Weaknesses of the ANEES and New Calibration Measures for Multivariate Predictions

Markus Walker, Marcel Reith-Braun, and Uwe D. Hanebeck

***Abstract*— Reliable quantification of uncertainty is crucial for trustworthy predictions in estimation theory and machine learning. However, existing credibility and calibration measures, such as the widely used averaged normalized estimation error squared (ANEES), often exhibit limitations when applied to biased model predictions. In this paper, we systematically review the ANEES and its alternatives, analyze their strengths and weaknesses, and highlight cases where standard measures fail to detect miscalibration. Building on recent advances in calibration measures, we propose two new measures: the generalized uncertainty calibration error and its normalized version. These measures unify and extend the concepts of estimator credibility and regression calibration to multivariate settings. Comprehensive experiments demonstrate the characteristics of credibility and calibration measures, including the proposed measures, and their applicability to regression models.**

***Index Terms*— Uncertainty quantification, credibility measures, calibration measures.**

I. Introduction

Quantifying uncertainty is a fundamental challenge in both estimation theory and machine learning. Reliable predictions are essential for decision-making in safety-critical applications, and autonomous systems [1], [2]. However, assessing the quality of these predictions—particularly their uncertainty estimates—remains a nontrivial task [3], [4].

In estimation theory, measures such as the averaged normalized estimation error squared (ANEES) [5] have long been used to evaluate the credibility of state estimators, such as Kalman filters. The ANEES assesses the consistency of the estimated states and their associated covariance matrices under the assumption of Gaussianity, and has been widely adopted in the literature [6]–[9]. Subsequent works [10]–[12] have further analyzed its properties, limitations, and proposed alternatives, such as the matrix norm relative error (MNRE) and the noncredibility index (NCI), to address specific shortcomings of the ANEES. When used for assessing the distribution of estimated states in state estimation, these measures are often referred to as *credibility measures* [11].

In the context of machine learning, particularly for regression problems, research typically focuses on *calibration measures* that evaluate the similarity between predicted probability distributions and observed data from a stochastic process. A prediction is called well-calibrated if it is unbiased and has the same uncertainty (e.g., variance) as the stochastic process [13], as shown in Fig. 1. Popular measures include the uncertainty calibration error (UCE) [14], expected normalized calibration error (ENCE) [15], and quantile calibration error (QCE) [4]. These measures are designed to evaluate the global calibration of predictive models. Furthermore, these calibration measures can be used to assess the local calibration of regression models when applied to local regions of the model's input space, as demonstrated in our previous work [16]–[18].

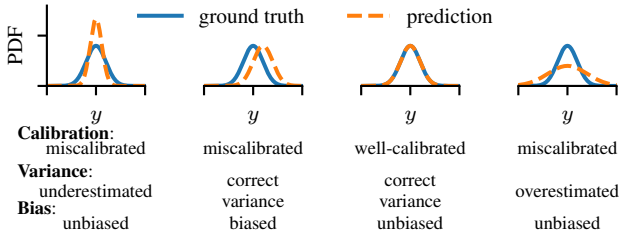


Fig. 1: Examples of different predictions and their characteristics in terms of calibration, confidence and bias (adapted version from [13]).

Despite the variety of available measures, there is no universal solution that is applicable across all domains and data modalities. Many existing measures are tailored to specific applications or assumptions, such as univariate or multivariate predictions, Gaussianity, or the availability of ground truth information. This diversity complicates the comparison of model performance and the interpretation of calibration results.

Motivation and Scope: This paper aims to bridge the gap between credibility measures from estimation theory and calibration measures from machine learning. Although both types of measures are widely used, they are often developed and applied in isolation, which can lead to inconsistent or incomplete assessments of model quality. A unified perspective is particularly important as predictive models are increasingly deployed in safety-critical and high-stakes domains, where both credible and well-calibrated uncertainty estimates are essential for reliable decision-making. We systematically review the ANEES and its alternatives, highlighting their strengths and weaknesses in both theory and applications. Building on [14], [15], we extend calibration measures for regression models to multivariate predictions, leveraging insights from estimator credibility measures.

Contribution: The main contributions of this paper are: First, we provide a review of the fundamental concepts of estimation and prediction quality assessment in estimation theory and machine learning. Second, the ANEES and its limitations are analyzed, with illustrative examples demonstrating cases where it fails to detect incredibility. Third, alternative credibility measures, such as the MNRE, log-matrix norm ratio (MNR), and NCI, are discussed and their properties compared. Fourth, state-of-the-art calibration measures for regression

This work is part of the German Research Foundation (DFG) AI Research Unit 5339 regarding the combination of physics-based simulation with AI-based methodologies for the fast maturation of manufacturing processes.

Markus Walker, Marcel Reith-Braun and Uwe D. Hanebeck are with the Intelligent Sensor-Actuator-Systems Laboratory (ISAS), Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany (e-mail: {markus.walker, marcel.reith-braun, uwe.hanebeck}@kit.edu).

models, including the UCE, ENCE, and QCE, are reviewed, and their relationships to credibility measures are examined. Finally, a new calibration measure is proposed that extends existing approaches to multivariate predictions in regression models, leveraging insights from estimator credibility.

Notation: In this paper, underlined letters, e.g., $\underline{x}$, denote vectors, boldface letters, such as $\underline{\boldsymbol{x}}$, represent random variables, while boldface capital letters, such as $\mathbf{A}$, indicate matrices. Sets are denoted by calligraphic letters, e.g., $\mathcal{D}$. Estimated or predicted quantities are denoted by the superscript e, for instance, $\underline{\boldsymbol{x}}^{\mathrm{e}}$, and the ground truth quantities are denoted by the superscript GT, e.g., $\underline{\boldsymbol{x}}^{\mathrm{GT}}$. Covariance matrices are denoted by $\mathbf{C}$, and mean squared error (MSE) matrices are given by $\mathbf{\Sigma}$. Expectation operators are denoted by $\mathrm{E}\{\cdot\}$; e.g., $\mathrm{E}_{p(\underline{x})}\{\underline{\boldsymbol{x}}\}$ denotes the expectation of $\underline{\boldsymbol{x}}$ under the density $p(\underline{x})$. When appropriate, the density argument is omitted for simplicity. The same convention applies to the variance operator, denoted by $\mathrm{Var}\{\cdot\}$.

II. Preliminaries

In this section, we introduce the fundamental concepts used throughout the paper and clarify the terminology related to the assessment of estimation and prediction quality. This includes *credibility*, as known from estimation theory [10], [11], and *calibration*, as it is common in recent machine learning applications [3], [4], [14], [15]. Credibility refers to how well a single estimate matches the true state (a random variable), typically assessed using ground truth information or Monte Carlo (MC) simulations, whereas calibration evaluates how well the predicted distribution of a regression model aligns with the true data-generating process. In the remainder of this paper, the term *estimations* is used in relation to credibility measures, and *predictions* in terms of regression models. This distinction guides the structure of the paper: following the preliminaries in this section, credibility measures are discussed first in Sec. III, followed by calibration measures in Sec. IV.

A. Credibility of Estimators

Credibility refers to how well an estimator, such as a parameter estimator or a filter, estimates the true random variable of interest, referred to as estimand or state. The moments of the true random variable $\underline{\boldsymbol{x}}^{\mathrm{GT}}$ are described by its mean $\underline{\mu}^{\mathrm{GT}}$ and its covariance matrix $\mathbf{C}^{\mathrm{GT}}$, whereas the estimand is given by its mean $\underline{\mu}^{\mathrm{e}}$ and covariance matrix $\mathbf{C}^{\mathrm{e}}$. For credibility assessment, the estimation error $\underline{\boldsymbol{e}}$ is used, which is defined as $\underline{\boldsymbol{e}} = \underline{\boldsymbol{x}}^{\mathrm{GT}} - \underline{\mu}^{\mathrm{e}}$. An optimal (and credible) estimator would match the true state (is unbiased), i.e., $\underline{\mu}^{\mathrm{e}} = \underline{\mu}^{\mathrm{GT}}$ and $\mathbf{C}^{\mathrm{e}} = \mathbf{C}^{\mathrm{GT}}$, with the optimal error $\underline{\boldsymbol{e}}^{*} = \underline{\boldsymbol{x}}^{\mathrm{GT}} - \underline{\mu}^{\mathrm{GT}}$, meaning that the estimation should be unbiased ($\mathrm{E}\{\underline{\boldsymbol{e}}\} = \underline{0}$), and the MSE matrix $\mathbf{\Sigma} = \mathrm{E}\{\underline{\boldsymbol{e}}\underline{\boldsymbol{e}}^{\top}\}$ should be equal to $\mathbf{C}^{\mathrm{e}}$. Assessing the credibility involves checking both expectations; however, in practice, calculating the expectations is not applicable since the true random variable is usually unknown, and therefore, the evaluation is typically performed using sample approximations based on multiple MC simulations [19]. Furthermore, there is no universally accepted scalar measure of the credibility of estimators. One of the most commonly used measures is the ANEES, whose weaknesses and alternatives are described in Sec. III.

B. Calibration of Regression Models

Measuring the calibration of regression models in the context of machine learning involves assessing the quality of predictions made by a model using available test data. In this context, data is generated from a true data-generating process that follows the distribution $\underline{\boldsymbol{y}} \sim p^{\mathrm{GT}}(\underline{y} \mid \underline{x})$, where $\underline{\boldsymbol{y}}$ is a random variable over the N_y-dimensional sample space $\underline{y} \in \mathbb{R}^{N_y}$, and $\underline{\boldsymbol{x}}$ is the input as a random variable over the N_x-dimensional sample space. The data-generating process can also be represented as $\underline{\boldsymbol{y}} = f^{\mathrm{GT}}(\underline{\boldsymbol{x}})$, with $f^{\mathrm{GT}} \colon \mathbb{R}^{N_x} \to \mathbb{R}^{N_y}$ being the true mapping from the input data $\underline{x}$ to the output data $\underline{y}$. In regression tasks, the goal is to learn a mapping $f^{\mathrm{approx}}(\cdot)$ approximating the true, but unknown data-generating process $f^{\mathrm{GT}}(\cdot)$ given a finite training data set $\mathcal{D}$ that consists of one or multiple realizations of $\underline{\boldsymbol{y}}$, given different input realizations of $\underline{\boldsymbol{x}}$. For example, this can be achieved by using approximation methods such as Bayesian neural networks, or other machine learning models.

Once training is complete, the learned model is used to predict the output distribution $p(\underline{y} \mid \underline{x}, \mathcal{D})$ for a given input $\underline{x}$. Since the estimated distribution is based on the finite training data set, it may not perfectly capture the true distribution. Therefore, model predictions are evaluated using the test data set $\mathcal{D}_{\mathrm{Test}} = (\underline{x}_n, \underline{y}_n)_{n=1}^{N_{\mathrm{Test}}}$, which consists of N_{Test} input-output pairs. Note that there could be multiple realizations of the output for the same input value. However, we assume that, in the standard case, there is only one realization per input value. For evaluation, calibration measures are used to assess the quality of the predicted probability density functions (PDFs) $p(\underline{y} \mid \underline{x}_n, \mathcal{D})$ in terms of how well they match the true PDFs $p(\underline{y} \mid \underline{x}_n)$, at the test points $\underline{x}_n$.

III. ANEES, its Weaknesses and Alternatives

Assessing the credibility of estimates can be done concerning different aspects, such as bias credibility ($\mathrm{E}\{\underline{\boldsymbol{e}}\} = \underline{0}$), MSE credibility ($\mathbf{\Sigma} = \mathbf{C}^{\mathrm{e}}$, assuming zero mean errors), or their joint evaluation [11]. In this section, we first review the ANEES, which is widely used in the literature [5], [11], [19], along with its weaknesses and alternatives proposed in the literature.

A. Definition and Properties of the ANEES

To assess the MSE credibility, the ANEES [5], [19] is commonly used, which is based on the squared Mahalanobis distance. The ANEES is defined by [19]

$$\mathrm{ANEES} = \frac{1}{N_{\mathrm{MC}} \cdot N_e} \sum_{n=1}^{N_{\mathrm{MC}}} \underline{e}_n^{\top} (\mathbf{C}_n^{\mathrm{e}})^{-1} \underline{e}_n \ , \tag{1}$$

where $\underline{e}_n \in \mathbb{R}^{N_e}$ is the N_e-dimensional error realization, $\mathbf{C}_n^{\mathrm{e}}$ is the estimated covariance matrix, and $d_{\mathrm{M},n}^2 = \underline{e}_n^{\top} (\mathbf{C}_n^{\mathrm{e}})^{-1} \underline{e}_n$ is the squared Mahalanobis distance, also known as normalized estimation error squared (NEES). The error realization is given by $\underline{e}_n = \underline{x}_n - \underline{\mu}_n^{\mathrm{e}}$, where $\underline{x}_n$ is the realization of the true state $\underline{\boldsymbol{x}}^{\mathrm{GT}}$ in the n-th out of N_{MC} MC runs, $\underline{\mu}_n^{\mathrm{e}}$ is the estimated mean, and $\mathbf{C}_n^{\mathrm{e}}$ is the estimated covariance matrix, both given by the estimator. Note that $\underline{\mu}_n^{\mathrm{e}}$ and $\mathbf{C}_n^{\mathrm{e}}$ can also remain constant across multiple MC runs for comparison with multiple error realizations. The usage

and interpretation of the ANEES is based on the fundamental assumption that is key to understanding its values.

Assumption 1: The errors are normal, have zero mean, and the MSE matrix $\mathbf{\Sigma} = \mathrm{E}\{\underline{\boldsymbol{e}}\underline{\boldsymbol{e}}^\top\}$ is equal to the true error covariance matrix $\mathbf{C}^{\mathrm{GT}}$, that is, $\underline{\boldsymbol{e}} \sim \mathcal{N}(\underline{e}; \underline{0}, \mathbf{\Sigma})$.

A direct consequence of Assumption 1 is that the squared Mahalanobis distance of the errors $d^2_{\mathrm{M},n}$ is chi-square distributed with N_e degrees of freedom [5], [19]. This is conveniently illustrated by the following example.

Example 1: Consider the case of an unbiased univariate estimate ($N_e = 1$). The average squared Mahalanobis distance is given by

$$\sum_{n=1}^{N_{\mathrm{MC}}} d^2_{\mathrm{M},n} = \sum_{n=1}^{N_{\mathrm{MC}}} \left(\frac{e_n}{\sigma^{\mathrm{GT}}}\right)^2 = \sum_{n=1}^{N_{\mathrm{MC}}} z_n^2$$

with zero mean normally distributed error $\boldsymbol{e} \sim \mathcal{N}(e; 0, (\sigma^{\mathrm{GT}})^2)$, and standard deviation $\sigma^{\mathrm{GT}} = \sigma^{\mathrm{e}}_n$. The squared Mahalanobis distance, therefore, is equal to the squared error divided by the estimated variance, i.e., $\boldsymbol{z} = \boldsymbol{e}^2/(\sigma^{\mathrm{e}})^2$. This can be seen as standardizing the error, so that each realization z_n follows the standard normal distribution. Summing over all normalized squared errors results, by definition, in a chi-squared distributed random variable with N_{MC} degrees of freedom, i.e., $\sum_{n=1}^{N_{\mathrm{MC}}} z_n^2 \sim \chi^2_{N_{\mathrm{MC}}}$ [20].

The univariate case can be generalized to multivariate random variables, resulting in the ANEES being chi-square distributed with $k = N_{\mathrm{MC}} \cdot N_e$ degrees of freedom [19]. A convenient property of the chi-square distribution is that its expected value equals its degrees of freedom, i.e., $\mathrm{E}\{\chi^2_k\} = k$. This is used to normalize the ANEES by dividing it by the degrees of freedom. Therefore, the expectation of the ANEES is equal to one, if Assumption 1 holds [11].

Following the definition of the ANEES and its distribution, three cases can be distinguished:

1) When the ANEES is equal to one, the estimates are MSE credible, which means that the estimated uncertainties, i.e., the covariance matrices $\mathbf{C}^{\mathrm{e}}_n$, match the true MSE matrix $\mathbf{\Sigma}$.
2) When the ANEES is lower than one, the estimated covariance matrices are dominating the errors, which means that the estimated uncertainty is too high.
3) When the ANEES is larger than one, the errors are dominating over the estimated uncertainty, which means that the estimated uncertainty is too low.

The latter two cases indicate that the estimated uncertainty is not MSE credible, i.e., the estimated covariance matrices $\mathbf{C}^{\mathrm{e}}_n$ do not match the true MSE matrix $\mathbf{\Sigma}$.

Furthermore, by utilizing the chi-square distribution, the ANEES can be used to perform a statistical test for the goodness-of-fit of the estimated distribution. For this, one compares the ANEES with the critical values of the chi-square distribution [19].

B. Weaknesses of the ANEES

Weaknesses of the ANEES have been examined in several papers [10]–[12]. Most prior work has evaluated the ANEES as a credibility measure for normally distributed estimators, typically by comparing results with ideal estimators and known MSE matrices [10], [11], or by applying statistical tests for the goodness-of-fit of the estimated distribution [12].

1) Use of the Arithmetic Mean: The primary limitation of the ANEES is its reliance on the arithmetic mean of the squared Mahalanobis distances $d^2_{\mathrm{M},n}$ in (1) [11]. The NEES itself is a ratio of the MSE matrix to the estimated covariance matrix, which is inherently asymmetric: underconfident estimates yield values between zero and one, while overconfident estimates can range from one to infinity. This asymmetry complicates the comparison of different estimators based solely on their ANEES values, as underconfidence and overconfidence are treated differently. Averaging of ratios can lead to unexpected results, where one NEES component of the ANEES that is close to zero can be compensated by another component that is larger than one, leading to a ANEES close to one, even though the estimated distribution does not match the true distribution [10].

Example 2: To illustrate the weakness of averaging ratios in the ANEES, consider the problem of estimating temperature in two separate MC runs, A and B. Run A is overly cautious: it reports a very wide uncertainty (variance much larger than the actual squared error). Run B is overconfident: it reports a very narrow uncertainty (variance much smaller than the actual squared error). Suppose both runs are unbiased, so their mean estimands are correct, but their reported variances differ greatly. In terms of the NEES, run A yields a value of $1/100$ (variance 100 times too large), while run B yields $199/100$ (variance about two times too small). The average of these two NEES values is one, which, according to the ANEES, suggests perfect MSE credibility.

However, this average is misleading: neither estimand is actually reliable—one is much too uncertain, the other much too confident. This example demonstrates how the arithmetic mean of ratios, as used in the ANEES, can obscure significant incredibility, since extreme over- and under-confidence can cancel each other out in the average.

2) Biased Estimates: A second caveat to the ANEES is that it is not designed to check the bias credibility of the estimated distribution. The behavior of the ANEES for biased estimates can be misleading, e.g., when the ANEES is used and it is not checked if an estimate is biased. The following theorem formalizes this.

Theorem 1: Let $\boldsymbol{x}^{\mathrm{GT}} \sim \mathcal{N}(x; \mu^{\mathrm{GT}}, (\sigma^{\mathrm{GT}})^2)$ be the true random variable. Suppose an estimate is given by $\boldsymbol{x}^{\mathrm{e}} \sim \mathcal{N}(x; \mu^{\mathrm{e}}, (\sigma^{\mathrm{e}})^2)$, where $\mu^{\mathrm{e}} = \mu^{\mathrm{GT}} + \Delta\mu$, $\Delta\mu \neq 0$ (biased mean) and $(\sigma^{\mathrm{e}})^2 = c \cdot (\sigma^{\mathrm{GT}})^2$, with $c > 1$ (overdispersed variance). Then, the ANEES, computed with respect to the estimated mean and variance can satisfy $\mathrm{ANEES} = 1$, even though the mean and variance of $\boldsymbol{x}^{\mathrm{e}}$ are not equal to the mean and variance of $\boldsymbol{x}^{\mathrm{GT}}$. For simplicity, the superscript GT is omitted in the following proof.

Proof: We start with the expected value of the squared Mahalanobis distance

$$\mathrm{ANEES} = \mathrm{E}\left\{\frac{(\boldsymbol{x} - (\mu + \Delta\mu))^2}{(\sigma^{\mathrm{e}})^2}\right\},$$

which, for univariate estimates, i.e., $N_e = 1$, can be viewed as the expectation-based counterpart of (1). For a perfectly

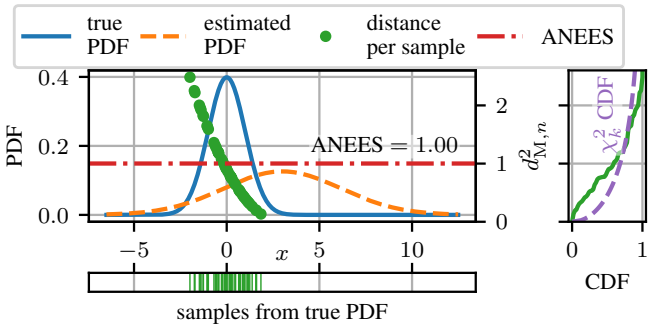


Fig. 2: Illustration of Example 3 for the weakness of the ANEES, where the ANEES is approximately one, but the estimated distribution is biased with a larger than true variance.

credible estimate, we have ANEES $= 1$. Therefore, for perfect credibility, it holds that

$$\mathrm{E}\{(\boldsymbol{x} - (\mu + \Delta\mu))^2\} = c \cdot \sigma^2 \ , \tag{2}$$

since $c \cdot (\sigma)^2$ is constant. Expanding the left-hand side

$$\underbrace{\mathrm{E}\{\boldsymbol{x}^2 - 2\boldsymbol{x} + \mu^2\}}_{=\mathrm{E}\{(\boldsymbol{x}-\mu)^2\}=\sigma^2} + \underbrace{\mathrm{E}\{2\mu\Delta\mu - 2\boldsymbol{x}\Delta\mu\}}_{=2\Delta\mu\mathrm{E}\{\mu-\boldsymbol{x}\}=0} + \underbrace{\mathrm{E}\{(\Delta\mu)^2\}}_{=(\Delta\mu)^2}$$

and solving for c yields

$$c = \frac{1}{\sigma^2}\left(\sigma^2 + (\Delta\mu)^2\right) = 1 + \frac{(\Delta\mu)^2}{\sigma^2} \ . \tag{3}$$

Thus, whenever (3) holds, ANEES $= 1$. Since the right-hand side of (3) is greater than one for $\Delta\mu \neq 0$, bias can be traded for a larger variance and the ANEES can be equal to one, even if the moments of $\boldsymbol{x}^{\mathrm{e}}$ are not equal to the moments of $\boldsymbol{x}$. ■

Note, however, that the considered scenario violates Assumption 1, since $\mathrm{E}\{\boldsymbol{e}\} = \Delta\mu \neq 0$, and the NEES and ANEES are no longer chi-square distributed. In fact, for biased estimates, the distribution follows a non-central chi-square distribution [21]. As a result, an ANEES value of one is not meaningful, since the expectation of the non-central chi-square distribution is not equal to one, and the ANEES can no longer be used to assess the MSE credibility. This issue is illustrated by the Example 3.

Example 3: In this example, the true distribution is a univariate normal distribution with $\mu^{\mathrm{GT}} = 0$ and $\sigma^{\mathrm{GT}} = 1$. The estimated distribution is a univariate normal distribution with $\mu^{\mathrm{e}} = \mu^{\mathrm{GT}} + \Delta\mu$ and $(\sigma^{\mathrm{e}})^2 = c \cdot (\sigma^{\mathrm{GT}})^2$, where the mean is shifted by $\Delta\mu = 3$ and the estimated variance is $c = 10$ times larger than the true variance. This corresponds to a situation in which the condition (3) holds. 50 samples are drawn from the true distribution, and the squared Mahalanobis distance of each sample to the estimated distribution is calculated. The ANEES is equal to one, even though the estimated distribution does not match the true distribution. However, the cumulative density function (CDF) of the Mahalanobis distances in Fig. 2 reveals that they are not chi-square distributed, which violates Assumption 1.

Even though the ANEES is widely used [6], [8], [22], [23], Assumption 1 is often not checked in practice. As demonstrated in Example 3, these results can be misleading. Therefore, the ANEES should be applied with caution. Violations of the underlying assumptions can be detected, e.g., by using the Kolmogorov–Smirnov test to check for significant differences between the empirical distribution of squared Mahalanobis distances and the chi-square distribution with N_e degrees of freedom, as demonstrated in [7], and should always accompany the application of the ANEES.

C. Alternatives to ANEES

To account for the weaknesses of the ANEES, several alternatives have been proposed in literature [10], [11]. All of these avoid using the arithmetic average of the squared Mahalanobis distances. Instead, the measures proposed in [10], [11] are based on ratios or differences of the matrix norms of the estimated covariance matrix and true MSE matrix, thereby overcoming the asymmetry inherent in the ANEES. E.g., the MNRE, and log-MNR, are defined by [10], [11]

$$\mathrm{MNRE} = \frac{\left\|\boldsymbol{\Sigma} - \mathbf{C}^{\mathrm{e}}\right\|}{\left\|\boldsymbol{\Sigma}\right\| + \left\|\mathbf{C}^{\mathrm{e}}\right\|} \ , \tag{4}$$

$$\text{log-MNR} = \log_{10}\left(\frac{\left\|\boldsymbol{\Sigma}\right\|}{\left\|\mathbf{C}^{\mathrm{e}}\right\|}\right) \ , \tag{5}$$

where $||\cdot||$ is an arbitrary matrix norm. The MNRE has the property that it is bounded between zero and one, where a value of zero indicates perfect MSE credibility. Note that the MNRE cannot distinguish between over- and underdispersed estimates. However, it is symmetric, meaning that it, in contrast to the ANEES, penalizes over- and underconfident estimates equally. The log-MNR is not bounded and symmetric. It can be used to distinguish between over- and underdispersed estimates, where $\boldsymbol{\Sigma} = \mathbf{C}^{\mathrm{e}}$ implies a value of zero, but not vice versa. If the norm of the estimated covariance matrix is larger than the norm of the MSE matrix, the log-MNR is negative, indicating an overdispersed estimate, while a positive value indicates an underdispersed estimate. Additional measures with similar properties are proposed in [10], [11], such as the log-mean squared error ratio and log-generalized error variance ratio, which are based on the trace and determinant of the covariance matrices, respectively. However, none of these measures, including the ANEES, check the bias credibility of the estimated distribution, i.e., they assume unbiased estimates. In fact, the weakness of the ANEES from Theorem 1 (relative comparison of the MSE and covariance matrix) also applies to difference-based comparisons, such as in the numerator of the MNRE. This can be seen by considering again the univariate case, and setting the difference $\mathrm{E}\{\boldsymbol{e}^2\} - (\sigma^{\mathrm{e}})^2$ to zero, which is equal to (2).

To check both the bias and the MSE credibility of the estimated distribution, [10], [11] proposed the NCI measure, which is a joint bias and covariance credibility measure. The NCI is given by [11, Eq. 4]

$$\begin{aligned}\mathrm{NCI} = \frac{10}{N}\sum_{n=1}^{N}\Bigg|&\log_{10}\left((\underline{e}_n - \underline{b}_n)^{\top}(\mathbf{C}^{\mathrm{e}}_n)^{-1}(\underline{e}_n - \underline{b}_n)\right) \\ &- \log_{10}\left((\underline{e}^{\mathrm{GT}}_n)^{\top}\boldsymbol{\Sigma}^{-1}\underline{e}^{\mathrm{GT}}_n\right)\Bigg| \ ,\end{aligned} \tag{6}$$

where $\underline{b}_n$ is the bias vector of the estimated distribution and $\underline{e}^{\mathrm{GT}}_n = \underline{x}_n - \underline{\mu}^{\mathrm{GT}}$ is the error between the n-th realization and the true mean. The NCI assumes that the bias and the

covariance matrix are provided by an estimator, and the true mean μ^{GT} is known, or can be approximated by multiple MC runs [11].

IV. Calibration Measures for Regression Models

In regression, assessing the calibration of predictions does not concern a single random variable or estimate, but rather evaluates how well multiple predicted distributions align with a typically unknown stochastic process. Furthermore, the true MSE matrix $\mathbf{\Sigma}$ for each input is generally unknown, and performing multiple MC simulations to estimate it is usually not feasible. Instead, one is typically given only a test data set $\mathcal{D}_{\mathrm{Test}}$ with only one sample per input. Therefore, each prediction cannot be evaluated individually, and measures such as the MNRE cannot be applied. However, several calibration measures have been proposed in literature to assess the quality of regression model predictions without requiring ground truth knowledge or multiple MC simulations; these are discussed below. First, however, we outline the general stochastic model on which the calibration measures are implicitly based.

A. Stochastic Model

For our derivations, let us w.l.o.g. assume a stochastic process with a single output dimension, and furthermore that the predictions are given by their moments, where $\mu^{\mathrm{e}}(\underline{\boldsymbol{x}})$ is the predicted mean and $(\sigma^{\mathrm{e}}(\underline{\boldsymbol{x}}))^2$ is the predicted variance for input $\underline{\boldsymbol{x}}$. For simplicity, we denote the predicted mean and variance as $\mu(\underline{\boldsymbol{x}})$ and $\sigma^2(\underline{\boldsymbol{x}})$, omitting the superscript e. To derive calibration measures for the predicted distributions, following the principles in [24], we first define the error as $\boldsymbol{e} = \boldsymbol{y} - \mu(\underline{\boldsymbol{x}})$. Using the law of total variance, the error variance is given by

$$\begin{aligned}\mathrm{Var}_{p(e)}\{\boldsymbol{e}\} = \mathrm{E}_{p(\underline{x})}\big\{\mathrm{Var}_{p(e)}\{\boldsymbol{e} \mid \underline{\boldsymbol{x}}\}\big\} \\ + \mathrm{Var}_{p(\underline{x})}\big\{\mathrm{E}_{p(e)}\{\boldsymbol{e} \mid \underline{\boldsymbol{x}}\}\big\} \ .\end{aligned} \tag{7}$$

The left-hand side can be rewritten as $\mathrm{Var}_{p(e)}\{\boldsymbol{e}\} = \mathrm{E}_{p(e)}\{\boldsymbol{e}^2\} - \mathrm{E}^2_{p(e)}\{\boldsymbol{e}\}$. If we assume that errors are unbiased, we obtain $\mathrm{E}_{p(e)}\{\boldsymbol{e} \mid \underline{\boldsymbol{x}}\} = 0$ and $\mathrm{E}_{p(e)}\{\boldsymbol{e}\} = 0$. Furthermore, the conditional variance can then be rewritten as

$$\begin{aligned}\mathrm{Var}_{p(e)}\{\boldsymbol{e} \mid \underline{\boldsymbol{x}}\} =& \mathrm{E}\Big\{(\boldsymbol{y} - \mu(\underline{\boldsymbol{x}}) - \mathrm{E}\{\boldsymbol{y} - \mu(\underline{\boldsymbol{x}})\})^2 \Big| \underline{\boldsymbol{x}}\Big\} \\ =& \mathrm{E}\Big\{(\boldsymbol{y} - \mu(\underline{\boldsymbol{x}}))^2 \Big| \underline{\boldsymbol{x}}\Big\} = \sigma^2(\underline{\boldsymbol{x}}) \ .\end{aligned}$$

Thus, (7) simplifies to

$$\mathrm{E}_{p(e)}\big\{\boldsymbol{e}^2\big\} = \mathrm{E}_{p(\underline{x})}\big\{\sigma^2(\underline{\boldsymbol{x}})\big\} \ . \tag{8}$$

Here, $\mathrm{E}_{p(e)}\big\{\boldsymbol{e}^2\big\}$ is the MSE, and $\mathrm{E}_{p(\underline{x})}\big\{\sigma^2(\underline{\boldsymbol{x}})\big\}$ is the expected predicted variance over the input space $\underline{x}$. Therefore, (8) suggests a procedure in which the calibration of the predicted distributions is assessed by comparing the MSE with the expected predicted variance. Since only realizations of the stochastic process are available, both expectations are replaced by their sample approximations, leading to

$$0 = \frac{1}{N_{\mathrm{Test}}} \sum_{n=1}^{N_{\mathrm{Test}}} (y_n - \mu(\underline{x}_n))^2 - \frac{1}{N_{\mathrm{Test}}} \sum_{n=1}^{N_{\mathrm{Test}}} \sigma^2(\underline{x}_n) \ ,$$

where N_{Test} is the number of test data points, $\underline{x}_n$ is the n-th input data point, and y_n is the corresponding output data point. Therefore, a general approach for assessing the calibration of the predicted distributions is to compare the (sample-based) MSE with the mean predicted variance; if both are equal, the predictions are considered well-calibrated. In the following, we will discuss different measures that are based on this idea.

B. Calibration Measures from Literature

The UCE [14] assesses the calibration of *univariate* predictions of regression models, where the differences between the MSE and the mean predicted variance are summed over S bins of partitioned test data. Binning partitions the test data into more homogeneous groups to localize calibration assessment and prevent the over- or underestimation of errors that may cancel each other out when aggregated globally [25][1]; therefore, binning is helpful in preventing cancellation of biases by large predicted variances. The UCE is defined as [14]

$$\mathrm{UCE} = \sum_{s=1}^{S} \frac{|\mathcal{B}_s|}{N_{\mathrm{Test}}} |\mathrm{MSE}(\mathcal{B}_s) - \mathrm{MV}(\mathcal{B}_s)| \ , \tag{9}$$

where $\mathcal{B}_s$ is the set of indices of test data points contained within the s-th bin, and $|\mathcal{B}_s|$ is the number of test data points within the s-th bin. The MSE and mean variance (MV) per bin are calculated as $\mathrm{MSE}(\mathcal{B}_s) = {}^1\!/_{|\mathcal{B}_s|} \sum_{i \in \mathcal{B}_s} (y_i - \mu_i^{\mathrm{e}})^2$, and $\mathrm{MV}(\mathcal{B}_s) = {}^1\!/_{|\mathcal{B}_s|} \sum_{i \in \mathcal{B}_s} (\sigma_i^{\mathrm{e}})^2$, respectively, where μ_i^{e} is the predicted mean, σ_i^{e} is the predicted standard deviation, and y_i is the observed output of the i-th test data point in the s-th bin.

Instead of absolute differences such as the UCE, [15] proposed the ENCE, which is a normalized measure for *univariate* predictions, using the root MSE and the root MV. It is defined as [15]

$$\mathrm{ENCE} = \frac{1}{S} \sum_{s=1}^{S} \frac{|\mathrm{RMSE}(\mathcal{B}_s) - \mathrm{RMV}(\mathcal{B}_s)|}{|\mathrm{RMV}(\mathcal{B}_s)|} \ . \tag{10}$$

Both the UCE and ENCE are only defined for univariate predictions. To overcome this limitation, [4] introduced the QCE to assess the calibration of multivariate predictions. It is defined as [4]

$$\begin{aligned}\mathrm{QCE}(\tau) &= \sum_{s=1}^{S} \frac{|\mathcal{B}_s|}{N_{\mathrm{Test}}} |\mathrm{freq}(\mathcal{B}_s) - \tau| \\ \mathrm{freq}(\mathcal{B}_s) &= \frac{1}{|\mathcal{B}_s|} \sum_{i \in \mathcal{B}_s} \mathbb{1}\big(d^2_{\mathrm{M},i} \leq a_\tau\big) \ ,\end{aligned}$$

where τ is the selected quantile level of interest, $\mathbb{1}(\cdot)$ is the indicator function, $a_\tau = F^{-1}_{\chi^2_k}(\tau)$ is the inverse CDF of the chi-square distribution, and k are the degrees of freedom. Similar to the ANEES, the QCE relies on Assumption 1—that the squared Mahalanobis distances $d^2_{\mathrm{M},i}$ within a bin of the predicted multivariate normal distributions follow a chi-square distribution. To check this assumption, the QCE compares the observed frequency of the squared Mahalanobis distances with the selected quantile of the

[1] E.g., the average of the errors $[-2, -2, 2, 2]$ is 0 (appears well-calibrated), but binning reveals systematic biases: The average for Bin 1 is -2 and the average for Bin 2 is 2. The absolute average over the bins is 2, which exposes the miscalibration.

chi-square distribution. To avoid relying on a single quantile level, [4] also proposed the mean QCE, defined as

$$\overline{\mathrm{QCE}} = \mathrm{E}\{\mathrm{QCE}(\tau)\} \approx \frac{1}{Q}\sum_{q=1}^{Q}\mathrm{QCE}(\tau_q) \ , \qquad (11)$$

where τ_q is the q-th quantile of interest, and Q is the number of quantiles.

Note that the UCE, ENCE, and QCE are designed for regression problems and are zero in the case of perfect calibration. However, by inserting the estimation error from Sec. III into the definitions of these measures, they can also be used to assess the credibility of estimators. This results in comparing the MSE and variance of the estimates, or, in the case of the QCE, the observed frequency of the squared Mahalanobis distances with a specific quantile of the chi-square distribution, which is similar to the ANEES and its alternatives. However, unlike credibility measures such as the ANEES, MNRE, or NCI, the presented calibration measures use binning strategies to partition the test data.

Additional methods to assess the quality of predictions include proper scoring rules such as the average negative log-likelihood (ANLL) of the predicted distributions. In the case of normally distributed predictions, the latter reads

$$\begin{aligned}\mathrm{ANLL} &= -\frac{1}{N_{\mathrm{Test}}}\sum_{n=1}^{N_{\mathrm{Test}}}\ln\Big(p(\underline{y}_n \,|\, \underline{x}_n, \mathcal{D})\Big) \\ &= \frac{1}{2N_{\mathrm{Test}}}\sum_{n=1}^{N_{\mathrm{Test}}}\Bigg(N_y\ln(2\pi)+\ln\Big(\det\big(\mathbf{C}_n^{\mathrm{e}}(\underline{x}_n)\big)\Big) \\ &\quad +\Big(\underline{y}_n-\underline{\mu}_n^{\mathrm{e}}(\underline{x}_n)\Big)^{\top}(\mathbf{C}_n^{\mathrm{e}}(\underline{x}_n))^{-1}\Big(\underline{y}_n-\underline{\mu}_n^{\mathrm{e}}(\underline{x}_n)\Big)\Bigg). \end{aligned} \qquad (12)$$

However, the ANLL has the disadvantage that there is no clearly defined best value, whereas UCE, ENCE, and QCE are zero if the predictions are perfectly calibrated. Furthermore, ANEES can serve as a calibration measure by averaging the Mahalanobis distances per test prediction. However, its drawbacks remain.

V. New Calibration Measures for Multivariate Predictions

Building up on the ideas of state-of-the-art measures for univariate predictions, we now propose an extended calibration measure for multivariate regression model predictions, using the same principles as in the derivation of the UCE.

Analogous to (7) and (8), we start with the law of total covariance, which, after inserting the assumption of unbiased predictions, simplifies to

$$\mathrm{E}\big\{\underline{\boldsymbol{e}}\underline{\boldsymbol{e}}^{\top}\big\} = \mathrm{E}_{p(\underline{\boldsymbol{x}})}\{\mathbf{C}(\underline{\boldsymbol{x}})\} \ . \qquad (13)$$

Here, $\underline{\boldsymbol{e}} = \underline{\boldsymbol{y}} - \underline{\mu}^{\mathrm{e}}(\underline{\boldsymbol{x}})$ is the estimation error, $\mathbf{C}(\underline{\boldsymbol{x}})$ is the predicted covariance matrix, and $\underline{\mu}^{\mathrm{e}}(\underline{\boldsymbol{x}})$ is the predicted mean vector for the input $\underline{\boldsymbol{x}}$. As in the univariate case, the MSE matrix and the expected predicted covariance matrix can be approximated using sample approximations.

To directly compare the difference between the averaged predicted covariance matrices and the approximated MSE matrix, we use an arbitrary matrix norm $||\cdot||$, combined with the binning scheme of the calibration measures, resulting in

$$\mathrm{GUCE} = \sum_{s=1}^{S}\frac{|\mathcal{B}_s|}{N_{\mathrm{Test}}}\big\|\bar{\boldsymbol{\Sigma}}(\mathcal{B}_s) - \bar{\mathbf{C}}^{\mathrm{e}}(\mathcal{B}_s)\big\| \ , \qquad (14)$$

which we call the generalized UCE (GUCE). Note that when only one bin is used, i.e., $S = 1$, the GUCE reduces to a numerator similar to that of the MNRE. Otherwise, the sample approximations are calculated using

$$\bar{\boldsymbol{\Sigma}}(\mathcal{B}_s) = \frac{1}{|\mathcal{B}_s|}\sum_{n\in\mathcal{B}_s}\underline{e}_n\underline{e}_n^{\top} \ , \quad \bar{\mathbf{C}}^{\mathrm{e}}(\mathcal{B}_s) = \frac{1}{|\mathcal{B}_s|}\sum_{n\in\mathcal{B}_s}\mathbf{C}(\underline{x}_n) \ .$$

This derivation is generic in the sense that other credibility measures can also be applied as calibration measures by replacing their MSE matrix $\boldsymbol{\Sigma}$ and their predicted covariance matrix $\mathbf{C}^{\mathrm{e}}$ with the sample approximations for (13). and applying a binning scheme. For example, using the idea of the MNRE, the GUCE can be normalized by the norms of $\bar{\boldsymbol{\Sigma}}$ and $\bar{\mathbf{C}}^{\mathrm{e}}$, leading to the normalized GUCE (NGUCE)

$$\mathrm{NGUCE} = \sum_{s=1}^{S}\frac{|\mathcal{B}_s|}{N_{\mathrm{Test}}}\frac{\big\|\bar{\boldsymbol{\Sigma}}(\mathcal{B}_s) - \bar{\mathbf{C}}^{\mathrm{e}}(\mathcal{B}_s)\big\|}{\big\|\bar{\boldsymbol{\Sigma}}(\mathcal{B}_s)\big\| + \big\|\bar{\mathbf{C}}^{\mathrm{e}}(\mathcal{B}_s)\big\|} \ . \qquad (15)$$

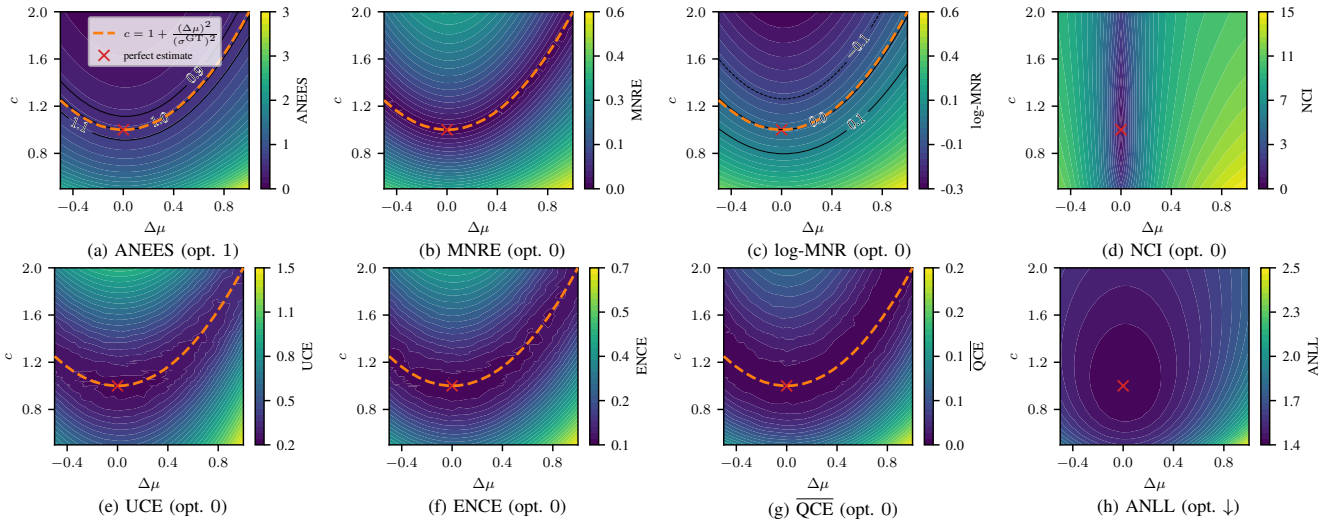
VI. Experiments

In this section, we numerically compare the presented measures in both estimation and regression settings. For this, where appropriate, we also apply credibility measures to the regression setting and calibration measures to estimation, to develop a unified picture of the strengths and limitations of the measures.

A. Comparison as Credibility Measure

We compare the behavior of the credibility measures for an estimate $\boldsymbol{x}^{\mathrm{e}}$ of the ground truth $\boldsymbol{x}^{\mathrm{GT}} \sim \mathcal{N}(x; 0, 1)$ that is slightly scaled and shifted, i.e., $\boldsymbol{x}^{\mathrm{e}} \sim \mathcal{N}\Big(x; \mu^{\mathrm{e}}, \big(\sigma^{\mathrm{e}}\big)^2\Big)$, with mean shift $\Delta\mu = \mu^{\mathrm{e}}$ and variance scaling factor $c = (\sigma^{\mathrm{e}})^2$. From the ground truth, we generate $10\,000$ random samples, which are used to estimate the MSE matrix $\boldsymbol{\Sigma}$ for the credibility measures. To visualize the characteristics of the measures, we consider multiple shifts of the mean $\Delta\mu \in [-0.5, 1]$ and multiple scaling factors $c \in [0.5, 2]$. For comparison, we consider the ANEES (1), the MNRE (4), the log-MNR (5), and the NCI (6), and the ANLL (12). Furthermore, we use the UCE (9), the ENCE (10), and the $\overline{\mathrm{QCE}}$ (11) (with ten equally spaced quantile levels in $(0, 1)$) as credibility measures, with 40 test samples per bin.

The results are displayed in Fig. 3. The parabola drawn in Fig. 3 depicts the curve where condition (3) holds, and the ANEES is one even for biased estimates. Accordingly, the ANEES reports values close to one in large areas even for largely biased estimates. The MNRE, the log-MNR, and the $\overline{\mathrm{QCE}}$ are also unable to assess the bias credibility of the estimated distribution and behave similarly to the ANEES. The binning of the UCE and ENCE measures leads to a slightly more robust behavior, with gradients pointing to the ground truth estimate, which is not the case for the ANEES, MNRE, and log-MNR measures. However, neither the UCE nor the ENCE reaches its optimal value due to the used binning.



(a) ANEES (opt. 1) (b) MNRE (opt. 0) (c) log-MNR (opt. 0) (d) NCI (opt. 0)

(e) UCE (opt. 0) (f) ENCE (opt. 0) (g) $\overline{\text{QCE}}$ (opt. 0) (h) ANLL (opt. ↓)

Fig. 3: Comparison of the ANEES, MNRE, log-MNR, and NCI measures, as well as the UCE, ENCE, and QCE for different shifts of the mean $\Delta\mu$ and different scaling factors c of the variance. Note that the estimate is perfect for zero bias and exact variance (i.e., $c = 1$ and $\Delta\mu = 0$). The optimal value for each measure is given in parentheses in the sub-captions.

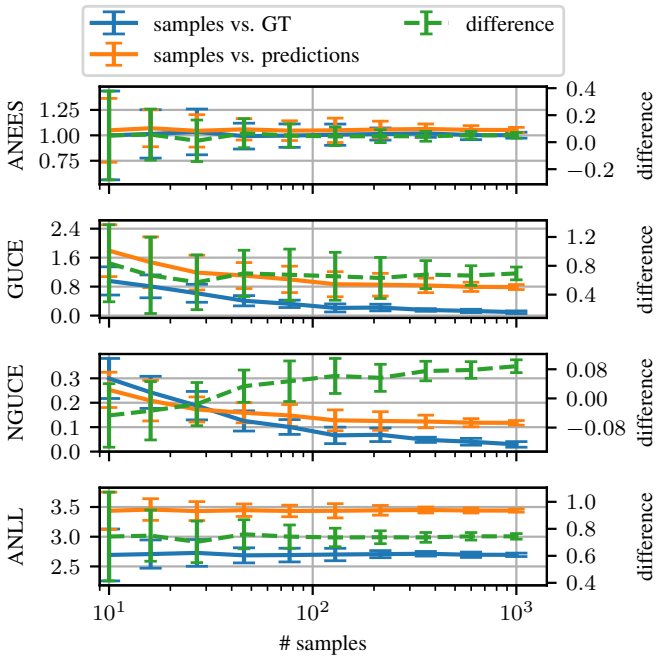


Fig. 4: Comparison of calibration measures for a synthetic regression example with known ground truth. The figure shows the behavior of the ANEES, UCE, GUCE, NGUCE, and ANLL as the number of ground truth samples increases. Each bar in the plot shows the average and the standard deviation over 20 runs. Lower values indicate better calibration, except for the ANEES, where a value close to one is optimal.

The NCI measure behaves differently and has its minimum at the perfect estimate; however, the NCI assumes that the bias is provided by the estimator, which is typically not the case in practice. The characteristics of the ANLL behave as expected, with a minimum when the estimate matches the ground truth. However, the value at this point cannot be used as an absolute measure of calibration and therefore must be compared with the ANLL of multiple estimates.

B. Comparison as Calibration Measure in Regression

To compare calibration measures, we consider a synthetic regression example, in which the ground truth is given by $\mathcal{N}(\underline{y}; \underline{\mu}^{\text{GT}}(x), \mathbf{C}^{\text{GT}}(x))$ with $\underline{\mu}^{\text{GT}}(x) = \begin{bmatrix} x & x^3 \end{bmatrix}^\top$, and x being the input variable. Each prediction is normally distributed with mean $\underline{\mu}^{\text{GT}}(x) + \Delta\underline{\mu}$ and $\Delta\underline{\mu} = \begin{bmatrix} 0 & 1.5 \end{bmatrix}^\top$. The covariance matrices for the ground truth and the regression model are given by

$$\mathbf{C}^{\text{GT}}(x) = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \quad \mathbf{C}^{\text{e}}(x) = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 3 \end{bmatrix},$$

respectively. Note that the choice of which variable contains the bias or incorrect variance is irrelevant when error structures are similar. As demonstrated in univariate examples, higher variances can compensate for biased predictions, allowing for the construction of arbitrary situations with such error structures. This makes it an appropriate test case for evaluating calibration measure properties. The calibration measures are evaluated using different numbers of ground truth samples, ranging from 10 to 1000, with equally spaced x values in $[-1, 1]$. Each experiment is repeated 20 times to account for the MC error when assessing the accuracy of the measures. As calibration measures, we consider the ANEES (1), the UCE (9), the GUCE (14), the NGUCE (15), and the ANLL (12). The Frobenius norm is used as the matrix norm. In preliminary experiments, we also tested the 1-norm and spectral norm, but observed no significant differences in the results.

The results are shown in Fig. 4. The ANEES is close to one, and therefore, unlike all other measures, is unable to detect the biased mean and scaled covariance matrix of the regression model properly. When comparing the ground truth realizations (test data) to the ground truth process rather than to the predictive distributions, one can see that, as desired, all measures approach their optimal value as the number of samples increases. When we compare the differences in the measures fed with predictions from the regression model vs. the ground truth process, we observe that these differences are always

positive for the GUCE and ANLL. This indicates that the regression model is always considered worse than the ground truth, which is a desirable property for a calibration measure. As expected, for low sample sizes, all measures exhibit high variance, which decreases as the sample size increases.

C. Discussion

Our experiments numerically demonstrated that the ANEES is not able to detect biased means and incorrect covariance matrices of the estimated distributions, and fails in specific cases, in particular if condition (3) is fulfilled. Moreover, also alternatives such as the MNRE and log-MNR suffer from the same weaknesses and are not able to detect the bias in the mean and the erroneously estimated covariance matrix. Alternatives such as the NCI require that the bias is provided by an estimator. If this is the case, the NCI works well. Applying calibration measures for credibility testing of estimates yields good results except for the $\overline{\text{QCE}}$. In particular, they show preferable behavior compared to the ANEES and its alternatives, since their gradients point to the ground truth estimate. We think that this is primarily due to the binning scheme employed in these measures. Our proposed GUCE and NGUCE measures both work well for regression models, and demonstrate robust behavior for examples where the ANEES fails. Furthermore, we showed that the ANLL has good properties, is able to detect biases and incorrect covariance matrices, but has the main disadvantage that there is no clear optimal value, such as for the GUCE.

VII. Conclusion

In this paper, we systematically reviewed the ANEES and its alternatives for assessing the credibility of estimators and calibration of regression models. While the ANEES remains a widely used and valuable measure, our analysis and experiments highlighted its limitations, particularly in detecting bias and miscalibration in certain scenarios. We demonstrated that alternative measures, such as the MNRE, log-MNR, and NCI, address some weaknesses but also suffer from similar limitations.

Building on recent advances, we extended calibration measures to multivariate settings and introduced new measures that unify concepts from estimator credibility and regression calibration. Our comprehensive experiments showed that using a combination of measures provides a more robust and informative assessment of model performance.

We recommend that practitioners employ multiple credibility and calibration measures, rather than relying solely on the ANEES, to ensure a comprehensive evaluation. This helps to avoid misleading conclusions and supports the development of more reliable and trustworthy predictive systems.

References

[1] R. McAllister, Y. Gal, A. Kendall, M. Van Der Wilk, A. Shah, R. Cipolla, and A. Weller, "Concrete problems for autonomous vehicle safety: advantages of Bayesian deep learning," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, ser. IJCAI'17. Melbourne, Australia: AAAI Press, 2017, pp. 4745–4753.

[2] E. Begoli, T. Bhattacharya, and D. Kusnezov, "The need for uncertainty quantification in machine-assisted medical decision making," *Nature Machine Intelligence*, vol. 1, no. 1, pp. 20–23, Jan. 2019.

[3] V. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 2018, pp. 2796–2804.

[4] F. Küppers, J. Schneider, and A. Haselhoff, "Parametric and multivariate uncertainty calibration for regression and object detection," in *Computer Vision – ECCV 2022 Workshops*, 2023, vol. 13805, pp. 426–442.

[5] Y. Bar-Shalom and K. Birmiwal, "Consistency and robustness of PDAF for target tracking in cluttered environments," *Automatica*, vol. 19, no. 4, pp. 431–437, 1983.

[6] Z. Chen, C. Heckman, S. Julier, and N. Ahmed, "Weak in the NEES?: Auto-tuning Kalman filters with Bayesian optimization," in *Proceedings of the 21st International Conference on Information Fusion (Fusion 2018)*, 2018, pp. 1072–1079.

[7] M. Walker, M. Reith-Braun, P. Schichtel, M. Knaak, and U. D. Hanebeck, "Identifying trust regions of Bayesian neural networks," in *Proceedings of the combined 2023 IEEE Symposium Sensor Data Fusion and International Conference on Multisensor Fusion and Integration (SDF-MFI 2023)*, Bonn, Germany, Nov. 2023, pp. 1–8.

[8] R. Forsling, B. Noack, and G. Hendeby, "A quarter century of covariance intersection: Correlations still unknown?" *IEEE Control Systems Magazine*, vol. 44, no. 2, pp. 81–105, 2024.

[9] M. Walker, H. Amirkhanian, M. F. Huber, and U. D. Hanebeck, "Trustworthy Bayesian perceptrons," in *Proceedings of the 27th International Conference on Information Fusion (Fusion 2024)*, Venice, Italy, Jul. 2024, pp. 1–8.

[10] X. R. Li, Z. Zhao, and V. P. Jilkov, "Estimator's credibility and its measures," in *Proceedings of the 15th Triennial IFAC World Congress*, Barcelona, Spain, Jul. 2002.

[11] X. R. Li and Z. Zhao, "Measuring estimator's credibility: Noncredibility index," in *Proceedings of the 9th International Conference on Information Fusion (Fusion 2006)*, Florence, Italy, 2006, pp. 1–8.

[12] ——, "Testing estimator's credibility - part I: Tests for MSE," in *Proceedings of the 9th International Conference on Information Fusion (Fusion 2006)*, Florence, Italy, 2006, pp. 1–8.

[13] E. Howerton *et al.*, "Context-dependent representation of within- and between-model uncertainty: aggregating probabilistic predictions in infectious disease epidemiology," *Journal of The Royal Society Interface*, vol. 20, no. 198, p. 20220659, Jan. 2023.

[14] M.-H. Laves, S. Ihler, J. F. Fast, L. A. Kahrs, and T. Ortmaier, "Well-calibrated regression uncertainty in medical imaging with deep learning," in *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, Jul. 2020, pp. 393–412.

[15] D. Levi, L. Gispan, N. Giladi, and E. Fetaya, "Evaluating and calibrating uncertainty prediction in regression tasks," *Sensors*, vol. 22, no. 15, p. 5540, 2022.

[16] M. Walker and U. D. Hanebeck, "Multi-scale uncertainty calibration testing for Bayesian neural networks using ball trees," in *Proceedings of the 2024 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2024)*, Plzeň, Czech Republic, Sep. 2024, pp. 1–7.

[17] M. Walker, P. S. Bien, and U. D. Hanebeck, "Voronoi trust regions for local calibration testing in supervised machine learning models," in *Proceedings of the 2024 IEEE Symposium Sensor Data Fusion: Trends, solutions, applications (SDF 2024)*, Bonn, Germany, Nov. 2024, pp. 1–8.

[18] M. Walker, M. Reith-Braun, and U. D. Hanebeck, "Local calibration testing in supervised machine learning models using input space kernels," in *Proceedings of the 28th International Conference on Information Fusion (Fusion 2025)*, Rio de Janeiro, Brazil, Jul. 2025, pp. 1–8.

[19] Y. Bar-Shalom, X.-R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation*. New York, USA: John Wiley & Sons, Inc., 2001.

[20] Y. Dodge, *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York, 2008.

[21] S. J. Press, "Linear combinations of non-central chi-square variates," *The Annals of Mathematical Statistics*, vol. 37, no. 2, pp. 480–487, Apr. 1966.

[22] J. Duník, O. Straka, O. Kost, S. Tang, T. Imbiriba, and P. Closas, "Noise identification for data-augmented physics-based state-space models," in *2024 IEEE Workshop on Signal Processing Systems (SiPS)*, 2024, pp. 101–106.

[23] F. Giraldo-Grueso, A. A. Popov, and R. Zanetti, "Gaussian mixture-based point mass filtering," in *Proceedings of the 27th International Conference on Information Fusion (Fusion 2024)*, 2024, pp. 1–8.

[24] P. Pernot, "Negative impact of heavy-tailed uncertainty and error distributions on the reliability of calibration statistics for machine learning regression tasks," *arXiv:2402.10043*, Aug. 2024.

[25] N. Posocco and A. Bonnefoy, "Estimating expected calibration errors," in *Artificial Neural Networks and Machine Learning – ICANN 2021*, I. Farkaš, P. Masulli, S. Otte, and S. Wermter, Eds. Cham: Springer International Publishing, 2021, vol. 12894, pp. 139–150.