = \mathbf{C} = \mathbf{I}$. *Let* $\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}$ *be the IMMSE estimate of* (16) *with* $a(\mathbf{y}) = \mathbf{y}'\mathbf{y}$. Then $\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}$ *strictly dominates the LS estimator* $\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = \mathbf{y}$ *for* $m \geqslant 4$ *where m is the length of* $\boldsymbol{\beta}$.

Proof. The proof is quite involved and is therefore relegated to Appendix A.

We now use the results of Proposition 1 to establish strict dominance for the general linear model (1). In this setting, $\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}$ is given by (16) and the LS estimate is given by (3).

Theorem 2. *Let* $\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}$ *be the IMMSE estimate of* (16) *for the model* $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{e}$ *with* $a(\mathbf{y})=\hat{\boldsymbol{\beta}}_{\mathrm{LS}}'\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$. *Let* $\mathbf{Q}=(\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}$, *and denote by* $\sigma_{\max}(\mathbf{Q})$ *the largest eigenvalue of the matrix* $\mathbf{Q}$. *Then* $\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}$ *strictly dominates the LS estimator* $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ *of* (3) *for* $d\geqslant 4$ *where is the effective dimension defined by*

$$d=\frac{\mathrm{tr}(\mathbf{Q})}{\sigma_{\max}(\mathbf{Q})}. \tag{20}$$

Proof. We begin by noting that in the general linear model the MSE of $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ is

$$E\{\|\hat{\boldsymbol{\beta}}_{\mathrm{LS}}-\boldsymbol{\beta}\|^2\}=\mathrm{tr}(\mathbf{Q})\triangleq\varepsilon_0. \tag{21}$$

Expressing the IMMSE estimate as $\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}=(1-\delta(a))\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ with

$$\delta(a)=(1-\zeta(a))I_{(4,\infty)}+I_{[0,4]} \tag{22}$$

and $\zeta(a)$ given by (19), the MSE can be computed as

$$\begin{aligned}E\{\|\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}-\boldsymbol{\beta}\|^2\}&=E\{\|(\hat{\boldsymbol{\beta}}_{\mathrm{LS}}-\boldsymbol{\beta})-\delta(a)\hat{\boldsymbol{\beta}}_{\mathrm{LS}}\|^2\}\\&=\varepsilon_0+E\{\hat{\boldsymbol{\beta}}_{\mathrm{LS}}'\hat{\boldsymbol{\beta}}_{\mathrm{LS}}\delta^2(a)\}+2E\{\delta(a)\hat{\boldsymbol{\beta}}_{\mathrm{LS}}'(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\mathrm{LS}})\},\end{aligned} \tag{23}$$

where we used (21). Denoting

$$\widetilde{\Delta}=-E\{\hat{\boldsymbol{\beta}}_{\mathrm{LS}}'\hat{\boldsymbol{\beta}}_{\mathrm{LS}}\delta^2(a)\}-2E\{\delta(a)\hat{\boldsymbol{\beta}}_{\mathrm{LS}}'(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}_{\mathrm{LS}})\}, \tag{24}$$

to prove the theorem we need to show that $\widetilde{\Delta}>0$ for $d\geqslant 4$.

In order to obtain a more convenient characterization of $\widetilde{\Delta}$ we use the eigendecomposition of $\mathbf{Q}$. Specifically, since $\mathbf{Q}$ is positive definite, it has an eigendecomposition of the form $\mathbf{Q}=\mathbf{V}\Sigma\mathbf{V}'$ where $\mathbf{V}$ is an $m\times m$ unitary matrix and Σ is a diagonal matrix with diagonal elements $\sigma_i>0$, $1\leqslant i\leqslant m$. Next, define

$$\mathbf{v}=\mathbf{V}\mathbf{Q}^{-1/2}\hat{\boldsymbol{\beta}}_{\mathrm{LS}},\quad \mathbf{u}=\mathbf{V}\mathbf{Q}^{-1/2}\boldsymbol{\beta}. \tag{25}$$

With these definitions we have

$$\begin{aligned}&\hat{\boldsymbol{\beta}}_{\mathrm{LS}}'\hat{\boldsymbol{\beta}}_{\mathrm{LS}}=\mathbf{v}'\Sigma\mathbf{v},\\&\hat{\boldsymbol{\beta}}_{\mathrm{LS}}'\boldsymbol{\beta}=\mathbf{v}'\Sigma\mathbf{u},\\&a(\mathbf{y})=\hat{\boldsymbol{\beta}}_{\mathrm{LS}}'\mathbf{Q}^{-1}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}=\mathbf{v}'\mathbf{v}.\end{aligned} \tag{26}$$

Using (26) we can express $\widetilde{\Delta}$ of (24) as

$$\widetilde{\Delta}=-E\{\delta^2(\mathbf{v}'\mathbf{v})\mathbf{v}'\Sigma\mathbf{v}\}+2E\{\delta(\mathbf{v}'\mathbf{v})\mathbf{v}'\Sigma\mathbf{v}\}-2E\{\delta(\mathbf{v}'\mathbf{v})\mathbf{v}'\Sigma\mathbf{u}\}. \tag{27}$$

To evaluate $\widetilde{\Delta}$ we rely on the following lemma (Judge and Bock, 1978, p. 322, Theorem 2).

Lemma 1. *Let* $\mathbf{y}\sim\mathcal{N}(\boldsymbol{\beta},\mathbf{I}_m)$. *Then* $E\{g(\mathbf{y}'\mathbf{y})\mathbf{y}'\Sigma\mathbf{y}\}=\mathrm{tr}(\Sigma)E\{g(\chi^2_{m+2,\lambda})\}+\boldsymbol{\beta}'\Sigma\boldsymbol{\beta}E\{g(\chi^2_{m+4,\lambda})\}$ *where* $\lambda=\boldsymbol{\beta}'\boldsymbol{\beta}/2$, *for any function g for which the expectations are defined.*

Using the fact that $\mathbf{v}\sim\mathcal{N}(\mathbf{u},\mathbf{I}_m)$ together with Lemmas 1 and 2 (see Appendix A), (27) becomes

$$\begin{aligned}\widetilde{\Delta}=&-\mathrm{tr}(\Sigma)E\{\delta^2(\chi^2_{m+2,\lambda})\}-\mathbf{u}'\Sigma\mathbf{u}E\{\delta^2(\chi^2_{m+4,\lambda})\}+2\mathrm{tr}(\Sigma)E\{\delta(\chi^2_{m+2,\lambda})\}\\&+2\mathbf{u}'\Sigma\mathbf{u}E\{(\chi^2_{m+4,\lambda})\}-2\mathbf{u}'\Sigma\mathbf{u}E\{\delta(\chi^2_{m+2,\lambda})\}\\=&\,\varepsilon_0E\{2\delta(\chi^2_{m+2,\lambda})-\delta^2(\chi^2_{m+2,\lambda})\}-\mathbf{u}'\Sigma\mathbf{u}E\{\delta^2(\chi^2_{m+4,\lambda})+2\delta(\chi^2_{m+2,\lambda})-2\delta(\chi^2_{m+4,\lambda})\}\\\triangleq&\,\varepsilon_0\xi_1(\lambda,m)-\mathbf{u}'\Sigma\mathbf{u}\xi_2(\lambda,m),\end{aligned} \tag{28}$$

where $\lambda = \mathbf{u}'\mathbf{u}/2$. Now, since $0 \leqslant \delta(x) \leqslant 1$, $2\delta(x) \geqslant \delta(x) \geqslant \delta^2(x)$ for all x, and consequently $\xi_1(\lambda, m) \geqslant 0$ for all λ. We also have that $\xi_2(\lambda, m) \geqslant 0$ due to the fact that $\delta(x)$ is decreasing with x so that $E\{\delta(\chi^2_{m+2,\lambda}) - \delta(\chi^2_{m+4,\lambda})\} \geqslant 0$. Therefore,

$$\begin{aligned}\widetilde{\Delta} &\geqslant \varepsilon_0\xi_1(\lambda, m) - \sigma_{\max}\mathbf{u}'\mathbf{u}\xi_2(\lambda, m)\\ &= \sigma_{\max}(d\xi_1(\lambda, m) - 2\lambda\xi_2(\lambda, m))\\ &\geqslant \sigma_{\max}\,(4\xi_1(\lambda, m) - 2\lambda\xi_2(\lambda, m)) \triangleq \sigma_{\max}\widetilde{\Delta}_0(\lambda, m),\end{aligned} \tag{29}$$

where $\sigma_{\max} = \sigma_{\max}(\mathbf{Q})$, and the last inequality is the result of the fact that $d = \varepsilon_0/\sigma_{\max} \geqslant 4$ and $\xi_1(\lambda, m) \geqslant 0$. Using the fact that $\mathrm{tr}(\mathbf{Q}) \leqslant m\sigma_{\max}(\mathbf{Q})$, we have immediately $d \leqslant m$, so that the requirement $d \geqslant 4$ implies $m \geqslant 4$. Therefore, to show that $\widetilde{\Delta} > 0$ for $d \geqslant 4$, it is sufficient to show that $\widetilde{\Delta}_0(\lambda, m) > 0$ for $m \geqslant 4$.

Proposition 2. *$\widetilde{\Delta}_0(\lambda, m) = 4\xi_1(\lambda, m) - 2\lambda\xi_2(\lambda, m) > 0$ for all λ and $m \geqslant 4$.*

Proof. See Appendix A.

Applying Proposition 2 completes the proof of the theorem.

5. Examples

To illustrate our method, we compare it with the conventional pre-test estimator of (5) and the Bock method (7), which is an extension of the positive-part James–Stein estimator to the general linear model (1).

It is well known that the Bock approach dominates LS, when the effective dimension $d > 2$, while the pre-test estimator does not for any value of d. Nonetheless, due to the test nature of $\hat{\boldsymbol{\beta}}_{\mathrm{PT}}$ it has been the estimator of choice in many applications, despite the fact that its MSE performance is inferior over a large range of values to that of $\hat{\boldsymbol{\beta}}_{\mathrm{JS}}$. The proposed estimator, can be seen as a compromise between these two approaches. On the one hand, we preserve the test nature of the pre-test method. But, at the same time, this strategy dominates LS and has performance closer to that of the Bock approach over a wide range of values.

In Fig. 3 we plot the MSE as a function of $\|\boldsymbol{\beta}\|^2/n$, where all components of $\boldsymbol{\beta}$ are equal, using $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}$, $\hat{\boldsymbol{\beta}}_{\mathrm{JS}}$ and $\hat{\boldsymbol{\beta}}_{\mathrm{PT}}$ for $\mathbf{C} = \mathbf{I}$, $n = 4$, and $\mathbf{X}$ chosen as a diagonal matrix with diagonal elements 1.3, 1.3, 0.7, 0.7. In computing $\hat{\boldsymbol{\beta}}_{\mathrm{PT}}$ we used $\alpha = 0.05$. For each value of $\|\boldsymbol{\beta}\|$, the MSE was computed by averaging over 25,000 noise realizations. It is apparent from the figure that for the parameters chosen the IMMSE estimate has lower MSE than the Bock method over the entire parameter range. Both of these methods have lower MSE than the pre-test approach for intermediate and high values of $\|\boldsymbol{\beta}\|$.

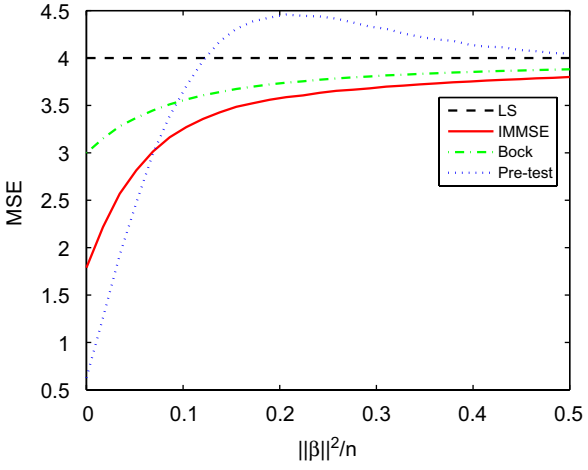

Fig. 3. MSE of the LS, IMMSE, Bock and pre-test estimators as a function of the norm of $\boldsymbol{\beta}$ for $n = 4$.

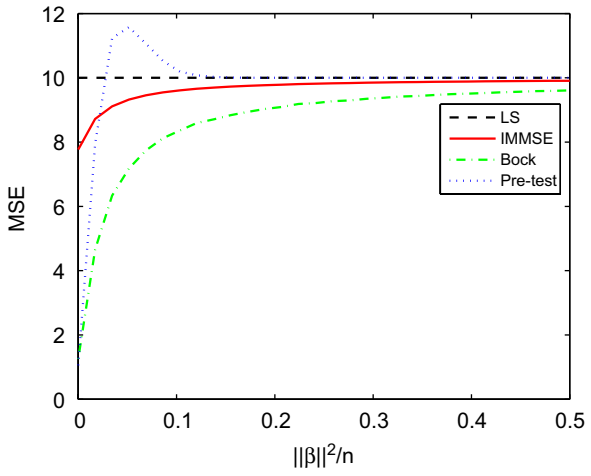

Fig. 4. MSE of the LS, IMMSE, Bock and pre-test estimators as a function of the norm of $\boldsymbol{\beta}$ for $n = 10$.

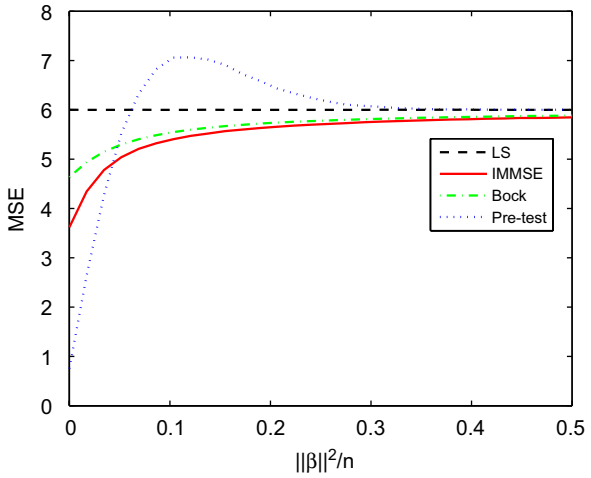

Fig. 5. MSE of the LS, IMMSE, Bock and pre-test estimators as a function of the norm of $\boldsymbol{\beta}$ for $n = 6$ and a random choice of $\mathbf{X}$.

In general, the IMMSE method does not dominate the Bock approach, as can be seen in Fig. 4. Here we repeat the simulations used to generate Fig. 3 with $n = 10$ and $\mathbf{X} = \mathbf{I}$. Evidently, the IMMSE has higher MSE than the Bock method in this setting, but still performs better than the pre-test method over most of the parameter range.

To evaluate the performance for more general choices of $\mathbf{X}$, in Fig. 5 we plot the MSE as a function of $\|\boldsymbol{\beta}\|^2/n$, with $n = 6$ and $\mathbf{X}$ chosen at random. Specifically, $\mathbf{X} = \mathbf{I} + \mathbf{R}$ where the elements of $\mathbf{R}$ are zero-mean, Gaussian, independent and identically distributed (iid) random variables with variance 0.04. The rest of the parameters remain as in Fig. 3. Evidently, even in the case in which $\mathbf{X}$ is random, the results are similar to those of Fig. 3. In particular, the IMMSE estimate has lower MSE than the Bock method.

The examples illustrate that the pre-test method has good performance only for very small values of $\|\boldsymbol{\beta}\|$. In most of the regime, its performance is inferior to that of the Bock and IMMSE approaches. Although the IMMSE method is not guaranteed to perform better than the Bock estimate, it preserves the test nature of the pre-test strategy while maintaining comparable performance to that of the Bock approach. Most importantly, it dominates LS for all values of $\boldsymbol{\beta}$. Therefore, it may be a good alternative to the conventional pre-test strategy in many settings.

6. Conclusion

We proposed a pre-test type estimator that dominates the traditional LS method when the effective dimension is large enough. The estimator is derived from iterations on the optimal linear MSE solution, with the LS as the initial condition. We then demonstrated via numerical examples that the IMMSE approach can be a good alternative to the pre-test estimate, since it preserves the test nature of the method, while resulting in MSE performance that is superior to that of the pre-test over most of the parameter range, and is comparable to that of the Bock (James–Stein) strategy.

The IMMSE estimator can be modified by changing the threshold value of 4. Initial experiments show that using an adaptive threshold (which depends on the signal dimension) can improve the performance. An interesting question for future research is whether the IMMSE solution can be improved by using a more sophisticated choice of threshold, rather than the choice of 4 which is an outcome of the iterations.

Acknowledgements

The authors would like to thank Dr. Arie Yeredor and Tsvi Dvorkind for many fruitful discussions. They would also like to thank the reviewers and associate editor for helpful comments and suggestions.

Appendix A.

Proof of Proposition 1. To prove the proposition we first note that the MSE of $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ is

$$E\{\|\mathbf{y}-\boldsymbol{\beta}\|^2\} = E\{\|\mathbf{e}\|^2\} = m, \tag{30}$$

where we used the fact that $\mathbf{y} = \boldsymbol{\beta} + \mathbf{e}$.

Now, the IMMSE estimate can be expressed as

$$\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}} = (1-\delta(a))\mathbf{y}, \tag{31}$$

where we defined

$$\delta(a) = (1-\zeta(a))I_{(4,\infty)} + I_{[0,4]}, \tag{32}$$

with $\zeta(a)$ given by (19). Using (31) the MSE of $\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}$ can be computed as

$$E\{\|\hat{\boldsymbol{\beta}}_{\mathrm{IMMSE}}-\boldsymbol{\beta}\|^2\} = E\{\|(\mathbf{y}-\boldsymbol{\beta})-\delta(a)\mathbf{y}\|^2\} = m + E\{a\delta^2(a)\} + 2E\{\delta(a)\mathbf{y}'(\boldsymbol{\beta}-\mathbf{y})\}, \tag{33}$$

where we used (30) and the fact that $a = \mathbf{y}'\mathbf{y}$. Therefore, to prove the proposition we need to show that $\Delta > 0$ where Δ is defined by

$$\Delta = -E\{a\delta^2(a)\} - 2E\{\delta(a)\mathbf{y}'(\boldsymbol{\beta}-\mathbf{y})\}. \tag{34}$$

To this end we first rely on the following lemma Judge and Bock (1978, p. 321, Theorem 1).

Lemma 2. *Let* $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\beta}, \mathbf{I}_m)$. *Then* $E\{g(\mathbf{y}'\mathbf{y})\mathbf{y}\} = \boldsymbol{\beta}E\{g(\chi^2_{m+2,\lambda})\}$ *where* $\lambda = \boldsymbol{\beta}'\boldsymbol{\beta}/2$, *for any function g for which the expectations are defined.*

Using Lemma 2 and the fact that $a = \mathbf{y}'\mathbf{y} \sim \chi^2_{m,\lambda}$ with $\lambda = \boldsymbol{\beta}'\boldsymbol{\beta}/2$, we can write Δ of (34) as

$$\begin{aligned}\Delta &= -E\{\chi^2_{m,\lambda}\delta^2(\chi^2_{m,\lambda})\} - 4\lambda E\{\delta(\chi^2_{m+2,\lambda})\} + 2E\{\chi^2_{m,\lambda}\delta(\chi^2_{m,\lambda})\}\\ &= E\{\chi^2_{m,\lambda}\delta(\chi^2_{m,\lambda})(1-\delta(\chi^2_{m,\lambda}))\} - E\{4\lambda\delta(\chi^2_{m+2,\lambda})\} + E\{\chi^2_{m,\lambda}\delta(\chi^2_{m,\lambda})\}\\ &\triangleq f_1 - f_2 + f_3.\end{aligned}$$

Let us first examine f_1. Note that $a\delta(a)(1-\delta(a)) = I_{(4,\infty)}(a)$. Therefore,

$$f_1 = E\{I_{(4,\infty)}(a)\} = P(a>4) = 1 - F_{\chi^2}(4|m,\lambda), \tag{35}$$

where $F_{\chi^2}(a|m,\lambda)$ is the cumulative distribution function (CDF) of $\chi^2_{m,\lambda}$, and is given by

$$F_{\chi^2}(a|m,\lambda)=\sum_{k=0}^{\infty}\frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}\frac{\gamma(m/2+k,a/2)}{\Gamma(m/2+k)}, \tag{36}$$

where $\gamma(m,a)$ is the lower incomplete Gamma function and $\Gamma(m)$ is the ordinary Gamma function (see Appendix B). Using relation (63) and the fact that $\sum_{k=0}^{\infty}\mathrm{e}^{-\lambda}\lambda^k/k!=1$, f_1 can be written as

$$f_1=\sum_{k=0}^{\infty}\frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}\frac{\Gamma(m/2+k,2)}{\Gamma(m/2+k)}\triangleq\mathrm{e}^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}h_1(k,m), \tag{37}$$

where we defined

$$h_1(k,m)=\frac{\Gamma(m/2+k,2)}{\Gamma(m/2+k)}. \tag{38}$$

We next evaluate f_2. To this end we use the probability density function (pdf) of the Noncentral Chi-square distribution

$$f_{\chi^2}(a|m,\lambda)=\sum_{k=0}^{\infty}\frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}f_{\chi^2}(a|m+2k), \tag{39}$$

where $f_{\chi^2}(a|m)$ is the pdf of the central chi-square distribution

$$f_{\chi^2}(a|m)=\frac{\mathrm{e}^{-a/2}a^{m/2-1}}{2^{m/2}\Gamma(m/2)}. \tag{40}$$

Evaluating the expectation using (39) we have

$$\begin{aligned}4\lambda E\{\delta(\chi^2_{m+2,\lambda})\}&=4\lambda\int_0^{\infty}\delta(x)f_{\chi^2}(x|m+2,\lambda)\,\mathrm{d}x\\&=\sum_{k=0}^{\infty}\frac{4\mathrm{e}^{-\lambda}\lambda^{k+1}}{k!}\int_0^{\infty}\frac{\delta(x)x^{m/2+k}\mathrm{e}^{-x/2}0.5^{m/2+k+1}}{\Gamma(m/2+k+1)}\,\mathrm{d}x\underset{x\to 2x,k\to k-1}{=}\\&=\mathrm{e}^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}4k\int_0^{\infty}\frac{\delta(2x)x^{m/2+k-1}\mathrm{e}^{-x}}{\Gamma(m/2+k)}\,\mathrm{d}x\\&\triangleq\mathrm{e}^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}h_2(k,m).\end{aligned} \tag{41}$$

Similarly, f_3 can be evaluated as

$$\begin{aligned}E\{\chi^2_{m,\lambda}\delta(\chi^2_{m,\lambda})\}&=\int_0^{\infty}x\delta(x)f_{\chi^2}(x|m,\lambda)\,\mathrm{d}x\\&=\sum_{k=0}^{\infty}\frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}\int_0^{\infty}\frac{\delta(x)x^{m/2+k}\mathrm{e}^{-x/2}0.5^{m/2+k}}{\Gamma(m/2+k)}\,\mathrm{d}x\underset{x\to 2x}{=}\\&=\mathrm{e}^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}2\int_0^{\infty}\frac{\delta(2x)x^{m/2+k}\mathrm{e}^{-x}}{\Gamma(m/2+k)}\,\mathrm{d}x\triangleq\mathrm{e}^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}h_3(k,m).\end{aligned} \tag{42}$$

Substituting (38), (41) and (42) into (34) we conclude that

$$\Delta = \mathrm{e}^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}[h_1(k,m)-h_2(k,m)+h_3(k,m)] \triangleq \mathrm{e}^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}h(k,m). \tag{43}$$

Therefore, to prove Proposition 1 it is sufficient to show that $h(k,m)\geqslant 0$ for all $k\in\mathbb{N}$, and that $h(0,m)>0$ for $m\geqslant 4$.

To this end, we first establish the following lemma.

Lemma 3. *Let* $h(k,m)=h_1(k,m)-h_2(k,m)+h_3(k,m)$ *with* $h_i(k,m)$, $1\leqslant i\leqslant 3$ *given by* (38), (41) *and* (42). *Then* $h(0,4)>0$ *and also* $h(k,4)\geqslant 0$ *for integers* $1\leqslant k\leqslant 5$ *and real values* $k\geqslant 6$.

Proof. We begin by bounding $\delta(a)$ from below (for $h_3(k,m)$) and above (for $h_2(k,m)$) using its Taylor series expansion:

$$\frac{1}{4}\left(\frac{2}{x}+\frac{1}{x^2}+\frac{1}{x^3}\right)\leqslant\delta(2x)\leqslant\frac{1}{4}\left(\frac{2}{x}+\frac{1}{x^2}+\frac{6}{x^3}\right)\quad\text{for } x\geqslant 2. \tag{44}$$

Substituting (44) into (41) and (42) will allow us to replace the integrals with Gamma functions, yielding a lower bound on $h(k,m)$. Specifically,

$$\begin{aligned}
h_2(k,m) &= 4k\int_0^{\infty}\frac{\delta(2x)x^{m/2+k-1}\mathrm{e}^{-x}}{\Gamma(m/2+k)}\,\mathrm{d}x\\
&\leqslant \frac{k}{\Gamma(m/2+k)}\left(4\int_0^{2}x^{m/2+k-1}\mathrm{e}^{-x}\,\mathrm{d}x+\int_2^{\infty}\mathrm{e}^{-x}(2x^{m/2+k-2}+x^{m/2+k-3}+6x^{m/2+k-4})\,\mathrm{d}x\right)\\
&= \frac{k}{\Gamma(m/2+k)}(4\gamma(m/2+k,2)+2\Gamma(m/2+k-1,2)+\Gamma(m/2+k-2,2)\\
&\quad +6\Gamma(m/2+k-3,2)).
\end{aligned}$$

Similarly,

$$\begin{aligned}
h_3(k,m) &\geqslant \frac{1}{2\Gamma(m/2+k)}(4\gamma(m/2+k+1,2)+2\Gamma(m/2+k,2)+\Gamma(m/2+k-1,2)\\
&\quad +\Gamma(m/2+k-2,2)).
\end{aligned}$$

Using (63) and (65) which can be found in Appendix B, we have for $k\geqslant 6$,

$$h(k,4)\geqslant(4-2k)+\left(2k-4+\frac{1-4/k}{(k+1)}\right)\frac{\Gamma(k+2,2)}{\Gamma(k+2)}-\frac{\mathrm{e}^{-2}2^{k+3}}{\Gamma(k+2)}. \tag{45}$$

We now show that $h(k,4)\geqslant 0$ for $k\geqslant 6$:

$$\begin{aligned}
h(k,4) &\geqslant (4-2k)+\left(2k-4+\frac{1-4/k}{(k+1)}\right)\left(1-\frac{\gamma(k+2,2)}{\Gamma(k+2)}\right)-\frac{\mathrm{e}^{-2}2^{k+3}}{\Gamma(k+2)}\\
&= \frac{1-4/k}{(k+1)}-\left(2k-4+\frac{1-4/k}{(k+1)}\right)\frac{\gamma(k+2,2)}{\Gamma(k+2)}-8\frac{\mathrm{e}^{-2}2^{k}}{\Gamma(k+2)}\\
&\geqslant \frac{1-4/k}{(k+1)}-\left(\frac{2k-3}{k+1}\right)\frac{\gamma(k+2,2)}{\Gamma(k+1)}-8\frac{\mathrm{e}^{-2}2^{k}}{(k+1)\Gamma(k+1)}\\
&= \frac{1}{(k+1)}\left(1-\frac{4}{k}-\frac{8\mathrm{e}^{-2}2^{k}}{\Gamma(k+1)}-\frac{(2k-3)\gamma(k+2,2)}{\Gamma(k+1)}\right)\triangleq\frac{1}{(k+1)}g(k),
\end{aligned}$$

where the first inequality follows from (63) applied to the ratio of Gamma functions, the second inequality is true since $(1-4/k)/(k+1)<1$, and the last equality follows from (61). Since $g(k)$ is monotonically increasing with k, and $g(6)>0$, we have that $h(k,4)\geqslant 0$ for $k\geqslant 6$. Note that the inequality is valid for all $k\geqslant 6$, not only integer values.

Unfortunately, applying the bounds (44) to $h_2(k,4)$ and $h_3(k,4)$ for $k\leqslant 5$ is not sufficient to establish $h(k,4)\geqslant 0$ for $k\leqslant 5$. In order to prove this we need to use higher order approximations of $\delta(2x)$. However, this will result in integrals of the form $\int x^{t-1}\mathrm{e}^{-x}\,\mathrm{d}x$ with $t<1$ which are no longer Gamma functions. To evaluate such integrals we rely on partial integration detailed in Appendix C which allows us to apply an approximation of any desired order to $\delta(2x)$, leading to a tighter lower bound on $h(k,4)$. Using this approach, we obtain the following lower bounds, where the number in the brackets indicates the order of approximation used in the Taylor expansion of $\delta(2x)$:

$$\begin{aligned}
&h(1,4)\geqslant 0.3421 \quad (N_{\mathrm{app}}=2),\\
&h(2,4)\geqslant 0.0085 \quad (N_{\mathrm{app}}=10),\\
&h(3,4)\geqslant 0.0065 \quad (N_{\mathrm{app}}=10),\\
&h(4,4)\geqslant 0.0543 \quad (N_{\mathrm{app}}=5),\\
&h(5,4)\geqslant 0.0912 \quad (N_{\mathrm{app}}=4).
\end{aligned} \tag{46}$$

Finally, for $k=0$, $h_2(0,m)=0$. Since $h_1(0,m)$ and $h_3(0,m)$ are positive, then $h(0,m)>0$, completing the proof. □

From Lemma 3 it follows that Proposition 1 holds true for $m=4$. We now use induction on m to show that the results can be extended to any $m>4$.

Lemma 4. *Let $h(k,m)\geqslant 0$ and $h(0,m)>0$ for some integer $m\geqslant 1$ and all integers $k\geqslant 1$. Then $h(k,m+2\ell)\geqslant 0$ and $h(0,m+2\ell)>0$ for all integers $k,\ell\geqslant 1$.*

Proof. To prove the lemma, note that

$$\begin{aligned}
h(k,m+1) &= h_1(k,m+1)-h_2(k,m+1)+h_3(k,m+1)\\
&= h_1(k+0.5,m)-h_2(k+0.5,m)+\frac{h_2(k,m+1)}{2k}+h_3(k+0.5,m)\\
&= h(k+0.5,m)+\frac{h_2(k,m+1)}{2k}\\
&\geqslant h(k+0.5,m),
\end{aligned} \tag{47}$$

where the first and second equalities follow from the definitions of $h_i(k,m)$, and the last inequality is a result of the fact that $h_2(k,m)\geqslant 0$. From (47) we conclude that $h(k,m+2)\geqslant h(k+0.5,m+1)\geqslant h(k+1,m)$, which proves the lemma. □

Using Lemma 4 it follows immediately that $h(k,m)\geqslant 0$ for all m even with $m\geqslant 4$. If we can prove that $h(k,5)\geqslant 0$ for all integers $k\geqslant 1$, then it will follow that $h(k,m)\geqslant 0$ for all m odd with $m\geqslant 5$, which in turn means that $h(k,m)\geqslant 0$ for all $m\geqslant 4$ completing the proof of the proposition.[2] It therefore remains to prove that $h(k,5)\geqslant 0$ for all integers $k\geqslant 1$. To this end we rely on the following lemma.

Lemma 5. *Let $h(k+\frac{1}{2},m)\geqslant 0$ for some m and all integer k, $k\geqslant N_0$. Then $h(k,m+2\ell-1)\geqslant 0$ for all integer k, $k\geqslant N_0$, and $\ell\geqslant 1$.*

Proof. The proof follows immediately from (47). □

Since we have shown in Lemma 3 that $h(k,4)\geqslant 0$ for all $k\geqslant 6$ (not only integer values) it follows immediately from Lemma 5 that $h(k,5)\geqslant 0$ for $k\geqslant 6$. However, we still need to prove that $h(k,5)\geqslant 0$ for integers k in the range $1\leqslant k\leqslant 5$

[2] Notice that for all m and $k=0$, the inequality is strict, i.e. $h(0,m)>0$.

(note that $h(0,5)>0$ since $h_2(0,5)=0$). As in (46), we use upper and lower bounds on $\delta(2x)$ to evaluate lower bounds on the required integrals, resulting in:

$$\begin{aligned}
&h(1,5)\geqslant 0.6902 \quad (N_{\text{app}}=2),\\
&h(2,5)\geqslant 0.2719 \quad (N_{\text{app}}=3),\\
&h(3,5)\geqslant 0.1868 \quad (N_{\text{app}}=3),\\
&h(4,5)\geqslant 0.2033 \quad (N_{\text{app}}=3),\\
&h(5,5)\geqslant 0.2173 \quad (N_{\text{app}}=3).
\end{aligned} \tag{48}$$

Evidently, Proposition 1 holds for $m=4,5$. Therefore, using Lemma 4 it follows that it is correct for all $m\geqslant 4$. □

Proof of Proposition 2. To prove the proposition, we begin by noting that the difference in MSE between the LS and IMMSE estimators in the iid case, Δ of (34), can be expressed as

$$\Delta(\lambda,m)=m\xi_1(\lambda,m)-2\lambda\xi_2(\lambda,m), \tag{49}$$

where we used the fact that in this case $\mathbf{Q}=\mathbf{I}$ and subsequently $d=\varepsilon_0=m$. Using (29) and (49), we have that

$$\widetilde{\Delta}_0(\lambda,m)=\Delta(\lambda,m)-(m-4)\xi_1(\lambda,m). \tag{50}$$

We next express $\xi_1(\lambda,m)$ as

$$\begin{aligned}
\xi_1(\lambda,m)&=E\{2\delta(\chi^2_{m+2,\lambda})-\delta^2(\chi^2_{m+2,\lambda})\}=\int_0^\infty (2\delta(x)-\delta^2(x))\,f_{\chi^2}(x|m+2,\lambda)\,\mathrm{d}x\\
&=\sum_{k=0}^\infty \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}\int_0^\infty \frac{(2\delta(x)-\delta^2(x))x^{m/2+k}\mathrm{e}^{-x/2}}{2^{m/2+k+1}\Gamma(m/2+k+1)}\,\mathrm{d}x\\
&\triangleq \sum_{k=0}^\infty \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}h_4(k,m),
\end{aligned} \tag{51}$$

where $f_{\chi^2}(x|m,\lambda)$ is the pdf of $\chi^2_{m,\lambda}$, defined by (39) and (40). Using (43), (50) and (51) we can write

$$\widetilde{\Delta}_0(\lambda,m)=\sum_{k=0}^\infty \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}[h_1(k,m)-h_2(k,m)+h_3(k,m)-(m-4)h_4(k,m)]. \tag{52}$$

Since $\lambda\geqslant 0$ by definition, to complete the proof of the proposition, it is sufficient to show that $q(k,m)\geqslant 0$ and $q(0,m)>0$ for all $m\geqslant 4$, where we defined

$$q(k,m)=h_1(k,m)-h_2(k,m)+h_3(k,m)-(m-4)h_4(k,m). \tag{53}$$

First let us show that $q(0,m)>0$. From (52) we have that $q(0,m)=\widetilde{\Delta}_0(0,m)$. Inserting $\lambda=0$ into (29) and using the fact that $\xi_1(\lambda,m)>0$ we get,

$$q(0,m)=\widetilde{\Delta}_0(0,m)=4\xi_1(0,m)>0. \tag{54}$$

Next we show that $q(k,m)\geqslant 0$ for $k\geqslant 1$, $m\geqslant 4$. Using (52) and the definitions of $h_i(k,m)$,

$$\begin{aligned}
q(k,m+1)&=h_1(k,m+1)-h_2(k,m+1)+h_3(k,m+1)-(m+1-4)h_4(k,m+1)\\
&=h_1(k+0.5,m)-h_2(k+0.5,m)+\frac{h_2(k,m+1)}{2k}+h_3(k+0.5,m)\\
&\quad-(m+1-4)h_4(k+0.5,m)\\
&=q(k+0.5,m)+\frac{h_2(k,m+1)}{2k}-h_4(k,m+1),
\end{aligned} \tag{55}$$

where h_1, h_2, h_3 and h_4 are defined in (37), (41), (42) and (51), respectively. From (41) and (51),

$$\begin{aligned}
&\frac{h_2(k,m+1)}{2k} - h_4(k,m+1)\\
&= \int_0^\infty \frac{2\delta(x)x^{m/2+k-0.5}\mathrm{e}^{-x/2}}{2^{m/2+k+0.5}\Gamma(m/2+k+0.5)}\,\mathrm{d}x - \int_0^\infty \frac{(2\delta(x)-\delta^2(x))x^{m/2+k+0.5}\mathrm{e}^{-x/2}}{2^{m/2+k+1.5}\Gamma(m/2+k+1.5)}\,\mathrm{d}x\\
&= \int_0^\infty 2\delta(x) f_{\chi^2}(x|m+1+2k)\,\mathrm{d}x - \int_0^\infty (2\delta(x)-\delta^2(x)) f_{\chi^2}(x|m+3+2k)\,\mathrm{d}x\\
&= E\{2\delta(\chi^2_{m+1+2k})\} - E\{2\delta(\chi^2_{m+3+2k}) - \delta^2(\chi^2_{m+3+2k})\}\\
&\geqslant E\{2\delta(\chi^2_{m+3+2k})\} - E\{2\delta(\chi^2_{m+3+2k}) - \delta^2(\chi^2_{m+3+2k})\}\\
&= E\{\delta^2(\chi^2_{m+3+2k})\} \geqslant 0,
\end{aligned} \tag{56}$$

where the inequality is due to the fact that $\delta(x)$ is monotonically decreasing in x, and consequently $E\{\delta(\chi^2_m)\}$ is monotonically decreasing in m. Substituting (56) into (55),

$$\begin{aligned}
&q(k,m+1) \geqslant q(k+0.5,m),\\
&q(k,m+2) \geqslant q(k+1,m).
\end{aligned} \tag{57}$$

From Lemma 3, $h(k,4) = q(k,4) \geqslant 0$ for all integer k as well as for all real $k \geqslant 6$. Combined with (57) this implies that $q(k,5) \geqslant 0$ for $k \geqslant 6$. For $m=5$ and $k=1,2,3,4,5$ as in (46), we use upper and lower bounds on $\delta(2x)$ to evaluate lower bounds on $q(k,5)$:

$$\begin{aligned}
&q(1,5) \geqslant 0.3055 \quad (N_{\mathrm{app}}=2),\\
&q(2,5) \geqslant 0.1005 \quad (N_{\mathrm{app}}=5),\\
&q(3,5) \geqslant 0.0843 \quad (N_{\mathrm{app}}=5),\\
&q(4,5) \geqslant 0.1296 \quad (N_{\mathrm{app}}=5),\\
&q(5,5) \geqslant 0.1551 \quad (N_{\mathrm{app}}=5).
\end{aligned} \tag{58}$$

Since we have shown that Proposition 2 holds for $m=4,5$, recursively using (57) it follows that the proposition is true for all $m \geqslant 4$. □

Appendix B. Gamma function identities

In this appendix we summarize several Gamma function identities that we use in various parts of the paper. Throughout, it is assumed that $x, y \geqslant 0$.

We begin by recalling the definition of the Gamma function:

$$\Gamma(y) = \int_0^\infty t^{y-1}\mathrm{e}^{-t}\,\mathrm{d}t. \tag{59}$$

The upper and lower Gamma functions are defined by

$$\Gamma(y,x) = \int_x^\infty t^{y-1}\mathrm{e}^{-t}\,\mathrm{d}t, \quad \gamma(y,x) = \int_0^x t^{y-1}\mathrm{e}^{-t}\,\mathrm{d}t. \tag{60}$$

A fundamental property of the Gamma function is the recursion

$$\Gamma(y+1) = y\Gamma(y). \tag{61}$$

For integer values we have

$$\Gamma(k+1) = k! \quad \text{for } k \in \mathbb{N}. \tag{62}$$

The following identities are used in several of the proofs in the paper:

$$\frac{\gamma(y,x)+\Gamma(y,x)}{\Gamma(y)} \equiv 1, \tag{63}$$

$$\frac{\Gamma(k,x)}{\Gamma(k)} = \mathrm{e}^{-x}\sum_{m=0}^{k-1}\frac{x^m}{m!} = 1-\mathrm{e}^{-x}\sum_{m=k}^{\infty}\frac{x^m}{\Gamma(m+1)} \quad \text{for } k\in\mathbb{N}, \tag{64}$$

$$\Gamma(y,x) = \frac{\Gamma(y+1,x)-x^y\mathrm{e}^{-x}}{y}, \quad \gamma(y,x) = \frac{\gamma(y+1,x)+x^y\mathrm{e}^{-x}}{y}, \quad \text{for} \quad y\neq 0. \tag{65}$$

Appendix C. Gamma "like" functions

In this appendix we show how to evaluate Gamma-like integrals with negative powers.

Suppose first that $\ell = 2k > 0$ is an even positive integer, and consider the integral $\int_x^\infty t^{-(\ell+1)}\mathrm{e}^{-t}\,\mathrm{d}t$ for $x>0$. Using integration by parts we have that

$$\begin{aligned}
\int_x^\infty \frac{\mathrm{e}^{-t}}{t^{\ell+1}}\,\mathrm{d}t &= \left[-\frac{\mathrm{e}^{-t}}{\ell t^\ell}\right]\Big|_{t=x}^{t=\infty} - \frac{1}{\ell}\int_x^\infty \frac{\mathrm{e}^{-t}}{t^\ell}\,\mathrm{d}t \\
&= \frac{\mathrm{e}^{-x}}{\ell x^\ell} - \frac{1}{\ell}\int_x^\infty \frac{\mathrm{e}^{-t}}{t^\ell}\,\mathrm{d}t = \frac{\mathrm{e}^{-x}}{\ell x^\ell} - \frac{\mathrm{e}^{-x}}{\ell(\ell-1)x^{\ell-1}} + \frac{1}{\ell(\ell-1)}\int_x^\infty \frac{\mathrm{e}^{-t}}{t^{\ell-1}}\,\mathrm{d}t = \cdots \\
&= \frac{\mathrm{e}^{-x}}{\ell!}\sum_{m=1}^{\ell}(-1)^m\frac{(m-1)!}{x^m} + \frac{1}{\ell!}\int_x^\infty \frac{\mathrm{e}^{-t}}{t}\,\mathrm{d}t \\
&= \frac{\mathrm{e}^{-x}}{\ell!}\sum_{m=1}^{\ell}(-1)^m\frac{(m-1)!}{x^m} + \frac{\mathrm{Ei}(x)}{\ell!},
\end{aligned}$$

where

$$Ei(x) \triangleq \int_x^\infty \frac{\mathrm{e}^{-t}}{t}\,\mathrm{d}t, \tag{66}$$

is the exponential integral.

In a similar way, we can show that for $\ell = 2k+1$ and odd positive integer,

$$\int_x^\infty \frac{\mathrm{e}^{-t}}{t^{\ell+1}}\,\mathrm{d}t = \frac{\mathrm{e}^{-x}}{\ell!}\sum_{m=1}^{\ell}(-1)^{m+1}\frac{(m-1)!}{x^m} - \frac{\mathrm{Ei}(x)}{\ell!}. \tag{67}$$

References

Aigner, D.J., Judge, G.G., 1977. Application of pre-test and Stein estimators to economic data. Econometrica 45 (5), 1279–1288.

Alam, K., 1973. A family of admissible minimax estimators of the mean of a multivariate normal distribution. Ann. Statist. 1, 517–525.

Bancroft, T.A., 1944. On biases in estimation due to the use of preliminary tests of significance. The Annals of Mathematical Statistics 15 (2), 190–204.

Baranchik, A.J., 1964. Multiple regression and estimation of the mean of a multivariate normal distribution. Technical Report 51, Stanford University.

Ben-Haim, Z., Eldar, Y.C., 2005. Blind minimax estimators: improving on least squares estimation. in: IEEE Workshop on Statistical Signal Processing (SSP'05), Bordeaux, France, July 2005.

Ben-Haim, Z., Eldar, Y.C., 2007. Blind minimax estimation. IEEE Trans. Inform. Theory 53, 3145–3157.

Berger, J.O., 1976. Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss. Ann. Statist. 4 (1), 223–226.

Bock, M.E., 1975. Minimax estimators of the mean of a multivariate normal distribution. Ann. Statist. 3 (1), 209–218.

Bock, M.E., Yancey, T.A., Judge, G.G., 1973. The statistical consequences of preliminary test estimators in regression. J. Amer. Statist. Assoc. 68, 109–116.

Efron, B., 1975. Biased versus unbiased estimation. Adv. in Math. 16, 259–277.
Eldar, Y.C., 2006. Universal weighted MSE improvement of the least-squares estimator. IEEE Trans. Signal Process., to appear.
Eldar, Y.C., Oppenheim, A.V., 2003. Covariance shaping least-squares estimation. IEEE Trans. Signal Process. 51, 686–697.
Eldar, Y.C., Ben-Tal, A., Nemirovski, A., 2004. Linear minimax regret estimation of deterministic parameters with bounded data uncertainties. IEEE Trans. Signal Process. 52, 2177–2188.
Eldar, Y.C., Ben-Tal, A., Nemirovski, A., 2005. Robust mean-squared error estimation in the presence of model uncertainties. IEEE Trans. Signal Process. 53, 168–181.
Giles, A., Giles, D.E.A., 1993. Pre-test estimation and testing in econometrics: recent developments. J. Econ. Surveys 7, 145–197.
Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67.
James, W., Stein, C., 1961. Estimation of quadratic loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, pp. 361–379.
Judge, G.G., Bock, M.E., 1978. The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics. North-Holland, Amsterdam.
Judge, G.G., Bock, M.E., 1983. Biased estimation. In: Griliches, Z., Intriligator, M.D. (Eds.), Handbook of Econometrics, vol. 1. North-Holland, Amsterdam, pp. 599–649 Chapter 10.
Kibria, B.M.G., Saleh, A.K.Md.E., 2006. Optimum critical value for pre-test estimators. Comm. Statist.—Simulation Comput. 35 (2), 309–319.
Lehmann, E.L., Casella, G., 1999. Theory of Point Estimation. second ed. Springer, New York.
Marquardt, D.W., 1970. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. Technometrics 12 (3), 592–612.
Mayer, L.S., Willke, T.A., 1973. On biased estimation in linear models. Technometrics 15, 497–508.
Mickens, R.E., 1990. Difference Equations: Theory and Applications. Chapman & Hall, New York, NY.
Scolve, S.L., Morris, C., Radhakrishnan, R., 1972. Non-optimality of preliminary-test estimators for the mean of a multivariate normal distribution. Ann. Math. Statist. 43, 1481–1490.
Stein, C., 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, pp. 197–206.
Strawderman, W.E., 1971. Proper Bayes minimax estimators of multivariate normal mean. Ann. Math. Statist. 42, 385–388.
Tikhonov, A.N., Arsenin, V.Y., 1977. Solution of Ill-Posed Problems. V.H. Winston, Washington, DC.
Wan, A.T.K., Zou, G.H., 2003. Optimal critical values of pre-tests when estimating the regression variance: analytical findings under a general loss structure. J. Econometrics 114 (1), 165–196.