Proceedings of the ASME 2018
Verification and Validation Symposium
VVS2018
May 16-18, 2018, Minneapolis, MN, USA

VVS2018-9327

AN EVALUATION OF VALIDATION METRICS FOR PROBABILISTIC MODEL OUTPUTS

Paul Gardner*, Charles Lord, and
Robert J. Barthorpe
Dynamics Research Group
Department of Mechanical Engineering
University of Sheffield
Mappin Street, Sheffield S1 3JD
Email: pagardner1@sheffield.ac.uk

ABSTRACT

Probabilistic modelling methods are increasingly being employed in engineering applications. These approaches make inferences about the distribution, or summary statistical moments, for output quantities. A challenge in applying probabilistic models is validating output distributions. An ideal validation metric is one that intuitively provides information on key divergences between the output and validation distributions. Furthermore, it should be interpretable across different problems in order to informatively select the appropriate statistical method. In this paper, two families of measures for quantifying differences between distributions are compared: f-divergence and integral probability metrics (IPMs). Discussions and evaluation of these measures as validation metrics are performed with comments on ease of computation, interpretability and quantity of information provided.

INTRODUCTION

Understanding and modelling uncertainty in a probabilistic manner is important in engineering applications. Quantifying output or parameter distributions provides insight into how uncertainties are propagated through a system in addition to understanding parameter sensitivities. This information allows for more robust decisions and a better understanding of risk in engineering applications. Probabilistic modelling approaches, namely uncertainty quantification techniques, are especially useful in producing robust predictions of damage states from computer models, part of a forward model driven structural health monitoring strategy [1–3], and are also vital in creating a digital twin of a structure. This increased utilisation in probabilistic modelling in an engineering context leads to a focus on standard methods for validating output or parameter distributions. In the past deterministic validation metrics, such as normalised mean square error, have allowed validation of model outputs in a simple and intuitive manner. Moving to a probabilistic framework requires a validation procedure with appropriate, informative validation metrics for assessing the performance of these methods. An additional challenge in an engineering context is that generally their is a lack of validation data, posing difficulties in accurately inferring the underlying statistical distribution of the data.

There are several approaches to validating probabilistic models applied in an engineering context. The first step in validation is often performing hypothesis testing, this involves trying to falsify the null hypothesis that the observational data could have been generated by the model. This can be approached in a frequentist or Bayesian manner [4, 5]. If the hypothesis test rejects the null hypothesis then other validation metrics are required in order to interrogate the causes of dissimilarity between the distributions of the model and the observational data. Often the ideal approach is to assess the utility of the probabilistic model, however this generally involves extensive knowledge and expertise from the modeller [6]. Distance/divergence measures that quan-

*Address all correspondence to this author.

tify the dissimilarities between two distributions are therefore useful in assessing the performance of the model. A common measure is the Kolmogorov distance and it's area alternative, the area metric, where distances/areas are calculated between cumulative frequency distributions (CDF) [5, 7–9]. However there are a broad range of distance/divergence measures available, these are divided into two main families, namely f-divergence and integral probability metrics (IPMs). This paper seeks to evaluate the performance of several specific forms of the two families of distance/divergence measures as validation metrics after hypothesis testing has been performed.

The author defines in this paper criteria for evaluating validation metrics for probabilistic engineering models as follows:

1. It should quantify the difference between the model predictions and observational data.
2. It should be interpretable and aid identifying improvements.
3. It should provide objective information and be consistent when applied to different probabilistic models or applications.
4. It should account for the complete form of the distributions (and not just statistical moments) - if the underlying distribution of the observational data is unknown it should have a non-parametric estimator.
5. It should be computationally efficient and where applicable, have appropriate convergence properties for engineering data sets.

For clarity of terminology a *validation metric* here refers to methods for quantifying the dissimilarities between predictions and observational data. A *metric*, where used on it's own, refers to the statistical distance definition; distance/divergence is a metric if it abides by four requirements:

1. Non-negative - $D(\mathbb{P},\mathbb{Q}) \geq 0$
2. Zero only when $\mathbb{P} = \mathbb{Q}$ - $D(\mathbb{P},\mathbb{Q}) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$
3. Symmetric - $D(\mathbb{P},\mathbb{Q}) = D(\mathbb{Q},\mathbb{P})$
4. Obeys the triangle inequality - $D(\mathbb{P},\mathbb{M}) \geq D(\mathbb{P},\mathbb{Q}) + D(\mathbb{Q},\mathbb{M})$

Where D is a distance *metric*; $\mathbb{P}$, $\mathbb{Q}$ and $\mathbb{M}$ are probability measures (for the general metric these quantities can be any non-negative real number).

The outline of this paper is as follows: f-divergence followed by the IPMs are defined mathematically with examples of specific forms from each family. A five storey building structure is then presented as a case study for which the distance/divergence measures are applied and evaluated. Finally a discussion and conclusions are highlighted, outlining areas for further research.

f-DIVERGENCE

The class of distances/divergences that depend on a ratio between probability measures are known as the *Csiszár's ϕ-divergence* or f-divergence. These measures are of the form defined in Eqn. (1).

$$D_\phi(\mathbb{P},\mathbb{Q}) = \int_M \phi\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P} \tag{1}$$

Where M is a measurable space and ϕ is a convex function. Eqn. (1) holds when $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$ and $-\infty$ otherwise. Different forms of the f-divergence depend on the choice of function ϕ with notable cases being the Kullback-Liebler (KL) divergence, $\phi(t) = t\log(t)$, Hellinger distance, $\phi(t) = \left(\sqrt{t}-1\right)^2$, and total variation distance, $\phi(t) = |t-1|$. The family of divergence measures is widely used throughout information theory and machine learning.

Kullback-Liebler Divergence

The KL-divergence is the most widely used f-divergence and has many applications. A notable example is in performing variational inference as it is a natural formulation of the ratio between two likelihood functions [10]. The KL-divergence of probability measures $\mathbb{P}$ and $\mathbb{Q}$ is shown in Eqn. (2).

$$D_{KL}(\mathbb{P},\mathbb{Q}) = \int p(x)\log\left(\frac{p(x)}{q(x)}\right) dx \tag{2}$$

Where $p(x)$ and $q(x)$ are probability distributions of the random variable x. The KL-divergence is a measure of relative entropy [11] taking the units nats, or bits depending on the base of the logarithm, exponential or two respectively. The divergence informs of the average number of extra nats (or bits) required to encode the data given that the distribution $\mathbb{Q}$ is used to model the 'true' distribution $\mathbb{P}$. This can be thought of as how well $\mathbb{Q}$ approximates $\mathbb{P}$. The KL-divergence can be difficult to estimate and often proves challenging when the dimension size of samples increases (i.e. in the instants where d increases when $M = \mathbb{R}^d$). On the other hand the divergence can be practical to compute between low-dimensional probability density functions and therefore is useful when the observational density function is known or can be accurately approximated.

The KL-divergence is not a metric as it does not meet two of the four requirements: it is neither symmetric nor obeys the triangle inequality. A smoothed and symmetrised form of the KL-divergence is the Jenson-Shannon divergence [12] which by taking the square root becomes a metric, known as the Jenson-Shannon distance defined in Eqn. (3).

$$D_{JSD}(\mathbb{P},\mathbb{Q}) = \sqrt{\frac{1}{2}D_{KL}(\mathbb{P},\mathbb{M}) + \frac{1}{2}D_{KL}(\mathbb{Q},\mathbb{M})} \tag{3}$$

Where $\mathbb{M} = \frac{1}{2}(\mathbb{P}+\mathbb{Q})$ and is the midpoint. This will always produce a finite result, unlike the KL-divergence as $\mathbb{P}$ and $\mathbb{Q}$ are always absolutely continuous with respect to $\mathbb{M}$. The computational overheads of the Jenson-Shannon distance are high due to the mixture distribution $\mathbb{M}$, which becomes prohibitive in high dimensional data. In addition it is less sensitive to scenarios when distribution $\mathbb{Q}$ contains sample values that are impossible in $\mathbb{P}$, unlike the KL-divergence.

Empirical estimation of the KL-divergence in a non-parametric manner for continuous distributions can be approximated using several approaches [13,14]. Here an non-parametric estimation method based on data-dependent partitions is used; which has been shown to be strongly consistent [13]. For the unidimensional case, assume independent and identically distributed (i.i.d.) samples from probability measures $\mathbb{P}$ and $\mathbb{Q}$; $X = \{x_1, x_2, \ldots x_m\}$ and $Y = \{y_1, y_2, \ldots y_m\}$. The algorithm orders Y so that $Y_{(1)} \leq Y_{(2)} \leq \ldots Y_{(n)}$, where $Y_{(i)}$ refers to the ith index of Y. A partition of empirically equivalent segments divides Y, called l_n spacings, as defined in Eqn. (4), with l_n points in each interval (expect possibly the final one).

$$I^n = \{(-\infty, Y_{(l_n)}], (Y_{(l_n)}, Y_{(2l_n)}], \ldots, (Y_{(l_n(T_n-1))}, +\infty)\} \tag{4}$$

Where brackets have interval notation meaning, $l_n \leq n$ and $T_n = \lfloor n/l_n \rfloor$. The empirical estimate of the KL-divergence can then be calculated from Eqn. (5).

$$\hat{D}_{KL}(\mathbb{P},\mathbb{Q}) = \sum_{i=1}^{T_n} \mathbb{P}_m(I_i^n) \log \frac{\mathbb{P}_m(I_i^n)}{\mathbb{Q}_n(I_i^n)} \tag{5}$$

Where $\mathbb{P}_m$ and $\mathbb{Q}_m$ are empirical probability measures. This can easily be adapted to multidimensional data. As the number of samples and partitions increase $\hat{D}_{KL}(\mathbb{P},\mathbb{Q})$ approaches $D_{KL}(\mathbb{P},\mathbb{Q})$ [13].

Figure 1 presents a convergence study of the empirical estimator for unidimensional samples drawn from two Gaussian distributions, $\mathbb{P} \sim \mathcal{N}(0,1)$ and $\mathbb{Q} \sim \mathcal{N}(1,1)$. 500 repeats were performed at each sample size in order to demonstrate the variance of the estimator. It is clearly presented that although the estimator will convergence, this can be slow and requires a large sample size. In most engineering applications it is often not possible to obtain even hundreds of samples at each input indicating a drawback with the estimator.

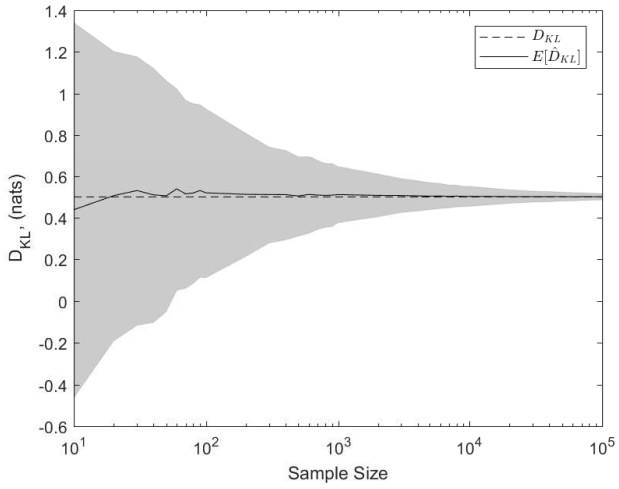

FIGURE 1. ESTIMATION OF KL-DIVERGENCE USING DATA-DEPENDENT PARTITIONS WHERE $\mathbb{P} \sim \mathcal{N}(0,1)$ AND $\mathbb{Q} \sim \mathcal{N}(1,1)$. $D_{KL}(\mathbb{P},\mathbb{Q}) = 0.5$.

Hellinger Distance

The Hellinger distance is analogous to the Euclidean distance for probability measures as it is an L_2 norm, defined in Eqn. (6).

$$D_H(\mathbb{P},\mathbb{Q}) = \sqrt{\frac{1}{2}\int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx} \tag{6}$$

Hellinger distance is a metric meeting all four requirements as well as having the property that $D_H(\mathbb{P},\mathbb{Q}) \leq 1$. This provides an intuitive interpretation of the distance where values close to zero mean very similar probability measures and a distance close to one indicates very dissimilar probability measures.

INTEGRAL PROBABILITY METRICS

IPMs differ from f-divergences as they depend on the difference rather than ratio of probability measures. These measures are defined as in Eqn. (7).

$$D_{\mathcal{F}}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} \left| \int_M f d\mathbb{P} - \int_M f d\mathbb{Q} \right| \tag{7}$$

Where $\mathcal{F}$ is a class of functions on M. The choice of $\mathcal{F}$ leads to various IPMs, such as the total variation distance, where $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$ and the Maximum Mean Discrepancy (MMD) where $\mathcal{F} = \{f : \|f\|_{\mathcal{H}} \leq 1\}$ (i.e. all f that are reproducing kernel Hilbert space (RKHS), $\mathcal{H}$).

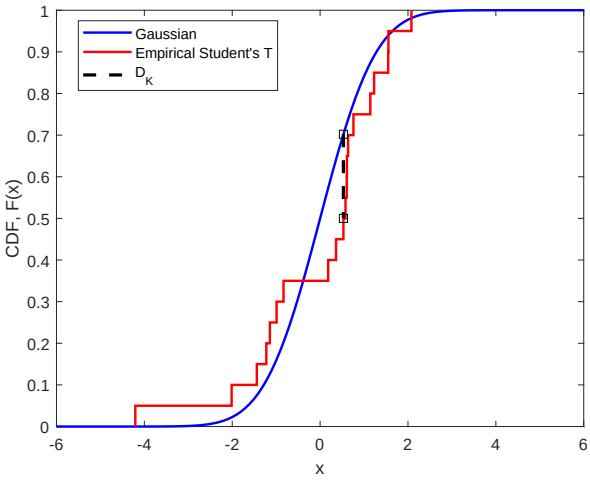

FIGURE 2. AN EXAMPLE OF KOLMOGOROV DISTANCE BETWEEN $\mathbb{P} = \mathcal{N}(0,1)$ AND $\mathbb{Q} \sim \mathcal{T}(5)$, $D_K = 0.25$. WHERE $\mathcal{T}$ IS A STUDENT'S T DISTRIBUTION.

Total Variation and Kolmogorov Distances

Total variation distance is the L_1-norm equivalent to the f-divergence Hellinger distance and is defined in Eqn. (8).

$$D_{TV}(\mathbb{P},\mathbb{Q}) = \sqrt{\frac{1}{2}\int |p(x)-q(x)|dx} \tag{8}$$

This is the only distance measure that can be classed as both an f-divergence and IPM [15]. Total variation distance, like the Hellinger distance, takes values in $[0,1]$ aiding interpretability.

The Kolmogorov distance is closely related to the total variation distance. The Kolmogorov distance is the L_1 norm between two cumulative density functions (CDF). The measure is also bounded by $[0,1]$. The Kolmogorov distance is as defined in Eqn. (9).

$$D_K(\mathbb{P},\mathbb{Q}) = \sup_{x\in\mathbb{R}} \left|F_p(x) - F_q(x)\right| \tag{9}$$

Where sup is the supremum, the least upper bound of pointwise differences and $F_p(x)$ is a CDF for the measure $\mathbb{P}$ over the random variable x. Simply the Kolmogorov is the largest vertical difference between the two CDFs. An empirical estimate can be easily formed by substituting empirical CDFs, $\hat{F}(x)$ into the equation. An illustration of this metric is shown in Fig. 2.

The Kolmogorov distance is useful in performing a Kolmogorov-Smirnov (K-S) test, a non-parametric test for comparing two one-dimensional CDFs. The premise of the K-S test is to identify whether the null hypothesis, $F_p(x) = F_q(x)$ is true for all x (usually for empirical CDFs). The Kolmogorov theorem states that as the number of samples tends to infinity, if the null hypothesis is true (that the samples were from the proposed distribution), then the $\sqrt{n}D_K$ tends to a Kolmogorov distribution that is not dependent on the hypothesised distribution. The K-S test therefore results in a comparison of $\sqrt{n}D_K$ against a critical value of the Kolmogorov distributions at significance level α_K, i,e, if $\sqrt{n}D_K > K_{\alpha_K}$ (where K is the Kolmogorov distribution) then the null hypothesis is rejected.

Maximum Mean Discrepancy

MMD assesses the difference between distributions by using a smooth function that produces large values when points are drawn from $\mathbb{P}$ and small when drawn from $\mathbb{Q}$. The function produces values for the two samples with the difference between the mean of these function values equating to the MMD, i.e. the maximum distance between the means once transformed through the smooth function f as defined in Eqn. (10).

$$D_{MMD}(\mathbb{P},\mathbb{Q}) = \sup_{f\in\mathcal{F}} \left|\mathbb{E}_x[f(x)] - \mathbb{E}_y[f(y)]\right| \tag{10}$$

Where x and y are samples from $\mathbb{P}$ and $\mathbb{Q}$ respectively. The function class $\mathcal{F}$ are those which produce a RKHS called reproducing kernels, $k(\cdot,\cdot)$. There is a choice of kernels that can be chosen in MMD with a popular choice being the radial basis kernel as defined in Eqn. (11). The parameter σ is often determined by the median pairwise distance among the joint data [16]. A difficulty with MMD is the heuristic nature of kernel selection.

$$k(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) \tag{11}$$

MMD can be empirically estimated in both unbiased and biased forms, depending on whether the U-statistics or V-statistics are used to calculated the sample means. These two forms are shown in Eqn. (12) and Eqn. (13).

$$D^2_{MMDu}(\mathbb{P},\mathbb{Q}) = \frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j\neq i}^{m} k(x_i,x_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(y_i,y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i,y_j) \tag{12}$$

$$D^2_{MMDb}(\mathbb{P},\mathbb{Q}) = \frac{1}{m^2}\sum_{i,j=1}^{m} k(x_i,x_j) + \frac{1}{n^2}\sum_{i,j=1}^{n} k(y_i,y_j) - \frac{2}{mn}\sum_{i,j=1}^{m,n} k(x_i,y_j) \quad (13)$$

Where m and n are the number of points in the samples X and Y respectively. These two forms of the statistic will both be small when $\mathbb{P} = \mathbb{Q}$ and large when the distributions are far apart.

An additional benefit of the MMD metric is that the kernel can be applied over a variable t, in order to visualise the behaviour of the MMD, producing the witness function, f^*. An empirical estimation of the witness function, outlined in Eqn. (14), can be formed to provide a method for visually determining the dissimilarities between two distributions.

$$f^*(t) \propto \frac{1}{m}\sum_{i=1}^{m} k(x_i,t) - \frac{1}{n}\sum_{i=1}^{n} k(y_i,t) \quad (14)$$

The witness function intuitively is zero where the two distributions are the same, positive when $\mathbb{P}$ is larger and negative when $\mathbb{Q}$ is greater, as far as the smoothness constraint allows. The example in Fig. 3 demonstrates the information gained from calculating the witness function. A radial basis kernel is used with $\sigma = 0.85$. It can be easily identified from Fig. 3 that more probability mass is located around zero from the Laplace distribution compared to the student's t; this is indicated by negative values in the witness function. In addition, the heavy tails of the student's t distribution contain more probability mass than the Laplace; this results in positive values at the tails of the witness function.

MMD can also be used for hypothesis testing. A test statistic, mD^2_{MMD} (either the biased or unbiased form can be used) is compared to a threshold $c_{\alpha_{MMD}}$ in order to determine whether the null hypothesis that $\mathbb{P} = \mathbb{Q}$ can be rejected, i.e. $mD^2_{MMD} > c_{\alpha_{MMD}}$. As MMD hypothesis testing is an empirical approach it is possible that due to finite samples an incorrect result may be returned. The test therefore allows the definition of an upper bound on the probability of Type I errors denoted as α_{MMD}. Type I errors are when $\mathbb{P} = \mathbb{Q}$ but the null hypothesis is rejected; due to the samples. Type II errors are when $\mathbb{P} \neq \mathbb{Q}$ but the null hypothesis is accepted. The MMD two sample test is shown to be consistent [17] i.e. it achieves Type I errors of α_{MMD} and Type II errors of zero in the limit of a large sample size. The threshold $c_{\alpha_{MMD}}$ is calculated via a bootstrap approach where a data-dependent threshold is estimated from calculating the test statistic from random permutations of the samples and finding the $(1-\alpha_{MMD})$th quantile [17, 18].

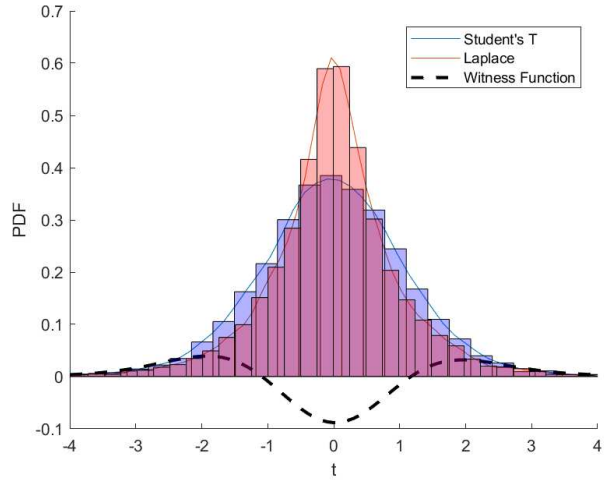

FIGURE 3. WITNESS FUNCTION BETWEEN $\mathbb{P} \sim \mathcal{L}(0, 0.71)$ AND $\mathbb{Q} \sim \mathcal{T}(8)$, D_{MMDb} = 0.11. WHERE $\mathcal{L}$ IS A LAPLACE DISTRIBUTION

TABLE 1. A COMPARISON OF DISTANCE/DIVERGENCES - f-DIVERGENCE.

$\mathbb{P}$	$\mathbb{Q}$	$D_{KL}(\mathbb{P},\mathbb{Q})$	$D_{KL}(\mathbb{Q},\mathbb{P})$	$D_H(\mathbb{P},\mathbb{Q})$
$\mathcal{N}(2,1)$	$\mathcal{N}(0,1)$	2	2	0.63
$\mathcal{N}(0,100)$	$\mathcal{N}(0,1)$	1.81	47.20	0.75
$\mathcal{L}(0,0.71)$	$\mathcal{N}(0,1)$	0.07	0.07	0.15
$\mathcal{T}(5)$	$\mathcal{N}(0,1)$	0.03	0.11	0.11

Numerical Examples

Table 1 and 2 demonstrate the distance/divergences applied to various combinations of distributions with varying parameters (presented in Fig. 4).

The results in Tab. 1 demonstrate the asymmetry of the KL-divergence. In addition, it is presented that KL-divergence is sensitive to scenarios where the proposal distribution ($\mathbb{Q}$) has little or no probability mass in areas where the target distribution is expected to see probability mass. The Hellinger distance gives a clear indication that, for the first two examples, the two distributions are far away. It also gives a better assessment of the differences between the Laplace and student's t distributions penalising these more than the KL-divergences.

Table 2 shows that the Kolmogorov distance is less sensitive to change in variance than the Hellinger distance and the MMD. It also provides smaller distances when comparing the Laplace

TABLE 2. A COMPARISON OF DISTANCE/DIVERGENCES - IPM. MMD IS CALCULATED USING 10000 SAMPLES FROM THE DISTRIBUTIONS SHOWN.

$\mathbb{P}$	$\mathbb{Q}$	$D_K(\mathbb{Q},\mathbb{P})$	$D_{MMDb}(\mathbb{Q},\mathbb{P})$
$\mathcal{N}(2,1)$	$\mathcal{N}(0,1)$	0.69	0.69
$\mathcal{N}(0,100)$	$\mathcal{N}(0,1)$	0.26	0.42
$\mathcal{L}(0,0.71)$	$\mathcal{N}(0,1)$	0.07	0.11
$\mathcal{T}(5)$	$\mathcal{N}(0,1)$	0.04	0.05

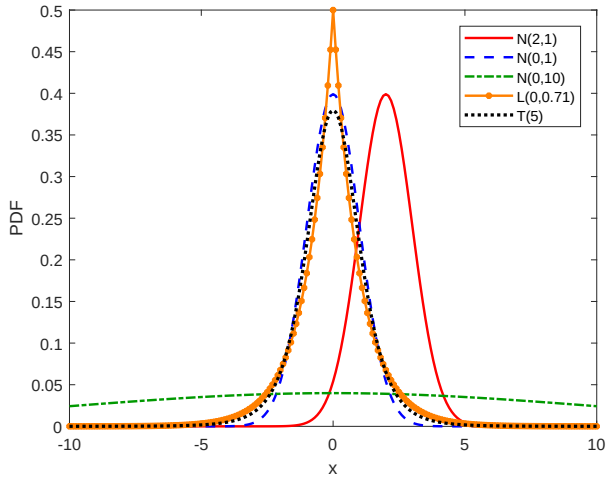

FIGURE 4. DISTRIBUTIONS USED IN COMPARISON OF DISTANCE/DIVERGENCES.

and student's t to a standard normal. Finally the MMD shows similar results to the Hellinger distance, whilst also being a empirical based metric.

CASE STUDY - BAYESIAN HISTORY MATCHING EXAMPLE

Bayesian history matching is an uncertainty quantification method that uses an implausibility metric in order to discard parts of the input space. The method aims to calibrate a statistical model of the form shown in Eqn. (15).

$$\mathbf{z}_j(\mathbf{x}) = \eta_j(\mathbf{x},\theta) + \delta_j + e_j \tag{15}$$

Where $\mathbf{z}_j(\mathbf{x})$ is the jth experimental output given inputs $\mathbf{x}$,

FIGURE 5. EXPERIMENTAL SETUP OF A REPRESENTATIVE FIVE STOREY BUILDING STRUCTURE.

$\eta_j(\mathbf{x},\theta)$ is the jth computer model output given $\mathbf{x}$ and parameters θ. The model discrepancy is δ and e is the observational uncertainty. Bayesian history matching calibrates the statistical model by assessing an implausibility metric in a likelihood free scheme in order to remove parameters that were unlikely to have generated the observational data. The implausibility metric is defined in Eqn. (16).

$$I_j(\mathbf{x},\theta) = \frac{|\mathbf{z}_j(\mathbf{x}) - \mathbb{E}^*\left[\mathcal{GP}_j(\mathbf{x},\theta)\right]|}{\left[V_{o,j} + V_{m,j} + V_{c,j}(\mathbf{x},\theta)\right]^{1/2}} \tag{16}$$

Where, V_o, V_m and $V_c(\mathbf{x},\theta)$ are the variances associated with the observational, model discrepancy and code uncertainties and $\mathbb{E}^*\left[\mathcal{GP}_j(\mathbf{x},\theta)\right]$ is the mean of a Gaussian process (GP) emulator. Due to limitations in the scope of the paper the reader is referred to [3, 19] for an overview of Bayesian history matching.

As the model discrepancy in Bayesian history matching is defined only as a variance it can be informative to infer it's functional form. In order to do this a GP regression model (with noise) can be fitted between the calibrated model outputs and the observational calibration data. This means that the model can predict across the input space whilst inferring any model discrepancy caused by missing physics or simplifications.

The combined Bayesian history matching and GP regres-

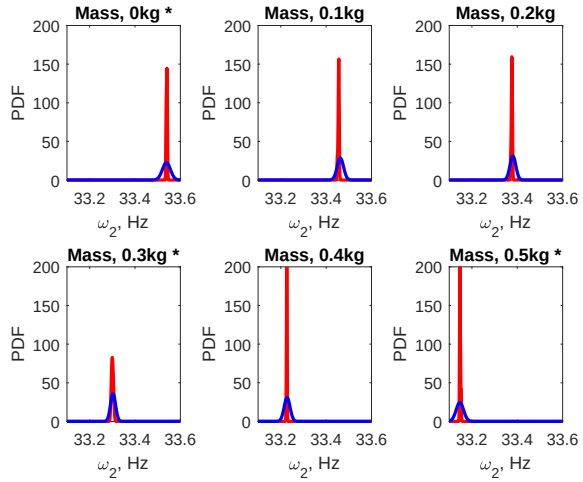

FIGURE 6. BAYESIAN HISTORY MATCHING PREDICTIVE DISTRIBUTIONS FOR ω_2 (BLUE) COMPARED TO KERNEL DENSITY ESTIMATES USING A GAUSSIAN KERNEL (RED) OF THE OBSERVATIONAL DATA. * DENOTES DATA USED TO CALIBRATE THE COMPUTER MODEL.

sion model approach was applied to a five story building structure, displayed in Fig. 5. The objective was to calibrate three material properties (θ) of a finite element computer model in order to predict the second and third natural frequencies (ω_2 and ω_3) of the structure under varying levels of mass ($\mathbf{x} = \{0, 0.1, \ldots, 0.5\}$kg). Three damage extents were used in calibration, $\mathbf{x} = \{0, 0.3, 0.5\}$kg whilst the remaining data was used for validation. Due to limitations in the scope of the paper the reader is referred to [3] for more details on the analysis.

Validation of Bayesian History Matching Predictions

The proposed validation metrics outlined in the previous sections were applied to the Bayesian history matching predictions. The output distributions when predicting the second natural frequency with respect to varying masses are presented in Fig. 6. It is noted that the normalised mean squared error for this prediction was 0.04 indicating an excellent fit between the predictive and observational means. Visually it can seen in Fig 6 that the mean predictions appear to show good agreement, however the variances of the predictions are larger than the observational data.

A good validation strategy will involve an initial hypothesis test. This allows the modeller to gain understanding of whether improvements to the model are required. Here the K-S test and MMD two sample test were applied to determine whether the null hypothesis, that the observational data could have been produced from the predictive distributions, could be rejected. Due to

TABLE 3. THE AVERAGE RESULTS FROM 100 REPEATS OF HYPOTHESIS TESTING USING K-S TEST AND MMD BOOTSTRAP METHODS. WHERE 1 = REJECTION OF NULL HYPOTHESIS; $\alpha_K = 0.05$ and $\alpha_{MMD} = 0.05$.

Test	0kg	0.1kg	0.2kg	0.3kg	0.4kg	0.5kg
K-S Test	1	1	1	0	1	1
MMD Test	0.99	0.91	0.95	0.64	0.98	0.97

the stochastic nature of the MMD hypothesis test (that samples must be drawn from the predictive distribution) 100 repeats were performed; the biased test statistic was also used. Table 3 shows the results from the hypothesis tests. Both tests show a rejection of the null hypothesis for all masses apart from at a mass of 0.3kg. The MMD two sample test also informs that the 0.3kg prediction is also unlikely to have produced the observational data in the majority of the hypothesis tests. This information provides a position to integrate the model further.

A valuable quality of the MDD metric is that the witness function can be calculated and plotted as demonstrated in Fig. 7. This information provides a quick method of diagnosing the issues associated with the predictive distributions. The asymmetry of witness functions across all the mass states indicate that mean of the distributions are different. The large positive spike at the centre of the witness function identifies that there is significant probability mass contained in the observational data that is not modelled by the predictions. In addition the negative witness function at both sides of the distribution clearly shows that the predictive distributions have more significant probability mass at the tails. This information helps infer that the predictive variances are greater than that of the observational data.

The next stage of the validation process is to calculate the distance/divergence measures. The measures outlined previously were applied to the case study and are displayed in Fig. 8. The KL divergence was calculated using the estimator outlined previously. The Hellinger distance was estimated using a kernel density estimate (with a Gaussian kernel) of the observational data. The Kolmogorov distance was calculated between the empirical CDFs of the observational data and predictive CDFs, whilst the MMD was calculated between the observational samples and samples drawn from the predictive distributions. Figure 8 displays that all of the distance/divergence measures follow a similar trend, indicating the smallest difference occurs between the distributions at 0.3kg; confirming the hypothesis test results. The Kolmogorov and Hellinger distances, being both bounded $[0, 1]$, show that all the mass states, apart from 0.3kg, are above 0.5, interpreting that they are significantly dissimilar. The Kolmogorov

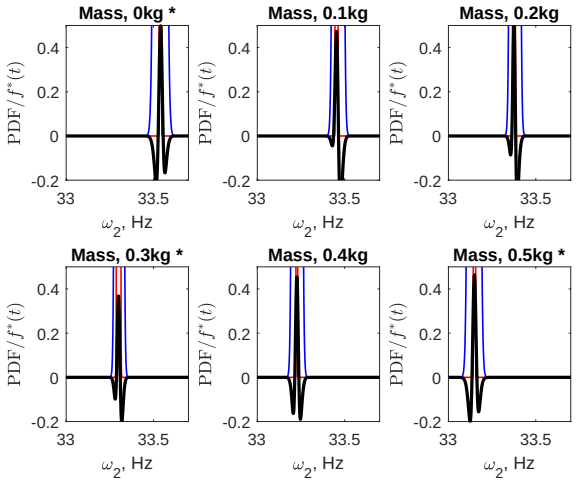

FIGURE 7. MMD WITNESS FUNCTION (BLACK) COMPARING BAYESIAN HISTORY MATCHING PREDICTIONS (BLUE) AGAINST OBSERVATIONAL DATA (RED) FOR ω_2. * DENOTES DATA USED TO CALIBRATE THE COMPUTER MODEL.

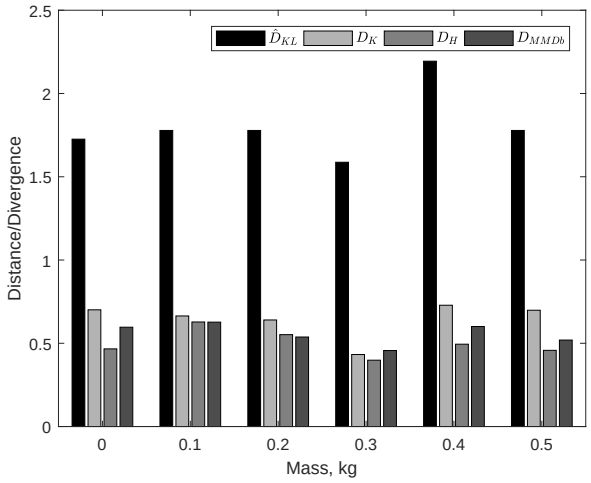

FIGURE 8. COMPARISON OF DISTANCE/DIVERGENCE MEASURES BETWEEN PREDICTIONS AND OBSERVATIONAL DATA FOR ω_2.

distance is more conservative for this case study when compared to the Hellinger distance; this may be due to the kernel density estimates used in the Hellinger distance. The KL-divergence is more challenging to interpret, and therefore the trend provides useful information rather than particular values. The MMD distance produced similar distances to that of the Kolmogorov and Hellinger distance, however because it does not have an upper bound and is harder to interpret.

DISCUSSION

The distance/divergence measures outlined in this paper have particular strengths in different application areas. Discussed below are comments on how well each of the distance/divergence measures meet the five criteria outlined in the introduction.

The KL-divergence is suited for problems comparing two likelihood functions where greater penalisation is required when a proposal distribution has little probability mass in areas of the target distribution. The KL-divergence is difficult to utilise as a validation metric because it is not a metric and has no upper bound. As a consequence the divergence is hard to comment on when comparing different applications. The Jenson-Shannon distance could be applied as an alternative but it smooths the performance of the KL-divergence removing the main benefits of the divergence. It is also complex to calculate due to the mixture distribution. The KL-divergence has well understood non-parametric estimators that have strongly consistent convergence properties. Nonetheless these properties exist in the limit, which for engineering applications can be a challenge due to the lack of numerous repeats of observational data.

The Hellinger distance is easy to interpret due to it being both a metric and bound $[0,1]$. This means that it can be compared across different probabilistic models and applications in a consistent manner. The metric here does not have a non-parameter estimator, without using a kernel density estimate, and therefore convergence properties are not defined. Kolmogorov distance provides similar qualities as the Hellinger distance, whilst also providing an empirical form and hypothesis test. These additions explain the metrics established position in an engineering context [7].

Finally, the MMD provides similar qualities as that of the Kolmogorov distance whilst also providing a highly interpretable witness function for assessing and diagnosing key differences. MMD is a non-parametric approach, but because it requires two i.i.d. samples it will be more computationally expensive in scenarios where the predictive distribution is known. The MMD is a comparable alternative to the Kolmogorov distance. The additional information provided by the witness function can be key in diagnosing key dissimilarities. Finally it is suggested that a validation procedure for probabilistic models in engineering applications would combine three key stages: hypothesis testing, applying the witness function and analysing distances.

CONCLUSION

Two families of distance/divergence measures were outlined and specific examples compared. It has been shown that a validation procedure that begins with hypothesis testing followed by using the MMD witness function as a diagnostic tool, with dis-

tances calculated in support of the prior analysis, is an effective validation strategy.

The KL-divergence, although suited for comparing two likelihoods, poses challenges when used as a validation metric. This is due to it being less interpretable and difficult to estimate in higher dimensional spaces. The other f-divergence in this paper, the Hellinger distance, has been shown to be highly interpretable, however the lack of a non-parametric estimator with known and well studied convergence properties mean that it's application as a validation metric in engineering applications may be sparse.

IPMs have been identified here as suitable metrics for use in validating probabilistic engineering models. Both the Kolmogorov and MMD have appropriate hypothesis tests and empirical estimators. The MMD has the additional strength of providing a witness function. This is a key diagnostic tool in identifying improvements and visually interpreting dissimilarities in between two distributions.

Further research should be performed in understanding the choice of kernel's in MMD and applying the metric in high dimensional scenarios. Additionally, both the K-S and MMD two sample tests should be compared with Bayesian hypothesis testing to identify the most appropriate hypothesis test for engineering applications.

REFERENCES

[1] Gardner, P., Barthorpe, R. J., and Lord, C., 2016. "The Development of a Damage Model for the use in Machine Learning Driven SHM and Comparison with Conventional SHM Methods". In Proceedings of ISMA2016 International Conference on Noise and Vibration Engineering, pp. 3333–3346.

[2] Gardner, P., Lord, C., and Barthorpe, R. J., 2017. "Bayesian calibration and bias correction for forward model-driven SHM". *Proceedings of the The 11th International Workshop on Structural Health Monitoring*, pp. 2019–2027.

[3] Gardner, P., Lord, C., and Barthorpe, R. J., 2018. "Bayesian History Matching for Forward Model-Driven Structural Health Monitoring". *Proceedings of IMAC XXXVI*.

[4] Sankararaman, S., and Mahadevan, S., 2015. "Integration of model verification, validation, and calibration for uncertainty quantification in engineering systems". *Reliability Engineering & System Safety,* ***138***, pp. 194–209.

[5] Liu, Y., Chen, W., Arendt, P., and Huang, H.-Z., 2011. "Toward a Better Understanding of Model Validation Metrics". *Journal of Mechanical Design,* ***133***(7), p. 071005.

[6] Lloyd, J. R., 2015. "Statistical Model Criticism using Kernel Two Sample Tests". In NIPS, pp. 1–9.

[7] Oberkampf, W. L., and Roy, C. J., 2010. "Verification and Validation in Scientific Computing". *Verification and Validation in Scientific Computing*(May), pp. 371–408.

[8] Xu, H., Jiang, Z., Apley, D. W., and Chen, W., 2015. "New Metrics for Validation of Data-Driven Random Process Models in Uncertainty Quantification". *Journal of Verification, Validation and Uncertainty Quantification,* ***1***(2), p. 021002.

[9] Wang, Z., Fu, Y., Yang, R.-J., Barbat, S., and Chen, W., 2016. "Validating Dynamic Engineering Models Under Uncertainty". *Journal of Mechanical Design,* ***138***(11), p. 111402.

[10] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D., 2017. Variational Inference: A Review for Statisticians.

[11] Murphy, K. P., 2012. *Machine Learning: A Probabilistic Perspective*.

[12] Lin, J., 1991. "Divergence Measures Based on the Shannon Entropy". *IEEE Transactions on Information Theory,* ***37***(1), pp. 145–151.

[13] Wang, Q., Kulkarni, S. R., and Verdú, S., 2005. "Divergence estimation of continuous distributions based on data-dependent partitions". *IEEE Transactions on Information Theory,* ***51***(9), pp. 3064–3074.

[14] Nguyen, X. L., Wainwright, M. J., and Jordan, M. I., 2007. "Nonparametric estimation of the likelihood ratio and divergence functionals". In IEEE International Symposium on Information Theory - Proceedings, pp. 2016–2020.

[15] Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. G., 2009. "On integral probability metrics, \phi-divergences and binary classification". pp. 1–18.

[16] Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J., 2008. "A kernel statistical test of independence". *Neural Information Processing Systems*, pp. 585–592.

[17] Gretton, A., 2012. "A Kernel Two-Sample Test". *Journal of Machine Learning Research,* ***13***, pp. 723–773.

[18] Dougal J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, A. G., 2017. "Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy". *Proceedings of ICLR 2017*, apr.

[19] Andrianakis, I., Vernon, I. R., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., Goldstein, M., and White, R. G., 2015. "Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda". *PLoS Computational Biology,* ***11***(1).