The Gaussian MLE versus the Optimally Weighted LSE

Mohamed R.-H. Abdalmoaty Håkan Hjalmarsson Bo Wahlberg

Postprint (August 21st, 2020)

In this lecture note, we derive and compare the asymptotic covariance matrices of two parametric estimators: the Gaussian Maximum Likelihood Estimator (MLE), and the optimally weighted Least-Squares Estimator (LSE). We assume a general model parameterization where the model's mean and variance are jointly parameterized, and consider Gaussian and non-Gaussian data distributions.

1 Relevance

In system identification and estimation theory, asymptotic covariance matrices are usually used to compare the accuracy of consistent and asymptotically normal parametric estimators for sufficiently large data records. If the data distribution is Gaussian and its mean and variance are independently parameterized, a well-known result is that the asymptotic covariance matrices of the Gaussian MLE and the optimally weighted LSE are equal and coincide with the asymptotic Cramér-Rao lower bound (CRLB). In the non-Gaussian case however, as we show in this note, the accuracy of these two estimators may differ. They depend on the parameterization and the shape of the data distribution in terms of the first four moments. The results are particularly useful when estimating parameters in general semiparametric models.

2 Prerequisites

This lecture note can be used in courses on system identification, statistical signal processing, or estimation theory. The necessary background that has been assumed is similar to the intersection of the prerequisites of those courses. In particular, an exposure to basic probability, stochastic process and linear algebra is required.

3 Problem Statement and Solution

The problem is to analyze and compare the asymptotic covariance matrices of the Gaussian MLE and the optimally weighted LSE for general semiparametric models.

The model

Suppose that the model is given by

$$y_t = \mu_t(\boldsymbol{\theta}) + e_t(\boldsymbol{\theta}), \quad t = 1, 2, \ldots, N,$$

where $\{y_t\} \subset \mathbb{R}$ is a sequence of observed data, N denotes the number of available data samples, $\boldsymbol{\theta} \in \mathbb{R}^d$, with a finite positive integer d, is the unknown parameter vector to be estimated, $\{\mu_t(\boldsymbol{\theta})\}$ is a sequence of known real-valued functions of $\boldsymbol{\theta}$, and $\{e_t(\boldsymbol{\theta})\}$ is an unobserved sequence of zero mean independent real-valued random variables with known *parameter-dependent variances* denoted as $\lambda_t(\boldsymbol{\theta})$; i.e., for all $\boldsymbol{\theta}$ and t it holds that $\mathbb{E}[e_t(\boldsymbol{\theta})] = 0$, and $\mathbb{E}[e_t^2(\boldsymbol{\theta})] = \lambda_t(\boldsymbol{\theta})$. Notice that the model does not specify the full distribution of the data. Therefore, the model is semiparametric where the parameter vector $\boldsymbol{\theta}$ *jointly* parameterizes the mean and the variance of the data. Let us denote the true parameter as $\boldsymbol{\theta}_\circ$.

Two estimators

We now consider two parameter estimation methods, given as special cases of the general framework described in [1, Chapter 7]. The Gaussian MLE, denoted as $\hat{\boldsymbol{\theta}}_1$, is defined as

$$\hat{\boldsymbol{\theta}}_1 = \arg\min_{\boldsymbol{\theta}} \; V_1(\boldsymbol{\theta}), \tag{1}$$

where

$$V_1(\boldsymbol{\theta}) = \frac{1}{N} \sum_{t=1}^{N} \frac{(y_t - \mu_t(\boldsymbol{\theta}))^2}{2\lambda_t(\boldsymbol{\theta})} + \frac{1}{2} \log \lambda_t(\boldsymbol{\theta}). \tag{2}$$

The optimally weighted LSE, denoted as $\hat{\boldsymbol{\theta}}_2$, is defined as

$$\hat{\boldsymbol{\theta}}_2 = \arg\min_{\boldsymbol{\theta}} \; V_2(\boldsymbol{\theta}), \tag{3}$$

where

$$V_2(\boldsymbol{\theta}) = \frac{1}{N} \sum_{t=1}^{N} \frac{(y_t - \mu_t(\boldsymbol{\theta}))^2}{2\lambda_t(\boldsymbol{\theta}_\circ)}. \tag{4}$$

These two estimators are instances of the general family of prediction error method estimators (see [1, Section 7.2]), defined as minimizers of criterion functions

$$V(\boldsymbol{\theta}) = \frac{1}{N} \sum_{t=1}^{N} \ell(\varepsilon_t(\boldsymbol{\theta}), \boldsymbol{\theta}, t)$$

where $\varepsilon_t(\boldsymbol{\theta}) := y_t - \mu_t(\boldsymbol{\theta})$ is the prediction error, and ℓ is a general scalar-valued function. In the Gaussian MLE case,

$$\ell(\varepsilon, \boldsymbol{\theta}, t) = \ell_1(\varepsilon, \boldsymbol{\theta}, t) = \frac{\varepsilon^2}{2\lambda_t(\boldsymbol{\theta})} + \frac{1}{2} \log \lambda_t(\boldsymbol{\theta})$$

which is both time- and parameter-dependent. In the optimally weighted LSE case,

$$\ell(\varepsilon, \boldsymbol{\theta}, t) = \ell_2(\varepsilon, t) = \frac{\varepsilon^2}{2\lambda_t(\boldsymbol{\theta}_\circ)}$$

which is independent of the parameter; however, it depends on the true value $\boldsymbol{\theta}_\circ$. In practice, the unknown $\boldsymbol{\theta}_\circ$ in the definition of ℓ_2 can be replaced by a consistent estimator of $\boldsymbol{\theta}$, without affecting the asymptotic covariance of the estimator. For instance, an unweighted LSE, defined using $\ell(\varepsilon, \boldsymbol{\theta}, t) = \frac{1}{2}\varepsilon^2$, may be used; an alternative is the Gaussian MLE defined above. Although different substitutions lead to estimators with different finite sample properties, their asymptotic covariance matrices coincide with that of the optimally weighted LSE (see for example [2]).

Asymptotic Covariance

When the scalar-valued function ℓ is both time- and parameter-independent, i.e., when $\ell(\varepsilon, \boldsymbol{\theta}, t) = \ell(\varepsilon)$, its form only acts as a scaling of the asymptotic covariance matrix, as explained in [1, page 286], and in [3] when ℓ corresponds to a probability density function.

In some cases, this also holds when ℓ is time-dependent (see problem 9T.1 in [1]). However, if ℓ is parameter-dependent, this property does not hold.

Notation: In what follows, we will use a prime symbol to indicate differentiation with respect to the parameter vector $\boldsymbol{\theta}$. For any real-valued function $V(\boldsymbol{\theta})$, the symbol $V'(\boldsymbol{\theta})$ denotes the gradient, defined as a d-dimensional column vector. The symbol $V''(\boldsymbol{\theta})$ denotes the derivative of the gradient vector $V'(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, which is a $d \times d$ matrix.

Suppose that the minimizers in (1) and (3) are sought over a closed and bounded subset $\Theta \subset \mathbb{R}^d$ such that $\boldsymbol{\theta}_\circ \in \Theta$. Furthermore, for $i = 1, 2$, assume that $\mathbb{E}[V_i(\boldsymbol{\theta})]$ converges as $N \to \infty$, uniformly over Θ, to a deterministic matrix $\overline{V}_i(\boldsymbol{\theta})$ such that $\sqrt{N}\,\overline{V}_i'(\boldsymbol{\theta}_\circ) \to 0$ as $N \to \infty$. Then, under some mild regularity conditions on the model (see [1, Chapter 9] or [4, Chapter 7]), it holds that $\sqrt{N}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_\circ)$ is asymptotically normally distributed with mean zero and covariance matrix

$$P_i = \left[\overline{V}_i''(\boldsymbol{\theta}_\circ)\right]^{-1} \left[\lim_{N\to\infty} N\mathbb{E}\left[V_i'(\boldsymbol{\theta}_\circ)[V_i'(\boldsymbol{\theta}_\circ)]^\top\right]\right] \left[\overline{V}_i''(\boldsymbol{\theta}_\circ)\right]^{-1},$$

where it is assumed that the limits and the matrix inverse exist. In the following, we will assume, under the same regularity conditions from above, that the interchange of limits and expectation is possible.

The Gaussian case

Although the computations in the case of Gaussian data distributions are known and may be found in classical textbooks (see for example [5, Appendix 3C]), we include them here to highlight the role of the third- and fourth-order moments of the data distribution when the mean and variance are jointly parameterized. We will also refer back to these computations when considering the non-Gaussian case.

Suppose that the true data distribution is Gaussian, and let us first consider the computations of P_1. From (2), using the chain rule, it holds that

$$NV_1'(\boldsymbol{\theta}) = \sum_{t=1}^{N} -\frac{(y_t - \mu_t(\boldsymbol{\theta}))}{\lambda_t(\boldsymbol{\theta})}\mu_t'(\boldsymbol{\theta}) - \frac{(y_t - \mu_t(\boldsymbol{\theta}))^2}{2\lambda_t^2(\boldsymbol{\theta})}\lambda_t'(\boldsymbol{\theta}) + \frac{1}{2\lambda_t(\boldsymbol{\theta})}\lambda_t'(\boldsymbol{\theta}).$$

Differentiating one more time, we get

$$
\begin{aligned}
NV_1''(\boldsymbol{\theta}) = \sum_{t=1}^{N} & -\frac{(y_t-\mu_t(\boldsymbol{\theta}))}{\lambda_t(\boldsymbol{\theta})}\mu_t''(\boldsymbol{\theta}) + \frac{1}{\lambda_t(\boldsymbol{\theta})}\mu_t'(\boldsymbol{\theta})[\mu_t'(\boldsymbol{\theta})]^\top + \frac{(y_t-\mu_t(\boldsymbol{\theta}))}{\lambda_t^2(\boldsymbol{\theta})}\mu_t'(\boldsymbol{\theta})[\lambda_t'(\boldsymbol{\theta})]^\top \\
& - \frac{(y_t-\mu_t(\boldsymbol{\theta}))^2}{2\lambda_t^2(\boldsymbol{\theta})}\lambda_t''(\boldsymbol{\theta}) + \frac{(y_t-\mu_t(\boldsymbol{\theta}))}{\lambda_t^2(\boldsymbol{\theta})}\lambda_t'(\boldsymbol{\theta})[\mu_t'(\boldsymbol{\theta})]^\top + \frac{(y_t-\mu_t(\boldsymbol{\theta}))}{\lambda_t^3(\boldsymbol{\theta})}\lambda_t'(\boldsymbol{\theta})[\lambda_t'(\boldsymbol{\theta})]^\top \\
& + \frac{1}{2\lambda_t(\boldsymbol{\theta})}\lambda_t''(\boldsymbol{\theta}) - \frac{1}{2\lambda_t^2(\boldsymbol{\theta})}\lambda_t'(\boldsymbol{\theta})[\lambda_t'(\boldsymbol{\theta})]^\top.
\end{aligned}
$$

Then, it holds that

$$
\mathbb{E}[V_1''(\boldsymbol{\theta}_\circ)] = \frac{1}{N}\sum_{t=1}^{N}\frac{1}{\lambda_t(\boldsymbol{\theta}_\circ)}\mu_t'(\boldsymbol{\theta}_\circ)[\mu_t'(\boldsymbol{\theta}_\circ)]^\top + \frac{1}{2\lambda_t^2(\boldsymbol{\theta}_\circ)}\lambda_t'(\boldsymbol{\theta}_\circ)[\lambda_t'(\boldsymbol{\theta}_\circ)]^\top. \tag{5}
$$

Moreover,

$$
\begin{aligned}
N^2V_1'(\boldsymbol{\theta})[V_1'(\boldsymbol{\theta})]^\top = \sum_{t=1}^{N}\sum_{s=1}^{N} & \frac{(y_t-\mu_t(\boldsymbol{\theta}))(y_s-\mu_s(\boldsymbol{\theta}))}{\lambda_t(\boldsymbol{\theta})\lambda_s(\boldsymbol{\theta})}\mu_t'(\boldsymbol{\theta})[\mu_s'(\boldsymbol{\theta})]^\top - \frac{(y_t-\mu_t(\boldsymbol{\theta}))}{\lambda_t(\boldsymbol{\theta})\lambda_s(\boldsymbol{\theta})}\mu_t'(\boldsymbol{\theta})[\lambda_s'(\boldsymbol{\theta})]^\top \\
& + \frac{1}{4\lambda_t(\boldsymbol{\theta})\lambda_s(\boldsymbol{\theta})}\lambda_t'(\boldsymbol{\theta})[\lambda_s'(\boldsymbol{\theta})]^\top + \frac{(y_t-\mu_t(\boldsymbol{\theta}))(y_s-\mu_s(\boldsymbol{\theta}))^2}{\lambda_t(\boldsymbol{\theta})\lambda_s^2(\boldsymbol{\theta})}\mu_t'(\boldsymbol{\theta})[\lambda_s'(\boldsymbol{\theta})]^\top \\
& + \frac{(y_t-\mu_t(\boldsymbol{\theta}))^2(y_s-\mu_s(\boldsymbol{\theta}))^2}{4\lambda_t^2(\boldsymbol{\theta})\lambda_s^2(\boldsymbol{\theta})}\lambda_t'(\boldsymbol{\theta})[\lambda_s'(\boldsymbol{\theta})]^\top - \frac{(y_t-\mu_t(\boldsymbol{\theta}))^2}{2\lambda_t^2(\boldsymbol{\theta})\lambda_s(\boldsymbol{\theta})}\lambda_t'(\boldsymbol{\theta})[\lambda_s'(\boldsymbol{\theta})]^\top.
\end{aligned}
$$

Taking the expectation on both sides and using the independence assumption of the model, we get that

$$
\begin{aligned}
N^2\mathbb{E}[V_1'(\boldsymbol{\theta}_\circ)[V_1'(\boldsymbol{\theta}_\circ)]^\top] = & \sum_{t=1}^{N}\frac{1}{\lambda_t(\boldsymbol{\theta}_\circ)}\mu_t'(\boldsymbol{\theta}_\circ)[\mu_t'(\boldsymbol{\theta}_\circ)]^\top + \sum_{t=1}^{N}\sum_{s=1}^{N}\frac{1}{4\lambda_t(\boldsymbol{\theta}_\circ)\lambda_s(\boldsymbol{\theta}_\circ)}\lambda_t'(\boldsymbol{\theta}_\circ)[\lambda_s'(\boldsymbol{\theta}_\circ)]^\top \\
& + \sum_{t=1}^{N}\frac{\alpha_t(\boldsymbol{\theta}_\circ)}{\lambda_t^3(\boldsymbol{\theta}_\circ)}\mu_t'(\boldsymbol{\theta}_\circ)[\lambda_t'(\boldsymbol{\theta}_\circ)]^\top \\
& + \sum_{t=1}^{N}\frac{\beta_t(\boldsymbol{\theta}_\circ)}{4\lambda_t^4(\boldsymbol{\theta}_\circ)}\lambda_t'(\boldsymbol{\theta}_\circ)[\lambda_t'(\boldsymbol{\theta}_\circ)]^\top + \sum_{t=1}^{N}\sum_{\substack{s=1\\s\neq t}}^{N}\frac{1}{4\lambda_t(\boldsymbol{\theta}_\circ)\lambda_s(\boldsymbol{\theta}_\circ)}\lambda_t'(\boldsymbol{\theta}_\circ)[\lambda_s'(\boldsymbol{\theta}_\circ)]^\top \\
& - \sum_{t=1}^{N}\sum_{s=1}^{N}\frac{1}{2\lambda_t(\boldsymbol{\theta}_\circ)\lambda_s(\boldsymbol{\theta}_\circ)}\lambda_t'(\boldsymbol{\theta}_\circ)[\lambda_s'(\boldsymbol{\theta}_\circ)]^\top
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha_t(\boldsymbol{\theta}_\circ) &= \mathbb{E}[(y_t-\mu_t(\boldsymbol{\theta}_\circ))^3], \\
\beta_t(\boldsymbol{\theta}_\circ) &= \mathbb{E}[(y_t-\mu_t(\boldsymbol{\theta}_\circ))^4]
\end{aligned}
$$

are the third- and fourth-order moments of the model when $\boldsymbol{\theta} = \boldsymbol{\theta}_\circ$, receptively. Further-

more, the sum of the three double sums evaluates to $-\frac{1}{4\lambda_t^2(\boldsymbol{\theta}_\circ)}\lambda_t'(\boldsymbol{\theta}_\circ)[\lambda_t'(\boldsymbol{\theta}_\circ)]^\top$ and therefore

$$\begin{aligned} N^2\mathbb{E}[V_1'(\boldsymbol{\theta}_\circ)[V_1'(\boldsymbol{\theta}_\circ)]^\top] = &\sum_{t=1}^{N}\frac{1}{\lambda_t(\boldsymbol{\theta}_\circ)}\mu_t'(\boldsymbol{\theta}_\circ)[\mu_t'(\boldsymbol{\theta}_\circ)]^\top - \frac{1}{4\lambda_t^2(\boldsymbol{\theta}_\circ)}\lambda_t'(\boldsymbol{\theta}_\circ)[\lambda_t'(\boldsymbol{\theta}_\circ)]^\top \\ &+\sum_{t=1}^{N}\frac{\alpha_t(\boldsymbol{\theta}_\circ)}{\lambda_t^3(\boldsymbol{\theta}_\circ)}\mu_t'(\boldsymbol{\theta}_\circ)[\lambda_t'(\boldsymbol{\theta}_\circ)]^\top + \sum_{t=1}^{N}\frac{\beta_t(\boldsymbol{\theta}_\circ)}{4\lambda_t^4(\boldsymbol{\theta}_\circ)}\lambda_t'(\boldsymbol{\theta}_\circ)[\lambda_t'(\boldsymbol{\theta}_\circ)]^\top. \end{aligned} \tag{6}$$

We now use the assumption that the data distribution is Gaussian. Under this assumption, the third- and fourth-order moments are $\alpha_t(\boldsymbol{\theta}_\circ) = 0$ and $\beta_t(\boldsymbol{\theta}_\circ) = 3\lambda_t^2(\boldsymbol{\theta}_\circ)$, respectively. Then, it is straightforward to see that in such a case the expression in (6) reduces to

$$N\mathbb{E}[V_1'(\boldsymbol{\theta}_\circ)[V_1'(\boldsymbol{\theta}_\circ)]^\top] = \frac{1}{N}\sum_{t=1}^{N}\frac{1}{\lambda_t(\boldsymbol{\theta}_\circ)}\mu_t'(\boldsymbol{\theta}_\circ)[\mu_t'(\boldsymbol{\theta}_\circ)]^\top + \frac{1}{2\lambda_t^2(\boldsymbol{\theta}_\circ)}\lambda_t'(\boldsymbol{\theta}_\circ)[\lambda_t'(\boldsymbol{\theta}_\circ)]^\top, \tag{7}$$

which is equal to $\mathbb{E}[V_1''(\boldsymbol{\theta}_\circ)]$. We conclude that

$$P_1 = \left[\overline{V}_1''(\boldsymbol{\theta}_\circ)\right]^{-1} = \left[\lim_{N\to\infty}\left(\frac{1}{N}\sum_{t=1}^{N}\frac{1}{\lambda_t(\boldsymbol{\theta}_\circ)}\mu_t'(\boldsymbol{\theta}_\circ)[\mu_t'(\boldsymbol{\theta}_\circ)]^\top + \frac{1}{2\lambda_t^2(\boldsymbol{\theta}_\circ)}\lambda_t'(\boldsymbol{\theta}_\circ)[\lambda_t'(\boldsymbol{\theta}_\circ)]^\top\right)\right]^{-1} \tag{8}$$

which is equal to the asymptotic Gaussian CRLB. Notice that here $\overline{V}_1''(\boldsymbol{\theta}_\circ)$ is the per sample Fisher information matrix, and that it is given as the sum of two terms: the first corresponds to the information from the mean, while the second is due to that from the variance.

Next, we compute P_2. From (4), using the chain rule, it holds that

$$NV_2'(\boldsymbol{\theta}) = \sum_{t=1}^{N} -\frac{(y_t - \mu_t(\boldsymbol{\theta}))}{\lambda_t(\boldsymbol{\theta}_\circ)}\mu_t'(\boldsymbol{\theta}),$$

and

$$NV_2''(\boldsymbol{\theta}) = \sum_{t=1}^{N}\frac{1}{\lambda_t(\boldsymbol{\theta}_\circ)}\mu_t'(\boldsymbol{\theta})[\mu_t'(\boldsymbol{\theta})]^\top - \frac{(y_t - \mu_t(\boldsymbol{\theta}))}{\lambda_t(\boldsymbol{\theta}_\circ)}\mu_t''(\boldsymbol{\theta}). \tag{9}$$

Moreover,

$$N^2V_2'(\boldsymbol{\theta})[V_2'(\boldsymbol{\theta})]^\top = \sum_{t=1}^{N}\sum_{s=1}^{N}\frac{(y_t - \mu_t(\boldsymbol{\theta}))(y_s - \mu_s(\boldsymbol{\theta}))}{\lambda_t(\boldsymbol{\theta}_\circ)\lambda_s(\boldsymbol{\theta}_\circ)}\mu_t'(\boldsymbol{\theta})[\mu_s'(\boldsymbol{\theta})]^\top.$$

Taking the expectation on both sides and using the independence assumption of the model,

$$N\mathbb{E}[V_2'(\boldsymbol{\theta}_\circ)[V_2'(\boldsymbol{\theta}_\circ)]^\top] = \frac{1}{N}\sum_{t=1}^{N}\frac{1}{\lambda_t(\boldsymbol{\theta}_\circ)}\mu_t'(\boldsymbol{\theta}_\circ)[\mu_t'(\boldsymbol{\theta}_\circ)]^\top.$$

Now, using (9), it holds that

$$\mathbb{E}[V_2''(\boldsymbol{\theta}_\circ)] = \frac{1}{N}\sum_{t=1}^{N}\frac{1}{\lambda_t(\boldsymbol{\theta}_\circ)}\mu_t'(\boldsymbol{\theta}_\circ)[\mu_t'(\boldsymbol{\theta}_\circ)]^\top = N\mathbb{E}[V_2'(\boldsymbol{\theta}_\circ)[V_2'(\boldsymbol{\theta}_\circ)]^\top],$$

and hence we conclude that

$$P_2 = \left[\overline{V}_2''(\boldsymbol{\theta}_\circ)\right]^{-1} = \left[\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\frac{1}{\lambda_t(\boldsymbol{\theta}_\circ)}\mu_t'(\boldsymbol{\theta}_\circ)[\mu_t'(\boldsymbol{\theta}_\circ)]^\top\right]^{-1}. \tag{10}$$

By comparing (8) and (10), and noting that the term $\frac{1}{2\lambda_t^2(\boldsymbol{\theta}_\circ)}\lambda_t'(\boldsymbol{\theta}_\circ)[\lambda_t'(\boldsymbol{\theta}_\circ)]^\top$ is positive, we see that $P_2 \succeq P_1$ for Gaussian data distributions; in other words $P_2 - P_1$ is a positive semidefinite matrix. This result is due to the joint parameterization of the mean and variance.

> **Conclusion 1:** For Gaussian data distributions, where the mean and variance are jointly parameterized, the Gaussian MLE achieves the CRLB and is therefore asymptotically efficient. The optimally weighted LSE, on the other hand, may not be asymptotically efficient.

The non-Gaussian case

Now suppose that the true data distribution is non-Gaussian and let us first consider the computations of P_1. Referring back to the computations leading to P_1 in the Gaussian case, we see that the expression in (5) is valid in the non-Gaussian case as well. This is because only the independence assumption of the model was required in the computations, and not the form of the data distribution. On the other hand, the expression in (7) is valid only for Gaussian data distributions, due to the specific substitutions used for the third- and fourth-order moments. Therefore, in the non-Gaussian case (7) is not valid. Instead, (6) has to be used.

Consequently, in the non-Gaussian case $\mathbb{E}[V_1''(\boldsymbol{\theta}_\circ)] \neq N\mathbb{E}[V_1'(\boldsymbol{\theta}_\circ)[V_1'(\boldsymbol{\theta}_\circ)]^\top]$, and the asymptotic covariance matrix of the Gaussian MLE is given by

$$P_1 = \left[\lim_{N\to\infty}\mathbb{E}[V_1''(\boldsymbol{\theta}_\circ)]\right]^{-1}\left[\lim_{N\to\infty}N\mathbb{E}\left[V_1'(\boldsymbol{\theta}_\circ)[V_1'(\boldsymbol{\theta}_\circ)]^\top\right]\right]\left[\lim_{N\to\infty}\mathbb{E}[V_1''(\boldsymbol{\theta}_\circ)]\right]^{-1} \tag{11}$$

where $\mathbb{E}[V_1''(\boldsymbol{\theta}_\circ)]$ is as in (5) and $N\mathbb{E}[V_1'(\boldsymbol{\theta}_\circ)[V_1'(\boldsymbol{\theta}_\circ)]^\top]$ is as in (6), where the third- and fourth-order moments appear.

Next, we compute P_2. Referring back to the computations leading to P_2 in the Gaussian case, we see that the expression in (10) is valid in the non-Gaussian case as well. This is because higher order moments were not used when evaluating the expression, and no assumption was made on the shape of the data distribution.

The scalar parameter case

A direct comparison between P_1 and P_2 in the non-Gaussian case is generally not possible. In order to get some insight regarding the effect of the third- and fourth-order moments on P_1 in the non-Gaussian case, we will assume in this part that $\theta \in \mathbb{R}$, and that the model is fourth-order stationary; i.e., the first four moments do not depend on t (and therefore we may drop the subscript t from the notations).

Notice that using (11) and (10) it holds that

$$P_1 = \frac{A+C}{(A+B)^2} \qquad \text{and} \qquad P_2 = \frac{1}{A} \tag{12}$$

where

$$A = \frac{(\mu'(\theta_\circ))^2}{\lambda(\theta_\circ)}, \quad B = \frac{1}{2}\left(\frac{\lambda'(\theta_\circ)}{\lambda(\theta_\circ)}\right)^2, \quad C = D + E - \frac{1}{2}B,$$

and

$$D = \frac{\beta(\theta_\circ)(\lambda'(\theta_\circ))^2}{4\lambda^4(\theta_\circ)}, \quad E = \frac{\alpha(\theta_\circ)\mu'(\theta_\circ)\lambda'(\theta_\circ)}{\lambda^3(\theta_\circ)}.$$

Then, $P_1 \leq P_2$ if and only if

$$A \leq \frac{(A+B)^2}{A+C} \iff C \leq \frac{B^2}{A} + 2B \iff D + E \leq \frac{B^2}{A} + \frac{5}{2}B, \tag{13}$$

or equivalently ([6])

$$\kappa(\theta_\circ) \leq \left[\frac{(\lambda'(\theta_\circ))^2}{\lambda(\theta_\circ)\,(\mu'(\theta_\circ))^2} - \frac{4}{\lambda(\theta_\circ)}\frac{\mu'(\theta_\circ)}{\lambda'(\theta_\circ)}\alpha(\theta_\circ) + 5\right] \tag{14}$$

where $\kappa(\theta_\circ) = \frac{\beta(\theta_\circ)}{\lambda^2(\theta_\circ)}$ is the kurtosis. Otherwise, $P_2 < P_1$ and the optimal LSE will have a smaller asymptotic variance. For symmetric data distributions, the third-order moment $\alpha(\theta_\circ) = 0$ and the condition (14) giving $P_1 \leq P_2$ reduces to

$$\kappa(\theta_\circ) \leq \left[\frac{(\lambda'(\theta_\circ))^2}{\lambda(\theta_\circ)\,(\mu'(\theta_\circ))^2} + 5\right].$$

This shows that in the non-Gaussian case, the Gaussian MLE is not always better than the optimally weighted LSE; this will depend on the model parameterization and the shape of the data distribution in terms of symmetry and kurtosis.

> **Conclusion 2:** For non-Gaussian data distributions where the mean and variance are jointly parameterized, the Gaussian MLE is not necessarily better than the optimally weighted LSE. In order to decide which estimator is better, the knowledge of the third- and fourth-order moments (as functions of $\boldsymbol{\theta}$) is required.

4 Illustrative Example

The following example, taken from [6, Sections 10.2 and 10.3], is used to illustrate the results. It is specifically chosen as it provides a case where the optimally weighted LSE coincides with the (correctly specified) asymptotically efficient MLE, while the Gaussian MLE is inefficient.

Suppose that the model is given by

$$y_t = \theta u_t^2 + \sqrt{2}\theta u_t^2 \varepsilon_t, \quad t = 1, 2, \ldots, N,$$

where $\{u_t\}$ is a known realization of independent standard Gaussian random variables, and $\{\varepsilon_t\}$ is a sequence of independent random variables with zero mean and unit variance. In this case,

$$\mu_t(\theta) = \theta u_t^2, \qquad \lambda_t(\theta) = 2\theta^2 u_t^4. \tag{15}$$

Now consider the following two data distributions

$$\begin{aligned} \text{Gaussian:} \quad & y_t \sim \mathcal{N}\left(\theta u_t^2, 2\theta^2 u_t^4\right), \\ \text{Non-Gaussian (Gamma):} \quad & y_t \sim \Gamma\left(\frac{1}{2}, 2\theta u_t^2\right), \end{aligned} \tag{16}$$

and notice that in both cases, the mean and variance of y_t are given by (15).

Suppose that the true parameter $\theta_\circ = 1$, and notice that $\mu'(\theta_\circ) = u_t^2$, and $\lambda_t'(\theta_\circ) = 4u_t^4$. Then, in the Gaussian case, by using (8) and (10), it holds that

$$P_1 = 0.4 \ \ (= \text{asymptotic Gaussian CRLB}), \qquad P_2 = 2.$$

Now, notice that the third- and fourth-order moments of the Gamma distribution in (16) when $\theta = \theta_\circ$ are $8u_t^6$ and $60u_t^8$, respectively. Then, in the Gamma case, by using (11) and (10), it holds that

$$P_1 = 2.96, \qquad P_2 = 2 \ \ (= \text{asymptotic Gamma CRLB}),$$

where the variance P_2 coincides with the asymptotic Gamma CRLB. To see this, recall that, by definition, the one sample log-likelihood function of the Gamma model in (16) is

$$\log p(y_t;\theta) = -\log\left(\Gamma\left(\frac{1}{2}\right)\right) - \frac{1}{2}\log(2\theta u_t^2) - \frac{1}{2}\log(y_t) - \frac{y_t}{2\theta u_t^2},$$

and its second derivative with respect to θ is

$$(\log p(y_t;\theta))'' = \frac{1}{2\theta^2} - \frac{y_t}{\theta^3 u_t^2}.$$

Consequently, the per sample Fisher information is $\mathbb{E}\left[-(\log p(y_t;\theta))''|_{\theta=\theta_\circ}\right] = 0.5$, and the asymptotic Gamma CRLB is 2.

Therefore, we have a case where the optimally weighted LSE coincides with the MLE and is (asymptotically) more accurate than the Gaussian MLE. Furthermore, by comparing the CRLB in both cases, we notice that the bound associated with the Gaussian assumption is smaller than that of the Gamma assumption. This is due to the joint parameterization of the mean and variance.

> **Conclusion 3:** The min-max optimality property of the Gaussian distribution (see [3]) may not hold when the mean and variance are jointly parameterized.

5 What we have learned

The results provided in the previous sections show that for non-Gaussian data distributions, with jointly parameterized mean and variance, the Gaussian MLE is not necessarily better than the optimally weighted LSE. We derived the expressions of the asymptotic covariance matrices, and established a condition, when $\theta \in \mathbb{R}$, under which one of the estimators may be preferred. Finally, using an example, we saw that when the mean and variance are jointly parameterized, the min-max property of the Gaussian distribution may not hold.

6 Acknowledgment

This research was supported by the Swedish Research Council via the projects 2016-06079 (NewLEADS), 2015-05285, and 2019-04956.

7 Authors

Mohamed R.-H. Abdalmoaty (abda@kth.se), *Håkan Hjalmarsson* (hjalmars@kth.se), and *Bo Wahlberg* (bo@kth.se) are with the Division of Decision and Control Systems, KTH Royal Institute of Technology, Stockholm, Sweden.

References

[1] L. Ljung, *System Identification: Theory for the User*, 2nd ed. Prentice Hall, 1999.

[2] L. Pronzato and A. Pázman, "Recursively re-weighted least-squares estimation in regression models with parameterized variance," in *2004 12th European Signal Processing Conference*. IEEE, 2004, pp. 621–624.

[3] P. Stoica and P. Babu, "The Gaussian Data Assumption Leads to the Largest Cramér-Rao Bound [Lecture Notes]," *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 132–133, May 2011.

[4] T. Söderström and P. Stoica, *System Identification*. Prentice Hall, 1989.

[5] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation theory*. PTR Prentice-Hall, 1993.

[6] M. Abdalmoaty and H. Hjalmarsson, "Identification of stochastic nonlinear models using optimal estimating functions," *Automatica*, vol. 119, p. 109055, 2020.