Relative Error Measures for Evaluation of Estimation Algorithms[*]

X. Rong Li
Department of Electrical Engineering
University of New Orleans
New Orleans, LA, USA
xli@uno.edu

Zhanlue Zhao
Department of Electrical Engineering
University of New Orleans
New Orleans, LA, USA
zzhao1@uno.edu

Abstract – *This paper is part of a series of publications that deal with evaluation of estimation algorithms. This series introduces and justifies a variety of metrics useful for evaluating various aspects of the performance of an estimation algorithm, among other things. This paper focuses on relative error measures, i.e., those with respect to some references, including the magnitude of the quantity to be estimated, its prior mean, and/or measurement error. It proposes several relative metrics that are particularly good for measuring different aspects of estimation performance. They often reveal the inherent error characteristics of an estimator better than widely used metrics of the absolute error. The metrics are illustrated via an example of target localization with radar measurements.*

Keywords: Performance measure, estimation, filtering.

1 Introduction

Parameter, signal, and state estimation algorithms are widely used in science and engineering. No matter how solid such an algorithm 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 others.

In spite of its importance in practice, theoretical work on performance measures for evaluation of estimation algorithms has been very limited, although substantial work has been done for target tracking and data fusion algorithms (see, e.g., [1, 5, 4, 9, 10, 6, 3]). This paper addresses only the *measures of performance for evaluating estimation algorithms*, not target tracking or data fusion algorithms—for example, no uncertainty in data origin is assumed. In our opinion, the available metrics are not sufficient to meet various needs in many applications.

This paper is part of a series of publications started with [7]. It focuses on performance measures of *relative error* in that they are normalized estimation errors with respect to some references. It proposes a variety of new performance measures. It also intends to stimulate further investigation on this important topic. "Absolute" measures, which are not relative to a reference, have been studied in [7].

2 Relative Error Measures

The following convention will be maintained throughout the paper. We refer to any quantity to be estimated as an ***estimatee***. It can be a time-invariant (or slowly varying) parameter, a (deterministic or random) process or signal, in particular, the state of a (deterministic or random) system. We will use the term ***estimator*** to mean both a parameter estimator and a filter. The measures are presented in a form suitable for parameter estimators directly, but are applicable to filters at any given time in a straightforward way. The estimatee, its estimate, and estimation error are denoted by x, $\hat{x}$, and $\tilde{x}$, respectively. They are assumed to be column vectors. In practice, performance evaluation is often done by the Monte-Carlo simulation, at least before the field tests or experiments are conducted. 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 $\tilde{x}_i$ and $\tilde{x}_j$ are independent for $i \neq j$. All default vectors are column vectors. The Euclidean norm of a vector a is denoted as

$$\|a\| = (a'a)^{1/2}$$

where a' stands for the transpose of the column vector a.

A relative error is one that is relative to some reference. There are many choices for the reference. Relative error often reveals the inherent error characteristics of an estimator better than the absolute error. For example, it is usually reasonable to expect that relative error of an estimator is less variant than the absolute error as the magnitude of the estimatee varies. Given two estimators and their performance for two different problems, *respectively*, it would be misleading to use any absolute error measure for their performance comparison, but relative error measures can be used. For such reasons, we recommend evaluating the performance of estimators in terms of relative error in most cases, although the literature is full of performance evaluation in terms of absolute error.

Clearly, estimation error $\tilde{x}$ depends on the magnitude of the estimatee x and the accuracy of the data z (as well as the prior distribution of x for a Bayesian estimator) as the input to the estimator $\hat{x}(z)$. As a result, probably the most natural relative error is $\tilde{x}/\|x\|$ or $\|\tilde{x}\|/\|x\|$. Another good choice of the reference is the measurement error; that is, the

[*]Research supported in part by ARO grant W911NF-04-1-0274 and NASA/LEQSF grant (2001-4)-01.

estimation error $\tilde{x}$ in terms of the measurement error. We deal with measures for such relative errors in this paper. A third class of relative errors is the estimation error of one estimator relative to that of another estimator. This will be treated in a subsequent part of the series.

3 Relative Error of Estimation

3.1 Estimation Error Relative to Estimatee

Measures of the absolute error $\tilde{x}$ are often not desirable. For instance, an absolute error of $\tilde{x} = 1$ is only 1% for an estimatee of $x = 100$ but 50% for an estimatee of $x = 2$. The relative error $\tilde{x}/\|x\|$ versions of the measures discussed in [7] are simply given by their corresponding formulas with $\tilde{x}_i$ replaced by $\tilde{x}_i/\|x_i\|$. The absolute error norm $\|\tilde{x}_i\|$ used in these measures are replaced by the relative error norm $\|\tilde{x}_i\|/\|x_i\|$. For example, the *RMS relative error* (RMSRE) and *average relative error* (ARE) are

$$\begin{aligned} \text{RMSRE}(\hat{x}) &= \left(\frac{1}{M}\sum_{i=1}^{M} \|\tilde{x}_i\|^2/\|x_i\|^2\right)^{1/2} && (1) \\ \text{ARE}(\hat{x}) &= \frac{1}{M}\sum_{i=1}^{M} \|\tilde{x}_i\|/\|x_i\| && (2) \end{aligned}$$

where x_i is the estimatee on run i (i.e., the ith realization of the estimatee x). Such measures are simple but limited. Relative measures presented below are more appealing.

3.2 Improvement of Posterior over Prior

We now introduce a metric of relative error for Bayesian (and recursive) estimation. We call it ***Bayesian estimation error quotient*** (***BEEQ***). It quantifies the improvement, in terms of error, of a Bayesian estimator $\hat{x}$ over the prior mean $\bar{x}$ or of the updated estimate $\hat{x}$ of a recursive estimator over the predicted estimate $\bar{x}$. We define it as

$$r^*(\hat{x}) = \frac{\text{AEE}(\hat{x})}{\text{AEE}(\bar{x})} = \frac{\sum_{i=1}^{M} \|x_i - \hat{x}_i\|}{\sum_{i=1}^{M} \|x_i - \bar{x}\|} \qquad (3)$$

where x_i and $\hat{x}_i$ are the ith realizations of x and $\hat{x}$, respectively, $\bar{x}$ is the prior mean (or prediction) of x, and the ***average Euclidean error*** (***AEE***) was introduced in [7] as

$$\text{AEE}(\hat{x}) = \frac{1}{M}\sum_{i=1}^{M} \|\tilde{x}_i\|, \quad \text{AEE}(\bar{x}) = \frac{1}{M}\sum_{i=1}^{M} \|x_i - \bar{x}\|$$

We do not recommend the use of root-mean-square error (***RMSE***) here, such as $r^*(\hat{x}) = \frac{\text{RMSE}^*(\hat{x})}{\text{RMSE}^*(\bar{x})} = \left(\sum_{i=1}^{M} \|x_i - \hat{x}_i\|^2 / \sum_{i=1}^{M} \|x_i - \bar{x}\|^2\right)^{1/2}$, because of the shortcomings of the RMSE (e.g., undue dominance of large terms and lack of a natural interpretation, see [7]) and the fact that merits of RMSE are largely irrelevant here.

One may be tempted to define BEEQ as $\frac{1}{M}\sum_{i=1}^{M} r_i$, that is, the *arithmetic average* of individual error quotients $r_i = \|x_i - \hat{x}_i\|/\|x_i - \bar{x}\|$. This definition is not appropriate. As for any arithmetic average of a positive quantity, particularly a ratio, the average will be dominated by its large terms (i.e., by the cases in which errors are amplified). In other words, good (small) terms should be counted but would be essentially ignored in this definition. What is much worse and in fact fatal is that, as a result of this drawback, this measure is significantly greater than 1 even for many optimal estimators, which is misleading and unacceptable since a good estimator should have a BEEQ significantly smaller than 1. Instead, the *geometric average* is much more appropriate here and thus BEEQ is better defined as $r(\hat{x}) = \left(\prod_{i=1}^{M} r_i\right)^{1/M}$, which is better computed through its logarithm for numerical reasons:

$$\log[r(\hat{x})] = \frac{1}{M}\sum_{i=1}^{M} \log r_i, \quad r_i = \frac{\|x_i - \hat{x}_i\|}{\|x_i - \bar{x}\|} \qquad (4)$$

Error amplification and error reduction are thus balanced.

BEEQ quantifies the contribution of the data to Bayesian (and recursive) estimation. As the error $\|x - \bar{x}\|$ increases, an approximately constant BEEQ indicates that data has an insignificant contribution to estimation, because the measurement error is in effect much larger than $\|x - \bar{x}\|$; the degree of drop in BEEQ reflects that of increase in the contribution of the data; BEEQ will not increase unless an increase in $\|x - \bar{x}\|$ will lead to an even larger increase in the measurement error.

It follows from the properties of arithmetic and geometric averages (see, e.g., [2]) that BEEQ is always bounded by the smallest and largest individual error quotients: $r_{\min} \le r(\hat{x}) \le r_{\max}$, $r_{\min} \le r^*(\hat{x}) \le r_{\max}$, where $r_{\min} = \min\{r_1, \ldots, r_M\}$ and $r_{\max} = \max\{r_1, \ldots, r_M\}$.

4 Estimation Error Relative to Measurement Error

Clearly, the estimation accuracy depends on the accuracy of the data (measurements) as the input to the estimator. In fact, a primary benefit of estimation is that estimates are more accurate than measurements (after they are converted to the same space). As a measure of this accuracy improvement, we introduce a metric, named ***estimate-measurement error ratio*** (***EMER***).

Assume that a (vector-valued) measurement z is related to the estimatee x by $z = g(x, v)$, where v is the measurement error. Let $h(x) = E[g(x, v)|x] = E[z|x]$. For the additive zero-mean noise case with $z = a(x) + v$, we have $h(x) = a(x)$. Then, EMER is defined by

$$\rho^*(\hat{x}) = \frac{\text{AEE}^*(h(\hat{x}))}{\text{AEE}^*(z)} = \frac{\sum_{i=1}^{M} \|h(x_i) - h(\hat{x}_i)\|}{\sum_{i=1}^{M} \|h(x_i) - z_i\|} \qquad (5)$$

where AEE* represents AEE in the measurement space:

$$\begin{aligned} \text{AEE}^*(h(\hat{x})) &= \frac{1}{M}\sum_{i=1}^{M} \|h(x_i) - h(\hat{x}_i)\| \\ \text{AEE}^*(z) &= \frac{1}{M}\sum_{i=1}^{M} \|h(x_i) - z_i\| \end{aligned}$$

That is, EMER is defined as—after converting x and $\hat{x}$ to the measurement space—the ratio of the average distance between x and $\hat{x}$ over the average distance between

x and z. Note that the use of AEE* is preferable to that of RMSE* since RMSE is too much dominated by the large error terms, although measurement errors are often given in terms of standard deviation.

Similar to BEEQ, the *geometric average* of individual estimate-measurement error ratios is more appealing and thus EMER $\rho(\hat{x})$ is better defined by:

$$\log[\rho(\hat{x})] = \frac{1}{M}\sum_{i=1}^{M}\log\rho_i, \quad \rho_i = \frac{\|h(x_i) - h(\hat{x}_i)\|}{\|h(x_i) - z_i\|} \tag{6}$$

Also similar to BEEQ, EMER is always bounded by the smallest and largest individual error ratios: $\rho_{\min} \le \rho(\hat{x}) \le \rho_{\max}$, $\rho_{\min} \le \rho^*(\hat{x}) \le \rho_{\max}$, where $\rho_{\min} = \min\{\rho_1, \ldots, \rho_M\}$ and $\rho_{\max} = \max\{\rho_1, \ldots, \rho_M\}$.

Clearly, we expect that EMER < 1 for a good estimator and the smaller the EMER the better the estimator.

The ultimate goal of estimation is usually to approximate the estimatee as closely as possible. The closeness is usually better measured in the estimatee space directly, rather than in the measurement space, as the above EMER does. In some situations, such as positioning or localization applications, the mapping $h(\cdot)$ defined above is invertible; that is, the mapping $h^{-1}(\cdot)$ from the measurement space to the estimatee space is known. Then, a better definition of EMER is

$$\log[\rho_x(\hat{x})] = \frac{1}{M}\sum_{i=1}^{M}\log\rho_i^x, \quad \rho_i^x = \frac{\|x_i - \hat{x}_i\|}{\|x_i - h^{-1}(z_i)\|} \tag{7}$$

or alternatively

$$\rho_x^*(\hat{x}) = \frac{\text{AEE}(\hat{x})}{\text{AEE}(h^{-1}(z))} = \frac{\sum_{i=1}^{M}\|x_i - \hat{x}_i\|}{\sum_{i=1}^{M}\|x_i - h^{-1}(z_i)\|} \tag{8}$$

where $\text{AEE}(h^{-1}(z)) = \text{AEE}(\hat{x})|_{\hat{x}=h^{-1}(z)}$. It quantifies the amount of improvement an estimator has on the estimate provided by the measurement directly.

If N (scalar- or vector-valued) measurements $z_{i1}, \ldots, z_{iN}$ are used to obtain the estimate $\hat{x}_i$ on run i, then z_i in the above is the stacked vector of these measurements, that is, $z_i = [z_{i1}', \ldots, z_{iN}']'$. For instance, z_i could include range, bearing, and elevation measurements (r, b, e). EMER decreases as more measurements are used in the estimator. Thus EMER should be compared only for estimators using the same size, $\dim(z_i) = \sum_{j=1}^{N}\dim(z_{ij})$, of measurements. To compare EMER of two recursive estimators, the estimators should have the same total measurement size, $\sum_k \dim(z_k)$, or the same measurement size at each recursion, $\dim(z_k)$. To compare EMER of a recursive estimator and a batch estimator, the estimators should have the same total measurement size. However, this can be partially resolved as explained below.

Consider the special case where N statistically identical and independently distributed (scalar- or vector-valued) measurements $z_{i1}, \ldots, z_{iN}$ are used to obtain the estimate $\hat{x}_i$ on run i, that is, $z_i = [z_{i1}', \ldots, z_{iN}']'$. For example, each measurement vector used to obtain $\hat{x}$ could include multiple triples of range, bearing, and elevation measurements (r, b, e), say, of a stationary target from multiple scans of a single radar or of a moving target from multiple radars. In this case, $h(x_i) = [h_1(x_i)', \ldots, h_N(x_i)']'$ and $y_i \overset{\triangle}{=} h_1(x_i) = \cdots = h_N(x_i)$. Then the above-defined EMER is better modified by replacing $\|h(x_i) - z_i\|$ above with $\|h(x_i) - z_i\|/\sqrt{N}$ so that

$$\begin{aligned} \bar{\rho}^*(\hat{x}) &= \frac{\sum_{i=1}^{M}\|h(x_i) - h(\hat{x}_i)\|}{\sum_{i=1}^{M}\|h(x_i) - z_i\|/\sqrt{N}} \\ \bar{\rho}_i &= \frac{\|h(x_i) - h(\hat{x}_i)\|}{\|h(x_i) - z_i\|/\sqrt{N}} \end{aligned}$$

Note that if $h(\hat{x}_i) = \hat{z}_i = (1/N)\sum_{j=1}^{N} z_{ij}$, we have $\text{cov}(\hat{z}_i|x_i) = \text{cov}(z_{ij}|x_i)/N$ for the sample mean $\hat{z}_i$ and, letting $y_i = E[z_{ij}|x_i] = E[\hat{z}_i|x_i]$,

$$\begin{aligned} \frac{\|h(x_i) - h(\hat{x}_i)\|}{\|h(x_i) - z_i\|} &= \frac{\left(\sum_{j=1}^{N}\|h_j(x_i) - \hat{z}_i\|^2\right)^{1/2}}{\left(\sum_{j=1}^{N}\|h_j(x_i) - z_{ij}\|^2\right)^{1/2}} \\ &= \frac{\left(\sum_{j=1}^{N}\|y_i - \hat{z}_i\|^2\right)^{1/2}}{\left(\sum_{j=1}^{N}\|y_i - z_{ij}\|^2\right)^{1/2}} \\ &\approx \frac{(N\text{tr}[\text{cov}(\hat{z}_i|x_i)])^{1/2}}{(N\text{tr}[\text{cov}(z_{ij}|x_i)])^{1/2}} = \frac{1}{\sqrt{N}} \end{aligned}$$

where the $\approx$ sign follows from the approximation of sample covariance and true covariance: $\text{cov}(\hat{z}_i|x_i) \approx (1/N)\sum_{j=1}^{N}(y_i - \hat{z}_i)(y_i - \hat{z}_i)'$, and likewise for the denominator. The modified definition thus leads to $\bar{\rho}(\hat{x}) \approx \bar{\rho}^*(\hat{x}) \approx 1$, which is independent of measurement size and preferable to the measurement-size dependent result $\rho(\hat{x}) \approx \rho^*(\hat{x}) \approx 1/\sqrt{N}$ that the above definitions would yield. As such, the modified definition makes EMER less variant with respect to the measurement size used in $\hat{x}$ and often below 1 for a good estimator since it should normally not be worse than the one with $h(\hat{x}) = \hat{z}$ for a small data size.[1] Consequently, the modified EMER of two or more estimators for problems with different measurement sizes can still be compared meaningfully.

EMERs $\rho(\hat{x})$ and $\rho_x(\hat{x})$ are the error ratios in the measurement space and estimate space, respectively. It is not appropriate to define an estimate-measurement error ratio using $\text{AEE}(\hat{x})/\text{AEE}^*(z)$ or $\|x_i - \hat{x}_i\|/\|h(x_i) - z_i\|$ because the ratio so-defined is a mixture of two spaces and has no physical interpretation. For example, if $z = 4x + v$, the estimator $\hat{x} = z/4$ clearly provides no error reduction. This improper definition gives the misleading result of $1/4$, while the EMER defined above gives the correct answer: $\rho(\hat{x}) = \rho_x(\hat{x}) = 1$.

5 Error Reduction of Bayesian Estimation

EMER quantifies the improvement of the estimator over the measurement and is more suitable for classical (non-Bayesian) estimation. BEEQ measures the improvement of

[1] For a large data size, it is very hard to beat $h(\hat{x}) = \hat{z}$ since the sample mean $\hat{z}$ is a consistent estimator of $h(x)$, that is, it converges to $h(x)$.

the estimate over the prior mean and is appropriate only for Bayesian or recursive estimation, which, however, depends also on the measurement error. It would be nice to quantify the overall improvement of a Bayesian (or recursive) estimator over both the prior mean (or prediction) and the measurement. For this purpose, we define the ***Bayesian error reduction factor*** (***BERF***) by

$$\begin{aligned}\eta^*(\hat{x}) &= \frac{r_z^*(\hat{x})/\text{AEE}^*(h(\bar{x})) + \rho^*(\hat{x})/\text{AEE}^*(z)}{1/\text{AEE}^*(h(\bar{x})) + 1/\text{AEE}^*(z)} \\ &= \frac{r_z^*(\hat{x}) + \beta\rho^*(\hat{x})}{1+\beta} \qquad (9)\end{aligned}$$

or (better)

$$\log[\eta(\hat{x})] = \frac{1}{M}\sum_{i=1}^{M} \log \eta_i, \quad \eta_i = \frac{r_i^z + \beta_i \rho_i}{1+\beta_i} \qquad (10)$$

where $\beta = \text{AEE}^*(h(\bar{x}))/\text{AEE}^*(z)$ and $\beta_i = \|h(x_i) - h(\bar{x})\|/\|h(x_i) - z_i\|$ are the error ratios of prior mean to measurement in the measurement space and $r_z^*(\hat{x})$ and r_i^z are the measurement-space versions of $r^*(\hat{x})$ and r_i, given by

$$\begin{aligned}r_z^*(\hat{x}) &= \frac{\text{AEE}^*(h(\hat{x}))}{\text{AEE}^*(h(\bar{x}))} = \frac{\sum_{i=1}^{M}\|h(x_i) - h(\hat{x}_i)\|}{\sum_{i=1}^{M}\|h(x_i) - h(\bar{x})\|} \\ r_i^z &= \frac{\|h(x_i) - h(\hat{x}_i)\|}{\|h(x_i) - h(\bar{x})\|}\end{aligned}$$

Note that $\eta^*(\hat{x})$ has the following desirable properties:

- It depends more on $\rho^*(\hat{x})$ if there is a larger error in the prior mean, in particular, $\eta^*(\hat{x}) = \rho^*(\hat{x})$ if the prior mean is unknown (which can be interpreted as error of prior mean being ∞), as a non-Bayesian would claim.

- It depends more on $r_z^*(\hat{x})$ for a larger measurement error, in particular, $\eta^*(\hat{x}) = r_z^*(\hat{x}) = 1$ if there is no measurement (i.e., measurement error is ∞).

- $\eta^*(\hat{x}) < 1$ if $r_z^*(\hat{x}) < 1$ and $\rho^*(\hat{x}) < 1$. It follows that $\eta^*(\hat{x})$ of a good Bayesian estimator $\hat{x}$ should be below 1, since we expect $r_z^*(\hat{x}) < 1$ and $\rho^*(\hat{x}) < 1$. Furthermore, $\min\{r_z^*(\hat{x}), \rho^*(\hat{x})\} \le \eta^*(\hat{x}) \le \max\{r_z^*(\hat{x}), \rho^*(\hat{x})\}$, that is, $\eta^*(\hat{x})$ is a compromise of $\rho^*(\hat{x})$ and $r_z^*(\hat{x})$.

Additionally, for the estimator that ignores the measurement (i.e., $\hat{x} = \bar{x}$), and the one that is equivalent to the measurement (i.e., any $\hat{x}_z$ such that $h(\hat{x}_z) = z$ even if $h^{-1}(\cdot)$ does not exist), we have

$$\eta^*(\bar{x}) = \frac{1+\beta^2}{1+\beta}, \quad \eta^*(\hat{x}_z) = \frac{1+\beta^{-2}}{1+\beta^{-1}}$$

which implies that $\min_{\hat{x}} \eta^*(\hat{x}) \le \min\{\eta^*(\bar{x}), \eta^*(\hat{x}_z)\} \le 1$ for all β. Since a good Bayesian estimator $\hat{x}$ should be better than the prior mean $\bar{x}$ and the measurements' equivalent $\hat{x}_z$, we expect that its BERF is below the smaller of $\frac{1+\beta^2}{1+\beta}$ and $\frac{1+\beta^{-2}}{1+\beta^{-1}}$, which is never larger than 1.

Clearly, all these properties hold true also for η_i with $\rho^*(\hat{x})$, $r_z^*(\hat{x})$, and β replaced by ρ_i, r_i^z, and β_i, respectively. For example,

$$\eta_i(\bar{x}) = \frac{1+\beta_i^2}{1+\beta_i}, \quad \eta_i(h^{-1}(z_i)) = \frac{1+\beta_i^{-2}}{1+\beta_i^{-1}}$$

If h^{-1} is available, it is often better to use the versions in the estimatee space:

$$\begin{aligned}\log[\eta_x(\hat{x})] &= \frac{1}{M}\sum_{i=1}^{M}\log\eta_i^x, \quad \eta_i^x = \frac{r_i + \beta_i^x \rho_i^x}{1+\beta_i^x} \\ \eta_x^*(\hat{x}) &= \frac{r^*(\hat{x})/\text{AEE}(\bar{x}) + \rho_x^*(\hat{x})/\text{AEE}(h^{-1}(z))}{1/\text{AEE}(\bar{x}) + 1/\text{AEE}(h^{-1}(z))} \\ &= \frac{r^*(\hat{x}) + \beta_x \rho_x^*(\hat{x})}{1+\beta_x}\end{aligned}$$

where $\beta_x = \text{AEE}(\bar{x})/\text{AEE}(h^{-1}(z))$ and $\beta_i^x = \|x_i - \bar{x}\|/\|x_i - h^{-1}(z_i)\|$ are the error ratios of prior mean to measurement in the estimatee space.

The above results for $\eta^*(\hat{x})$ and η_i also hold for $\eta_x^*(\hat{x})$ and η_i^x.

6 Discussions

Generally speaking, a good estimator $\hat{x}$ should have an EMER smaller than 1, reflecting the requirement that the estimate should be more accurate than the measurement (after they are converted to the same space), since the trivial choice $\hat{x} = \hat{x}_z$ such that $h(\hat{x}_z) = z$ guarantees EMER = 1. A good Bayesian estimator $\hat{x}$ should have BEEQ ≤ 1 and BERF ≤ 1, otherwise one may choose $\hat{x} = \bar{x}$ to guarantee BEEQ $r(\hat{x}) = 1$ or choose either $\hat{x} = \bar{x}$ or $\hat{x} = \hat{x}_z$ to guarantee BERF $\eta^*(\hat{x}) \le 1$. The smaller these measures the better.

For the two versions of BEEQ, $r(\hat{x})$ is in general preferable to $r^*(\hat{x})$ because $r^*(\hat{x})$ is unduely dominated by the large $\|x_i - \hat{x}_i\|$ and/or $\|x_i - \bar{x}_i\|$ terms, but $r(\hat{x})$ is not (i.e., it is balanced). Likewise, ρ, ρ_x, η, η_x are preferable to their starred versions ρ^*, ρ_x^*, η^*, η_x^*, respectively. However, the starred versions are slightly simpler and more convenient to use.

Each measure presented in this section has two versions, in the measurement space and in the estimate space, respectively. The measurement-space version is almost always obtainable, while the estimate-space version requires the availability of the inverse $h^{-1}(z)$. Unless the purpose of estimation is to fit a data model, however, the estimate-space version is normally preferable to the measurement-space version since the former is more generic in that its results have a wider applicability.

BEEQ, EMER, and BERF remain unchanged whether the absolute error or the relative error is used because they are all *relative* measures. For example,

$$\begin{aligned}\rho_z^*(\hat{x}) &= \frac{\sum_{i=1}^{M}\|h(x_i) - h(\hat{x}_i)\|}{\sum_{i=1}^{M}\|h(x_i) - z_i\|} \\ &= \frac{\sum_{i=1}^{M}\|h(x_i) - h(\hat{x}_i)\|/\|h(x_i)\|}{\sum_{i=1}^{M}\|h(x_i) - z_i\|/\|h(x_i)\|}\end{aligned}$$

All measures discussed in this paper use the Euclidean norm. To avoid its two drawbacks discussed in [7], it is better to split these measures into, e.g., two parts corresponding to position and velocity, respectively.

7 Examples: Target Localization

In this section we illustrate the above measures via an example. Consider target localization using a radar or active sonar with range r and bearing b measurements:

$$r = \sqrt{\mathsf{x}^2 + \mathsf{y}^2} + v_r, \quad b = \tan^{-1}\frac{\mathsf{y}}{\mathsf{x}} + v_b$$

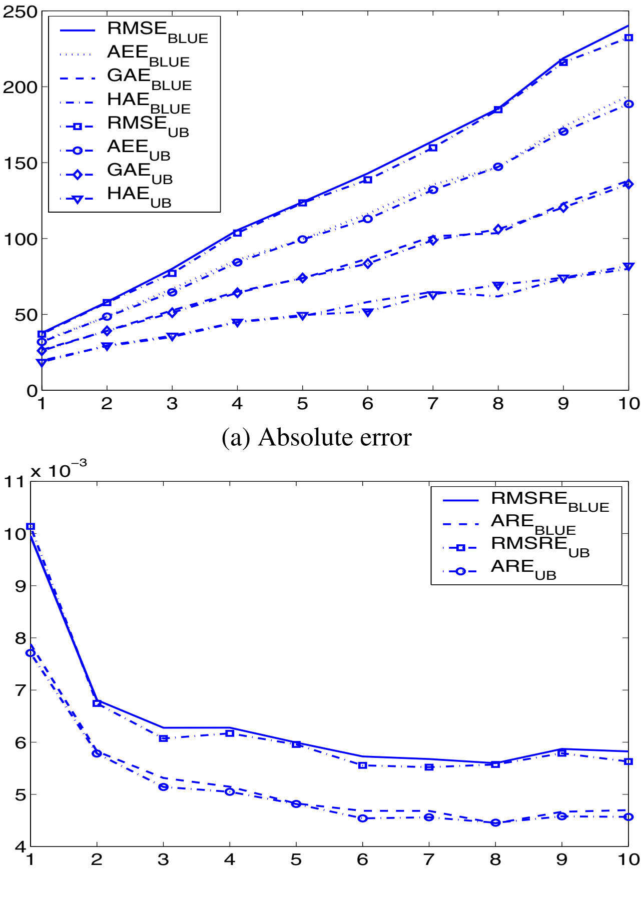
where (x, y) is the true location of a stationary target in the Cartesian coordinates, $v = [v_r, v_b]' \sim \mathcal{N}(0, R)$ is the zero-mean Gaussian measurement noise with covariance $R = \text{diag}(\sigma_r^2, \sigma_b^2)$.

Suppose that estimates of the target location in the Cartesian coordinates are needed. The known prior information of the target location $x = [\mathsf{x}, \mathsf{y}]'$ is: $x \sim \mathcal{N}(\bar{x}, C_x)$, that is, x is Gaussian distributed with mean $\bar{x}$ and covariance C_x. We consider two estimators: $\hat{x}_{\text{UB}}$ and $\hat{x}_{\text{BLUE}}$, based on the unbiased measurement conversion method of [8], combined with the Kalman filter, and the recursive best linear unbiased estimation (BLUE) method of [11], respectively. Note that while $\hat{x}_{\text{BLUE}}$ outperforms $\hat{x}_{\text{UB}}$ significantly for tracking a moving target [11], their performance difference is not significant for this stationary target localization example. However, use of two estimators with close performance serves to check the sensitivity of the metrics and the difference between the two versions of each relative metric proposed.

A simulation with 500 Monte Carlo runs was conducted in which the target location was generated as a random variable $x \sim \mathcal{N}(\bar{x}, C_x)$. Unless otherwise stated, all plots were obtained from $\hat{x}_{\text{UB}}$ and $\hat{x}_{\text{BLUE}}$, each using four independent range-bearing measurement pairs.

Fig. 1(a) shows plots of RMSE, AEE, geometric average error (GAE), and harmonic average error (HAE) for $\hat{x}_{\text{UB}}$ and $\hat{x}_{\text{BLUE}}$ vs. expected target locations $\bar{x} = \alpha[-5000, 20000]'$ (where x-axis is α) with $C_x = 2000^2 I$, $R = \text{diag}(40^2, 0.01^2)$. These metrics were studied in [7]. As expected, they all increase considerably as α increases, and RMSE $>$ AEE $>$ GAE $>$ HAE. Fig. 1(b) shows plots of RMSRE of (1) and ARE of (2) for the same setting. Compared with the absolute RMSE and AEE of Fig. 1(a), these relative error metrics are less variant with respect to the target location and hence better for performance comparison, provided the error of the prior mean is much larger than the measurement error. Note that for this parameter setting the cross-range error $r\sigma_b = 0.01\alpha\sqrt{5000^2 + 20000^2} = 206\alpha$ is significantly smaller than the error of the prior mean, which is 2828.

Fig. 2(a) shows plots of two versions (r and r^*) of BEEQ, given by (4) and (3), for $\hat{x}_{\text{UB}}$ and $\hat{x}_{\text{BLUE}}$ vs. error in the prior mean with $\bar{x} = [-5000, 20000]'$, $C_x = (50\alpha)^2 I$, $R = \text{diag}(40^2, 0.25^2)$. It can be seen that BEEQ is quite stable (total changes are within 20%) as the error in the prior mean increases by 10 times. This should be the case whenever the prior mean is much more accurate than the measurements. An absolute error metric would change greatly over such a large interval. Fig. 2(b) shows the two versions (η and η^*) of BERF, given by (10) and (9), for the same setting. They are also quite stable. When the prior mean is much more accurate than the measurements, BERF should be close to BEEQ (in the measurement space) since its weight for EMER is small. In general, BEEQ is calculated in the estimate space, but BERF and EMER are in the measurement space. For our localization example they are in the Cartesian and polar coordinates, respectively.



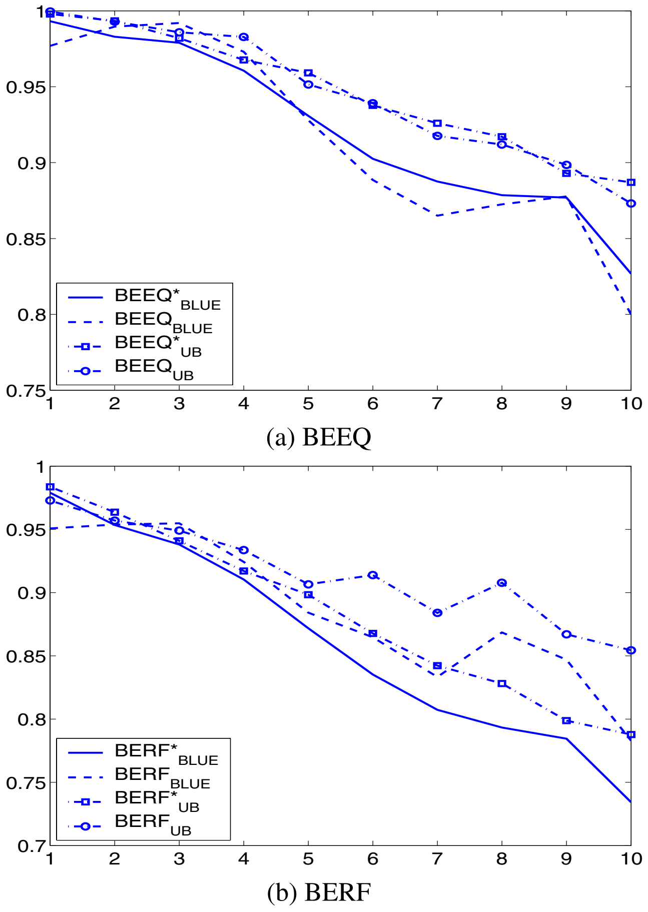
(a) Absolute error

(b) Relative error

Fig. 1: Absolute and relative errors vs. target location $\bar{x} = \alpha[-5000, 20000]'$ (x-axis is α) for $C_x = 2000^2 I$, $R = \text{diag}(40^2, 0.01^2)$.

Fig. 3 shows plots of EMER of $\hat{x}_{\text{UB}}$ and $\hat{x}_{\text{BLUE}}$ vs. measurement size for $\bar{x} = [-5000, 20000]'$, $C_x = 2000^2 I$, $R = \text{diag}(40^2, 0.01^2)$. The plots in (a) and (b) are the modified EMER of (9) with the scaling $\sqrt{N}$ and the original EMER of (6) without the scaling, respectively, where N is the number of statistically repeated measurements used in the estimators. As intended, scaling makes EMER largely invariant (i.e., flat) with respect to the measurement size, provided measurements are significantly more accurate than the prior mean.

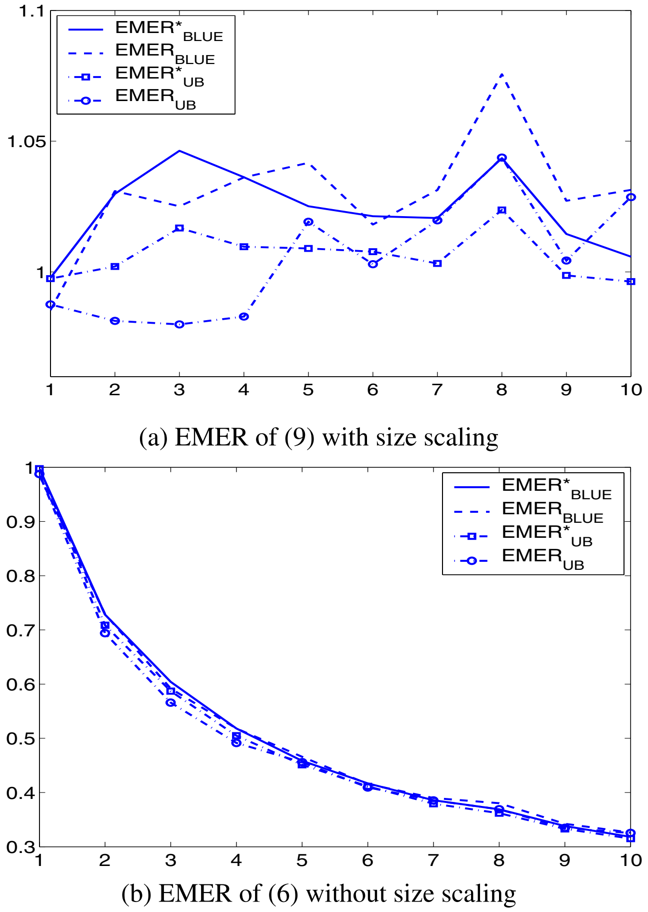
Fig. 4(a) shows plots of two versions (ρ and ρ^*) of EMER, given by (6) and (5), for $\hat{x}_{\text{UB}}$ and $\hat{x}_{\text{BLUE}}$ vs. measurement error with $\bar{x} = [-5000, 20000]'$, $C_x = 2000^2 I$, $R = \text{diag}(40^2, (0.005\alpha)^2)$. It can be seen that EMER is very stable (i.e., flat) as the measurement error increases by



(a) BEEQ

(b) BERF

Fig. 2: BEEQ and BERF vs. error of prior mean (x-axis is α) for $\bar{x} = [-5000, 20000]'$, $C_x = (50\alpha)^2 I$, $R = \text{diag}(40^2, 0.25^2)$.



(a) EMER of (9) with size scaling

(b) EMER of (6) without size scaling

Fig. 3: EMER with and without data size scaling vs. data size for $\bar{x} = [-5000, 20000]'$, $C_x = 2000^2 I$, $R = \text{diag}(40^2, 0.01^2)$.

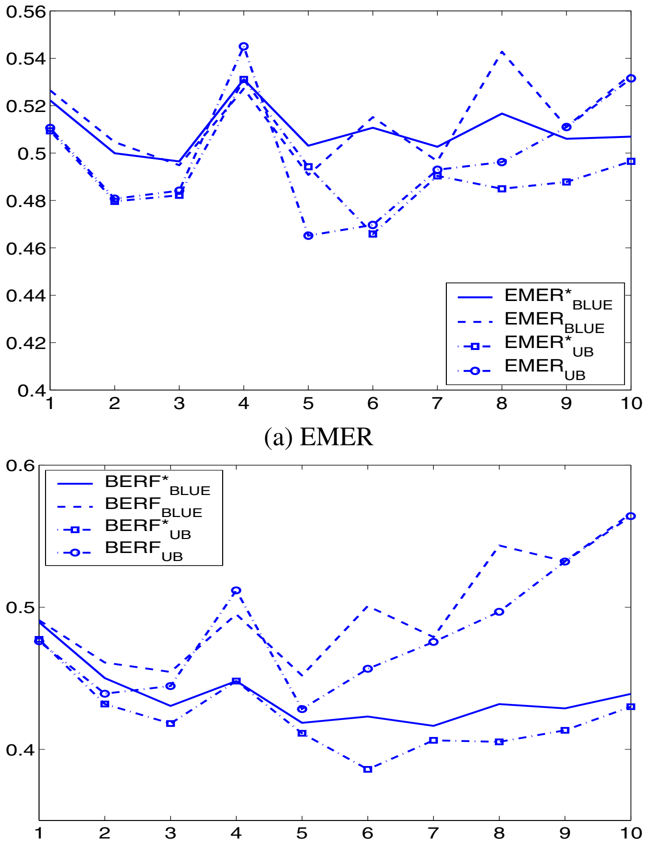
10 times. This should be the case whenever the measurements are much more accurate than the prior mean. This demonstrates that EMER is ideally suited to a classical estimator, which is equivalent to a Bayesian estimator having a prior mean with an infinitely large error. Fig. 4(b) shows the two versions (η and η^*) of BERF, given by (10) and (9), for the same setting. They are also quite stable. In this case BERF is close to EMER since they are both in the measurement space and measurements are much more accurate than the prior mean.

The BERF curves in Figs. 2 and 4 correspond to two fairly extreme cases, where they are largely flat, roughly matching to BEEQ and EMER, respectively. Fig. 5 shows BERF of (10) for $\hat{x}_{\text{BLUE}}$ vs. errors of prior mean and measurements in a more typical case with $\bar{x} = [-5000, 20000]'$, $C_x = \sigma_0^2 I$, $R = \text{diag}(40^2, \sigma_b^2)$. It can be seen that BERF does not fluctuate much as the errors of the prior mean and measurements each increases by 10 times. The other version of BERF for $\hat{x}_{\text{BLUE}}$ and the two versions of BERF for $\hat{x}_{\text{UB}}$ are almost indistinguishable from the one shown here.

All figures showed so far demonstrate that the relative error metrics proposed are rather stable (flat) over a large region of parameter setting. This manifests that they are good indicators of the inherent performance characteristics of an estimator, and thus are useful for performance evaluation and comparison.

Fig. 6 shows the empirical probability density function (pdf) of the Euclidean error norm $\|x - \hat{x}\|$ for $\hat{x}_{\text{UB}}$ and $\hat{x}_{\text{BLUE}}$ with $\bar{x} = [-5000, 20000]'$, $C_x = 2000^2 I$, $R = \text{diag}(40^2, 0.01^2)$, obtained from $10,000$ runs. From Fig. 1(a) at $\alpha = 5$, which corresponds to the parameter setting here, RMSE $\approx$ 120, AEE $\approx$ 100, GAE $\approx$ 75, HAE $\approx$ 50 for both estimators. It is clear from Fig. 5(a) and Fig. 1(a) at $\alpha = 5$ that use of RMSE to represent the typical or average error is not really reasonable due to its undue dominance by large errors, which occur infrequently. The dominance of large errors in AEE is weaker. For this example, error mode $\approx$ 30 and is even smaller than the optimistic HAE. Whether to use GAE, AEE, or error median (which is around 72) is a matter of choice, depending on the particular applications.

Fig. 6 also verifies that the two estimators $\hat{x}_{\text{UB}}$ and $\hat{x}_{\text{BLUE}}$ have virtually the same performance (error distribution) for this example. As such, Figs. 2 through 5 thus serve to check whether the difference between the two versions of BEEQ, EMER, and BERF (i.e., $\{r, \rho, \eta\}$ and $\{r^*, \rho^*, \eta^*\}$) is larger than that of $\hat{x}_{\text{UB}}$ and $\hat{x}_{\text{BLUE}}$. Figs. 2(a), 3(a), and 4(a) indicate that the former difference is slightly smaller than the latter for the BEEQ and EMER; while Figs. 2(b) and 4(b) appear to suggest that the former difference is either comparable to or slightly larger than the latter for the BERF.



(a) EMER

(b) BERF

Fig. 4: EMER and BERF vs. measurement error (x-axis is α) for $\bar{x} = [-5000, 20000]'$, $C_x = 2000^2 I$, $R = \text{diag}(40^2, (0.005\alpha)^2)$.

Overall, it seems reasonable to conclude for this example that the difference between the two versions is insignificant.

8 Summary

A variety of new practical metrics for measuring *relative* error of estimation algorithms has been proposed. These metrics are *relative* in that they are normalized estimation errors with respect to three references, respectively: magnitude of the estimatee, error of prior mean, and measurement error. They include Bayesian estimation error quotient (BEEQ), estimate-measurement error ratio (EMER), Bayesian error reduction factor (BERF). These metrics are useful for measuring different aspects of the performance of an estimation algorithm.

It is an illusion that performance evaluation can be done completely fairly and impartially. This is partly because simple metrics cannot capture a complete picture of the performance of an estimation algorithm and those that are more complete are more complex and subject to subjective interpretations. Also, use of any metric in performance evaluation implicitly favors the estimator that tries to optimize this same metric. Nevertheless, all is not lost. What one should do is to choose the metrics that are more relevant to the application at hand. That is also the value of having a wide spectrum of metrics available, along with a good understanding of them. While it is certainly superior in theory to have a unified metric for all applications, to be discussed in forthcoming papers, specific metrics can be significantly more useful in practice for particular applications.

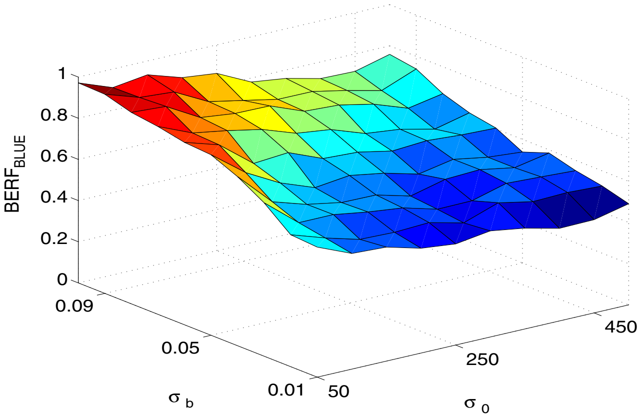


Fig. 5: BERF vs. errors of prior mean and measurements for $\bar{x} = [-5000, 20000]'$, $C_x = (50C)^2 I$, $R = \text{diag}(40^2, (0.01R_2)^2)$.

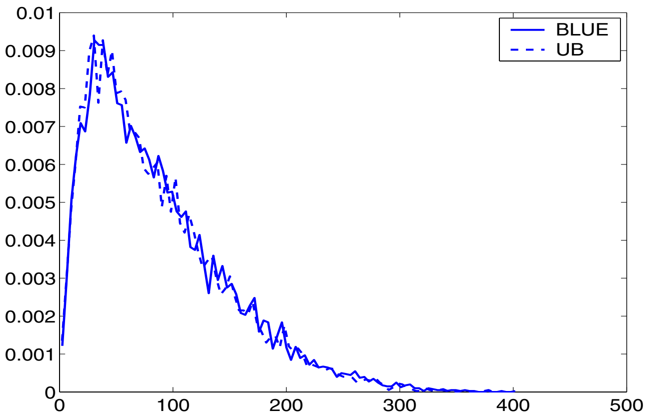


Fig. 6: Empirical pdf for $\bar{x} = [-5000, 20000]'$, $C_x = 500^2 I$, $R = \text{diag}(40^2, 0.01^2)$.

References

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

[2] P. S. Bullen. *Handbook of Means and Their Inequalities*. Kluwer Academic, Dordrescht, The Netherlands, 2003.

[3] C. Y. Chong. Problem Characterization in Tracking/Fusion Algorithm Evaluation. *IEEE Aerospace and Electronic Systems Magazine*, 16:12–17, July 2001.

[4] O. E. Drummond. Methodologies for Performance Evaluation of Multitarget Multisensor Tracking. In *Proc. 1999 SPIE Conf. on Signal and Data Processing of Small Targets,* vol. 3809, pages 355–369, Denver, CO, USA, July 1999.

[5] 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, pages 510–527, Apr. 1998.

[6] A. I. El-Fallah, R. P. Mahler, T. Zajic, E. Sorensen, M. G. Alford, and R. K. Mehra. Scientific Performance Evaluation for Sensor Menagement. In *Signal Processing, Sensor Fusion, and Target Recognition IX*, volume SPIE 4052, pages 183–194, 2000.

[7] X. R. Li and Z.-L. Zhao. Measures of Performance for Evaluation of Estimators and Filters. In *Proc. 2001 SPIE Conf. on Signal and Data Processing of Small Targets,* vol. 4473, pages 530–541, San Diego, CA, USA, July-August 2001.

[8] L. Mo, X. Song, Y. Zhou, Z. Sun, and Y. Bar-Shalom. Unbiased Converted Measurements for Tracking. *IEEE Trans. Aerospace and Electronic Systems*, AES-34(3):1023–1026, 1998.

[9] R.L. Rothrock and O.E. Drummond. Performance Metrics for Multiple-Sensor, Multiple-Target Tracking. In *Proc. SPIE Conf. on Signal and Data Processing of Small Targets 2000,* vol. 4048, pages 521–531, Apr. 2000.

[10] T. Zajic, J. L. Hoffman, and R. P. Mahler. Scientific Performance Metrics for Data Fusion: New Results. In *Signal Processing, Sensor Fusion, and Target Recognition IX*, volume SPIE 4052, pages 172–182, 2000.

[11] Z.-L. Zhao, X. R. Li, and V. P. Jilkov. Optimal Linear Unbiased Filtering with Nonlinear Radar Measurements for Target Tracking. *IEEE Trans. Aerospace and Electronic Systems*, 40(4), Oct. 2004.