15th Triennial World Congress, Barcelona, Spain

ESTIMATOR'S CREDIBILITY AND ITS MEASURES

X. Rong Li Zhanlue Zhao Vesselin P. Jilkov

University of New Orleans, Department of Electrical Engineering
New Orleans, LA 70148, USA
Phone: 504-280-7416, Fax: 504-280-3950, E-mail: xli@uno.edu

Abstract: Most estimators/filters provide assessments of their own estimation errors. Are these self-assessments trustable? To what degree are they trustable? This paper provides practical answers to such questions, referred to as the credibility of the estimators/filters. It formulates the concept of credibility and proposes practical measures of credibility with justifications. Numerical examples are provided to illustrate the use of the measures.

Keywords: Estimation, Filtering, Performance Evaluation

1. INTRODUCTION

Algorithms for parameter, signal, and state estimation are widely used in science and engineering. No matter how solid such an estimation algorithm, or estimator for short, is in theory, its performance and characteristics must be evaluated in practice to serve a number of purposes, such as verification of its validity, demonstration of its performance, and comparison with other estimators.

More specifically, estimators are almost always derived based on more or less restrictive assumptions. These assumptions are often not transparent to practitioners. Even if they are, in many practical situations, it is not easy to verify the validity of these assumptions directly. For a practitioner, the validity of these assumptions per se is of little concern. What is most important is whether the estimator works well for the application under consideration. This can be evaluated by stochastic simulation using a number of measures, particularly those discussed in (Li and Zhao, 2001).

This paper deals with a closely related issue—the credibility of an estimator. Many estimators provide self-assessments of estimation errors based on some simplifying assumptions. These self-assessments carry useful information about the estimation errors and the capability of the estimators. It would be ashamed to waste such information. Similar to those used to derive the estimator itself, however, these assumptions are usually even less transparent to practitioners and harder to verify. Even worse, the self-assessments could be quite misleading when the underlying assumptions are not adequately accurate. Then important questions for practitioners include: Can we trust these self-assessments and by how much amount? If not, are the estimators too optimistic or pessimistic?

Albeit very important in practice, work on this issue has been scarce. Although limited treatments of this topic can be found in publications, e.g., (Bar-Shalom and Birmiwal, 1983; Drummond *et al.*, 1998; Bar-Shalom *et al.*, 2001), in our opinion, it has received attention far less than what it deserves. As a result, it is virtually impossible for a practitioner to answer the above important questions satisfactorily.

The purpose of this paper is three-fold. First, it provides a formal definition of the credibility of an estimator to facilitate further studies. Second, it proposes practical metrics to *measure* the credibility of an estimator. Use of these metrics will enable a practitioner to answer questions like "By how much amount an estimator is noncredible?" Furthermore, the credibilities of two or more estimators can be compared using these metrics in a meaningful way. Finally, it intends to stimulate further studies of this important topic.

Terminology and notation. The following convention will be maintained throughout the paper. We refer to the quantity to be estimated as **estimatee**. It can be a time-invariant (or slowly varying) parameter, a (deterministic or random) process or signal, or in particular, the state of a (deterministic or random) system. We will use the term **estimator** to mean both parameter estimator and filter (in particular, state estimator). Let the n-dimensional estimatee, its estimate, and estimation error be denoted by x, $\hat{x}$, and $\tilde{x}$, respectively. We emphasize that n is reserved throughout the paper to denote the **dimension** of the estimatee. We denote the *actual* **bias** and mean-square error (**MSE**) matrix of $\hat{x}$ (i.e., mean and mean-square (not covariance) matrix of $\tilde{x}$) by μ and Σ, respectively. By **self-assessment** of an estimator, we mean the bias and MSE matrix of $\hat{x}$ *given by the estimator*. By **error covariance**, we always mean the MSE matrix given by the estimator, denoted by P, as opposed to the actual MSE matrix Σ. We always assume that the estimator-computed bias is (approximately) zero; otherwise the computed bias should be added to $\hat{x}$ to make $\hat{x}$ unbiased. Subscript i stands for quantities pertaining to the ith run of a Monte-Carlo simulation. It is always assumed that a total of M Monte-Carlo *independent* runs are conducted, and thus $\tilde{x}_i$ and $\tilde{x}_j$ are independent because $(x_i, \hat{x}_i)$ and $(x_j, \hat{x}_j)$ are independent. All default vectors are column vectors.

2. CREDIBILITY

As explained, self-assessments of an estimator contain valuable information about the estimation errors

[1] Research Supported in part by ONR grant N00014-00-1-0677, NSF grant ECS-9734285, and NASA/LEQSF grant (2001-4)-01.

and the capability of the estimator, and should be utilized properly. However, this information is not always reliable and it may even be misleading. An important issue is then how reliable an estimator's self-assessment is and how to determine this reliability both qualitatively and quantitatively. We refer to this issue as *credibility* issue of an estimator. Evidently, it amounts to the evaluation of the self-assessments.

Clearly, the credibility issue has two related sides. On the qualitative side, it addresses whether an estimator's self-assessment is credible. The answer should be "yes" or "no." Unfortunately, like many other decision problems, there is not a clear line that separates the two answers in most cases. An estimator accepted as credible by one user may be rejected by another user for the same case because the required levels of credibility may differ. Similarly, two noncredible estimators may have vastly different levels of noncredibility, which may lead to completely different actions. Therefore, it would be desirable if we could quantify the amount by which an estimator is credible or noncredible. This will enable us to compare quantitatively the credibility levels of estimators. In this sense, the quantitative side is probably more important than the qualitative side.

Definitions. An estimator is said to be **credible** at a level α ($0 \leq \alpha \leq 1$) if the difference between its actual estimation error (in particular, bias and MSE matrix) and its self-assessment (in particular, calculated bias and error covariance) is statistically insignificant at level α in the sense that the two errors can be treated as equal statistically. The maximum level α at which an estimator is rejected as being noncredible is referred to as the **noncredibility level** of the estimator. An estimator is said to be **optimistic** at a level α if its self-assessing estimation error is statistically smaller than the actual error at the α level. It is **pessimistic** in the opposite situations.

We will consider only the first two moments of the estimation error since most estimators are not able to provide information about the higher moments. We emphasize that the word "difference" above should not be interpreted literally. For example, a small difference in A and B could actually mean that A/B is close to 1, as well as $A - B$ is small.

To our knowledge, the most notable prior publications with a considerable amount of treatment of the issue of credibility are (Bar-Shalom and Birmiwal, 1983; Bar-Shalom *et al.*, 2001), referred to as *(finite-sample) consistency*. They address the issue and present a test for determining whether a filter should be accepted as credible, although without a formal definition of the (finite-sample) consistency. The term "credibility" is recommended because consistency is an extremely well-established concept widely used in statistics, which differs very much from the concept of credibility, and furthermore, "credible/credibility" has been used in statistics as a technical term, such as the "credible region" and the "degree of credibility," in reference to the commensurability of a hypothesis (model) relative to data or evidence.

The above definition makes it explicit that rigorously speaking, when we are speaking of the credibility (or noncredibility) of an estimator, the corresponding level should also be specified. However, we emphasize that it is *not* appropriate to define the (minimum) level at which an estimator is accepted as credible as the *credibility level* of the estimator.

3. CREDIBILITY MEASURES

The *normalized estimation error squared* (*NEES*) in the ith run is defined by $\epsilon_i = (x_i - \hat{x}_i)' P_i^{-1} (x_i - \hat{x}_i)$. The **average normalized estimation error squared** (**ANEES**), defined by (Bar-Shalom and Birmiwal, 1983; Drummond *et al.*, 1998; Bar-Shalom *et al.*, 2001)

$$\bar{\epsilon} = \frac{1}{nM} \sum_{i=1}^{M} \epsilon_i \tag{1}$$

or its equivalents are often used as a measure of an estimator's credibility: The closer to 1 the ANEES is, the more credible the estimator.

3.1 *Drawbacks of ANEES as Credibility Measure*

If the estimation error is indeed Gaussian distributed, whether an estimator is credible can be determined reasonably well by a chi-square test based on ANEES. However, this does not imply that ANEES is a good metric for measuring the credibility of an estimator (i.e., how credible or noncredible the estimator is). In fact, it can be easily seen from the definition that ANEES has the following drawbacks as a metric.

Firstly, ANEES penalizes optimism much more severely than pessimism. Note first that NEES is in essence equal to the actual MSE over the estimator-calculated error covariance. An estimator is too optimistic (or pessimistic) if NEES is substantially greater (or smaller) than n, where $n = \dim x$ is equal to the mean of NEES if the estimator is perfectly credible. Consider two cases, NEES $= 100n$ and NEES $= n/100$. In the first case, actual MSE is 100 times the calculated one, while the calculated one is 100 times the actual one in the second case. Since in both cases the two MSEs differ by a factor of 100, the two cases are *equally* noncredible [2]. However, a case with ten NEES's of $100n$ and one NEES of $n/100$ will have a much worse ANEES than a case with ten NEES's of $n/100$ and one NEES of $100n$. This drawback stems from the fact that in the ideal Gaussian case $\tilde{x} \sim \mathcal{N}(0, P)$, ANEES is chi-square distributed, which is highly asymmetrical around its mean—although it is more likely to have a small value, the possibility of a few large terms makes the mean significantly larger than the mode (the location of the peak of the density). In fact, the mean is n but the mode is $n - 2$ for $n > 2$.

Secondly, the use of ANEES is not convenient for comparing credibilities of different estimators. Consider two estimators with ANEES of 2.2 and 0.5, respectively. Most practitioners will be confused as which estimator is more credible. From the discussion above, it should be clear that 0.5 is equivalent to 2.0 in the ideal case because $2.0/1 = 1/0.5$. However, due to the drawback discussed above, the first estimator is

[2] Of course, the first case is worse than the second in terms of estimation accuracy, but they are equivalent as far as credibility is concerned.

probably more credible because ANEES = 0.5 indicates that the NEES is significantly smaller than unity in virtually all runs while ANEES = 2.2 could have been resulted from only a few large terms.

These drawbacks arise from using *arithmetic* average of a ratio—the NEES. As elaborated in (Li and Zhao, 2001), the *geometric* average would be much more appropriate for a ratio. Quite often it is more desirable to have more accurate measures.

3.2 *Noncredibility Indices*

Assume the estimator is unbiased. Then the credibility issue is concerned with the difference between or relative "ratio" of the actual MSE Σ and the estimator's error covariance P. However, Σ and P are in general matrices and cannot be compared directly—there is no generally accepted measure of difference between two matrices.

A difference between Σ and P is "equivalent" to that between Σ^{-1} and P^{-1}. One of the simplest and most widely used ways for the comparison of Σ^{-1} and P^{-1} is to compare $\tilde{x}'P^{-1}\tilde{x}$ and $\tilde{x}'\Sigma^{-1}\tilde{x}$, where $\tilde{x}$ is the estimation error. This is particularly appealing in the context of measuring credibility. The commonest quantity that quantifies the difference between $\tilde{x}'P^{-1}\tilde{x}$ and $\tilde{x}'\Sigma^{-1}\tilde{x}$ is $y = \tilde{x}'P^{-1}\tilde{x} - \tilde{x}'\Sigma^{-1}\tilde{x}$. Since y is random and $\tilde{x}_i'\Sigma_i^{-1}\tilde{x}_i \sim \chi_n^2$ under the assumption $\tilde{x}_i \sim \mathcal{N}(0, P)$, a natural idea is to use its sample average $\frac{1}{M}\sum_{i=1}^{M} y_i = n\times\text{ANEES} - \frac{1}{M}\sum_{i=1}^{M}\tilde{x}_i'\Sigma_i^{-1}\tilde{x}_i \approx n(\text{ANEES}-1)$ as a measure of the difference. However, this is directly proportional to ANEES, which has serious flaws as a measure, as discussed above.

An equally natural yet probably better idea is to use

$$\rho = \frac{\tilde{x}'P^{-1}\tilde{x}}{\tilde{x}'\Sigma^{-1}\tilde{x}} \tag{2}$$

to quantify the difference between Σ and P. In fact, ρ can be called the **credibility variable**. It is in general a function of the random error $\tilde{x}$. For a vector $\tilde{x}$, it is a *NEES ratio*—the actual NEES normalized by the ideal NEES. For a scalar $\tilde{x}$, it is actually a constant, independent of $\tilde{x}$, and is equal to the ratio of the true mean-square error over the estimator's error variance—this ratio is the most convincing measure of the credibility in the scalar case. This NEES ratio has an intimate relationship with the *relative deviation* of NEES: $\frac{\tilde{x}'P^{-1}\tilde{x}-\tilde{x}'\Sigma^{-1}\tilde{x}}{\tilde{x}'P^{-1}\tilde{x}+\tilde{x}'\Sigma^{-1}\tilde{x}}$.

For a vector $\tilde{x}$, ρ is not a good credibility measure because it is highly dependent on the random $\tilde{x}$. To remove (reduce) the uncertainty in ρ (a ratio), as elaborated in (Li and Zhao, 2001), geometric average

$$\left[\prod_{i=1}^{M}\rho_i\right]^{1/M} = \left[\prod_{i=1}^{M}\frac{\tilde{x}_i'P_i^{-1}\tilde{x}_i}{\tilde{x}_i'\Sigma_i^{-1}\tilde{x}_i}\right]^{1/M} = \left[\prod_{i=1}^{M}\frac{\epsilon_i}{\epsilon_i^*}\right]^{1/M}$$

is much more preferable to arithmetic average, where $\epsilon_i = \tilde{x}_i'P_i^{-1}\tilde{x}_i$ is the NEES of the estimator and $\epsilon_i^* = \tilde{x}_i'\Sigma_i^{-1}\tilde{x}_i$ is the NEES of a perfectly credible estimator. Often Σ_i is not known but can be approximated by its sample value $\frac{1}{M}\sum_{i=1}^{M}\tilde{x}_i\tilde{x}_i'$. From numerical and other considerations, we use logarithm and define the **noncredibility index** (**NCI**) by

$$\text{NCI} = \frac{10}{M}\sum_{i=1}^{M}\log_{10}(\epsilon_i) - \frac{10}{M}\sum_{i=1}^{M}\log_{10}(\epsilon_i^*) \tag{3}$$

The extra constant 10 is an amplification factor, as in the definition of the signal-to-noise ratio (SNR) in terms of power. For scalar $\tilde{x}$ with data-independent P, NCI is not random and is equal to $10\log_{10}(\Sigma/P)$ and is thus a perfect measure. Note that ANEES as a measure is flawed even in the scalar case. For a vector $\tilde{x}$, NCI is the sample average of 10 times the logarithm of the NEES ratio, $10\log_{10}(\rho)$, in analogy to average SNR.

Note that $\epsilon/E[\epsilon_*] = \tilde{x}'P^{-1}\tilde{x}/E[\epsilon_*]$ can be called the normalized NEES since it has a unity mean if the estimator is perfectly credible (ϵ_* stands for its ideal NEES). Then another idea is simply to use the geometric average of this normalized NEES: $\left[\prod_{i=1}^{M}(\epsilon_i/E[\epsilon_*])\right]^{1/M}$, which is free of the above drawbacks of the ANEES. As for NCI, we use its logarithm and define an alternative **noncredibility index** (**NCI-2**) by

$$\text{NCI-2} = \frac{10}{M}\sum_{i=1}^{M}\log_{10}(\epsilon_i) - 10E[\log_{10}(\epsilon_*)] \tag{4}$$

Subtraction of the term $E[\log(\epsilon_*)]$ makes the NCI-2 of a perfectly credible estimator close to zero since in this case $E[\text{NCI-2}] = 0$. Clearly, NCI-2 turns out to be a simplified version of NCI in that the sample average $\frac{1}{M}\sum_{i=1}^{M}\log_{10}(\epsilon_i^*)$ is replaced by the theoretical mean $E[\log_{10}(\epsilon_*)]$. As such, use of the sample mean here is more accurate than the theoretical mean because the latter relies on not necessarily accurate assumptions on the distribution of ϵ_*.

$P^{-1/2}\tilde{x}$ can be viewed as a normalized estimation error and $\delta/E[\delta_*]$ as one-dimensional equivalent normalized estimation error squared, where δ_* is the δ of a perfectly credible estimator, $\delta = \left[\text{sum}(P^{-1/2}\tilde{x})\right]^2 = \left(\sum_{j=1}^{n}u_j\right)^2$, and u_j is the jth element of $P^{-1/2}\tilde{x}$. Similar to NCI-2, we propose another measure, **noncredibility index-3** (**NCI-3**),

$$\text{NCI-3} = \frac{10}{M}\sum_{i=1}^{M}\log_{10}(\delta_i) - 10E[\log_{10}(\delta_*)] \tag{5}$$

NCIs are in essence the average ratio of the one-dimensional-equivalent true estimation error power to the calculated estimation error power in logarithm (dB), similar to the SNR definition.

For NCI-2 and NCI-3 to be useful, we need to know $E[\log_{10}(\epsilon_*)]$ and $E[\log_{10}(\delta_*)]$, respectively. In view of the central limit theorem it can be assumed in most cases that for a perfectly credible estimator, $\tilde{x}_* \sim \mathcal{N}(0, P)$. Then $\epsilon_* = \sum_{j=1}^{n}u_j^2 \sim \chi_n^2$ is standard chi-square with n degrees of freedom and $\delta_* \sim \chi_1^2(n)$ is chi-square with 1 degree of freedom and parameter $\sigma^2 = n$. It can be shown then that

$$10E[\log_{10}(\epsilon_*)] = \frac{10}{\ln 10}[\ln 2 + \psi(n/2)] \tag{6}$$

$$10E[\log_{10}(\delta_*)] = \frac{10}{\ln 10}[\ln(n/2) - \gamma] \tag{7}$$

where $\psi(x) = \frac{d}{dx}\ln\Gamma(x)$ is the Euler psi function, which has the recursion

$$\psi(m+1) = -\gamma + \sum_{k=0}^{m-1}\frac{1}{k+1}$$

$$\psi\left(m+\frac{1}{2}\right) = -\gamma + 2\left[\sum_{k=1}^{m}\frac{1}{2k-1} - \ln 2\right]$$

$\gamma = -\psi(1) = 0.57721566490$ is the Euler constant and $\psi\left(\frac{1}{2}\right) = -\gamma - 2\ln 2$. In the case where $\tilde{x}_* \not\sim \mathcal{N}(0, P)$ but $\tilde{x}_*$ has a known distribution, $E[\log_{10}(\epsilon_*)]$ and $E[\log_{10}(\delta_*)]$ may be obtained analytically or numerically. In any case, they may be obtained by simulation for the estimator under consideration by setting Σ in the computation of ϵ_* and δ_* with its sample value: $\Sigma = \frac{1}{M}\sum_{i=1}^{M}\tilde{x}_i\tilde{x}_i'$, where $\tilde{x}_i$ is the actual estimation error in run i. As seen from above, NCI-2 so obtained turns out to be NCI. Similarly, NCI-3 so obtained turns out to be equal to an alternative NCI: $\frac{10}{M}\sum_{i=1}^{M}\log_{10}(\delta_i/\delta_i^*)$. Under the above Gaussian assumption, the sample averages and the theoretical means can be expected to be close, otherwise the former may be significantly more accurate than the latter.

All NCIs are free of the drawbacks of the ANEES discussed above. For instance, (a) optimism and pessimism are penalized to the same degree; (b) the levels of noncredibility of different estimators can be compared simply by comparing the absolute values of their NCIs—the larger the worse; and (c) a positive and negative NCI represents optimism and pessimism, respectively. NCI is most accurate and NCI-3 is least accurate, but NCI requires the extra work of computing the sample MSE $\frac{1}{M}\sum_{i=1}^{M}\tilde{x}_i\tilde{x}_i'$. All NCIs are presented here because it is possible that one is easier to use than the others for a particular case. If the Gaussian assumption $\tilde{x} \sim \mathcal{N}(0, P)$ turns out to be valid, then NCI ≈ 0 ($= 0$ if the precise Σ_i^{-1} is used), $\epsilon_i \sim \chi_n^2$ and $\delta_i \sim \chi_1^2(n)$. In this case NCI-2 is more reliable (i.e., less uncertain) than NCI-3 because it can be shown that var[NCI-2] $\le$ var[NCI-3], where the equality holds if and only if $n = 1$, and in fact,

$$\text{var[NCI-2]} = \begin{cases} 9.31/M, & n = 1 \\ \dfrac{9.31}{3M} - \displaystyle\sum_{k=0}^{n/2-1}\frac{1.8861}{k^2M}, & n = 2m \ge 2 \\ \dfrac{9.31}{M} - \displaystyle\sum_{k=0}^{(n-1)/2-1}\frac{1.8861}{M[k+(1/2)]^2} & \\ \qquad n = 2m+1 \ge 3 & \end{cases}$$

$$\text{var[NCI-3]} = 9.31/M$$

That is, for a perfectly credible estimator, all NCIs are around zero and the standard deviation of NCI-2 is upper bounded by (approximately) $3/\sqrt{M}$, which is the standard deviation of NCI-3.

Finally, we also propose log-ANEES: **LNEES** $= 10\log(\text{ANEES})$ as a heuristic measure. By taking logarithm, it is hoped that the drawback of the ANEES that large errors are amplified can be corrected.

3.3 *Noncredibility Matrix and Its Scalar Measures*

Definitions. The *noncredibility matrix* (**NCM**) Δ of an estimator is: $\Delta = \Sigma - P$; the *normalized noncredibility matrix* (**NNCM**) of an estimator with nonsingular P is: NNCM $= P^{-1/2}\Delta P^{-1/2} = P^{-1/2}\Sigma P^{-1/2} - I$; the *normalized noncredibility matrix w.r.t.* I of an estimator with nonsingular P is: $\Gamma = P^{-1/2}\Sigma P^{-1/2}$; an estimator is said to be **optimistic definite** (or semidefinite) if NCM positive definite (or semidefinite), or **pessimistic definite** (or semidefinite) if NCM negative definite (or semidefinite).

With these definitions, an estimator is perfectly credible if and only if $\Delta = 0$. The "size" of NNCM (not NCM) measures the level of noncredibility. However, the direct use of a matrix norm (or some other scalar measure) of NNCM or Γ suffers from drawbacks similar to those of ANEES as a measure of noncredibility. Better scalar measures should be used. For this purpose, consider the following measures: log matrix norm ratio (*log-MNR*), matrix norm relative error (*MNRE*), log mse ratio (*log-MSER*) [3], the mse relative error (*MSERE*), log generalized error variance ratio (*log-GEVR*) [4], and the generalized error variance relative error (*GEVRE*), defined by

$$\text{MNRE1} = \frac{\|\Sigma - P\|}{\|\Sigma\| + \|P\|}, \quad \text{MNRE2} = \frac{\|\Sigma\| - \|P\|}{\|\Sigma\| + \|P\|}$$

$$\text{log-MSER} = \log\frac{\text{tr}(\Sigma)}{\text{tr}(P)}, \quad \text{MSERE} = \frac{\text{tr}(\Sigma) - \text{tr}(P)}{\text{tr}(\Sigma) + \text{tr}(P)}$$

$$\text{log-MNR} = \log\frac{\|\Sigma\|}{\|P\|}, \quad \text{GEVRE} = \frac{\det(\Sigma) - \det(P)}{\det(\Sigma) + \det(P)}$$

$$\text{log-GEVR} = \log\frac{\det(\Sigma)}{\det(P)} = \log[\det(\Gamma)]$$

where $\|A\|$ stands for any matrix norm of A. These measures are free of the drawbacks mentioned above and enjoy some nice properties. For example, (a) they are all symmetric w.r.t. Σ and P (so that optimism and pessimism are treated equally); (b) $0 \le$ MNRE ≤ 1, $0 \le$ GEVRE ≤ 1, $0 \le$ MSERE ≤ 1; (c) optimistic/pessimistic (semi)definite implies positive/negative (nonnegative/nonpositive) MNRE2, GEVRE, MSERE, log-GEVR, and log-MSER, (d) an estimator is perfectly credible (i.e., $P = \Sigma$) if and only if MNRE1 $= 0$; $P = \Sigma$ implies MNRE2 $= 0$ but the converse is not true. These measures have their own pros and cons and hence are complementary to each other. For instance, log-MSER and MSERE have a clear physical interpretation of power but depend only on the diagonal elements, while log-GEVR and GEVRE have a clear geometric interpretation of volume but it is zero when the parallelotope is collapsed.

Different estimation errors $\tilde{x}$ may excite different "modes" of the credibility variable ρ defined by (2). As a result, average ρ (and hence NCI) carry useful information about the credibility of an estimator that is dependent on the distribution of the estimation error (beyond the second moment). NCI, however, is a

---

[3] The MSE matrix is MSE $= E[(x-\hat{x})(x-\hat{x})']$ and the scalar mean-square error is mse $= E[(x-\hat{x})'(x-\hat{x})] = \text{tr(MSE)}$.

[4] The determinant of a covariance matrix is called a *generalized variance* in statistics, and is proportional to the volume of the error ellipsoid.

scalar measure. On the other hand, NCM is a matrix measure that depends only on the actual and calculated second moments of the estimation error. These two classes of measures are related. For example, it can be shown that ρ and NCI are related to the eigenvalues of Γ by

$$0 \le \lambda_{\min}(\Gamma) \le \rho \le \lambda_{\max}(\Gamma)$$
$$10\log_{10}\lambda_{\min}(\Gamma) \le \text{NCI} \le 10\log_{10}\lambda_{\max}(\Gamma)$$

As such, $[\lambda_{\min}(\Gamma), \lambda_{\max}(\Gamma)]$ can be called the **credibility interval** of an estimator.

4. EVALUATION OF VARIOUS MEASURES' EFFECTIVENESS

To better explain the concepts, consider a scalar estimatee. It is better to recognize that for a large mse relative error $\gamma = (\Sigma - P)/\Sigma$, a log scale is more appropriate than a linear scale, while for a small γ, a linear scale is more appropriate. For example, $\gamma = 8$ and $\gamma = 8.2$ have a negligible difference. On the other hand, while $\gamma = 0.1$ differs significantly from $\gamma = 0.3$, it is usually not fair to think that an estimator with $\gamma = 0.3$ is three times more noncredible than an estimator with $\gamma = 0.1$. On the other hand, it is clearly more reasonable to use a log scale for the mse ratio P/Σ.

4.1 *Scalar Case*

As explained in Sec. 3.2, in the scalar case, NCI is a perfect measure because it is equal to $10\log_{10}(\Sigma/P)$, which is not random at all. However, ANEES is flawed, as explained before and illustrated below.

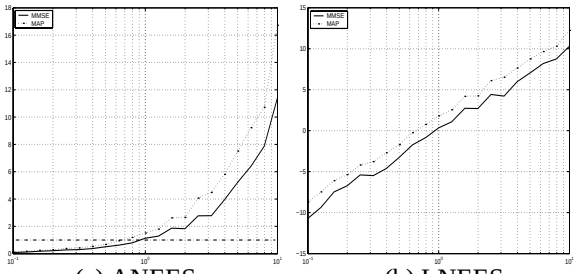
Consider the same example as the one considered in (Li and Zhao, 2001). The estimation errors of a MAP estimator and an MMSE estimator of x with density $f(x) = e^{-x}1(x)$ using a single measurement $z = x + v$ are

$$\tilde{x}^{\text{MAP}} = x - \max(x + v - 1, 0) = \begin{cases} 1 - v, & z > 1 \\ x, & z \le 1 \end{cases}$$
$$\tilde{x}^{\text{MMSE}} = 1 - v - (\sqrt{2\pi}[1 - \Phi(1 - z)])^{-1} e^{-(z-1)^2/2}$$

where $v \sim \mathcal{N}(0, 1)$ and $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function. $\tilde{x}^{\text{MAP}}$ is highly non-Gaussian while $\tilde{x}^{\text{MMSE}}$, albeit non-Gaussian, is not far from Gaussian.

Fig. 1 shows ANEES and LNEES versus Σ/P from 500 Monte Carlo runs. The horizontal axes are obtained by varying P while holding Σ equal to their sample values $\frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{x}_i)^2$ for MAP and MMSE estimators, respectively. The corresponding 95% probability interval for ANEES is $(0.8799, 1.1277)$.

It turns out that the chi-square test based on ANEES accepted the MMSE estimators $\hat{x}^{\text{MMSE}}$ with $\Sigma/P \approx 0.9$ and rejected all other MMSE estimators. This result is acceptable since $\hat{x}^{\text{MMSE}}(\Sigma/P \approx 0.9)$ almost perfectly credible. However, the same chi-square test incorrectly accepted a noncredible MAP estimators $\hat{x}^{\text{MAP}}(\Sigma/P \approx 0.65)$ and rejected all other MAP estimators, including the credible one, $\hat{x}^{\text{MAP}}(P = \Sigma)$.



(a) ANEES (b) LNEES

Fig. 1. ANEES and LNEES versus Σ/P.

This mistake arises from the incorrect assumption that $\tilde{x}^{\text{MAP}}(P = \Sigma) \sim \mathcal{N}(0, P)$. As such, NEES and thus ANEES of $\hat{x}^{\text{MAP}}(P = \Sigma)$ are far from chi-square distributed. The Gaussian assumption is not bad for the MMSE estimators in this example and thus the results of credibility tests are fine. This example demonstrates that the chi-square test based on ANEES is sensitive to the validity of the Gaussian assumption. For the scalar case, Σ/P is the most convincing measure of the credibility. It is also evident that the ANEES amplifies the region over which the estimator is optimistic and suppresses the pessimistic region.

NCI is not shown because NCI for this scalar case is never random at all and is a perfect straight line connecting $(10^{-1}, -10)$ and $(10^{1}, 10)$, which passes through the origin $(10^{0}, 0)$ because $\rho_i = \tilde{x}_i' P_i^{-1}\tilde{x}_i/(\tilde{x}_i'\Sigma^{-1}\tilde{x}_i) = \Sigma/P_i$. This is perfect: (a) optimism and pessimism are treated symmetrically; (b) everybody is treated fairly—no particular region is suppressed or amplified; (c) estimators of different types (e.g., MAP and MMSE) have identical NCI curve and thus NCI values from different estimators can be directly compared in a meaningful way. As such, the credibility of different estimators in different cases may be compared by their NCIs. NCI-2 is exactly NCI shifted upward by the amount $\frac{10}{M}\sum_{i=1}^{M}\log_{10}(\tilde{x}_i^2/P) - 10E[\log_{10}(\epsilon_*)]$ and thus is also a perfect straight line. For the MMSE estimator, it almost coincides with NCI, while for the MAP estimator, it has a small upward shift. The LNEES curves are close to a straight line but with some variation. The LNEES of the MAP estimator deviates significantly from the ideal one [i.e., the one that passes through the origin $(10^{0}, 0)$].

4.2 *Vector Case*

For a vector x, care should be taken to evaluate the effectiveness of credibility measures. Since our NCIs are based on the NEES ratio, it would be unfair to evaluate the effectiveness of various credibility measures by checking how close they are to the direct proportion of $\log(\rho)$ or $\log(\delta/\delta_*)$. Albeit meaningful, such evaluation is in favor of our NCI. We devise below a more impartial evaluation.

Let $A = [a_{ij}]$ be nonrandom and $\tilde{A} = [\tilde{a}_{ij}] = A + B$, where $B = [b_{ij}]$ is the difference between A and $\tilde{A}$. Generate b_{ij} randomly with a $\mathcal{N}[0, (\beta a_{ij})^2]$ distribution. Then $\text{var}(\tilde{a}_{ij}) = \text{var}(b_{ij}) = (\beta a_{ij})^2$. Note that while $E[\tilde{A}] = A$, in *most* runs $\tilde{A}$ is substantially

different than A unless β^2 is quite small, and the difference increases with β^2. Note that $-B$ and B do not cancel each other as far as the difference between $\tilde{A}$ and A is concerned. In this sense, the scalar β^2 quantifies the difference between $\tilde{A}$ and A. As such, we may think the following holds in a statistical sense [5]: $\beta A \approx B$. Now let $A = \Sigma^{-1/2}$ and $\tilde{A} = P^{-1/2}$. It thus follows that since $B = \tilde{A} - A$,

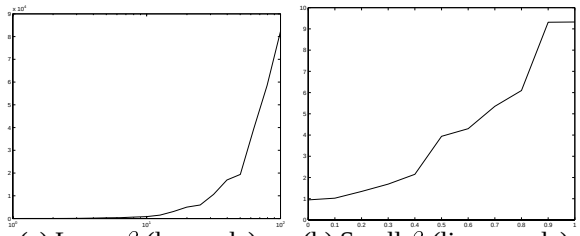
$$\begin{aligned} \beta I &\approx \Sigma^{1/2}(P^{-1/2} - \Sigma^{-1/2}) = \Sigma^{1/2}P^{-1/2} - I \\ \beta I &\approx (P^{-1/2} - \Sigma^{-1/2})\Sigma^{1/2} = P^{-1/2}\Sigma^{1/2} - I \end{aligned} \tag{8}$$

that is, β quantifies the credibility in some statistical sense. Evidently, β can also be used as a credibility measure.

Fig. 2 shows ANEES versus β while Fig. 3 shows NCI, NCI-2, NCI-3, and LNEES versus β, obtained from $M = 100$ Monte-Carlo runs. As argued above, linear and log scales are used for β over $[0, 1)$ and $[1, 100]$, respectively. The ANEES, NCIs, and LNEES were computed as follows. We set

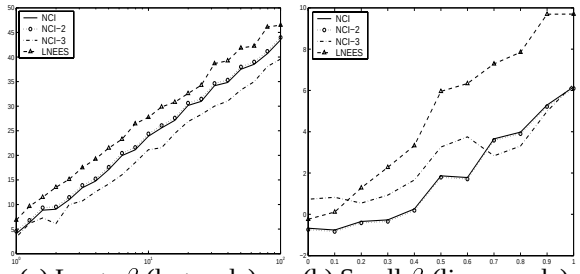
$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \quad \Sigma^{-1} = AA'$$

For each β value, we generated $\tilde{x}_i \sim \mathcal{N}(0, \Sigma)$, $b_{nm}^{(i)} \sim \mathcal{N}[0, (\beta a_{nm})^2]$, $B_i = [b_{nm}^{(i)}]$, $\tilde{A}_i = A + B_i$, $P_i^{-1} = \tilde{A}_i \tilde{A}_i'$, $i = 1, \ldots, M$. Then, ANEES, NCIs, and LNEES were computed from their formulas using $\tilde{x}_i$, P_i^{-1}, and Σ^{-1}.



(a) Large β (log scale) (b) Small β (linear scale)

Fig. 2. ANEES versus β.



(a) Large β (log scale) (b) Small β (linear scale)

Fig. 3. Noncredibility indices and LNEES vs. β.

Clearly, ANEES amplifies large β severely while NCIs and LNEES are approximately proportional to β, which is desirable. NCI and NCI-2 are essentially the same and are most reliable (i.e., with least variation). LNEES exhibits a larger variation but NCI-3 is even less reliable. The variation of the ANEES increases significantly as β increases.

[5] This is similar to viewing a zero-mean random variable as having a length equal to its standard deviation, which has a solid theoretical foundation.

NCI increases by 20 dB when β is increased 10 times. This is in agreement with the scalar case where NCI $= 10\log_{10}(\Sigma/P)$. For example, NCI $= 6$ dB corresponds to $\beta \simeq 1$, or equivalently [from (8)] $\Sigma^{1/2}P^{-1/2} \approx 2I$ (i.e., NCI $= 6$ dB corresponds to $\Sigma P^{-1} \approx 4I$). On the other hand, NCI $= 6$ dB corresponds to $\Sigma/P = 4$ because NCI $= 10\log_{10}(\Sigma/P)$ in the scalar case. Similarly, NCI $= 10$ dB corresponds to $\beta \simeq 2$, or equivalently $\Sigma P^{-1} \approx 9I$, which agrees with NCI $= 10\log_{10}(\Sigma/P) = 10\log_{10} 9 \simeq 10$ dB in the scalar case. This comparison demonstrates that NCI enjoys a nice property that it is invariant with respect to the dimension of the estimator. Note that LNEES does not possess this nice property.

In summary, these two examples demonstrate that NCI is the most accurate measure of the credibility of an estimator with many nice features. In view of this, NCI can be used as a universal measure for credibility of estimators.

5. SUMMARY

The problem of the credibility of an estimator, that is, whether and how much an estimator's self-assessment of the estimation errors can be trusted, has been formulated. A number of credibility measures, ranging from the most accurate noncredibility index (NCI) to the simple, intuitively appealing mse ratio, have been presented, along with justifications. The pros and cons of these measures have been explained. It has been shown via numerical examples that the proposed credibility measures are fairly accurate in that they provide fairly good indication of the level of (non)credibility. It is concluded that NCI, with its superior accuracy and nice properties, can be used as a universal measure of credibility of estimators. It has also be shown that the statistic, ANEES, while valid for credibility tests, has fundamental flaws as a measure of credibility.

6. REFERENCES

Bar-Shalom, Y. and K. Birmiwal (1983). Consistency and Robustness of PDAF for Target Tracking in Cluttered Environments. *Automatica* **19**, 431–437.

Bar-Shalom, Y., X. R. Li and T. Kirubarajan (2001). *Estimation with Applications to Tracking and Navigation: Theory, Algorithms, and Software*. Wiley. New York.

Drummond, O. E., X. R. Li and C. He (1998). Comparison of Various Static Multiple-Model Estimation Algorithms. In: *Proc. 1998 SPIE Conf. on Signal and Data Processing of Small Targets*, vol. 3373. pp. 510–527.

Li, X. R. and Z.-L. Zhao (2001). Measures of Performance for Evaluation of Estimators and Filters. In: *Proc. 2001 SPIE Conf. on Signal and Data Processing of Small Targets*, vol. 4473. San Diego, CA, USA. pp. 530–541.