Practical Measures and Test for Credibility of an Estimator

X. Rong Li Zhanlue Zhao Vesselin P. Jilkov

Department of Electrical Engineering
University of New Orleans
New Orleans, LA 70148, USA

ABSTRACT

Most estimators and filters provide assessments of their own estimation errors. Are these self-assessments trustable? What is the degree to which they are trustable? This paper provides practical answers to such questions, referred to as the (level of) credibility of the estimators/filters. It formulates the concept of credibility, proposes several practical measures of credibility, and discusses a credibility test. A special effort is made to explain the underlying rationales as clearly as possible. Numerical examples are provided to illustrate the pros and cons of the measures and test.

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 on the basis of 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. 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 proposed or discussed in [7].

This paper deals with a closely related issue—the credibility of an estimator. Many estimators provide self-assessments of their 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, usually these assumptions are 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 would be: Can we trust these self-assessments? If not, are the estimators too optimistic or pessimistic? By how much amount?

Albeit very important in practice, work on this issue has been scarce for some unknown reason. Although limited treatments of this topic can be found in publications, e.g., [1,2,4,3], 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 four-fold. First, it provides a formal definition of the credibility of an estimator to facilitate further studies of this topic. 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

Further author information: (Send correspondence to X. R. Li)
X. R. Li: E-mail: xli@uno.edu; Phone: 504-280-7416, Fax: 504-280-3950

metrics in a meaningful way. Third, it discusses how credibility can be *tested* properly, along with relevant theoretical results; that is, how to properly answer such questions as "Is the self-assessment of an estimator trustable?" and "Is the estimators too optimistic or pessimistic?". Finally, it intends to stimulate further studies of this important topic.

The remainder of the paper is organized as follows. Section 2 formulates the problem and defines the credibility of an estimator. Section 3 discusses a test concerning whether the self-assessment of an estimator should be rejected based on normalized estimation error squared, along with some relevant theoretical results. The main results are presented in Section 4, where several credibility measures are proposed, including so-called noncredibility indices. They serve as practical metrics to measure the amount by which an estimator is not credible. Numerical examples are provided in Section 5 to illustrate the utility of the metrics and test for estimator credibility. The paper is concluded by a summary in Section 6. Mathematical details are given in Appendix.

The following convention will be maintained throughout the paper.

Terminology and notation. We refer to the quantity to be estimated as **estimatee**. It can 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 saying the **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. We always assume that the estimator-computed bias is (approximately) zero; otherwise the computed bias should be added to $\hat{x}$ to make $\tilde{x}$ 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. All default vectors are column vectors. The P^{-1}-norm of a vector a is defined as $\|a\|_{P^{-1}} = (a'P^{-1}a)^{1/2}$, where a' stands for the transpose of the column vector a.

2. CREDIBILITY

As explained in the introduction, the self-assessment of an estimator contains valuable information about the estimation errors 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 in practice 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.

Since an estimator is data-driven and its self-assessment is data-dependent in general, the evaluation of its self-assessment should be done only in a statistical sense. In practice, this is almost always done by the Monte-Carlo method via computer simulations.

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.

We will consider only the first two moments of the estimation error since most estimators will not be able to provide any information about the higher moments.

With the above considerations, we define the credibility of an estimator as follows.

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

We emphasize that the word "difference" above should not be interpreted literally. In fact, we mean both (or either) additive difference and/or multiplicative difference. 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. In fact, the ratio is used in both credibility test and measures.

To the authors knowledge, the most notable prior publications that provide a considerable amount of treatment of the issue of credibility are [1,2], 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 relative to data or evidence.

The self-assessment of a noncredible estimator is not reliable at the level considered. This is not to be confused with the reliability of estimates provided by the estimator. One type of reliability does not necessarily imply the other.

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. In fact, we do not propose any definition of the credibility level in this paper due to a number of difficulties. A detailed explanation is given in Subsection 3.3.

3. CREDIBILITY TEST

3.1. ANEES

Recall that n is the dimension of x and P is the estimator-provided error covariance.

The *normalized estimation error squared* (*NEES*) in the ith run is defined by

$$\epsilon_i = (x_i - \hat{x}_i)' P_i^{-1} (x_i - \hat{x}_i) = \|\tilde{x}_i\|^2_{P_i^{-1}} \tag{1}$$

The **average normalized estimation error squared** (**ANEES**), defined by

$$\bar{\epsilon} = \frac{1}{nM} \sum_{i=1}^{M} \epsilon_i = \frac{1}{M} \sum_{i=1}^{M} \|\tilde{x}_i\|^2_{P_i^{-1}} / n \tag{2}$$

is recommended for testing if an estimator should be rejected as not credible or is optimistic or pessimistic. The closer to 1 the ANEES is, the more credible the estimator.

NEES can be interpreted as the distance squared between x and $\hat{x}$ in terms of the P^{-1}-norm $\|\tilde{x}\|_{P^{-1}}$. In this sense, ANEES is a one-dimensional-equivalent average distance in terms of the P^{-1}-norm.

3.2. Mean and Variance of ANEES

Denote the mean and covariance of estimation error $\tilde{x}$ as μ and Σ. Then it can be shown [3] that, regardless of the distribution of $\tilde{x}$,

$$\begin{aligned} E[\epsilon_i] &= E[E(\epsilon_i|P_i)] = E\left[\text{tr}(P_i^{-1}\Sigma) + \mu' P_i^{-1}\mu\right] \\ E[\bar{\epsilon}] &= \frac{1}{nM}\sum_{i=1}^{M} E[\epsilon_i] = \frac{1}{nM}\sum_{i=1}^{M} E\left[\text{tr}(P_i^{-1}\Sigma) + \mu' P_i^{-1}\mu\right] = \frac{1}{n} E\left[\text{tr}(P_i^{-1}\Sigma) + \mu' P_i^{-1}\mu\right] \end{aligned}$$

where the outside expectation in the first equation is over the random P_i (in general P_i is random due to its dependence on the measurements). If P_i does not depend on i (e.g., for linear estimators), then

$$E[\bar{\epsilon}] = [\text{tr}(P^{-1}\Sigma) + \mu' P^{-1}\mu]/n \tag{3}$$

In addition, assume $\tilde{x} \sim \mathcal{N}(\mu, \Sigma)$, which is justifiable by the central limit theorem. Then, it can be shown that

$$\begin{aligned} \text{var}(\epsilon_i) &= 2\text{tr}(P^{-1}\Sigma P^{-1}\Sigma) + 4\mu' P^{-1}\Sigma P^{-1}\mu - (\mu' P^{-1}\mu)^2 \\ \text{var}(\bar{\epsilon}) &= \frac{1}{(nM)^2}\sum_{i=1}^{M} \text{var}(\epsilon_i) = [2\text{tr}(P^{-1}\Sigma P^{-1}\Sigma) + 4\mu' P^{-1}\Sigma P^{-1}\mu - (\mu' P^{-1}\mu)^2]/(n^2 M) \end{aligned}$$

If in addition the estimation error $\tilde{x}$ has zero mean, then it follows from the above that

$$E[\bar{\epsilon}] = \text{tr}(P^{-1}\Sigma)/n, \qquad \text{var}(\bar{\epsilon}) = 2\text{tr}(P^{-1}\Sigma P^{-1}\Sigma)/(n^2 M) \tag{4}$$

We emphasize that these results hold true even if the estimator-calculated error covariance P is very different from the true MSE matrix Σ.

It is clear from the above that loosely speaking, the mean of the ANEES is larger (or smaller) than 1 if the actual MSE matrix Σ is larger (or smaller) than the estimator calculated error covariance P. In view of this, we may conclude that *an estimator is optimistic (or pessimistic) if its ANEES is significantly larger (or smaller) than* 1. Note, however, that this simple judgment is not reliable. See Subsection 4.1 for more details.

It can also be shown by Schwarz inequality that $\text{var}(\epsilon_i) \geq 2(E[\epsilon_i])^2/n$ and $\text{var}(\bar{\epsilon}) \geq 2(E[\bar{\epsilon}])^2/(n^3 M)$.

3.3. Test for Credibility

In the remaining parts of this paper, we assume that $\tilde{x}_i = x_i - \hat{x}_i$ has zero mean. If $\tilde{x}_i$ has a known bias b_i, then $\tilde{x}_i$ should be replaced by $\tilde{x}_i - b_i$.

Assume $\tilde{x} \sim \mathcal{N}(0, \Sigma)$. If the assumptions (models and approximations) based on which an estimator is obtained are valid, then the estimator-calculated error covariance P should be close to the actual MSE matrix Σ. As such, the NEES ϵ as a quadratic form in $\tilde{x}$ should have a (standard) χ^2_n distribution approximately. Then, ANEES $\bar{\epsilon} = \frac{1}{nM}\sum_{i=1}^{M} \epsilon_i$ should be (approximately) a sum of M independent χ^2_n random variables scaled down by nM, that is, a chi-square $\chi^2_{nM}(\frac{1}{nM})$ random variable, with nM degrees of freedom and parameter $\sigma^2 = \frac{1}{nM}$, which has mean 1 and variance $\frac{2}{nM}$.

A standard (i.e., $\sigma^2 = 1$) chi-square distribution with m degrees of freedom can be approximated by a Gaussian distribution

$$F_{\chi^2_m}(x) = \Phi\left(\sqrt{\frac{9m}{2}}\left[\left(\frac{x}{m}\right)^{1/3} - 1 + \frac{2}{9m}\right]\right)$$

and the percentile points can be calculated accordingly. As shown in [6], this Wilson-Hilferty approximation is usually accurate enough for $m > 5$ and is much more accurate than the more widely used Fisher approximation

$$F_{\chi^2_m}(x) = \Phi\left(\sqrt{2x} - \sqrt{2m-1}\right)$$

Since $nM \gg 5$ in virtually all practical cases, several typical (approximate) two-sided probability (not confidence) intervals for ANEES assuming the estimator is credible [more precisely, assuming $\tilde{x} \sim \mathcal{N}(0, P)$] are given in Table 1 based on the above Wilson-Hilferty approximation. Note that the intervals change little over a large range of probability and are not symmetric about 1.

Table 1. Examples of approximate two-sided probability intervals for ANEES.

Probability	Probability Interval	Probability Interval with $nM = 400$
99%	$\left(\left[1 - \frac{2}{9nM} - 2.576\sqrt{\frac{2}{9nM}}\right]^3, \left[1 - \frac{2}{9nM} + 2.576\sqrt{\frac{2}{9nM}}\right]^3\right)$	$(0.8272, 1.192)$
95%	$\left(\left[1 - \frac{2}{9nM} - 1.96\sqrt{\frac{2}{9nM}}\right]^3, \left[1 - \frac{2}{9nM} + 1.96\sqrt{\frac{2}{9nM}}\right]^3\right)$	$(0.8662, 1.143)$
90%	$\left(\left[1 - \frac{2}{9nM} - 1.645\sqrt{\frac{2}{9nM}}\right]^3, \left[1 - \frac{2}{9nM} + 1.645\sqrt{\frac{2}{9nM}}\right]^3\right)$	$(0.8866, 1.119)$
80%	$\left(\left[1 - \frac{2}{9nM} - 1.282\sqrt{\frac{2}{9nM}}\right]^3, \left[1 - \frac{2}{9nM} + 1.282\sqrt{\frac{2}{9nM}}\right]^3\right)$	$(0.9105, 1.092)$

Note that when $\tilde{x} \sim \mathcal{N}(0, \Sigma)$ and $P \neq \Sigma$, NEES and thus ANEES are actually not chi-square distributed because by the Ogasawara-Takahashi theorem [8] $\tilde{x}' P^{-1} \tilde{x}$ is chi-square if and only if $\Sigma P^{-1} \Sigma P^{-1} \Sigma = \Sigma P^{-1} \Sigma$.

The rejection of an estimator as credible is based on the following principle. The occurrence of an event of an extremely small probability on a single trial has a profound implication—the model (or assumptions) based on which the probability is calculated is incorrect and should be abandoned. This is sometimes referred to as the *principle of small-probability events*. It is this principle that underlies the acceptance (or rejection) of a theory as scientific. Specifically, a theory, no matter how exotic or bizarre, is *confirmed* to be correct and generally accepted if its predictions are verified by experiments or observations. Good examples include Einstein's general theory of relativity and the Big-Bang theory in cosmology. What is particularly amazing is that these theories were accepted by most physicists in the same field right after only *a few* bold predictions were verified. Why was this the case? The answer lies in the principle of small-probability events. Assuming these theories are not correct, a verification of any of their unexpected *predictions* would be highly improbable. Given such verifications, the underlying model, that is, the assumption that they are incorrect, has to be abandoned. That is the rationale for the following quote from Kitaigordski: *A first-rate theory predicts; a second-rate theory forbids and a third-rate theory explains after the event*. This principle is also applied in many areas outside science. For example, it is also based on this principle that DNA evidence can be used to convict a defendant in a criminal court if a match is found between the DNA structures of the two samples (one found at the crime scene and the other from the defendant): Assuming the defendant is not guilty, a match would have an extremely small probability (one in millions). Since a match is found, the assumption that the defendant is not guilty ought to be abandoned.

According to this small-probability-event principle, the self-assessment of an estimator (i.e., the assumption that the estimator is credible) should be rejected only if a very small-probability event occurred on a single trial, that is, when ANEES is outside the interval of a very high probability (say, 95% or 99%). Experience indicates that ANEES is usually outside the 95% probability (not confidence, since ANEES is random) interval if an estimator is indeed not credible as a result of a major breach of the assumption (see numerical examples in Section 5). When ANEES is outside the 95% probability interval, we can say that the estimator is not credible with at least 95% confidence or the noncredibility level is at least 95%.

ANEES is used to judge whether an estimator is credible or not in the sense of having a self-assessment of the MSE that is statistically acceptable. It is usually a good indication of whether the assumptions (and approximations) used by the estimator are sufficiently accurate. If ANEES is much greater than 1, the actual

estimation error is much larger than what the estimator believes (i.e., the estimator is too optimistic); if ANEES is much smaller than 1, the actual estimation error is much smaller than what the estimator believes (i.e., the estimator is too pessimistic).

To avoid unnecessary confusion, care should be exercised when using noncredibility levels. For instance, if ANEES of an estimator is outside the 90% interval but inside the 95% interval, then the estimator may be deemed *not* credible with a ("confidence") level in between 90% and 95%, that is, the *noncredibility level* is in between 90% and 95%. However, it is better to use such *noncredibility* levels only when we have rejected the estimator as being credible and the level is high (say, on or above 90%). In other words, we better use such noncredibility levels only for noncredible estimators because a credible estimator may have a seemingly "high" noncredibility level, which can be confusing to most practitioners. When ANEES is outside an interval of a not high probability (say, 70%), rejection of the credibility of an estimator and treating it with a noncredibility level of 70% lacks solid ground and is in fact quite questionable because the small-probability-event principle is not applicable here. In short, a noncredibility level is meaningful only if it is quite high.

On the other hand, in the case when ANEES falls inside its 95% probability interval, although we may say the estimator is accepted as credible at 95% level, it is *not* appropriate to think that the estimator has a 95% credibility level; otherwise all estimators are credible at the 100% level because every ANEES falls inside the 100% probability interval, which is $[0, \infty)$! Given the above definition of noncredibility levels, if we require that the credibility and noncredibility levels sum up to unity, then the credibility level of many credible estimators are very low (e.g., 10% or lower). This is highly undesirable. To define the credibility level as the minimum level at which an estimator is accepted as credible is seriously flawed. For instance, suppose that the ANEES is on an end point of its 90% probability interval (say, equal to 1.119 for $nM = 400$). This definition states that the estimator has a 90% *credibility* level. However, from our definition of noncredibility level, it also has a 90% *noncredibility* level. Clearly, the estimator should be defined as having a 90% *noncredibility* level because the ANEES is supposed to be in the interval with 90% probability. Otherwise the credibility level is higher (say 99%) if the ANEES is further away from 1 (say, 1.192 for $nM = 400$)! These examples illustrates that it is not easy to define properly a simple, quantitative credibility level for an estimator.

The above difficulty associated with the concept of credibility level stems from the inherent difficulty in interpreting the confidence level of a decision for hypothesis testing based on probability (or confidence) intervals. No simple, general, and good solution is available within this framework. A better solution is to use credibility measures, presented in the next section, rather than test-based confidence levels.

The credibility test discussed here is essentially the same as that presented in [1] and discussed in [2,3]. The difference is the introduction of the $\frac{1}{n}$ factor in the ANEES that makes the mean of the ANEES invariant of the dimension of x. This makes it more convenient to use. Also, we believe that the underlying principle is better explained, and the use of the replacement of the Fisher approximation by Wilson-Hilferty approximation is beneficial when nM is not large.

4. CREDIBILITY MEASURES

4.1. 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, as discussed in the previous section. 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:

- 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 is the dimension of x, which 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 calculate one is 100 times the actual one in the second case. Since in both cases the two MSEs differ with a factor of 100, the two cases are *equally* noncredible*. 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 case $\tilde{x} \sim \mathcal{N}(0, P)$, ANEES is chi-square distributed, which is highly nonsymmetrical 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, for $n > 2$, the mean is n but the mode is $n - 2$.

- 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 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 arises from using arithmetic average of a ratio—the NEES. As elaborated in [7] in the context of several measures, the geometric average would be much more appropriate as an average of a ratio.

4.2. Simple Measures of Credibility

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

We start with considering simple and intuitively appealing measures first. Two most widely used scalar measures of a matrix are its trace and determinant. Fortunately, they have clear physical interpretations for an MSE (or error covariance) matrix. With this in mind, the first such measure we propose is called **mse ratio** (**MSER**)†, defined by

$$\text{MSER} = \frac{\text{estimator's assessment of mse}}{\text{actual mse}} = \frac{\text{tr}(P)}{\text{tr}(\Sigma)} \tag{5}$$

and the second is called **mse relative error** (**MSERE**), defined by

$$\text{MSERE} = \frac{\text{actual mse} \ - \ \text{estimator's assessment of mse}}{\text{actual mse}} = \frac{\text{tr}(\Sigma) - \text{tr}(P)}{\text{tr}(\Sigma)} = \frac{\text{tr}(\Sigma - P)}{\text{tr}(\Sigma)} \tag{6}$$

Clearly MSER and MSEE depend only on the diagonal elements of the MSE matrix and the estimator's error covariance matrix. This is not quite desirable. The following two simple measures do not have this shortcoming, but they are slightly less intuitive: **generalized error variance ratio** (**GEVR**)‡, defined by

*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.

†Recall that for an estimator $\hat{x}$, the scalar mean-square error is mse $= E[(x - \hat{x})'(x - \hat{x})]$ and the MSE matrix is MSE $= E[(x - \hat{x})(x - \hat{x})']$.

‡Note that the determinant of a covariance matrix is called a *generalized variance* in statistics.

$$\text{GEVR} = \frac{\det(\text{estimator's error covariance})}{\det(\text{actual MSE})} = \frac{\det(P)}{\det(\Sigma)} \tag{7}$$

and the **generalized error variance relative error** (**GERVE**), defined by

$$\text{GERVE} = \frac{\det(\text{actual MSE}) - \det(\text{estimator's error covariance})}{\det(\text{actual MSE})} = \frac{\det(\Sigma) - \det(P)}{\det(\Sigma)} \tag{8}$$

Note that GERVE is not equal to

$$\frac{\det(\text{actual MSE} - \text{estimator's error covariance})}{\det(\text{actual MSE})} = \frac{\det(\Sigma - P)}{\det(\Sigma)}$$

which is a meaningful measure, but not a relative error in the generalized variance in the strict sense.

Of course, measures based on matrix norms may also be proposed, but they are not recommended as general measures due to a lack of good physical interpretation of these measures. However, this does not prevent their use in a particular application. For example, the ratio $\|\Sigma - P\|_F / \|\Sigma\|_F$ may be used as a measure, where $\|A\|_F = \left(\sum_{i,j=1}^n a_{ij}^2\right)^{1/2}$ is the Frobenius norm of matrix A, which amounts to the Euclidean norm of A by treating A as a vector. This is perhaps the most thorough measure with which no difference between Σ and P may escape without being detected, but it ignores the matrix structure completely.

4.3. Noncredibility Indices

The above measures are simple and intuitively appealing, but may not serve the purpose well in many cases. Quite often it is more desirable to have more accurate, albeit more sophisticated, measures.

Note first that a difference between Σ and P is "equivalently" 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, and is thus not recommended.

An equally natural yet probably better idea is to use

$$\rho = \frac{\tilde{x}'P^{-1}\tilde{x}}{\tilde{x}'\Sigma^{-1}\tilde{x}} \tag{9}$$

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 unquestionably the most convincing measure of the credibility in the scalar case. Note that this NEES ratio has an intimate relationship with the *relative deviation* of NEES: $\frac{\tilde{x}'P^{-1}\tilde{x} - \tilde{x}'\Sigma^{-1}\tilde{x}}{\tilde{x}'\Sigma^{-1}\tilde{x}}$.

For a vector $\tilde{x}$, ρ cannot be used as a credibility measure because it is highly dependent on the random $\tilde{x}$. To remove (reduce) the uncertainty in ρ (a ratio), as elaborated in [7], 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. Usually Σ_i is not known but can be approximated by its sample value $\frac{1}{M}\sum_{i=1}^{M} \tilde{x}_i \tilde{x}_i'$. For better numerical properties, we use the logarithm and define the **noncredibility index** (**NCI**) by

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

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 $\tilde{x}$, NCI is not random and is directly proportional to $\log_{10}(\Sigma/P)$ 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 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. For better numerical properties, we use its logarithm and define an alternative **noncredibility index** (**NCI-2**) by

$$\text{NCI-2} = 10\left[\frac{1}{M}\sum_{i=1}^{M} \log_{10}\left(\frac{\epsilon_i}{E[\epsilon_*]}\right) - E\left[\log_{10}\left(\frac{\epsilon_*}{E[\epsilon_*]}\right)\right]\right] = \frac{10}{M}\sum_{i=1}^{M} \log_{10}(\epsilon_i) - 10E[\log_{10}(\epsilon_*)] \tag{11}$$

Subtraction of the term $E[\log(\epsilon_*)]$ is introduced such that NCI-2 of a perfectly credible estimator is always around 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 $E[\log_{10}(\epsilon_*)]$. Note that use of the sample average here is more accurate than the theoretical mean because the latter relies on not necessarily accurate assumptions on the distribution of ϵ_*.

Note that $P^{-1/2}\tilde{x}$ can be viewed a normalized estimation error. Let $\delta/E[\delta_*] = \left[\text{sum}(P^{-1/2}\tilde{x})\right]^2/E[\delta_*] = \left(\sum_{j=1}^{n} u_j\right)^2/E[\delta_*]$ be called one-dimensional equivalent normalized estimation error squared, where $\text{sum}(a) = a_1 + \cdots + a_n$ stands for the sum of all elements of a vector, and u_j is the jth element of $P^{-1/2}\tilde{x}$. Similar to NCI-2, we propose another measure, called **noncredibility index-3** (**NCI-3**), defined by

$$\text{NCI-3} = \frac{10}{M}\sum_{i=1}^{M} \log_{10}(\delta_i) - 10E[\log_{10}(\delta_*)] \tag{12}$$

Clearly, NCI-3 has a similar physical interpretation as NCI-2 since $\delta/E[\delta_*]$ is an alternative normalized NEES.

NCI-3 is justified as follows. Let $y_i = \text{sum}(P_i^{-1/2}\tilde{x}_i)/\sqrt{E[\delta_*]}$. Then if the estimator is perfectly credible y_i has zero mean and unity variance, and thus we have $E[\text{NCI-3}] = 0$ since δ_i's are i.i.d. and thus $E[\log_{10}(\delta_i)] = E[\log_{10}(\delta_*)], \forall i$. That is, if the estimator is perfectly credible, NCI-3 should be around zero. If the estimator is optimistic (or pessimistic), then normally $E[\log_{10}(\delta_i)] > E[\log_{10}(\delta_*)]$ (or $E[\log_{10}(\delta_i)] < E[\log_{10}(\delta_*)]$) and thus NCI-3 will be significantly larger (or smaller) than zero.

The above definitions of the 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 (see Appendix) under this Gaussian assumption that

$$10E[\log_{10}(\epsilon_*)] = \frac{10}{\ln 10}[\ln 2 + \psi(n/2)] \tag{13}$$

$$10E[\log_{10}(\delta_*)] = \frac{10}{\ln 10}[\ln(n/2) - \gamma] \tag{14}$$

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

$$\psi(m+1) = -\gamma + \sum_{k=0}^{m-1}\frac{1}{k+1}, \qquad \psi\left(m+\frac{1}{2}\right) = -\gamma + 2\left[\sum_{k=1}^{m}\frac{1}{2k-1} - \ln 2\right]$$

$\gamma = -\psi(1) = 0.57721566490$ 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:

$$\frac{10}{M}\sum_{i=1}^{M}\log_{10}(\delta_i/\delta_i^*) = \frac{10}{M}\sum_{i=1}^{M}\log_{10}(\delta_i) - \frac{10}{M}\sum_{i=1}^{M}\log_{10}(\delta_i^*)$$

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, as shown in Appendix, var[NCI-2] $\leq$ var[NCI-3], where the equality holds if and only if $n = 1$, and in fact,

$$\text{var[NCI-2]} = \begin{cases} 9.3076/M & n = 1 \\ \frac{1.8861}{M}\left[\frac{\pi^2}{6} - \sum_{k=0}^{n/2-1}\frac{1}{k^2}\right] & n = 2m \geq 2 \\ \frac{1.8861}{M}\left[\frac{\pi^2}{2} - \sum_{k=0}^{(n-1)/2-1}\frac{1}{[k+(1/2)]^2}\right] & n = 2m+1 \geq 3 \end{cases} \tag{15}$$

$$\text{var[NCI-3]} = 9.3076/M \tag{16}$$

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.

In addition, the two NCIs have the following relation:

$$\text{NCI-2} + 10E[\log_{10}(\epsilon_*)] \geq \frac{1}{n}\text{NCI-3} + 10E[\log_{10}(\delta_*)]$$

which follows from the well-known inequality $\sum_{j=1}^{n}u_j^2 \geq \frac{1}{n}\left(\sum_{j=1}^{n}u_j\right)^2$ and the monotonic property of logarithm.

Note that use of arithmetic average, rather than geometric average, of $\epsilon_i/E[\epsilon_*]$ in the NCI-3 would make it suffer from the first two drawbacks of the ANEES.

Finally, we also propose log-ANEES, called **LNEES**, defined by

$$\text{LNEES} = 10\log(\text{ANEES})$$

as a heuristic measure. By taking logarithm, it is hoped that the drawback of the ANEES that large errors are amplified can be corrected.

5. 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 (MSERE) $\gamma = (\Sigma - P)/\Sigma$, 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 (MSER) P/Σ.

5.1. Scalar Case

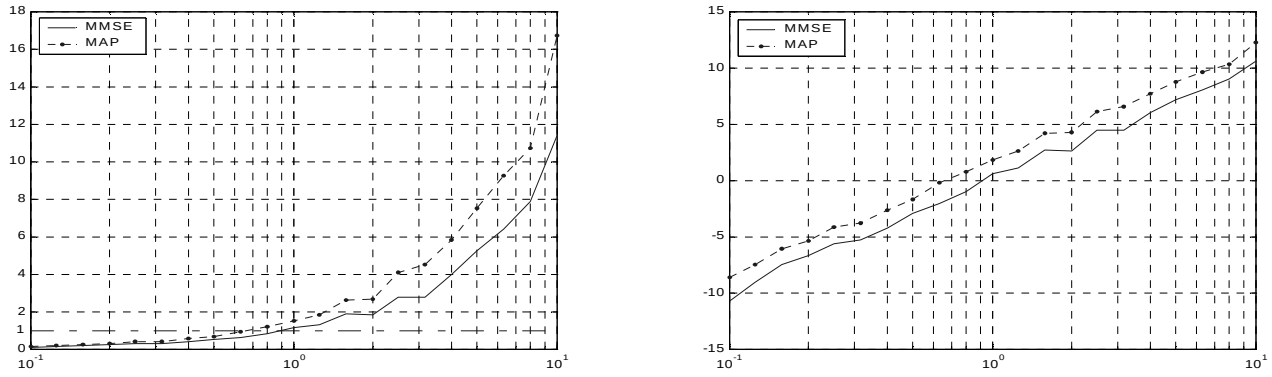
As explained in Subsection 4.3, in the scalar case, NCI is a perfect measure because it is not random at all and is directly proportional to $\log_{10}(\Sigma/P)$. However, ANEES is flawed, as illustrated below.

Consider the same example as the one considered in [7]. 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, respectively,

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

where $v \sim \mathcal{N}(0, 1)$ and $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function. Note that $\tilde{x}^{\text{MAP}}$ 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, where LNEES $= 10\log(\text{ANEES})$. The horizontal axes are obtained by varying P while holding Σ equal to its sample value $\frac{1}{M}\sum_{i=1}^{M}(x_i - \hat{x}_i)^2$ for MAP and MMSE estimators, respectively. The corresponding 95% probability interval for ANEES is $(0.8799, 1.1277)$, computed from Table 1.



(a) ANEES (b) LNEES

Fig. 1. ANEES and LNEES versus Σ/P.

It can be seen 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, $\hat{x}^{\text{MAP}}(P = \Sigma)$. 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 unquestionably 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 are not shown because NCI for the scalar case is never random at all and is always 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 $\frac{1}{M}\sum_{i=1}^{M} \tilde{x}_i^2 - E[\log_{10}(\epsilon_*)]$ 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 significant variations. The LNEES of the MAP estimator deviates significantly from the ideal one [i.e., the one that passes through the origin $(10^{0}, 0)$].

5.2. 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

$$\rho = \frac{\tilde{x}' P^{-1} \tilde{x}}{\tilde{x}' \Sigma^{-1} \tilde{x}} \quad \text{or} \quad \frac{\delta}{\delta_*} = \left[\frac{\text{sum}(P^{-1/2}\tilde{x})}{\text{sum}(\Sigma^{-1/2}\tilde{x})} \right]^2$$

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 $A = [a_{ij}]$ 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 $\mathcal{N}(0, \beta^2 a_{ij}^2)$ distribution. Then $\text{var}(\tilde{a}_{ij}) = \text{var}(b_{ij}) = \beta^2 a_{ij}^2$. 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[§]:

$$\beta A \approx B$$

Now let $A = \Sigma^{-1/2}$ and $\tilde{A} = P^{-1/2}$. It thus follows that since $B = \tilde{A} - A$,

$$\begin{aligned} \beta I &\approx \Sigma^{1/2}(P^{-1/2} - \Sigma^{-1/2}) = \Sigma^{1/2} P^{-1/2} - I \\ \beta I &\approx (P^{-1/2} - \Sigma^{-1/2})\Sigma^{1/2} = P^{-1/2}\Sigma^{1/2} - I \end{aligned} \tag{17}$$

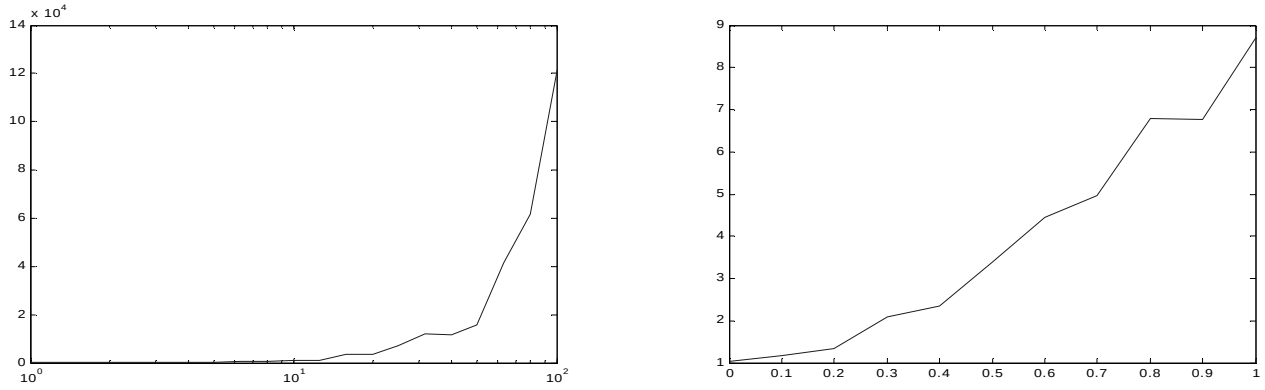
that is, β quantifies the credibility in some statistical sense. Evidently, β can also be used as a credibility measure.

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

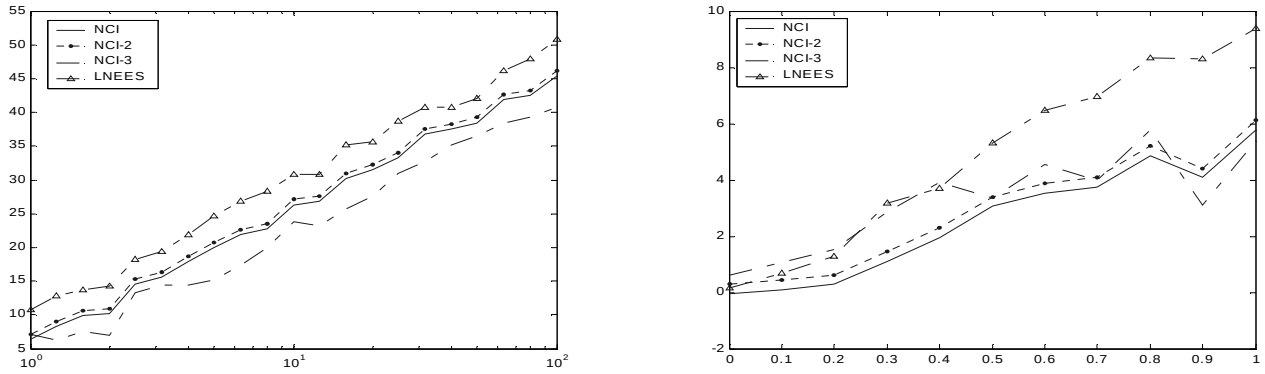
[§]This is similar to viewing a zero-mean random variable as having a length equal to its standard deviation, which has solid theoretical foundation.



(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 versus 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. LNEES exhibits more variations but NCI-3 is even less reliable. The variation of the ANEES increases significantly as β increases.

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 (17)] $\Sigma^{1/2}P^{-1/2} \approx 2I$, that is, NCI $= 10$ dB corresponds to $\Sigma P^{-1} \approx 4I$. On the other hand, NCI $= 6$ dB corresponds precisely 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 has a nice property that it is invariant with respect to the dimension of the estimator. Note that LNEES does not have such a 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.

6. 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 (MSER), have been presented, along with justifications. The pros and cons of these measures have been explained. A chi-square test based credibility test has been discussed. Some associated potential problems and confusions have been clarified. It has been shown via numerical examples that the proposed credibility measures are fairly accurate in the sense 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, used for the credibility test is not good as a measure of credibility.

APPENDIX A. PROOFS

A.1. Derivation of (13) and (14)

Assume that $q \sim \chi^2_n(\sigma^2)$. Then

$$\begin{aligned} E[\ln(q)] &= \int_0^\infty \ln(q) \frac{q^{n/2-1}e^{-q/(2\sigma^2)}}{(2\sigma^2)^{n/2}\Gamma(n/2)} dq = \frac{1}{\Gamma(n/2)} \int_0^\infty (\ln(2\sigma^2) + \ln x) x^{n/2-1} e^{-x} dx \\ &= \ln(2\sigma^2) + \frac{1}{\Gamma(n/2)} \int_0^\infty x^{n/2-1} e^{-x} \ln(x) dx = \ln(2\sigma^2) + \psi(n/2) \end{aligned}$$

where the last equation above follows from 4.352-1 of [5] or Mathematica that

$$\int_0^\infty x^{n/2-1} e^{-x} \ln(x) dx = \Gamma(n/2)\psi(n/2)$$

(13) and (14) then follows from the relationship $\log_{10} x = \ln x / \ln 10$.

A.2. Derivation of (15) and (16)

Assume that $q \sim \chi^2_n(\sigma^2)$. Then

$$\begin{aligned} \text{var}[\ln(q)] &= \int_0^\infty (\ln(q) - E[\ln(q)])^2 \frac{q^{n/2-1}e^{-q/(2\sigma^2)}}{(2\sigma^2)^{n/2}\Gamma(n/2)} dq = \frac{1}{\Gamma(n/2)} \int_0^\infty [\ln x - \psi(n/2)]^2 x^{n/2-1} e^{-x} dx \\ &= \frac{1}{\Gamma(n/2)} \int_0^\infty [(\ln x)^2 + \psi^2(n/2) - 2\psi(n/2) \ln x] x^{n/2-1} e^{-x} dx \\ &= \psi^2(n/2) - \frac{2\psi(n/2)}{\Gamma(n/2)} \int_0^\infty x^{n/2-1} e^{-x} \ln(x) dx + \frac{1}{\Gamma(n/2)} \int_0^\infty (\ln x)^2 x^{n/2-1} e^{-x} dx \\ &= -\psi^2(n/2) + [\psi^2(n/2) + \zeta(2, n/2)] = \zeta(2, n/2) = \sum_{k=0}^\infty \frac{1}{(k + n/2)^2} \end{aligned}$$

where the last equation above follows from 4.358-2 of [5] or Mathematica that

$$\int_0^\infty (\ln x)^2 x^{n/2-1} e^{-x} dx = \Gamma(n/2)[\psi^2(n/2) + \zeta(2, n/2)]$$

where $\zeta(z, y)$ is the Riemann zeta function, which is equal to

$$\zeta(z, r) = \sum_{k=0}^\infty \frac{1}{(k + r)^z}, \qquad \text{Re}(z) > 1, \ r \neq 0, -1, -2, \ldots$$

It is interesting to note that (a) var$[\ln(q)]$ is independent of σ^2 and (b) var$[\ln(q)]$ decreases as n increases. Consequently, we have, for $n = 1$,

$$\text{var}[\ln(q)] = \sum_{k=0}^{\infty} \frac{1}{(k+1/2)^2} = 4\sum_{k=0}^{\infty} \frac{1}{(2k+1)^2} = 4(\pi^2/8) = \frac{\pi^2}{2}$$

for $n = 2m \geq 2$,

$$\text{var}[\ln(q)] = \sum_{k=0}^{\infty} \frac{1}{(k+n/2)^2} = \sum_{k=0}^{\infty} \frac{1}{(k+m)^2} = \sum_{r=m}^{\infty} \frac{1}{r^2} = \sum_{r=0}^{\infty} \frac{1}{r^2} - \sum_{r=0}^{n/2-1} \frac{1}{r^2} = \frac{\pi^2}{6} - \sum_{r=0}^{n/2-1} \frac{1}{r^2}$$

and for $n = 2m + 1 \geq 3$,

$$\begin{aligned} \text{var}[\ln(q)] &= \sum_{k=0}^{\infty} \frac{1}{(k+n/2)^2} = \sum_{k=0}^{\infty} \frac{1}{(k+m+1/2)^2} = \sum_{r=0}^{\infty} \frac{1}{(r+1/2)^2} - \sum_{r=0}^{m-1} \frac{1}{(r+1/2)^2} \\ &= \frac{\pi^2}{2} - \sum_{r=0}^{(n-1)/2-1} \frac{1}{(r+1/2)^2} \end{aligned}$$

Now note that under the stated assumption, both ϵ_i's and δ_i's are i.i.d. and have the same distribution as ϵ_* and δ_*, respectively. (15) and (16) thus follows from the definitions of the two NCIs and the above:

$$\begin{aligned} \text{var}[\text{NCI-2}] &= \frac{10}{M}\text{var}[\log_{10}(\epsilon_*)] = \frac{10}{M(\ln 10)^2} \sum_{k=0}^{\infty} \frac{1}{(k+n/2)^2} \\ \text{var}[\text{NCI-3}] &= \frac{10}{M}\text{var}[\log_{10}(\delta_*)] = \frac{10}{M(\ln 10)^2} \sum_{k=0}^{\infty} \frac{1}{(k+1/2)^2} \end{aligned}$$

It is clear that var[NCI-2] $\leq$ var[NCI-3] holds with equality if and only if $n = 1$.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support of the research in part by ONR via grant N00014-00-1-0677 and NSF via Grant ECS-9734285.

REFERENCES

1. Y. Bar-Shalom and K. Birmiwal, "Consistency and Robustness of PDAF for Target Tracking in Cluttered Environments," *Automatica*, vol. 19, pp. 431–437, July 1983.
2. Y. Bar-Shalom and X. R. Li, *Estimation and Tracking: Principles, Techniques, and Software*. Boston, MA: Artech House, 1993. (Reprinted by YBS Publishing, 1998).
3. Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with Applications to Tracking and Navigation*. New York: Wiley, 2001.
4. O. E. Drummond, X. R. Li, and C. He, "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, 1998.
5. I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series and Products*. San Diego, CA: Academic Press, 2000.
6. N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*. Vol. 1, New York: Wiley, 2nd ed., 1994.
7. X. R. Li and Z. Zhao, "Practical Measures for Performance Evaluation of Estimators and Filters," in *Proc. Workshop on Estimation, Tracking, and Fusion — A Tribute to Yaakov Bar-Shalom*, (Monterey, CA), May 2001.
8. C. R. Rao, *Linear Statistical Inference and Its Applications*. New York: Wiley, 2nd ed., 1973.