CALIBRATION TESTS BEYOND CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most supervised machine learning tasks are subject to irreducible prediction errors. Probabilistic predictive models address this limitation by providing probability distributions that represent a belief over plausible targets, rather than point estimates. Such models can be a valuable tool in decision-making under uncertainty, provided that the model output is meaningful and interpretable. Calibrated models guarantee that the probabilistic predictions are neither over- nor under-confident. In the machine learning literature, different measures and statistical tests have been proposed and studied for evaluating the calibration of classification models. For regression problems, however, research has been focused on a weaker condition of calibration based on predicted quantiles for real-valued targets. In this paper, we propose the first framework that unifies calibration evaluation and tests for general probabilistic predictive models. It applies to any such model, including classification and regression models of arbitrary dimension. Furthermore, the framework generalizes existing measures and provides a more intuitive reformulation of a recently proposed framework for calibration in multi-class classification.

1 INTRODUCTION

We consider the general problem of modelling the relationship between a feature X and a target Y in a probabilistic setting, i.e., we focus on models that approximate the conditional probability distribution $\mathbb{P}(Y|X)$ of target Y for given feature X. The use of probabilistic models that output a probability distribution instead of a point estimate demands guarantees on the predictions beyond accuracy, enabling meaningful and interpretable predicted uncertainties. One such statistical guarantee is calibration, which has been studied extensively in metereological and statistical literature (DeGroot & Fienberg, 1983; Murphy & Winkler, 1977).

A calibrated model ensures that almost every prediction matches the conditional distribution of targets given this prediction. Loosely speaking, in a classification setting a predicted distribution of the model is called calibrated (or reliable), if the empirically observed frequencies of the different classes match the predictions in the long run, if the same class probabilities would be predicted repeatedly. A classical example is a weather forecaster who predicts each day if it is going to rain on the next day. If she predicts rain with probability 60% for a long series of days, her forecasting model is calibrated *for predictions of 60%* if it actually rains on 60% of these days.

If this property holds for almost every probability distribution that the model outputs, then the model is considered to be calibrated. Calibration is an appealing property of a probabilistic model since it provides safety guarantees on the predicted distributions even in the common case when the model does not predict the true distributions $\mathbb{P}(Y|X)$. Calibration, however, does not guarantee accuracy (or refinement)—a model that always predicts the marginal probabilities of each class is calibrated but probably inaccurate and of limited use. On the other hand, accuracy does not imply calibration either since the predictions of an accurate model can be too over-confident and hence miscalibrated, as observed, e.g., for deep neural networks (Guo et al., 2017).

In the field of machine learning, calibration has been studied mainly for classification problems (Bröcker, 2009; Guo et al., 2017; Kull et al., 2017; 2019; Kumar et al., 2018; Platt, 2000; Vaicenavicius et al., 2019; Widmann et al., 2019; Zadrozny, 2002) and for quantiles and confidence intervals of models for regression problems with real-valued targets (Fasiolo et al., 2020; Ho & Lee, 2005; Kuleshov et al., 2018; Rueda et al., 2006; Taillardat et al., 2016). In our work, however, we do not restrict ourselves to these problem settings but instead consider calibration for arbitrary predictive models. Thus, we generalize the common notion of calibration as:

Definition 1. Consider a model $P_X := P(Y|X)$ of a conditional probability distribution $\mathbb{P}(Y|X)$. Then model P is said to be calibrated if and only if

$$\mathbb{P}(Y|P_X) = P_X \qquad \text{almost surely.} \tag{1}$$

If P is a classification model, Definition 1 coincides with the notion of (multi-class) calibration by Bröcker (2009); Kull et al. (2019); Vaicenavicius et al. (2019). Alternatively, in classification some authors (Guo et al., 2017; Kumar et al., 2018; Naeini et al., 2015) study the strictly weaker property of confidence calibration (Kull et al., 2019), which only requires

$$\mathbb{P}\left(Y = y | \arg\max P_X = y\right) = \max P_X \qquad \text{almost surely} \tag{2}$$

to hold for every target y.

For real-valued targets, Definition 1 coincides with the so-called distribution-level calibration by Song et al. (2019). Distribution-level calibration implies that the predicted quantiles are calibrated, i.e., the outcomes for all real-valued predictions of the, e.g., 75% quantile are actually below the predicted quantile with 75% probability (Song et al., 2019, Theorem 1). Conversely, although quantile-based calibration is a common approach for real-valued regression problems (Fasiolo et al., 2020; Ho & Lee, 2005; Kuleshov et al., 2018; Rueda et al., 2006; Taillardat et al., 2016), it provides weaker guarantees on the predictions. For instance, the linear regression model in Fig. 1 empirically shows quantiles that appear close to being calibrated albeit being uncalibrated according to Definition 1.

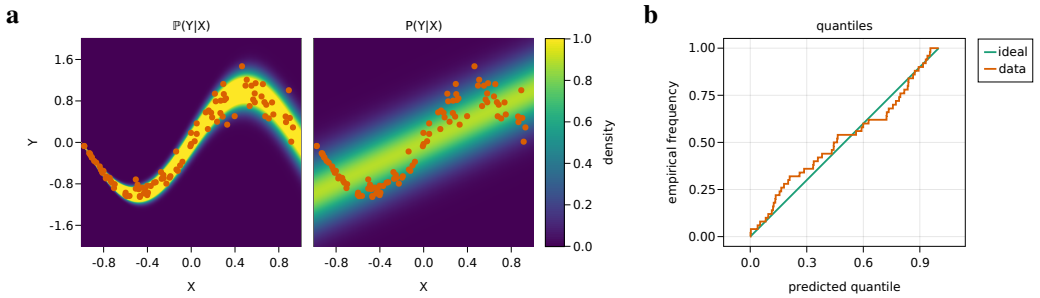

Figure 1: Illustration of a conditional distribution $\mathbb{P}(Y|X)$ with scalar feature and target. We consider a Gaussian predictive model P, obtained by ordinary least squares regression with 100 training data points (orange dots). Empirically the predicted quantiles on 50 validation data points appear close to being calibrated, although model P is uncalibrated according to Definition 1. Using the framework in this paper, on the same validation data a statistical test allows us to reject the null hypothesis that model P is calibrated at a significance level of $\alpha = 0.05$ ($p < 0.025$). See Appendix A.1 for details.

Figure 1 also raises the question of how to assess calibration for general target spaces in the sense of Definition 1, without having to rely on visual inspection. In classification, measures of calibration such as the commonly used expected calibration error (ECE) (Guo et al., 2017; Kull et al., 2019; Naeini et al., 2015; Vaicenavicius et al., 2019) and the maximum calibration error (MCE) (Naeini et al., 2015) try to capture the average and maximal discrepancy between the distributions on the left hand side and the right hand side of Eq. (1) or Eq. (2), respectively. These measures can be generalized to other target spaces (see Definition B.1), but unfortunately estimating these calibration errors from observations of features and corresponding targets is problematic. Typically, the predictions are different for (almost) all observations, and hence estimation of the conditional probability $\mathbb{P}\left(Y|P_X\right)$, which is needed in the estimation of ECE and MCE, is challenging even for low-dimensional target spaces and usually leads to biased and inconsistent estimators (Vaicenavicius et al., 2019).

Kernel-based calibration errors such as the maximum mean calibration error (MMCE) (Kumar et al., 2018) and the kernel calibration error (KCE) (Widmann et al., 2019) for confidence and multi-class calibration, respectively, can be estimated without first estimating the conditional probability and hence avoid this issue. They are defined as the expected value of a weighted sum of the differences of the left and right hand side of Eq. (1) for each class, where the weights are given as a function of the predictions (of all classes) and chosen such that the calibration error is maximized. A reformulation with matrix-valued kernels (Widmann et al., 2019) yields unbiased and differentiable estimators without explicit dependence on $\mathbb{P}(Y|P_X)$, which simplifies the estimation and allows to explicitly

account for calibration in the training objective (Kumar et al., 2018). Additionally, the kernel-based framework allows the derivation of reliable statistical hypothesis tests for calibration in multi-class classification (Widmann et al., 2019).

However, both the construction as a weighted difference of the class-wise distributions in Eq. (1) and the reformulation with matrix-valued kernels require finite target spaces and hence cannot be applied to regression problems. To be able to deal with general target spaces, we present a new and more general framework of calibration errors without these limitations.

Our framework can be used to reason about and test for calibration of *any probabilistic predictive model*. As explained above, this is in stark contrast with existing methods that are restricted to simple output distributions, such as classification and *scalar-valued* regression problems. A *key contribution* of this paper is a new framework that is applicable to *multivariate* regression, as well as situations when the output is of a different (e.g. discrete ordinal) or more complex (e.g, graph-structured) type, with clear practical implications.

Within this framework a KCE for general target spaces is obtained. We want to highlight that for multi-class classification problems its formulation is more intuitive and simpler to use than the measure proposed by Widmann et al. (2019) based on matrix-valued kernels. To ease the application of the KCE we derive several estimators of the KCE with subquadratic sample complexity and their asymptotic properties in tests for calibrated models, which improve on existing estimators and tests in the two-sample test literature by exploiting the special structure of the calibration framework. Using the proposed framework, we numerically evaluate the calibration of neural network models and ensembles of such models.[1]

2 Calibration error: A general framework

In classification, the distributions on the left and right hand side of Eq. (1) can be interpreted as vectors in the probability simplex. Hence ultimately the distance measure for ECE and MCE (see Definition B.1) can be chosen as a distance measure of real-valued vectors. The total variation, Euclidean, and squared Euclidean distances are common choices (Guo et al., 2017; Kull et al., 2019; Vaicenavicius et al., 2019). However, in a general setting measuring the discrepancy between $\mathbb{P}(Y|P_X)$ and P_X cannot necessarily be reduced to measuring distances between vectors. The conditional distribution $\mathbb{P}(Y|P_X)$ can be arbitrarily complex, even if the predicted distributions are restricted to a simple class of distributions that can be represented as real-valued vectors. Hence in general we have to resort to dedicated distance measures of probability distributions.

Additionally, the estimation of conditional distributions $\mathbb{P}(Y|P_X)$ is challenging, even more so than in the restricted case of classification, since in general these distributions can be arbitrarily complex. To circumvent this problem, we propose to use the following construction: We define a random variable $Z_X \sim P_X$ obtained from the predictive model and study the discrepancy between the *joint* distributions of the two pairs of random variables (P_X, Y) and (P_X, Z_X), respectively, instead of the discrepancy between the *conditional* distributions $\mathbb{P}(Y|P_X)$ and P_X. Since

$$(P_X, Y) \stackrel{d}{=} (P_X, Z_X) \qquad \text{if and only if} \qquad \mathbb{P}(Y|P_X) = P_X \quad \text{almost surely},$$

model P is calibrated if and only if the distributions of (P_X, Y) and (P_X, Z_X) are equal.

The random variable pairs (P_X, Y) and (P_X, Z_X) take values in the product space $\mathcal{P} \times \mathcal{Y}$, where $\mathcal{P}$ is the space of predicted distributions P_X and $\mathcal{Y}$ is the space of targets Y. For instance, in classification, $\mathcal{P}$ could be the probability simplex and $\mathcal{Y}$ the set of all class labels, whereas in the case of Gaussian predictive models for scalar targets $\mathcal{P}$ could be the space of normal distributions and $\mathcal{Y}$ be $\mathbb{R}$.

The study of the joint distributions of (P_X, Y) and (P_X, Z_X) motivates the definition of a generally applicable calibration error as an integral probability metric (Müller, 1997; Sriperumbudur et al., 2009; 2012) between these distributions. In contrast to common f-divergences such as the Kullback-Leibler divergence, integral probability metrics do not require that one distribution is absolutely continuous with respect to the other, which cannot be guaranteed in general.

[1] The code of the experiments is available at `https://github.com/anonymous/forpeerreview`.

Definition 2. Let $\mathcal{Y}$ denote the space of targets Y, and $\mathcal{P}$ the space of predicted distributions P_X. We define the calibration error with respect to a space of functions $\mathcal{F}$ of the form $f\colon \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$ as

$$\mathrm{CE}_{\mathcal{F}} \coloneqq \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{P_X,Y} f(P_X, Y) - \mathbb{E}_{P_X,Z_X} f(P_X, Z_X) \right|. \tag{3}$$

By construction, if model P is calibrated, then $\mathrm{CE}_{\mathcal{F}} = 0$ regardless of the choice of $\mathcal{F}$. However, the converse statement is not true for arbitrary function spaces $\mathcal{F}$. From the theory of integral probability metrics (see, e.g., Müller, 1997; Sriperumbudur et al., 2009; 2012), we know that for certain choices of $\mathcal{F}$ the calibration error in Eq. (3) is a well-known metric on the product space $\mathcal{P} \times \mathcal{Y}$, which implies that $\mathrm{CE}_{\mathcal{F}} = 0$ if and only if model P is calibrated. Prominent examples include the maximum mean discrepancy[2] (MMD) (Gretton et al., 2007), the total variation distance, the Kantorovich distance, and the Dudley metric (Dudley, 1989, p. 310).

As pointed out above, Definition 2 is a generalization of the definition for multi-class classification proposed by Widmann et al. (2019)—which is based on vector-valued functions and only applicable to finite target spaces—to *any probabilistic predictive model*. In Appendix E we show this explicitly and discuss the special case of classification problems in more detail. Previous results (Widmann et al., 2019) imply that in classification MMCE and, for common distance measures $d(\cdot,\cdot)$ such as the total variation and squared Euclidean distance, ECE_d and MCE_d are special cases of $\mathrm{CE}_{\mathcal{F}}$. In Appendix G we show that our framework also covers natural extensions of ECE_d and MCE_d to countably infinite discrete target spaces, which to our knowledge have not been studied before and occur, e.g., in Poisson regression.

The literature of integral probability metrics suggests that we can resort to estimating $\mathrm{CE}_{\mathcal{F}}$ from i.i.d. samples from the distributions of (P_X, Y) and (P_X, Z_X). For the MMD, the Kantorovich distance, and the Dudley metric tractable strongly consistent empirical estimators exist (Sriperumbudur et al., 2012). Here the empirical estimator for the MMD is particularly appealing since compared with the other estimators "it is computationally cheaper, the empirical estimate converges at a faster rate to the population value, and the rate of convergence is independent of the dimension d of the space (for $S = \mathbb{R}^d$)" (Sriperumbudur et al. (2012)).

Our specific design of (P_X, Z_X) can be exploited to improve on these estimators. If $\mathbb{E}_{Z_x \sim P_x} f(P_x, Z_x)$ can be evaluated analytically for a fixed prediction P_x, then $\mathrm{CE}_{\mathcal{F}}$ can be estimated empirically with reduced variance by marginalizing out Z_X. Otherwise $\mathbb{E}_{Z_x \sim P_x} f(P_x, Z_x)$ has to be estimated, but in contrast to the common estimators of the integral probability metrics discussed above the artificial construction of Z_X allows us to approximate it by numerical integration methods such as (quasi) Monte Carlo integration or quadrature rules with arbitrarily small error and variance. Monte Carlo integration preserves statistical properties of the estimators such as unbiasedness and consistency.

3 Kernel calibration error

For the remaining parts of the paper we focus on the MMD formulation of $\mathrm{CE}_{\mathcal{F}}$ due to the appealing properties of the common empirical estimator mentioned above. We derive calibration-specific analogues of results for the MMD that exploit the special structure of the distribution of (P_X, Z_X) to improve on existing estimators and tests in the MMD literature. To the best of our knowledge these variance-reduced estimators and tests have not been discussed in the MMD literature.

Let $k\colon (\mathcal{P} \times \mathcal{Y}) \times (\mathcal{P} \times \mathcal{Y}) \to \mathbb{R}$ be a measurable kernel with corresponding reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, and assume that

$$\mathbb{E}_{P_X,Y}\, k^{1/2}\big((P_X, Y), (P_X, Y)\big) < \infty \quad \text{and} \quad \mathbb{E}_{P_X,Z_X}\, k^{1/2}\big((P_X, Z_X), (P_X, Z_X)\big) < \infty.$$

We discuss how such kernels can be constructed in a generic way in Section 3.1 below.

Definition 3. Let $\mathcal{F}_k$ denote the unit ball in $\mathcal{H}$, i.e., $\mathcal{F} \coloneqq \{f \in \mathcal{H} | \|f\|_{\mathcal{H}} \leq 1\}$. Then the kernel calibration error (KCE) with respect to kernel k is defined as

$$\mathrm{KCE}_k \coloneqq \mathrm{CE}_{\mathcal{F}_k} = \sup_{f \in \mathcal{F}_k} \left| \mathbb{E}_{P_X,Y} f(P_X, Y) - \mathbb{E}_{P_X,Z_X} f(P_X, Z_X) \right|.$$

[2] As we discuss in Section 3, the MMD is a metric if and only if the employed kernel is characteristic.

As known from the MMD literature, a more explicit formulation can be given for the squared kernel calibration error $\mathrm{SKCE}_k \coloneqq \mathrm{KCE}_k^2$ (see Lemma B.2). A similar explicit expression for SKCE_k was obtained by Widmann et al. (2019) for the special case of classification problems. However, their expression relies on $\mathcal{Y}$ being finite and is based on matrix-valued kernels over the finite-dimensional probability simplex $\mathcal{P}$. A key difference to the expression in Lemma B.2 is that we instead propose to use real-valued kernels defined on the product space of predictions and targets. This construction is applicable to arbitrary target spaces and does not require $\mathcal{Y}$ to be finite.

3.1 Choice of kernel

The construction of the product space $\mathcal{P} \times \mathcal{Y}$ suggests the use of tensor product kernels $k = k_{\mathcal{P}} \otimes k_{\mathcal{Y}}$, where $k_{\mathcal{P}} \colon \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ and $k_{\mathcal{Y}} \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ are kernels on the spaces of predicted distributions and targets, respectively.[3]

By definition, so-called characteristic kernels guarantee that $\mathrm{KCE} = 0$ if and only if the distributions of (P_X, Y) and (P_X, Z_X) are equal (Fukumizu et al., 2004; 2008). Many common kernels such as the Gaussian and Laplacian kernel on $\mathbb{R}^d$ are characteristic (Fukumizu et al., 2008).[4] Szabó & Sriperumbudur (2018, Theorem 4) showed that a tensor product kernel $k_{\mathcal{P}} \otimes k_{\mathcal{Y}}$ is characteristic if $k_{\mathcal{P}}$ and $k_{\mathcal{Y}}$ are characteristic, continuous, bounded, and translation-invariant kernels on $\mathbb{R}^d$, but the implication does not hold for general characteristic kernels (Szabó & Sriperumbudur, 2018, Example 1). For calibration evaluation, however, it is sufficient to be able to distinguish between the conditional distributions $\mathbb{P}(Y|P_X)$ and $\mathbb{P}(Z_X|P_X) = P_X$. Therefore, in contrast to the regular MMD setting, it is *sufficient that kernel $k_{\mathcal{Y}}$ is characteristic and kernel $k_{\mathcal{P}}$ is non-zero almost surely*, to guarantee that $\mathrm{KCE} = 0$ if and only if model P is calibrated. Thus it is suggestive to construct kernels on general spaces of predicted distributions as

$$k_{\mathcal{P}}(p, p') = \exp\big(-\lambda d_{\mathcal{P}}^{\nu}(p, p')\big), \tag{4}$$

where $d_{\mathcal{P}}(\cdot, \cdot)$ is a metric on $\mathcal{P}$ and $\nu, \lambda > 0$ are kernel hyperparameters. The Wasserstein distance is a widely used metric for distributions from optimal transport theory that allows to lift a ground metric on the target space and possesses many important properties (see, e.g., Peyré & Cuturi, 2018, Chapter 2.4). In general, however, it does not lead to valid kernels $k_{\mathcal{P}}$, apart from the notable exception of elliptically contoured distributions such as normal and Laplace distributions (Peyré & Cuturi, 2018, Chapter 8.3).

In machine learning, common probabilistic predictive models output parameters of distributions such as mean and variance of normal distributions. Naturally these parameterizations give rise to injective mappings $\phi \colon \mathcal{P} \to \mathbb{R}^d$ that can be used to define a Hilbertian metric

$$d_{\mathcal{P}}(p, p') = \|\phi(p) - \phi(p')\|_2.$$

For such metrics, $k_{\mathcal{P}}$ in Eq. (4) is a valid kernel for all $\lambda > 0$ and $\nu \in (0, 2]$ (Berg et al., 1984, Corollary 3.3.3, Proposition 3.2.7). In Appendix D.3 we show that for many mixture models, and hence model ensembles, Hilbertian metrics between model components can be lifted to Hilbertian metrics between mixture models. This construction is a generalization of the Wasserstein-like distance for Gaussian mixture models proposed by Chen et al. (2019); Delon & Desolneux (2019).

3.2 Estimation

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a data set of features and targets which are i.i.d. according to the law of (X, Y). Moreover, for notational brevity, for $(p, y), (p', y') \in \mathcal{P} \times \mathcal{Y}$ we let

$$\begin{aligned} h\big((p, y), (p', y')\big) \coloneqq\ & k\big((p, y), (p', y')\big) - \mathbb{E}_{Z \sim p}\, k\big((p, Z), (p', y')\big) \\ & - \mathbb{E}_{Z' \sim p'}\, k\big((p, y), (p', Z')\big) + \mathbb{E}_{Z \sim p, Z' \sim p'}\, k\big((p, Z), (p', Z')\big). \end{aligned}$$

Note that in contrast to the regular MMD we marginalize out Z and Z'. Similar to the MMD, there exist consistent estimators of the SKCE, both biased and unbiased.

[3] As mentioned above, our framework rephrases and generalizes the construction used by Widmann et al. (2019). The matrix-valued kernels that they employ can be recovered by setting $k_{\mathcal{P}}$ to a Gaussian kernel on the probability simplex and $k_{\mathcal{Y}}(y_1, y_2) = \delta_{y_1, y_2}$.

[4] For a general discussion about characteristic kernels and their relation to universal kernels we refer to the paper by Sriperumbudur et al. (2011).

Lemma 1. *The plug-in estimator of* SKCE_k *is non-negatively biased. It is given by*

$$\widehat{\mathrm{SKCE}}_k = \frac{1}{n^2} \sum_{i,j=1}^{n} h\big((P_{X_i}, Y_i), (P_{X_j}, Y_j)\big).$$

Inspired by the block tests for the regular MMD (Zaremba et al., 2013), we define the following class of unbiased estimators. Note that in contrast to $\widehat{\mathrm{SKCE}}_k$ they do not include terms of the form $h\big((P_{X_i}, Y_i), (P_{X_i}, Y_i)\big)$.

Lemma 2. *The block estimator of* SKCE_k *with block size* $B \in \{2, \ldots, n\}$*, given by*

$$\widehat{\mathrm{SKCE}}_{k,B} \coloneqq \left\lfloor \frac{n}{B} \right\rfloor^{-1} \sum_{b=1}^{\lfloor n/B \rfloor} \binom{B}{2}^{-1} \sum_{(b-1)B+1 \leq i < j \leq bB} h\big((P_{X_i}, Y_i), (P_{X_j}, Y_j)\big),$$

is an unbiased estimator of SKCE_k.

The extremal estimator with $B = n$ is a so-called U-statistic of SKCE_k (Hoeffding, 1948; van der Vaart, 1998), and hence it is the minimum variance unbiased estimator. All presented estimators are consistent, i.e., they converge to SKCE_k almost surely as the number n of data points goes to infinity. The sample complexity of $\widehat{\mathrm{SKCE}}_k$ and $\widehat{\mathrm{SKCE}}_{k,B}$ is $O(n^2)$ and $O(Bn)$, respectively.

3.3 Calibration tests

A fundamental issue with calibration errors in general, including ECE, is that their empirical estimates do not provide an answer to the question if a model is actually calibrated. Even if the measure is guaranteed to be zero if and only if the model is calibrated, usually the estimates of calibrated models are non-zero due to randomness in the data and (possibly) the estimation procedure. In classification, statistical hypothesis tests of the null hypothesis

$$H_0\colon \text{model } P \text{ is calibrated},$$

so-called calibration tests, have been proposed as a tool for checking rigorously if P is calibrated (Bröcker & Smith, 2007; Vaicenavicius et al., 2019; Widmann et al., 2019). For multi-class classification, Widmann et al. (2019) suggested calibration tests based on the asymptotic distributions of estimators of the previously formulated KCE. Although for finite data sets the asymptotic distributions are only approximations of the actual distributions of these estimators, in their experiments with 10 classes the resulting p-value approximations seemed reliable whereas p-values obtained by so-called consistency resampling (Bröcker & Smith, 2007; Vaicenavicius et al., 2019) underestimated the p-value and hence rejected the null hypothesis too often (Widmann et al., 2019).

For fixed block sizes $\sqrt{\lfloor n/B \rfloor}\big(\widehat{\mathrm{SKCE}}_{k,B} - \mathrm{SKCE}_k\big) \xrightarrow{d} \mathcal{N}\big(0, \sigma_B^2\big)$ as $n \to \infty$, and, under H_0, $n\widehat{\mathrm{SKCE}}_{k,n} \xrightarrow{d} \sum_{i=1}^{\infty} \lambda_i (Z_i - 1)$ as $n \to \infty$, where Z_i are independent χ_1^2 distributed random variables. See Appendix B for details and definitions of the involved constants. From these results one can derive calibration tests that extend and generalize the existing tests for classification problems, as explained in Remarks B.1 and B.2. Our formulation illustrates also the close connection of these tests to different two-sample tests (Gretton et al., 2007; Zaremba et al., 2013).

4 Alternative approaches

For two-sample tests, Chwialkowski et al. (2015) suggested the use of the so-called unnormalized mean embedding (UME) to overcome the quadratic sample complexity of the minimum variance unbiased estimator and its intractable asymptotic distribution. As we show in Appendix C, there exists an analogous measure of calibration, termed unnormalized calibration mean embedding (UCME), with a corresponding calibration mean embedding (CME) test.

As an alternative to our construction based on the joint distributions of (P_X, Y) and (P_X, Z_X), one could try to directly compare the conditional distributions $\mathbb{P}(Y|P_X)$ and $\mathbb{P}(Z_X|P_X) = P_X$. For instance, Ren et al. (2016) proposed the conditional MMD based on the so-called conditional kernel mean embedding (Song et al., 2009; 2013). However, as noted by Park & Muandet (2020),

its common definition as operator between two RKHS is based on very restrictive assumptions, which are violated in many situations (see, e.g., Fukumizu et al., 2013, Footnote 4) and typically require regularized estimates. Hence, even theoretically, often the conditional MMD is "not an exact measure of discrepancy between conditional distributions" (Park & Muandet (2020)). In contrast, the maximum conditional mean discrepancy (MCMD) proposed in a concurrent work by Park & Muandet (2020) is a random variable derived from much weaker measure-theoretical assumptions. The MCMD provides a local discrepancy conditional on random predictions whereas KCE is a global real-valued summary of these local discrepancies.[5]

5 Experiments

In our experiments we evaluate the computational efficiency and empirical properties of the proposed calibration error estimators and calibration tests on both calibrated and uncalibrated models. By means of a classic regression problem from statistics literature, we demonstrate that the estimators and tests can be used for the evaluation of calibration of neural network models and ensembles of such models. This section contains only an high-level overview of these experiments to conserve space but all experimental details are provided in Appendix A.

5.1 Empirical properties and computational efficiency

We evaluate error, variance, and computation time of calibration error estimators for calibrated and uncalibrated Gaussian predictive models in synthetic regression problems. The results empirically confirm the consistency of the estimators and the computational efficiency of the estimator with block size $B = 2$ which, however, comes at the cost of increased error and variance.

Additionally, we evaluate empirical test errors of calibration tests at a fixed significance level $\alpha = 0.05$. The evaluations, visualized in Fig. 2 for models with ten-dimensional targets, demonstrate empirically that the percentage of incorrect rejections of H_0 converges to the set significance level as the number of samples increases. Moreover, the results highlight the computational burden of the calibration test that estimates quantiles of the intractable asymptotic distribution of $\widehat{\mathrm{SKCE}}_{k,n}$ by bootstrapping. As expected, due to the larger variance of $\widehat{\mathrm{SKCE}}_{k,2}$ the test with fixed block size $B = 2$ shows a decreased test power although being computationally much more efficient.

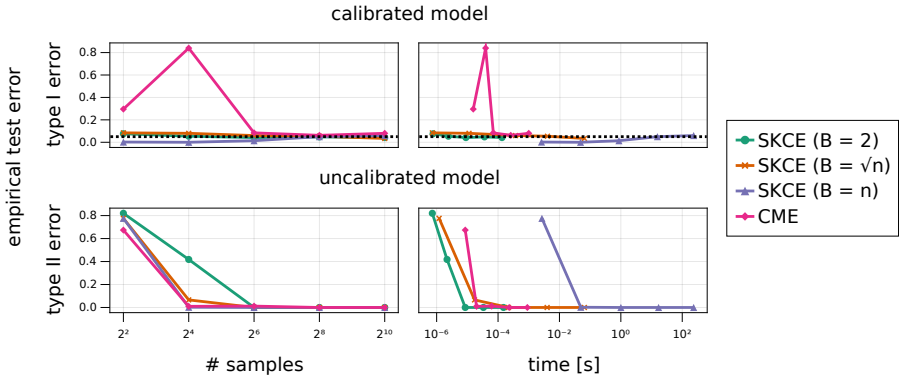

Figure 2: Empirical test errors for 500 data sets of $n \in \{4, 16, 64, 256, 1024\}$ samples from models with targets of dimension $d = 10$. The dashed black line indicates the set signficance level $\alpha = 0.05$.

[5]In our calibration setting, MMCD is almost surely equal to $\sup_{f \in \mathcal{F}_\mathcal{Y}} \left| \mathbb{E}_{Y|P_X}\left(f(Y)|P_X\right) - \mathbb{E}_{Z_X|P_X}\left(f(Z_X)|P_X\right)\right|$, where $\mathcal{F}_\mathcal{Y} := \{f \colon \mathcal{Y} \to \mathbb{R} | \|f\|_{\mathcal{H}_\mathcal{Y}} \leq 1\}$ for an RKHS $\mathcal{H}_\mathcal{Y}$ with kernel $k_\mathcal{Y} \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. If kernel $k_\mathcal{Y}$ is characteristic, $\mathrm{MCMD} = 0$ almost surely if and only if model P is calibrated (Park & Muandet, 2020, Theorem 3.7). Although the definition of MCMD only requires a kernel $k_\mathcal{Y}$ on the target space, a kernel $k_\mathcal{P}$ on the space of predictions has to be specified for the evaluation of its regularized estimates.

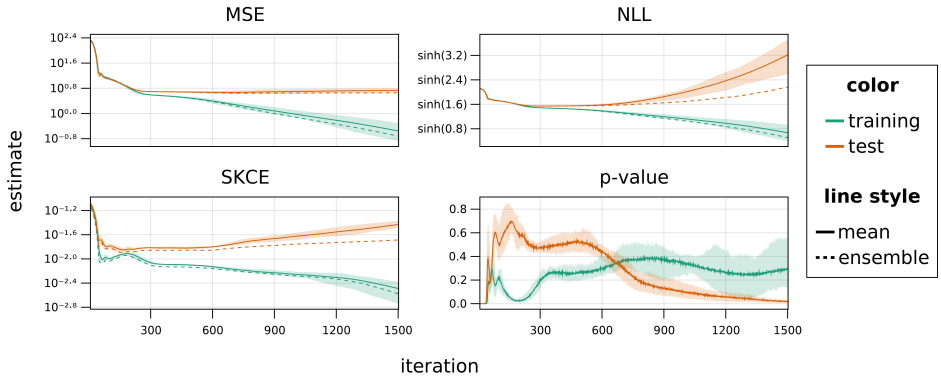

Figure 3: Mean squared error (MSE), average negative log-likelihood (NLL), $\widehat{\mathrm{SKCE}}_k$ (SKCE), and p-value approximation (p-value) of ten Gaussian predictive models for the Friedman 1 regression problem versus the number of training iterations. Evaluations on the training data set (100 samples) and the test data set (50 samples) are displayed in green and orange, respectively. The solid line and its surrounding band represent the mean and the range of the 10% to 90% quantile of the evaluations of the ten models, respectively. The dashed line visualizes the evaluations of the ensemble models.

5.2 Friedman 1 regression problem

The Friedman 1 regression problem (Friedman, 1979; 1991; Friedman et al., 1983) is a classic non-linear regression problem with ten-dimensional features and real-valued targets with Gaussian noise. We train a Gaussian predictive model whose mean is modelled by a shallow neural network and a single scalar variance parameter (consistent with the data-generating model) ten times with different initial parameters. Figure 3 shows estimates of the mean-squared error (MSE), the average negative log-likelihood (NLL), SKCE_k, and a p-value approximation for these models and their ensemble on the training and a separate test data set. All estimates indicate consistently that the models are overfit after 1500 training iterations. The estimations of SKCE_k and the p-values allow to focus on calibration specifically, whereas MSE indicates accuracy only and NLL, as any proper scoring rule (Bröcker, 2009), provides a summary of calibration and accuracy. The estimation of SKCE_k in addition to NLL could serve as another source of information for early stopping and model selection.

6 Conclusion

We presented a framework of calibration estimators and tests for any probabilistic model that captures both classification and regression problems of arbitrary dimension as well as other predictive models. We successfully applied it for measuring calibration of (ensembles of) neural network models.

Our framework highlights connections of calibration to two-sample tests and optimal transport theory which we expect to be fruitful for future research. For instance, the power of calibration tests could be improved by heuristics and theoretical results about suitable kernel choices or hyperparameters (cf. Jitkrittum et al., 2016). It would also be interesting to investigate alternatives to KCE captured by our framework, e.g., by exploiting recent advances in optimal transport theory (cf. Genevay et al., 2016).

Since the presented estimators of SKCE_k are differentiable, we imagine that our framework could be helpful for improving calibration of predictive models, during training (cf. Kumar et al., 2018) or post-hoc. Currently, many calibration methods (see, e.g., Guo et al., 2017; Kull et al., 2019; Song et al., 2019) are based on optimizing the log-likelihood since it is a strictly proper scoring rule and thus encourages *both* accurate and reliable predictions. However, as for any proper scoring rule, "Per se, it is impossible to say how the score will rank unreliable forecast schemes [...]. The lack of reliability of one forecast scheme might be outbalanced by the lack of resolution of the other" (Bröcker (2009)). In other words, if one does not use a calibration method such as temperature scaling (Guo et al., 2017) that keeps accuracy invariant[6], it is unclear if the resulting model is trading off calibration for accuracy when using log-likelihood for re-calibration. Thus hypothetically flexible calibration methods might benefit from using the presented calibration error estimators.

[6]Temperature scaling can be defined and applied for general probabilistic predictive models, see Appendix F.

REFERENCES

Miguel A Arcones and Evarist Giné. On the bootstrap of U and V statistics. *The Annals of Statistics*, 20(2):655–674, 1992.

Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups*. Springer New York, 1984. doi: 10.1007/978-1-4612-1128-0.

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519, July 2009. doi: 10.1002/qj.456.

Jochen Bröcker and Leonard A. Smith. Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22(3):651–661, June 2007. doi: 10.1175/waf993.1.

Yongxin Chen, Tryphon T. Georgiou, and Allen Tannenbaum. Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278, 2019. doi: 10.1109/access.2018.2889838.

Yukun Chen, Jianbo Ye, and Jia Li. Aggregated Wasserstein distance and state registration for hidden Markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(9): 2133–2147, September 2020. doi: 10.1109/tpami.2019.2908635.

Kacper Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pp. 1981–1989, Cambridge, MA, USA, 2015. MIT Press.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32(1/2):12, March 1983. doi: 10.2307/2987588.

Charles-Alban Deledalle, Shibin Parameswaran, and Truong Q. Nguyen. Image denoising with generalized gaussian mixture model patch priors. *SIAM Journal on Imaging Sciences*, 11(4): 2568–2609, January 2018. doi: 10.1137/18m116890x.

Julie Delon and Agnes Desolneux. A Wasserstein-type distance in the space of Gaussian mixture models. 2019.

Richard M. Dudley. *Real analysis and probability*. Wadsworth & Brooks/Cole Pub. Co, Pacific Grove, Calif, 1989.

Matteo Fasiolo, Simon N. Wood, Margaux Zaffran, Raphaël Nedellec, and Yannig Goude. Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, pp. 1–11, March 2020. doi: 10.1080/01621459.2020.1725521.

Jerome H. Friedman. A tree-structured approach to nonparametric multiple regression. In *Lecture Notes in Mathematics*, pp. 5–22. Springer Berlin Heidelberg, 1979. doi: 10.1007/bfb0098488.

Jerome H. Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991. ISSN 00905364.

Jerome H. Friedman, Eric Grosse, and Werner Stuetzle. Multidimensional additive spline approximation. *SIAM Journal on Scientific and Statistical Computing*, 4(2):291–301, June 1983. doi: 10.1137/0904023.

Kenji Fukumizu, Francis R Bach, and Michael I Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan): 73–99, 2004.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis (eds.), *Advances in Neural Information Processing Systems 20*, pp. 489–496. Curran Associates, Inc., 2008.

Kenji Fukumizu, Le Song, and Arthur Gretton. Kernel Bayes' rule: Bayesian inference with positive definite kernels. *Journal of Machine Learning Research*, 14(82):3753–3783, 2013. URL `http://jmlr.org/papers/v14/fukumizu13a.html`.

Matthias Gelbrich. On a formula for the l^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990. doi: 10.1002/mana.19901470121.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems 29*, pp. 3440–3448. 2016.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 249–256. PMLR, 5 2010.

E. Gómez, M.A. Gomez-Viilegas, and J.M. Marín. A multivariate generalization of the power exponential family of distributions. *Communications in Statistics - Theory and Methods*, 27(3): 589–600, January 1998. doi: 10.1080/03610929808832115.

E. Gómez-Sánchez-Manzano, M. A. Gómez-Villegas, and J. M. Marín. Multivariate exponential power distributions as mixtures of normal distributions with Bayesian applications. *Communications in Statistics - Theory and Methods*, 37(6):972–985, February 2008. doi: 10.1080/03610920701762754.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J. Smola. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19*, pp. 513–520. 2007.

Arthur Gretton, Kenji Fukumizu, Zaïd Harchaoui, and Bharath K. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems 22*, pp. 673–681. 2009.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 8 2017.

Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schön. Evaluating scalable Bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

Yvonne H. S. Ho and Stephen M. S. Lee. Calibrated interpolated confidence intervals for population quantiles. *Biometrika*, 92(1):234–241, March 2005. doi: 10.1093/biomet/92.1.234.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, September 1948. doi: 10.1214/aoms/1177730196.

Harold Hotelling. The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3): 360–378, August 1931. doi: 10.1214/aoms/1177732979.

Michael Innes, Elliot Saba, Keno Fischer, Dhairya Gandhi, Marco Concetto Rudilosso, Neethu Mariya Joy, Tejan Karmali, Avik Pal, and Viral Shah. Fashionable modelling with Flux, 2018.

Mike Innes. Flux: Elegant machine learning with Julia. *Journal of Open Source Software*, 3(25):602, May 2018. doi: 10.21105/joss.00602.

Wittawat Jitkrittum, Zoltán Szabó, Kacper P Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *Advances in Neural Information Processing Systems 29*, pp. 181–189. 2016.

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous univariate distributions: Vol. 1*. Wiley, New York, 2nd edition, 1994.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804. PMLR, 7 2018.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 623–631. PMLR, 4 2017.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems 32*, pp. 12316–12326. 2019.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2805–2814. PMLR, 7 2018.

Arak M. Mathai and Serge B. Provost. *Quadratic forms in random variables: Theory and applications*, volume 126. M. Dekker, New York, 1992.

Charles A. Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, January 2005. doi: 10.1162/0899766052530802.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, June 1997. doi: 10.2307/1428011.

Allan H. Murphy and Robert L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, 26(1):41, 1977. doi: 10.2307/2346866.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI Conference on Artificial Intelligence*, 2015.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings, 2020.

Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2018.

J. Platt. *Probabilities for SV Machines*, pp. 61–73. MIT Press, 2000.

Yong Ren, Jun Zhu, Jialian Li, and Yucen Luo. Conditional generative moment-matching networks. In *Advances in Neural Information Processing Systems 29*, pp. 2928–2936. 2016.

M. Rueda, S. Martínez-Puertas, H. Martínez-Puertas, and A. Arcos. Calibration methods for estimating quantiles. *Metrika*, 66(3):355–371, December 2006. doi: 10.1007/s00184-006-0116-1.

Robert J. Serfling (ed.). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, Inc., November 1980. doi: 10.1002/9780470316481.

Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5897–5906. PMLR, 6 2019.

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML 09*. ACM Press, 2009. doi: 10.1145/1553374.1553497.

Le Song, Kenji Fukumizu, and Arthur Gretton. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, July 2013. doi: 10.1109/msp.2013.2252713.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On integral probability metrics, ϕ-divergences and binary classification, 2009.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R.G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(70): 2389–2410, 2011.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6(0):1550–1599, 2012. doi: 10.1214/12-ejs722.

Zoltán Szabó and Bharath K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

Maxime Taillardat, Olivier Mestre, Michaël Zamo, and Philippe Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6):2375–2393, June 2016. doi: 10.1175/mwr-d-15-0260.1.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3459–3467. PMLR, 4 2019.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, October 1998. doi: 10.1017/cbo9780511802256.

Cédric Villani. *Optimal Transport*. Springer Berlin Heidelberg, 2009. doi: 10.1007/978-3-540-71050-9.

David Widmann, Fredrik Lindsten, and Dave Zachariah. Calibration tests in multi-class classification: A unifying framework. In *Proceedings of the 32th International Conference on Neural Information Processing Systems*, pp. 12236–12246. 2019.

Sidney J. Yakowitz and John D. Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, February 1968. doi: 10.1214/aoms/1177698520.

Bianca Zadrozny. Reducing multiclass to binary by coupling probability estimates. In *Advances in Neural Information Processing Systems 14*, pp. 1041–1048. MIT Press, 2002.

Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in Neural Information Processing Systems 26*, pp. 755–763. 2013.