Petre Stoica and
Prabhu Babu

The Gaussian Data Assumption Leads to the Largest Cramér-Rao Bound

The Gaussian data assumption is sometimes criticized as being unrealistic in certain applications. While this is a valid criticism, it is also true that the Gaussian assumption is a natural choice when nothing is known about the exact data distribution. The argument typically used to motivate this choice relies on the fact that the Gaussian distribution leads to the largest Cramér-Rao bound (CRB) in quite a general class of data distributions and for a significant set of parameter estimation problems (see [1]–[3], and the references therein). Consequently, the Gaussian CRB (i.e., the CRB that holds under the Gaussian assumption) is the worst-case one (over the distribution class), and therefore any optimal design based on attaining or minimizing it, including the parameter estimation operation itself, can be interpreted as being min-max optimal.

In this lecture note, we provide a simple and yet quite general proof of the aforesaid fact that the Gaussian assumption yields the largest CRB. This fact, which is sometimes considered to be a "folk theorem," is possibly known to many (see the cited works); however finding a proof of it in the literature, of comparable generality to that presented here, has eluded us.

PREREQUISITES

This lecture note can be used in courses on spectral estimation, on statistical signal processing, as well as on system identification and parameter estimation. The prerequisites for following and understanding this note are the same as for the said courses.

Digital Object Identifier 10.1109/MSP.2011.940411
Date of publication: 19 April 2011

THE MAIN RESULT

Assume that the sampled data under consideration satisfy the following equation:

$$y(t) = z(t, \boldsymbol{\theta}) + \epsilon(t) \quad t = 1, \ldots, N, \quad (1)$$

where $\{y(t)\}$ is an observed sequence, N denotes the number of available data samples, $\boldsymbol{\theta}$ is the unknown parameter vector to be estimated, $\{z(t, \boldsymbol{\theta})\}$ is a signal term that depends on $\boldsymbol{\theta}$ in a known fashion, and $\{\epsilon(t)\}$ is a white noise term (i.e., an unobserved sequence of independent and identically distributed random variables) that has mean equal to zero and variance denoted by σ^2. Furthermore, assume that $z(t, \boldsymbol{\theta})$ may depend on the past of $\epsilon(t)$ (i.e. on $\epsilon(t-1)$, $\epsilon(t-2)$, and so on) but not on its future, which implies that

$$z(s, \boldsymbol{\theta}) \text{ is independent of } \epsilon(t), \quad \text{for } s \le t. \quad (2)$$

The class of signals and systems described by (1), (2) is quite large. In particular, it encompasses all linear rational stochastic models employed in spectral estimation, time-series analysis and forecasting, and system identification such as ARMA (autoregressive moving average), ARMAX (ARMA with eXogenous inputs), and transfer function models (see e.g., [2] and [4]). As an example, consider the following transfer function description for a linear rational stochastic sampled data system:

$$y(t) = \frac{B(z^{-1})}{A(z^{-1})} u(t) + \frac{C(z^{-1})}{D(z^{-1})} \epsilon(t), \quad (3)$$

where $A(z^{-1})$, $B(z^{-1})$, $C(z^{-1})$, and $D(z^{-1})$ are polynomials in the unit-delay operator z^{-1}, $\{y(t)\}$ and $\{\epsilon(t)\}$ are as defined above, $\{u(t)\}$ is an observed input signal, $B(z^{-1})/A(z^{-1})$ is the input-to-output transfer function, and $C(z^{-1})/D(z^{-1})$ is the noise spectrum shaping transfer function. A standard assumption on (3), which we also make, is that $A(z^{-1})$, $D(z^{-1})$, and $C(z^{-1})$ have all their zeros outside the unit circle, and $\{u(t)\}$ is independent of $\{\epsilon(t)\}$, or at least the past of $\{u(t)\}$ is independent of $\epsilon(t)$ – a relaxed assumption that is useful when $u(t)$ is generated via a causal feedback from $y(t)$. Let $C(0)/D(0) = 1$ (which is no restriction), let $B(0)/A(0) = 0$ (i.e., the transfer function from $u(t)$ to $y(t)$ contains a delay), and let

$$z(t, \boldsymbol{\theta}) = \left[1 - \frac{D(z^{-1})}{C(z^{-1})}\right] y(t) + \frac{B(z^{-1})D(z^{-1})}{A(z^{-1})C(z^{-1})} u(t) \quad (4)$$

where θ is a vector made from the unknown coefficients of $A(z^{-1})$, etc. Then we can write $y(t)$ as in (1) with $z(t, \boldsymbol{\theta})$ satisfying (2). Furthermore, for any function $p(t)$ that depends on the past of $y(t)$ and $u(t)$, i.e., $\{y(t-1), u(t-1), y(t-2), u(t-2);$ and so on$\}$, we have that

$$E[p(t) - y(t)]^2 = E[p(t) - z(t, \boldsymbol{\theta})]^2 + E[\epsilon(t)]^2. \quad (5)$$

From (5) we deduce that $z(t, \boldsymbol{\theta})$ is the minimum-variance one-step predictor of $y(t)$ [sometimes denoted by $\hat{y}(t|t-1)$]. Therefore, not only is the description in (1) general enough to include the transfer function model in (3), but $z(t, \boldsymbol{\theta})$ in this description has the interesting interpretation of being the optimal one-step predictor of $y(t)$.

Let $f(\epsilon)$ denote the pdf (probability density function) of the noise term in (1), and assume that $f(\epsilon) > 0$ (for any ϵ) and also that $f(\epsilon)$ satisfies the regularity conditions required in the standard

derivation of the CRB, which essentially guarantee the interchangeability of the integration and differentiation operations on $f(\epsilon)$; see, e.g., [5]. Then, the log-likelihood function associated with (1) is given by (conditional on the necessary initial conditions)

$$g(\boldsymbol{\theta}) = \ln \prod_{t=1}^{N} f(y(t) - z(t, \boldsymbol{\theta})) = \sum_{t=1}^{N} \ln f(y(t) - z(t, \boldsymbol{\theta})). \quad (6)$$

It is well known that the CRB matrix for the parameter vector $\{\boldsymbol{\theta}, \sigma^2\}$ is block diagonal, with the block corresponding to $\boldsymbol{\theta}$ being equal to $\boldsymbol{F}^{-1}$, where $\boldsymbol{F}$ is the following Fisher information matrix (see, e.g., [4]):

$$\boldsymbol{F} = \boldsymbol{E}\left[\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T}\right]. \quad (7)$$

Observe from (6) that

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\sum_{t=1}^{N} \frac{f'(\epsilon(t))}{f(\epsilon(t))} z'(\mathrm{t}, \boldsymbol{\theta}) = -\sum_{t=1}^{N} \rho(t) z'(t), \quad (8)$$

where $f'(\epsilon) = \partial f(\epsilon)/\partial \epsilon$ $z'(t) = \partial z(t, \boldsymbol{\theta})/\partial \boldsymbol{\theta}$, and

$$\rho(t) = \frac{f'(\epsilon(t))}{f(\epsilon(t))}. \quad (9)$$

Next, use (7) and (8) along with the fact that $E[f'(\epsilon)/f(\epsilon)] = \int_{-\infty}^{\infty} f'(\epsilon) d\epsilon = \left[\int_{-\infty}^{\infty} f(\epsilon) d\epsilon\right]' = 0$ to verify the following equalities:

$$\begin{aligned} \boldsymbol{F} &= \left[E \sum_{t=1}^{N} \sum_{s=1}^{N} \rho(t)\rho(s) z'(t) z'^T(s)\right] \\ &= \sum_{t=1}^{N} E\left[\rho^2(t)\right] E\left[z'(t) z'^T(t)\right] \\ &\quad + \sum_{t=1}^{N} \sum_{s=1}^{t-1} \underbrace{E[\rho(t)]}_{0} E\left[\rho(s) z'(t) z'^T(s)\right] \\ &\quad + \sum_{t=1}^{N} \sum_{s=t+1}^{N} \underbrace{E[\rho(s)]}_{0} E\left[\rho(t) z'(t) z'^T(s)\right] \\ &= \gamma \boldsymbol{R} \end{aligned} \quad (10)$$

where

$$\gamma = E[\rho^2(t)] \quad (11)$$

does not depend on t, and

$$\boldsymbol{R} = \sum_{t=1}^{N} E[z'(t) z'^T(t)], \quad (12)$$

(we keep both the sum and the expectation in (12) to account for the possible presence of both deterministic terms and stochastic terms in $z'(t)$; see e.g., [4]).

From (10), we see that *the CRB depends on the data distribution only via the scalar* γ. Under the Gaussian distribution assumption we have

$$f(\epsilon) = \text{const e}^{-\frac{1}{2\sigma^2}\epsilon^2} \quad (13)$$

and therefore

$$\gamma = \int_{-\infty}^{\infty} \left[\frac{f'(\epsilon(t))}{f(\epsilon(t))}\right]^2 f(\epsilon) d\epsilon = \int_{-\infty}^{\infty} \frac{\epsilon^2}{\sigma^4} f(\epsilon) d\epsilon = 1/\sigma^2. \quad (14)$$

For an arbitrary distribution (satisfying the assumptions made, including the zero mean assumption) we can readily verify the following implications:

$$\begin{aligned} E(\epsilon) = \int_{-\infty}^{\infty} \epsilon f(\epsilon) d\epsilon = 0 \Rightarrow \int_{-\infty}^{\infty} f(\epsilon) d\epsilon \\ + \int_{-\infty}^{\infty} \epsilon f'(\epsilon) d\epsilon = 0 \\ \Rightarrow \int_{-\infty}^{\infty} \epsilon f'(\epsilon) d\epsilon = -1 \Rightarrow \end{aligned}$$

$$\begin{aligned} 1 &= \left[\int_{-\infty}^{\infty} \epsilon f'(\epsilon) d\epsilon\right]^2 \\ &= \left[\int_{-\infty}^{\infty} \epsilon f^{1/2}(\epsilon) \frac{f'(\epsilon)}{f^{1/2}(\epsilon)} d\epsilon\right]^2 \\ &\le \left[\int_{-\infty}^{\infty} \epsilon^2 f(\epsilon) d\epsilon\right] \\ &\quad \times \left[\int_{-\infty}^{\infty} \frac{f'^2(\epsilon)}{f(\epsilon)} d\epsilon\right] = \sigma^2 \gamma \end{aligned} \quad (15)$$

where the last line follows from the Cauchy-Schwarz inequality. Furthermore, the equality in (15) holds if and only if

$$f'(\epsilon) = \text{const } \epsilon f(\epsilon). \quad (16)$$

The only solution of this differential equation is known to be the Gaussian probability distribution function, (13). To prove this fact, rewrite (16) in the form $f'(\epsilon)/f(\epsilon) = \text{const } \epsilon$ and integrate the latter equation to obtain $\ln f(\epsilon) - \ln(\text{const}) = (\text{const}/2)/\epsilon^2$ or, equivalently, $f(\epsilon) = (\text{const}) e^{(\text{const}/2)/\epsilon^2}$, which confirms with (13).

Based on the above calculations we conclude that

$$\gamma \ge 1/\sigma^2, \quad (17)$$

where the equality holds only for the Gaussian distribution.

WHAT WE HAVE LEARNED

The direct implication of the result proved in the previous section is that the CRB matrix associated with the general parameter estimation problem under discussion takes on its largest (unique) value under the Gaussian assumption. This means that the parameter estimation methods or the experiment designs that are optimized based on the Gaussian CRB are min-max optimal in the sense that they yield the best CRB-related performance in the worst case (over a large class of data distributions that satisfy the stated regularity assumptions). Finally, note that the general expression for the Fisher information matrix presented in the previous section, [see (10)] together with results on γ in the literature can be used to derive the largest CRB matrix for other classes of functions $f(\epsilon)$ such as distributions with restricted support; in particular if the support of $f(\epsilon)$ is confined to the positive real line then the worst case $f(\epsilon)$ can be shown (see [3]) to be a scaled chi-squared distribution.

ACKNOWLEDGEMENT

This research was supported in part by the Swedish Research Council (VR). We thank Dr. Wenyi Zhang for a number of useful comments on a former version of this note and for pointing out [3].

AUTHORS

Petre Stoica (petre.stoica@it.ee.se) and ***Prabhu Babu*** (prabhu.babu@it.uu.se) are with the Department of Information Technology, Uppsala University, Uppsala, Sweden.

REFERENCES

[1] J. Delmas and H. Abeida, "Stochastic Cramér-Rao bound for noncircular signals with application to DOA estimation," *IEEE Trans. Signal Processing*, vol. 52, no. 11, pp. 3192–3199, 2004.

[2] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Upper Saddle River, NJ: Prentice-Hall, 2005.

[3] J. Bercher and C. Vignat, "On minimum Fisher information distributions with restricted support and fixed variance," *Inform. Sci.*, vol. 179, no. 22, pp. 3832–3842, 2009.

[4] T. Söderström and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[5] H. Cramér, "A contribution to the theory of statistical estimation," *Skand. Aktuarietidskr.*, vol. 29, pp. 85–94, 1946.

[SP]