Measuring Estimator’s Credibility: Noncredibility Index

X. Rong Li    Zhanlue Zhao
Department of Electrical Engineering
University of New Orleans
New Orleans, LA 70148, USA
xli@uno.edu, zhanluezhao@gmail.com

Abstract - Many estimators and filters provide assessments (e.g., MSE matrices) of their own estimation errors. They are, however, obtained based on simplifying assumptions that are not necessarily valid. Then the questions are: Are these self-assessments trustworthy? How trustworthy are they? We referred to these problems as the credibility of the estimators/filters. Solid technical answers to the first question are provided in two companion papers for this conference based on statistical hypothesis testing. Complementary to those, we answer the second question in this paper by proposing a family of metrics, called noncredibility indices (NCI) and inclination indicators (I), that measure how credible various self-assessments are. We show that the NCI and I have many desirable properties and are more appropriate than a bunch of possible alternatives and by far superior to a heuristic measure currently in use explicitly or implicitly. We also provide simple numerical examples to illustrate the application of the metrics proposed.

Keywords: Performance evaluation, estimation, credibility measure.

1 Introduction

Parameter, signal, and state estimation algorithms 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.

Many estimators provide self-assessments, such as MSE matrices, 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. However, usually these assumptions are not transparent to practitioners and are hard to verify. Even worse, they are not always reliable and may even be misleading when the underlying assumptions are not valid. An important issue in practice as well as theory is how credible these self-assessments are.

Albeit very important, work on this issue has been scarce. Although limited treatments of this topic can be found in such publications as [1, 3, 2], in our opinion, it has received attention far less than what it deserves. As a result, it is virtually impossible for a practitioner to resolve this important issue satisfactorily. In view of this, we have started to tackle this problem since a few years ago. Based on a closer examination, we elaborated in [6, 5] that the normalized estimation error squared (NEES) widely used for credibility testing is not good as a credibility measure, although it has often been so used explicitly or implicitly. Our preliminary results reported in [6, 5] include introduction of a few preliminary versions of a noncredibility index and simple, deterministic metrics based on the actual and estimator-provided MSE matrices for measuring the credibility of the latter. In this paper, we study this problem further. We propose a family of credibility metrics for bias, error covariance, and joint bias-covariance as well as MSE matrix. We elaborate on their properties and argue for their superiority to previously introduced metrics as well as some potential alternatives. In two companion papers [9, 10], we consider the closely related problem of whether the self-assessments are credible, rather than how noncredible they are.

Terminology and notation. The following convention will be maintained throughout the paper. We use the term estimator to mean both a parameter estimator and a filter. We always use \( \hat{x} \) to denote the \( n \)-dimensional estimation error and \( n \) is reserved for the dimension of the estimate. We denote the actual bias, error covariance, and mean-square error (MSE) matrix by \( \mu^* = E[\hat{x}], C^* = \text{cov}(\hat{x}), \) and \( P^* = E[\hat{x}\hat{x}'] \), respectively. By self-assessment of an estimator, we mean the bias, error covariance, and/or MSE matrix provided by the estimator, denoted by \( \mu_0, C, \) and \( P \). Subscript \( i \) stands for quantities pertaining to the \( i \)th independent run out of a total of \( N \) Monte-Carlo runs. All default vectors are column vectors. The \( A^{-1} \)-norm of a vector \( a \) is defined as \( \|a\|_{A^{-1}} = (a' A^{-1} a)^{1/2} \), where \( a' \) stands for the transpose of the column vector \( a \). The unique positive definite square-root matrix of \( A^{-1} \) is denoted by \( A^{-1/2} \).

We use \( G(y) \triangleq \exp(E[|y|]) \) to denote the geometric mean of random variable \( y \), while \( E[y] \) is the (arithmetic) mean of \( y \). We use overline to denote the sample average, for example, \( \overline{yy'} = \frac{1}{N} \sum_{i=1}^{N} y_i y_i' \) and the average normalized estimation error squared (ANEES) is \( \langle \overline{\varepsilon^2} \rangle = \frac{\overline{\varepsilon^2}}{\overline{P}^{1/2}} \). We use \( \bar{\mu}, \bar{C}, \) and \( \bar{P} \) to denote the credibility evaluator’s (not estimator’s) estimates of the bias, error covariance, and MSE matrix, respectively, and they can usually be taken for simplicity as the sample mean, sample covariance, and sample
MSE matrix, given by
\[ x = \frac{1}{N} \sum_{i=1}^{N} \tilde{x}_i, \quad \tilde{C} = (\tilde{x} - x)(\tilde{x} - x)', \quad \tilde{P} = \tilde{xx}' \]

2 Proposed Credibility Measures

We consider credibility measures for four separate cases in which the self-assessments are, respectively: (a) MSE matrix only, (b) bias only, (c) error covariance matrix only, and (d) bias and error covariance matrix. Consider the case with MSE matrix first.

2.1 Measure for MSE-Credibility

Here the key is the “difference” between the actual MSE \( P^* \) and the estimator-provided \( P \). However, \( P^* \) and \( P \) are matrices in general and cannot be compared directly—there is no generally accepted method of quantifying the difference between two matrices. Note that a difference between \( P^* \) and \( P \) is “equivalently” to that between \( (P^*)^{-1} \) and \( P^{-1} \) assuming the inverses exist. One of the simplest and most widely used ways of comparing \( (P^*)^{-1} \) and \( P^{-1} \) is to compare \( \hat{\nu}^2 P^{-1} \hat{x} \) and \( \hat{\nu}^2 (P^*)^{-1} \hat{x} \), where \( \hat{x} \) is the estimation error. This is particularly appealing in the context of measuring credibility. The most natural quantity that quantifies the difference between \( \hat{\nu}^2 P^{-1} \hat{x} \) and \( \hat{\nu}^2 (P^*)^{-1} \hat{x} \) is \( y = \hat{\nu}^2 P^{-1} \hat{x} - \hat{\nu}^2 (P^*)^{-1} \hat{x} \). Since \( y \) is random, naturally we may use its sample average \( \frac{1}{N} \sum_{i=1}^{N} y_i \) to reduce the randomness, which turns out to be equal to \( \sum_{i=1}^{N} \frac{\tilde{x}_i^n}{\tilde{P}^*_i \tilde{x}_i} \), under the assumption \( \tilde{x}_i \sim N(0, \tilde{P}^*_i) \) since \( \tilde{x}_i^n / \tilde{P}^*_i \tilde{x}_i \) is chi-square under this assumption. However, this is directly proportional to ANEES, which has serious flaws as a measure, as discussed in [6, 5].

An equally natural yet probably better idea is to use
\[ \rho^* = \frac{\tilde{x}^P P^{-1} \tilde{x}}{\tilde{x}^P (P^*)^{-1} \tilde{x}} = \frac{\epsilon}{\epsilon^*} \]

to quantify the difference between \( P^* \) and \( P \), where \( \epsilon = \tilde{x}^P P^{-1} \tilde{x} \) is the NEES of the estimator and \( \epsilon^* = \tilde{x}^P (P^*)^{-1} \tilde{x} \) is the NEES of a perfectly credible estimator. In fact, \( \rho^* \) can be called the credibility ratio. It is in general a function of the random error \( \tilde{x} \). For a vector \( \tilde{x} \), it is a ratio of the NEES—the actual NEES normalized by the ideal NEES. For a scalar \( \tilde{x} \), it is actually constant, independent of \( \tilde{x} \), and is equal to the ratio of the true mean-square error (mse) to the estimator-provided mse—this ratio is unquestionably a most convincing measure of the credibility in the scalar case. Note that it is intimately related with the relative deviation of NEES:

\[ \rho^* = \frac{\tilde{x}^P P^{-1} \tilde{x} - \tilde{x}^P (P^*)^{-1} \tilde{x}}{\tilde{x}^P (P^*)^{-1} \tilde{x}} = \rho^* - 1 \]

For a vector \( \tilde{x} \), \( \rho^* \) cannot be used as a credibility measure directly because it is highly dependent on the random \( \tilde{x} \). The drawbacks of ANEES arise from using arithmetic average of a ratio—the NEES. To reduce the randomness in \( \rho^* \) (a ratio), the geometric mean
\[ G(\rho^*) = \exp\left[ E[\ln \rho^*]\right] = 10^{E[\ln \rho^*]} \]
as a balanced measure that are dominated by neither large terms nor small terms, is much more preferable to the much more commonly used arithmetic mean, as elaborated in [7, 8]. Its finite-sample approximation is the geometric average
\[ \left[ \prod_{i=1}^{N} \rho_i \right]^{1/N} = \left[ \prod_{i=1}^{N} \frac{\tilde{x}_i^n P_i^{-1} \tilde{x}_i}{\tilde{x}^P_i (P^*_i)^{-1} \tilde{x}_i} \right]^{1/N} \]

This is clearly a relative measure and the use of its logarithm version is more appropriate than itself, especially when it might be much larger or smaller than 1. As such, we use the logarithm to compute \( \rho^* = 10 \log(\rho^*) \), which also has better numerical properties. By the same token, the signal-to-noise ratio (SNR) is defined in logarithm. The extra constant 10 is an amplification factor, as in the definition of the SNR in terms of power. We say the estimator’s self-assessment is optimistic inclined (or optimistic) for short, pessimistic inclined (or pessimistic for short), or balanced in MSE if the inclination indicator \( \nu^* > 0 \), \( \nu^* < 0 \), or \( \nu^* = 0 \), respectively, since it describes optimism, pessimism, or neither on average. However, use of \( \nu^* \) as a credibility measure alone has a serious flaw. Note that \( P^* - P > 0 \) if and only if \( \rho^* > 1 \), \( \forall \tilde{x} \) and so there is no term cancellation in \( E[\log(\rho^*)] \); likewise if \( P^* - P < 0 \). However, if \( P^* - P \) is indefinite, then \( E[\log(\rho^*)] \) would have both positive and negative terms and hence cancellations. In other words, a highly noncredible case with indefinite \( P^* - P \) may still have \( E[\log(\rho^*)] \) close to or equal to zero. This problem can be overcome by the use of the absolute value:
\[ \gamma^* = 10 E[|\log(\rho^*)|] = 10 E[|\log(\epsilon/\epsilon^*)|] \]

Usually \( P^* \) needed in the above is not known in practice. So it is replaced by an estimate \( \hat{P} \), the most common one being the sample MSE matrix: \( \hat{P} = \frac{1}{N} \sum_{i=1}^{N} \tilde{x}_i \tilde{x}_i' \). Consequently, we define the noncredibility index (NCI) by
\[ \gamma = 10 \left| \log_{10} (\rho) \right| = 10 \left| \sum_{i=1}^{N} \log_{10} (\rho_i) \right| \]

and the associated inclination indicator (I2) by
\[ \nu = 10 \log_{10} (\rho) = 10 \left| \sum_{i=1}^{N} \log_{10} (\rho_i) \right| \]

where
\[ \rho = \frac{\tilde{x}^P P^{-1} \tilde{x}}{\tilde{x}^P (P^*)^{-1} \tilde{x}} = \frac{\tilde{x}_i^n P_i^{-1} \tilde{x}_i}{\tilde{x}^P_i (P^*_i)^{-1} \tilde{x}_i} \]

For a scalar \( \tilde{x} \), the NCI and I2 are equal to
\[ \frac{10}{N} \sum_{i=1}^{N} \log_{10} (\tilde{P}_i / P_i) \] and \( \frac{10}{N} \sum_{i=1}^{N} \log_{10} (\tilde{P}_i / P_i) \) and hence nearly perfect. (Note that ANEES as a measure is flawed even in the scalar case.) For a vector \( \tilde{x} \), they are sample averages of 10 times (the absolute value of) the logarithm of the NEES ratio, in analogy to average SNR. Note that the roles of I2 and the NCI are somewhat similar to those of mean and mean-deviation of a distribution.

The above credibility ratios, noncredibility indices, and inclination indicators for the MSE will be denoted by \( \rho^*_P \), \( \rho_P \), \( \gamma_P \), \( \gamma_P \), \( \nu_P \), \( \nu_P \) in the sequel.
2.2 Credibility Measures for Other Cases

Similarly, if only the error covariance \( C \), rather than the MSE matrix, provided by the estimator is concerned, we define the corresponding credibility ratios by

\[
\rho_C = \frac{(\hat{x} - \mu)^\prime (\hat{x} - \mu)}{(\hat{x} - \bar{x})^\prime (\hat{x} - \bar{x})}
\]

where \( \bar{x} = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_i \) is the sample average of the estimation error (or some other estimate of the error mean) and \( \hat{C} \) is sample error covariance (or some other estimate of the error covariance).

Consider now the case in which only the bias \( \mu_0 \) provided by the estimator is concerned. For this case, we propose the use of

\[
\rho_\mu = \frac{(\hat{x} - \mu_0) (C^*)^{-1} (\hat{x} - \mu_0)}{(\bar{x} - \mu)^\prime (C^*)^{-1} (\hat{x} - \mu)} = \frac{||\hat{x} - \mu_0||_{(C^*)^{-1}}^2}{||\bar{x} - \mu||_{(C^*)^{-1}}^2}, \tag{3}
\]

This measure clearly makes good sense: It is the \( (C^*)^{-1} \)-weighted average deviation (squared) of \( \hat{x} \) from \( \mu_0 \) normalized by the deviation from \( \mu \). Replacing the unknown mean \( \mu^* \) and covariance \( C^* \) by their estimates \( \bar{x} \) and \( \hat{C} \), we define the corresponding credibility ratio by

\[
\rho_{\mu,C} = \frac{(\hat{x} - \mu_0) (\hat{C}^*)^{-1} (\hat{x} - \mu_0)}{(\bar{x} - \mu)^\prime (\hat{C}^*)^{-1} (\hat{x} - \mu)} = \frac{||\hat{x} - \mu_0||_{(\hat{C}^*)^{-1}}^2}{||\bar{x} - \mu||_{(\hat{C}^*)^{-1}}^2}, \tag{4}
\]

It is the basis of a measure of the noncredibility of the bias \( \mu_0 \) and the error covariance \( C \) jointly.

Note that the definitions of \( \rho_{\mu,C} \), \( \rho_C \), and \( \rho_\mu \) are consistent: If \( \mu_0 = \mu^* \) then \( \rho_{\mu,C} = \rho_C, \forall \bar{x} \); if \( \hat{C} = C^* \) then \( \rho_{\mu,C} = \rho_\mu, \forall \bar{x} \). Also, they are all normalized by the same quantity: \( (\bar{x} - \mu)^\prime (C^*)^{-1} (\bar{x} - \mu) \). Note, however, that \( \rho_{\mu,C} \), \( \rho_C \), and \( \rho_\mu \) are the same only when \( \mu_0 = \bar{x} \) and \( C = \hat{C} \), respectively, which are virtually impossible due to the randomness of \( \bar{x} \) and \( \hat{C} \). Consequently, it is better to use \( \rho_{\mu,C} \) only when both bias and error covariance (or MSE) matrix are provided by the estimator. When only one of them is provided, use \( \rho_\mu \) or \( \rho_C \) directly.

The NCIs \( \gamma \) and the inclination indicators \( \nu \) as well as their theoretical versions \( \gamma^* \) and \( \nu^* \) corresponding to the above credibility ratios \( \rho \) and \( \rho^* \) are defined by

\[
\gamma_\rho = 10E[|\log \rho_\rho|], \quad \gamma_\rho = 10|\log_{10} (\rho_\rho)|
\]

\[
\nu_\rho = 10E[|\log \rho_\rho|], \quad \nu_\rho = 10|\log_{10} (\rho_\rho)|
\]

where \( \rho = \mu, C, P \), or \( (\mu, C) \).

One of the two most common practical situations is that the MSE matrix \( P \) or the error covariance \( C \) is provided by the estimator under the assumption of zero bias \( \mu_0 = 0 \). In this case, \( \gamma_{\mu,C} \) along with \( \nu_{\mu,C} \) should be used with \( \mu_0 = 0 \) to measure the credibility and inclination of the claim of \( \mu = 0 \) and \( P \) or \( C \) jointly; \( \gamma_\mu, \gamma_C, \) and \( \gamma_P \), along with their inclination indicators, should be used to measure the credibility and inclinations of the claim of a bias \( \mu_0 = 0 \), error covariance \( C \), and MSE matrix \( P \), respectively. In the other most common case where no error covariance or MSE matrix is provided by the estimator but zero bias \( \mu_0 = 0 \) is assumed, only \( \gamma_\mu \) and \( \nu_\mu \) are used to measure the credibility and inclination of the claim of a zero bias.

All the credibility ratios above can be interpreted physically as the ratio of the actual normalized error power to the estimator-computed one. They can be called actual-to-computed error ratio (ACER) and thus the NCIs and inclination indicators are ACER in dB, in analogy to SNR.

2.3 Properties of Proposed Measures

The following concepts are important for evaluation of estimators and for discussing the properties of the NCI.

Definition. As defined in \([7,8]\), an average error measure or metric is said to be pessimistic (or optimistic) if it is dominated by the large (or small) estimation errors, or equivalently, it pays more attention to large (or small) errors in the sense that it consistently over (or under) estimates the actual error. It is balanced if it is neither optimistic nor pessimistic. As explained in \([7,8]\), the root-mean-square error and the average Euclidean error are pessimistic since they are dominated by the large terms; the harmonic average error is optimistic since it is dominated by the small terms; the geometric average error, median error, and error mode are balanced. The same concepts apply to credibility measures. A credibility measure or metric is said to be pessimistic (or optimistic) if it is dominated by the large (or small) estimation errors, or balanced if it is neither optimistic nor pessimistic.

Similar concepts apply to estimator’s self-assessments. An estimator is said to be pessimistic (or optimistic) in MSE if its computed MSE matrix is larger (or smaller) in some sense than the actual MSE matrix. Further, let

\[
\Delta_P = P^* - P, \quad \Delta_C = C^* - C
\]

where \( P^* \) and \( P \) are the actual and estimator-provided MSE matrices. Likewise for \( C^* \) and \( C \). We call \( \Delta \) the noncredibility matrix.

Definition. An estimator is said to be perfectly credible in MSE (or error covariance) matrix if \( \Delta_P = 0 \) (or \( \Delta_C = 0 \)); optimistic definite (or semidefinite) in MSE (or error covariance) matrix if \( \Delta_P \) (or \( \Delta_C \)) is positive definite (or semidefinite); pessimistic definite (or semidefinite) in
MSE (or error covariance) matrix if \( \Delta_P \) (or \( \Delta_C \)) is negative definite (or semidefinite).

The proposed credibility measures enjoy the following nice properties:

- NCIs and inclination indicators are balanced measures: This follows from that all credibility ratios take the form with a clear geometric interpretation

\[
\rho = \frac{||\hat{x} - a||_{P^{-1}}}{||\hat{x} - b||_{P^{-1}}}
\]

which is not dominated by large or small estimation errors. Inclination indicators are more balanced than NCIs because the former are the logarithms of the geometric average—a balanced average measure—of credibility ratios. NCIs are balanced w.r.t. estimation errors but dominated by both large and small credibility ratios \( \rho \), meaning that they are balanced w.r.t. \( \hat{x} \) yet sensitive to noncredibility—a very desirable property.

- The inclination indicator for MSE manifests optimistic/pessimistic definiteness well:
  
  - The estimator is optimistic definite in MSE if and only if the inclination indicator for MSE is always positive: \( \nu_p > 0, \forall \hat{x} \). In this case, \( \nu_p = \gamma_p, \forall \hat{x} \). It follows from
    \[
    \Delta_p > 0 \iff \tilde{x}'P^{-1}\tilde{x} > (P^*)^{-1}\tilde{x}, \forall \tilde{x}
    \]
    \[
    \iff \log \tilde{x}'P^{-1}\tilde{x} > \log (P^*)^{-1}\tilde{x}, \forall \tilde{x}
    \]
    \[
    \iff \nu_p > 0, \forall \tilde{x}
    \]
  
  - The estimator is pessimistic definite in MSE if and only if the inclination indicator for MSE is always negative: \( \nu_p < 0, \forall \hat{x} \). In this case, \( \nu_p = -\gamma_p, \forall \hat{x} \).
  
  - The estimator is perfectly credible in MSE if and only if the inclination indicator (or equivalently, NCI) for MSE is always zero: \( \nu_p = \gamma_p = 0, \forall \hat{x} \).

Parallel conclusions hold true for semidefiniteness and for error covariance.

- Inclination indicators are symmetric measures—optimism and pessimism are treated equally and they differ only by the sign of the inclination indicators:

  \[
  \begin{align*}
  \text{optimistic inclined} & \iff G(\rho_p^*) > 1 \iff \nu_p^* > 0 \\
  \text{pessimistic inclined} & \iff G(\rho_p^*) < 1 \iff \nu_p^* < 0 \\
  \text{balanced} & \iff G(\rho_p^*) = 1 \iff \nu_p^* = 0
  \end{align*}
  \]

where \( G(y) \) is the geometric mean of \( y \). This follows from the fact that inclination indicators are symmetric relative to (normalized by) the perfectly credible: \( \nu_p^* = 10E[\log(\epsilon)] - 10E[\log(\epsilon^*)] \). It is also strongly justified by the property immediately above. If the estimator’s self-assessment fails the credibility test, it is deemed optimistic/pessimistic (inclined) when the corresponding inclination indicator is positive/negative. This also provides a connection between the NCIs and the noncredibility levels of [9, 10].

- NCIs and inclination indicators are relative measures (with the ideal values 0). For example,

\[
\frac{\tilde{x}'P^{-1}\tilde{x} - \tilde{x}'(P^*)^{-1}\tilde{x}}{\tilde{x}'(P^*)^{-1}\tilde{x}} = \rho_p - 1
\]

- Credibility ratios, NCIs, and inclination indicators are dimension-normalized in that they are invariant w.r.t. dimension \( n \) of \( \tilde{x} \) (i.e., they are one-dimensional equivalent quantities). It can be shown that

\[
\begin{align*}
E[\rho_p^\ast] & \approx \frac{\text{tr}[P^*P^{-1}]}{n} \quad E[\rho^\ast] \approx \frac{\text{tr}[C^*C^{-1}]}{n} \\
E[\rho^{\ast}_{\mu,C}] & \approx 1 + \frac{||\mu^* - \mu_0||_{(C^*)^{-1}}^2}{n}
\end{align*}
\]

2.4 Superiority to Alternatives

In this section, we compare the proposed NCI with potential alternatives and explain the superiority of the NCI. The same conclusions also hold true for inclination indicators, but are omitted.

NCI vs. ANEES. The ANEES does not possess any of the NCI’s nice properties discussed above, except the one of dimension normalization. We are not aware of a single nice property of the ANEES that the NCI does not have. More importantly, the NCI is free of the serious drawbacks of the ANEES discussed in [6, 5].

NCI vs. LNEES. LNEES was introduced in [6] to alleviate the serious drawbacks of the ANEES. By taking the logarithm, it is hoped that the large-error dominance drawback of the ANEES can be corrected. But it is not a cure; neither does it have many nice properties of the NCI, such as being symmetric, balanced, and relative. The NCI is free of those drawbacks and can be thought of as a cure. This should be clear from the definitions:

\[
\begin{align*}
\text{NCI} & = 10 \log(\tilde{x}'P^{-1}\tilde{x}) \quad \text{LNEES} = 10 \log(\tilde{x}'P^{-1}\tilde{x})/n
\end{align*}
\]

Here logarithm of average is worse than average of logarithm and the NCI is normalized better than the LNEES.

NCI vs. simple deterministic measures. A group of deterministic measures of MSE-credibility was introduced in [6] based on the matrix norm, determinant, and trace of the MSE matrices \( P^* \) and \( P \), including

\[
\begin{align*}
\text{log-MSER} & = \frac{\text{tr}(P^*)}{\text{tr}(P)}, \quad \text{MNRE} = \frac{||P^* - P||}{||P||} \\
\text{log-GEVR} & = \frac{\text{det}(P^*)}{\text{det}(P)}, \quad \text{log-MNR} = \frac{||P^*||}{||P||}
\end{align*}
\]

where \( ||A|| \) stands for a matrix norm of \( A \). Similar measures can be introduced for error covariance matrices. They are all free of the drawbacks of the ANEES and enjoy some nice properties. Although some of them even possess a unity-range property that NCIs do not have, for example, \( 0 \leq \text{MNRE} \leq 1 \), overall NCIs have more and better desirable properties, discussed above, and the following two additional advantages. First, NCIs apply to bias-credibility
and joint bias-MSE credibility, but our attempts to extend those simple measures to these cases have not been successful. Second and probably more importantly, different estimation errors \( \hat{x} \) may excite different “modes” of the credibility ratio \( \rho \) or NCI \( \gamma \). As a result, the average \( \rho \) (and hence the NCI) carry useful information about the credibility of an estimator that is dependent on the distribution of the estimation error (beyond the second moments), whereas those simple measures are deterministic and dependent only on the second moments.

It follows from Theorem 1 below that the credibility ratio \( \rho^* \), and the NCI \( \gamma^* \) are related to the largest and smallest eigenvalues, \( \lambda_{\text{max}} \) and \( \lambda_{\text{min}} \), of the normalized matrix \( P_y = P^{-1/2} \rho \) by

\[
0 \leq \lambda_{\text{min}} \leq \rho^* \leq \lambda_{\text{max}}
\]

\[
0 \leq \lambda_{\text{min}} \leq \lambda_{\text{max}}
\]

\[
10 \log \lambda_{\text{min}} \leq \nu^* \leq 10 \log \lambda_{\text{max}}
\]

\[
10 \log \lambda_{\text{min}} \leq \gamma^* \leq 10 \log \lambda_{\text{max}}
\]

where \( \xi_{\text{min}} = \min\{\log \lambda_{\text{min}}, \log \lambda_{\text{max}}\} \) and \( \xi_{\text{max}} = \max\{\log \lambda_{\text{min}}, \log \lambda_{\text{max}}\} \). Note that \( P_y^* \) is the key matrix for MSE-credibility testing [9], where the test statistics depend only on the eigenvalues of \( P_y^* \) and in fact only on \( \lambda_{\text{max}} \) and \( \lambda_{\text{min}} \) for the union-intersection test.

The same relationships hold between \( \rho_C^*, u_C^* \), and the normalized matrix \( \hat{P}_y = C^{-1/2} \rho \) and the following alternative version of the NCI was introduced and justified in [6] for MSE-credibility

\[
\gamma^*_C = \frac{10}{N} \sum_{i=1}^{N} \log_{10}(\delta_i) - 10 \log(\delta_x) \quad (5)
\]

Clearly, it has a similar physical interpretation as the NCI since \( \delta/E[\delta^*] \) is an alternative normalized NEES.

The NCI and \( \gamma_C \) have many similar nice properties and are both free of the drawbacks of the ANEES. But the NCI is significantly more accurate than \( \gamma_C \). If the Gaussian assumption \( \tilde{x} \sim \mathcal{N}(0, P) \) is valid, then \( \nu^* = 0 \), \( \epsilon \sim \chi^2_1 \) and \( \delta \sim \chi^2_1(n) \). In this case \( \nu^* \) is more reliable (i.e., less uncertain) than \( \gamma^*_C \) because, as shown in [5], \( \text{var}(\nu^*) \leq \text{var}(\gamma^*_C) \), where the equality holds if and only if \( n = 1 \), and in fact,

\[
\text{var}(\nu^*) = \begin{cases} 
3.141/N, & n = 1 \\
3.141 \sum_{k=0}^{n/2-1} \frac{1}{k+1/2} - \sum_{k=0}^{n/2-1} \frac{1}{k+1/2}, & n = 2m + 2 \\
3.141 \sum_{k=0}^{(n-1)/2} \frac{1}{k+1/2} - \sum_{k=0}^{(n-1)/2} \frac{1}{k+1/2}, & n = 2m + 1 \geq 3 
\end{cases}
\]

\[
\text{var}(\gamma^*_C) = 3.141/N
\]

That is, for a perfectly credible estimator, \( \nu^* \) is around zero with a standard deviation upper bounded by (approximately) \( 3/\sqrt{N} \), which is the standard deviation of \( \gamma^*_C \).

**NCI vs. ratio of averages.** The NCI is based on geometric average of the NEES ratio. The following ratio of arithmetic averages of NEES

\[
r_P = \frac{x_P P^{-1/2}}{x_P P^{-1/2} + \sum_{i=1}^{N} \frac{x_i P^{-1/2} x_i}{x_i}} = \frac{1}{n} \sum_{i=1}^{N} \frac{x_i P^{-1/2} x_i}{x_i}
\]

is another potential alternative, where \( \tilde{P}_i \) is the sample MSE matrix (or some other estimate of the MSE matrix) used on run \( i \). However, \( r_P \) is dominated by the large errors \( \tilde{x} \) and thus not a balanced measure since both the numerator and the denominator are dominated by the large errors because they are both arithmetic averages. In addition, as the following theorem shows, it has a larger range than that of the NCI, which suggests that it is more random than the NCI. In other words, geometric average of ratios of quadratic forms is more concentrated than ratio of arithmetic averages of quadratic forms.

**Theorem 1.** Assume that \( P_i \) and \( \tilde{P}_i \) are nonsingular and let \( \lambda_{\text{max}}^* \) and \( \lambda_{\text{min}}^* \) be the largest and smallest eigenvalues of \( \tilde{P}_{y,i} = \tilde{P}_i^{1/2} \tilde{P}_i^{-1/2} \). Then \( \nu_i, \gamma^*_P, r_P \) have the ranges

\[
\rho_i \in [\lambda_{\text{min}}^*, \lambda_{\text{max}}^*]
\]

\[
u_P \in \left[ \frac{10}{N} \sum_{i=1}^{N} \log \lambda_{\text{min}}^*, \frac{10}{N} \sum_{i=1}^{N} \log \lambda_{\text{max}}^* \right]
\]

\[
\gamma_P \in \left[ \frac{10}{N} \sum_{i=1}^{N} \log \lambda_{\text{min}}^*, \frac{10}{N} \sum_{i=1}^{N} \log \lambda_{\text{max}}^* \right]
\]

\[
r_P \in \left[ \min_{i \leq i \leq N} \lambda_{\text{min}}^* \lambda_{\text{min}}^*, \max_{i \leq i \leq N} \lambda_{\text{max}}^* \right]
\]

where \( \lambda_{\text{min}}^* = \min\{\log \lambda_{\text{min}}^*, \log \lambda_{\text{max}}^*\} \) and \( \lambda_{\text{max}}^* = \max\{\log \lambda_{\text{min}}^*, \log \lambda_{\text{max}}^*\} \). Note that the estimator is credible when and only when \( \gamma_P \) is between the smallest of the smallest eigenvalues and the largest of the largest eigenvalues. In other words, \( \gamma_P \) has a range much smaller than that of \( r_P \).

### 3 Illustrative Examples

Two examples have been given in [6, 5] that demonstrate the superiority of a preliminary version of the NCI (in fact, inclination indicator) to the ANEES and LNEES as credibility measures. Our main goal in this section is to gain working knowledge as how large the value of the NCI is really large; in other words, we want to learn what the critical values of the NCI are which indicate marginal credibility of the estimator’s self-assessments. For this purpose, we adopt the examples in the two companion papers [9, 10] so that we can use credibility testing results therein to learn the critical values of the NCI. Our description of the examples is sketchy. The reader is referred to [9, 10] for more details.

For convenience and without loss of generality, we always assume that the estimator provided MSE matrix \( P = I \), error covariance \( C = I \), and bias \( \mu = 0 \). All results are based on multiple Monte Carlo runs and each run uses data of size \( N = 1000 \).

**Example 1.** We consider two cases of the truth \( \tilde{x} \sim \mathcal{N}(\mu^*, I) \) with (a) \( \mu^* = \eta [1, 1]^T \) and (b) \( \mu^* = \eta [1/2, 2]^T \), where \( 0 \leq \eta \leq 0.1 \). Fig. 1 shows the average NCI and inclination indicator (I2) values (over 100 runs) vs. \( \eta \) for both cases. Note that the estimator is credible when and only when \( \eta = 0 \). Clearly case (b) is less credible than case (a) for the same \( \eta \). Interestingly, I2 has values that are a small portion of the NCI because it is hard to say the estimator is...
optimistic or pessimistic (indeed the concepts of optimism and pessimism do not work well for this example).

**Example 2.** We consider the truth \( \tilde{x} \sim N(0, P^*) \) with (a) \( P^* = \sigma^2 I \) and (b) \( P^* = \sigma^2 \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} \), where \( 0 \leq \sigma^2 \leq 2 \). Fig. 2 shows the average NCI and \( I^2 \) values (over 20 runs) vs. \( \sigma^2 \) for both cases. Note that \( P = P^* \) is true when and only when \( \sigma^2 = 1 \) in case (a) and never true in case (b). So, as \( \sigma^2 \) increases from 0 to 2 in case (a), the NCI should decrease first towards zero, reach zero at \( \sigma^2 = 1 \), and then increase; while \( I^2 \) should increase from negative to positive and pass through zero at \( \sigma^2 = 1 \). In case (b), the NCI should never be close to zero. These predictions are indeed confirmed in Fig. 2. It is interesting to note that the NCI reach the minimum and \( I^2 \) is close to zero at \( \sigma^2 \approx 0.5 \) in case (b).

For these two examples, NCI \( \approx 0.5 \) corresponds to a noncredibility probability of above 0.9 by the tests of [9, 10].

**Example 3.** We consider the truth \( \tilde{x} \sim N(\mu^*, C^*) \) with (a) \( \mu^* = \eta \begin{bmatrix} 1 \\ 1 \end{bmatrix} \), \( C^* = \sigma^2 I \) and (b) \( \mu^* = \eta \begin{bmatrix} 2 \\ 2 \end{bmatrix} \), \( C^* = \sigma^2 \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} \), where \(-1 \leq \eta \leq 1\), \( 0 \leq \sigma^2 \leq 2 \). Note that the estimator is bias-cov credible when and only when \((\eta, \sigma) = (0, 1)\) in case (a) and never credible in case (b). Fig. 3 shows the average NCI values as well as constant contours of the NCI and \( I^2 \) (all over 20 runs) vs. \( \eta \) and \( \sigma^2 \) for case (a). They make good sense. The vast difference between the NCI and \( I^2 \) contours indicates that the case is far from optimistic definite or pessimistic definite, although it is optimistic inclined for \( \eta > 0 \) and mostly pessimistically inclined for \( \eta < 0 \). Results for case (b) have a similar pattern. The points of \((\eta, \sigma^2)\) on a contour of the NCI are equally noncredible in terms of the NCI.

**Example 4.** The above examples use Gaussian distributed data. Here we consider non-Gaussian data \( \tilde{x} \) having a Gaussian mixture distribution, given by

\[ f(\tilde{x}) = pN(\tilde{x}; \mu_1, \Sigma_1) + (1-p)N(\tilde{x}; \mu_2, \Sigma_2), \quad 0 \leq p \leq 1 \]

which has mean \( \mu^* = p\mu_1 + (1-p)\mu_2 \) and covariance \( C^* = p(\Sigma_1 + \mu_1\mu_1') + (1-p)(\Sigma_2 + \mu_2\mu_2') \). We consider two cases: (a) \( \mu_1 = [1, 1]' \), \( \mu_2 = 0 \) and \( \Sigma_1 = \Sigma_2 = \sigma^2 I \) and (b) \( \mu_1 = -\mu_2 = [1, 1]' \) and \( \Sigma_1 = \Sigma_2 = \sigma^2 \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix} \), where \( 0 \leq \sigma^2 \leq 1 \). Fig. 4 shows the average NCI values as well as constant contours of the NCI and \( I^2 \) (all over 100 runs) vs. \( p \) and \( \sigma^2 \) for case (a). Note that the distribution becomes more and more non-Gaussian as \( p \) increases from 0 to 0.5. The estimator is bias-cov credible in case (a) when and only when \((p, \sigma^2)\) is close to \((0, 1)\) and never credible in case (b). The closeness between the NCI and \( I^2 \) contour levels (in magnitude) indicates that the case is largely pessimistic “definite” (i.e., consistently pessimistic). They make good sense and are consistent with the testing results of [9]. Interestingly and not surprisingly, the highly undesirable valleys of Figs. 1(b) and 2(b) of [9] in the chi-square significance test corresponds roughly to the cases where \( I^2 = 0 \).

**Example 5.** Consider now a simple filtering example, that is, tracking a target of a nearly constant-velocity one-dimension motion using position-only measurements under the linear-Gaussian assumption of the Kalman filter. The system is given by

\[
x_{k+1} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} x_k + \begin{bmatrix} T^2/2 \\ T \end{bmatrix} w_k, \quad x_k = [x_k, \dot{x}_k]'\\
\dot{x}_k = [1, 0]x_k + v_k\\
\bar{w}_k = 0, \quad \text{cov}(w_k) = \sigma^2, \quad \bar{v}_k = 0, \quad \text{cov}(v_k) = 20^2
\]
where $T = 1$ and the true value of $\sigma$ is always 3. Consider three cases where the Kalman filter assumes $\sigma = 2, 3, 4$, respectively. The filter with $\sigma = 3$ is matched exactly, but with $\sigma = 2$ or $4$ it is mismatched—it is optimistic for $\sigma = 2$ and pessimistic for $\sigma = 4$. The filter was initialized by a weighted least-squares fit to the first two measurements, known as the two-point differencing in the target tracking community.

Fig. 5 shows the NCI and $I^2$ computed from 100 Monte-Carlo runs. We should not pay much attention on the first 10 or so time steps because of the transient due to non-ideal initialization. It can be seen that the matched/credible filter’s NCI is most of the time below 1, while that of the mismatched filter with $\sigma = 4$ fluctuates around 2 and the one with $\sigma = 2$ around 1. Note that in terms of NCI, the filter with $\sigma = 2$ is about twice noncredible than the one with $\sigma = 4$. Also, the filter with $\sigma = 2$ is indeed deemed optimistic since its $I^2$ is always positive and approximately equal to its NCI, while the one with $\sigma = 4$ is indeed deemed pessimistic since its $I^2$ is always negative and approximately equal to its NCI in magnitude. These results are consistent with those of credibility tests, reported in [9]. For example, as discussed in [9], the difference between the cases with $\sigma = 2$ and with $\sigma = 4$ makes good sense: It is better to be more conservative in choosing the process noise covariance (see, e.g., [4]). Also, the ratio of $\sigma = 3$ to $\sigma = 2$ is larger than that of $\sigma = 4$ to $\sigma = 3$ (the true ratio that matters most is those between the MSE matrices, not process noise covariances).

4 Summary

We have proposed a family of credibility metrics, called noncredibility indices and inclination indicators, to measure the credibility of an estimator, that is, how much an estimator’s self-assessments of the estimation errors can be trusted. These metrics are applicable to not only MSE-credibility, but also bias-credibility and joint bias-covariance credibility. They are shown to have many desirable properties that the heuristic measure currently in use lacks. Arguments have been provided to justify their adoption and convincing evidence has been shown for their superiority to several previously proposed or currently used alternatives. Simple numerical examples have been given to demonstrate their usage. Knowledge of their concrete values have been gained by comparing with the credibility testing results reported in the companion papers [9, 10]. We conclude that an estimator is not credible if the NCI is significantly larger than 1 and is consistently optimistic (or pessimistic) if the inclination indicator is positive (or negative) and close to the NCI in magnitude.

The results of this paper can be applied to filters in a straightforward way, e.g., by computing the NCI and inclination indicator as a function of time as well as to the credibility check of the initial estimate and covariance.

A Proof of Theorem 1

From the Rayleigh-Ritz theorem, $\lambda_{\min}(A) \leq a^t A a / (a^t a) \leq \lambda_{\max}(A), \forall a$, and rewriting $\rho$, as

$$\rho_i = \frac{\bar{x}^t P^{-1} \bar{x}}{\bar{x}^t \bar{x}} = \frac{u^t \hat{P}^{1/2} P^{-1} \hat{P}^{1/2} u}{u^t u}$$

it follows that

$$\lambda_{\min}(\hat{P}^{1/2} P^{-1} \hat{P}^{1/2}) \leq \rho_i \leq \lambda_{\max}(\hat{P}^{1/2} P^{-1} \hat{P}^{1/2})$$

By definition, the eigenvalues $\lambda$ satisfy $\hat{P}^{1/2} P^{-1} \hat{P}^{1/2} x = \lambda x$, or equivalently, $P^{-1/2} \hat{P} \hat{P}^{-1/2} x = \lambda P^{-1/2} x$, that is, $P^{-1/2} \hat{P} y = \lambda y$. This is in turn equivalent to $P^{-1/2} \hat{P} P^{-1/2} P^{-1/2} y = \lambda P^{1/2} y$, that is, $P^{-1/2} \hat{P} P^{-1/2} z = \lambda z$. This shows that $P^{-1/2} \hat{P} P^{-1/2}$ and $\hat{P}^{1/2} P^{-1} \hat{P}^{1/2}$ have the same set of eigenvalues:

$$\lambda(\hat{P}^{1/2} P^{-1} \hat{P}^{1/2}) = \lambda(P^{-1/2} \hat{P} P^{-1/2})$$

Note that $\hat{P}$ is positive definite; so is $\hat{P}_{y,i} = P^{-1/2} \hat{P} P^{-1/2}$ since $P^{-1/2}$ is nonsingular. It thus follows that

$$0 < \lambda_{\min}(\hat{P}_{y,i}) \leq \rho_i \leq \lambda_{\max}(\hat{P}_{y,i})$$

With $\gamma_P = \frac{10}{N} \sum_{i=1}^N |\log \rho_i|$ this leads to (6). Rewrite $r_P$ as $X^t P^{-1} X / (X^t \hat{P}^{-1} X)$, where $X = [\bar{x}_1, \ldots, \bar{x}_N]$ and

$$P = \text{diag}(P_1, \ldots, P_N), \quad \hat{P} = \text{diag}(\hat{P}_1, \ldots, \hat{P}_N)$$

Following the same steps as above, we have

$$\lambda_{\min}(\hat{P}_y) \leq r_P \leq \lambda_{\max}(\hat{P}_y)$$
where

$$\hat{P}_y = \hat{P}^{-1/2} \hat{P} \hat{P}^{-1/2} = \text{diag}(\hat{P}_{y,1}, \ldots, \hat{P}_{y,N})$$

These eigenvalues are the solutions of $|\hat{P}_y - \lambda I| = 0$ and

$$|\hat{P}_y - \lambda I| = |\text{diag}(\hat{P}_{y,1}, \ldots, \hat{P}_{y,N}) - \lambda I|$$

$$= |\hat{P}_{y,1} - \lambda I| \cdots |\hat{P}_{y,N} - \lambda I|$$

Thus, we have

$$\lambda_{\text{min}}(\hat{P}_y) = \min(\lambda_{\text{min}}(\hat{P}_{y,1}), \ldots, \lambda_{\text{min}}(\hat{P}_{y,N}))$$

$$= \min_{1 \leq i \leq N} \lambda_{\text{min}}(\hat{P}_{y,i})$$

$$\lambda_{\text{max}}(\hat{P}_y) = \max(\lambda_{\text{max}}(\hat{P}_{y,1}), \ldots, \lambda_{\text{max}}(\hat{P}_{y,N}))$$

$$= \max_{1 \leq i \leq N} \lambda_{\text{max}}(\hat{P}_{y,i})$$

that is,

$$\min_{1 \leq i \leq N} \lambda_{\text{min}}(\hat{P}_{y,i}) \leq r_P \leq \max_{1 \leq i \leq N} \lambda_{\text{max}}(\hat{P}_{y,i})$$

References


