Testing and Learning on Distributions with Symmetric Noise Invariance

Ho Chung Leon Law [1] Christopher Yau [1 2] Dino Sejdinovic [1]

Abstract

Kernel embeddings of distributions and the Maximum Mean Discrepancy (MMD), the resulting distance between distributions, are useful tools for fully nonparametric two-sample testing and learning on distributions. However, it is rarely that all possible differences between samples are of interest – discovered differences can be due to different types of measurement noise, data collection artefacts or other irrelevant sources of variability. We propose distances between distributions which encode invariance to additive symmetric noise, aimed at testing whether the assumed true underlying processes differ. Moreover, we construct invariant features of distributions, leading to learning algorithms robust to the impairment of the input distributions with symmetric additive noise. Such features lend themselves to a straightforward neural network implementation and can thus also be learned given a supervised signal.

1. Introduction

There are many sources of variability in data, and not all of them are pertinent to the questions that a data analyst may be interested in. Consider, for example, a nonparametric two-sample testing problem, recently attracting significant research interest, especially in the context of kernel embeddings of distributions (Gretton et al., 2012; Chwialkowski et al., 2015; Jitkrittum et al., 2016). We observe samples $\{X_{1j}\}_{j=1}^{N_1}$ and $\{X_{2j}\}_{j=1}^{N_2}$ from two data generating processes P_1 and P_2, respectively, and would like to test the null hypothesis that $P_1 = P_2$ without making any parametric assumptions on these distributions. With a large sample-size, the minutiae of the two data generating processes are uncovered (e.g. slightly different calibration of the data collecting equipment, different numerical precision, different conventions of dealing with edge-cases across the two processes), and we ultimately reject the null hypothesis, even if the sources of variation across the two samples may be irrelevant for the analysis. Similarly, we may be interested in *learning on distributions* (Muandet et al., 2012; Szabó et al., 2015; Sutherland et al., 2016), where the appropriate level of granularity in the data is distributional. For example, each label y_i in supervised learning is associated to a whole bag of observations $B_i = \{X_{ij}\}_{j=1}^{N_i}$ – assumed to come from a probability distribution P_i, or we may be interested in clustering such bags of observations. Again, nonparametric distances used in such contexts to facilitate a learning algorithm on distributions, such as Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), can be sensitive to irrelevant sources of variation and may lead to suboptimal or even misleading results. Moreover, even if the training data is not impaired by measurement noise, the testing data may be, i.e. we may be in a situation of *covariate shift* (Quinonero-Candela et al., 2009) on distribution inputs, and again building predictors which are invariant to noise is of interest.

While it may be tempting to revert back to a parametric setup and work with simple, easy to interpret models, we argue that a different approach is possible: we stay within a nonparametric framework, but *encode invariances* to sources of variation assumed to be irrelevant. In this contribution, we focus on *invariances to symmetric additive noise* on each of the data generating distributions. Namely, assume that the i-th sample $\{X_{ij}\}_{j=1}^{N_i}$ we observe does not follow the distribution P_i of interest but instead its convolution $P_i \star \mathcal{E}_i$ with some unknown noise distributions $\mathcal{E}_i$ assumed to be symmetric about 0 (we will additionally require that the noise distribution has a positive characteristic function). We would like to assess the differences between P_i and $P_{i'}$ while allowing $\mathcal{E}_i$ and $\mathcal{E}_{i'}$ to differ in an arbitrary way. We investigate two approaches to this problem: (1) measuring the degree of asymmetry of the paired differences $\{X_{ij} - X_{i'j}\}$, and (2) comparing the *phase functions* of the corresponding samples. While the first approach is simpler and presents a sensible solution for the two-sample testing problem, we demonstrate that phase functions give a much better gauge on the *relative comparisons* between bags of observations, as required for learning on distributions. We construct a neural network formulation of the latter approach, allowing backpropagation for learning dis-

[1] Department of Statistics, University Of Oxford, UK [2] Centre for Computational Biology, University of Birmingham, UK. Correspondence to: Ho Chung Leon Law <ho.law@spc.ox.ac.uk>.

criminative distribution features based on phase functions.

The paper is outlined as follows. In section 2, we provide a brief overview of the background and notation. In section 3, we provide details of the construction of phase features, before discussing learning of discriminative frequencies in phase and Fourier features in section 4. In section 5, we discuss the approach based on asymmetry in paired differences for two sample testing with invariances. Section 6 provides experimental details on synthetic and real data, before concluding in section 7.

2. Background and Setup

We will say that a random vector E on $\mathbb{R}^d$ is a *symmetric positive definite (SPD) noise component* if its characteristic function is positive, i.e. $\varphi_E(\omega) = \mathbb{E}_{X\sim E}\left[\exp(i\omega^\top E)\right] > 0$, $\forall\omega \in \mathbb{R}^d$. This means that E is (1) symmetric about zero, i.e. E and $-E$ have the same distribution and (2) if it has a density, this density must be a positive definite function (Rossberg, 1995). Note that many distributions used to model additive noise, including the spherical zero-mean Gaussian distribution, as well as multivariate Laplace, Cauchy or Student's t (but not uniform), are all SPD noise components.

Following the terminology similar to that of Delaigle & Hall (2016), we will say that a random vector X on $\mathbb{R}^d$ is *decomposable* if its characteristic function can be written as $\varphi_X = \varphi_{X_0}\varphi_E$, with $\varphi_E > 0$. Thus, if X can be written in the form $X = X_0 + E$, where X_0 and E are independent and E is an SPD noise component, then X is decomposable. We will say that X is *indecomposable* if it is not decomposable. In this paper, we will assume that only the indecomposable components of distributions are of interest and we will construct tools to directly measure differences between these indecomposable components, encoding invariance to other sources of variability. The class of Borel Probability measures on $\mathbb{R}^d$ will be denoted $\mathcal{M}_+^1(\mathbb{R}^d)$, while the class of indecomposable probability measures will be denoted by $\mathcal{I}(\mathbb{R}^d) \subseteq \mathcal{M}_+^1(\mathbb{R}^d)$.

2.1. Kernel Embeddings and Fourier Features

For any positive definite function $k\colon \mathcal{X}\times\mathcal{X} \mapsto \mathbb{R}$, there exists a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ of real-valued functions on $\mathcal{X}$. Function $k(\cdot, x)$ is an element of $\mathcal{H}_k$ and represents evaluation at x, i.e. $\langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$, $\forall f \in \mathcal{H}_k$, $\forall x \in \mathcal{X}$. The kernel mean embedding (cf. Muandet et al. (2016) for a recent review) of a probability measure P is defined by $\mu_P = \mathbb{E}_{X\sim P}[k(\cdot, X)] = \int_{\mathcal{X}} k(\cdot, x)dP(x)$. The Maximum Mean Discrepancy (MMD) between probability measures P and Q is then given by $\|\mu_P - \mu_Q\|_{\mathcal{H}_k}$. For shift-invariant kernels on $\mathbb{R}^d$, using Bochner's characterisation of positive definiteness (Wendland, 2004, 6.2), the squared MMD can be written as a weighted L_2-distance between characteristic functions (Sriperumbudur et al., 2010, Corollary 4)

$$\|\mu_P - \mu_Q\|_{\mathcal{H}_k}^2 = \int_{\mathbb{R}^d} \left|\varphi_P(\omega) - \varphi_Q(\omega)\right|^2 d\Lambda(\omega), \quad (1)$$

where Λ is the non-negative spectral measure (inverse Fourier transform) of kernel k as a function of $x - y$, while $\varphi_P(\omega)$ and $\varphi_Q(\omega)$ are the characteristic functions of probability measures P and Q.

Bochner's theorem is also used to construct random Fourier features (Rahimi & Recht, 2007) for fast approximations to kernel methods in order to approximate a pre-specified shift-invariant kernel by a finite dimensional explicit feature map. If we can draw samples from its spectral measure Λ, we can approximate k by[1]

$$\begin{aligned}\tilde{k}(x,y) &= \frac{1}{m}\sum_{j=1}^{m}\left[\cos(\omega_j^T x)\cos(\omega_j^T y) + \sin(\omega_j^T x)\sin(\omega_j^T y)\right] \\ &= \langle\phi(x), \phi(y)\rangle_{\mathbb{R}^{2m}}\end{aligned}$$

where $\omega_1, \ldots, \omega_m \sim \Lambda$, giving an explicit feature map $\phi(x) := \sqrt{\frac{1}{m}}\left[\cos\left(\omega_1^\top x\right), \sin\left(\omega_1^\top x\right)\ldots, \cos\left(\omega_m^\top x\right), \sin\left(\omega_m^\top x\right)\right]$, whereby the explicit computation of the kernel matrix is not needed and the computational complexity is reduced. This also allows computation with the approximate, finite-dimensional embeddings $\tilde{\mu}_P = \Phi(P) = \mathbb{E}_{X\sim P}\phi(X) \in \mathbb{R}^{2m}$, which can be understood as the evaluations (real and complex part stacked together) of the characteristic function φ_P at frequencies $\omega_1, \ldots, \omega_m$. We will refer to the approximate embeddings $\Phi(P)$ as Fourier features of distribution P.

2.2. Learning on Distributions

Kernel embeddings can be used for supervised learning on distributions. Assume we have a training set $\{B_i, y_i\}_{i=1}^n$, where input $B_i = \{x_{ij}\}_{j=1}^{N_i}$ is a bag of samples taking values in $\mathcal{X}$, and y_i is a response. Given a kernel $k\colon \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, we first map each B_i to the empirical embedding $\mu_{\hat{P}_i} = \frac{1}{N_i}\sum_{j=1}^{N_i} k(\cdot, x_{ij}) \in \mathcal{H}_k$ and then can apply any positive definite kernel on $\mathcal{H}_k$ as the kernel on bag inputs, e.g. linear kernel $K(B_i, B_i') = \langle\mu_{\hat{P}_i}, \mu_{\hat{P}_{i'}}\rangle_{\mathcal{H}_k}$, in order to perform classification (Muandet et al., 2012) or regression (Szabó et al., 2015). Approximate kernel embeddings have also been applied in this context (Sutherland et al., 2016). Distribution regression has recently been applied to Approximate Bayesian Computation (ABC) in order to construct optimal summary statistics for posterior inference of model parameters (Mitrovic et al., 2016). In

[1] a *complex feature map* $\phi(x) = \sqrt{\frac{1}{m}}\left[\exp\left(i\omega_1^\top x\right), \ldots, \exp\left(i\omega_m^\top x\right)\right]$ can also be used, but we follow the convention of real-valued Fourier features, since kernels of interest are typically real-valued.

the Appendix, we discuss a similar application to ABC of invariant distribution representations developed in this paper.

3. Phase Discrepancy and Phase Features

While MMD and kernel embeddings are related to characteristic functions, and indeed the same connection forms a basis for fast approximations to kernel methods using random Fourier features (Rahimi & Recht, 2007), the relevant notion in our context is the *phase function* of a probability measure, recently used for nonparametric deconvolution by Delaigle & Hall (2016). In this section, we overview this formalism. Based on the empirical phase functions, we will then derive and investigate hypothesis testing and learning framework using *phase features of distributions*.

In nonparametric deconvolution (Delaigle & Hall, 2016), the goal is to estimate the density function f_0 of a univariate random variable X_0, but in general we only have noisy data samples $X_1, \ldots, X_n \overset{iid}{\sim} X = X_0 + E$, where E denotes an independent noise term. Even though the distribution of E is unknown, making the assumption that E has a positive characteristic function (i.e. it is an SPD noise component), and that X_0 is indecomposable, i.e. X_0 itself does not contain any SPD noise components, Delaigle & Hall (2016) show that it is possible to obtain consistent estimates of f_0. They distinguish between the symmetric noise and the underlying indecomposable component by matching phase functions, defined as

$$\rho_X(\omega) = \frac{\varphi_X(\omega)}{|\varphi_X(\omega)|} = \exp\left(i\tau_X(\omega)\right) \tag{2}$$

where $\varphi_X(\omega)$ denotes the characteristic function of X. We follow the terminology of Delaigle & Hall (2016) but note that τ_X may also be called the phase function. Observe that $|\rho_X(\omega)| = 1$, and thus we are effectively removing the amplitude information from the characteristic function. For a SPD noise component E, the phase function is $\rho_E(\omega) \equiv 1$. But then since $\varphi_X = \varphi_{X_0}\varphi_E$, we have that $\rho_{X_0} = \rho_X = \varphi_X / |\varphi_X|$, i.e. the phase function is invariant to additive SPD noise components. This motivates us to construct explicit feature maps of distributions with the same property. In analogy to the MMD, we first define the phase discrepancy (PhD) as a weighted L_2-distances between the phase functions:

$$\mathrm{PhD}(X,Y) = \int_{\mathbb{R}^d} |\rho_X(\omega) - \rho_Y(\omega)|^2 \, d\Lambda(\omega)$$

for some non-negative measure Λ (w.l.o.g. a probability measure). Now suppose we write $X = X_0 + U$, $Y = Y_0 + V$, where U and V are SPD noise components. This then implies $\rho_X = \rho_{X_0}$ and $\rho_Y = \rho_{Y_0}$ Λ-everywhere, so that $\mathrm{PhD}(X,Y) = \mathrm{PhD}(X_0,Y_0)$. It is clear then that the PhD is not affected by additive SPD noise components, so it captures desired invariance. A natural next question is whether PhD for Λ supported everywhere is in fact a proper metric on the indecomposable probability measures $\mathcal{I}(\mathbb{R}^d)$, which are assumed to be of interest. Unfortunately, the answer is no, i.e. one can find indecomposable random variables X and Y s.t. $\rho_X = \rho_Y$ and thus $\mathrm{PhD}(X,Y) = 0$. An example is given in the Appendix. While such cases appear contrived, we will henceforth need to restrict attention to a subset of indecomposable probability measures $\mathcal{P}(\mathbb{R}^d) \subset \mathcal{I}(\mathbb{R}^d)$, which are uniquely determined by phase functions, i.e. $\forall P, Q \in \mathcal{P}(\mathbb{R}^d) : \rho_P = \rho_Q \Rightarrow P = Q$.
We now have the two following propositions (proofs are given in the Appendix).

Proposition 1.

$$PhD(X,Y) = 2 - 2\int \left(\tfrac{\mathbb{E}\xi_\omega(X)}{\|\mathbb{E}\xi_\omega(X)\|}\right)^\top \left(\tfrac{\mathbb{E}\xi_\omega(Y)}{\|\mathbb{E}\xi_\omega(Y)\|}\right) d\Lambda(\omega)$$

where $\xi_\omega(x) = \left[\cos\left(\omega^\top x\right), \sin\left(\omega^\top x\right)\right]^\top$ *and* $\|\cdot\|$ *denotes the standard* L_2 *norm.*

Moreover,

Proposition 2.

$$K(P_X, P_Y) = \int \left(\frac{\mathbb{E}\xi_\omega(X)}{\|\mathbb{E}\xi_\omega(X)\|}\right)^\top \left(\frac{\mathbb{E}\xi_\omega(Y)}{\|\mathbb{E}\xi_\omega(Y)\|}\right) d\Lambda(\omega)$$

is a positive definite kernel on probability measures.

Now, we can construct an approximate explicit feature map for kernel K. Taking a sample $\{\omega_i\}_{i=1}^m \sim \Lambda$, we define $\Psi : P_X \mapsto \mathbb{R}^{2m}$ given by:

$$\Psi(P_X) = \sqrt{\frac{1}{m}} \left[\frac{\mathbb{E}\xi_{\omega_1}(X)}{\|\mathbb{E}\xi_{\omega_1}(X)\|}, \ldots, \frac{\mathbb{E}\xi_{\omega_m}(X)}{\|\mathbb{E}\xi_{\omega_m}(X)\|}\right]$$

We will refer to $\Psi(\cdot)$ as the *phase features*. Note that these are very similar to Fourier features, but the $\cos, \sin$-pair corresponding to each frequency is normalised to have unit L_2 norm. In other words, $\Psi(\cdot)$ can be thought of as evaluations of the phase function at the selected frequencies. By construction, phase features are invariant to additive SPD noise components. For an empirical measure, we simply have the following:

$$\Psi(\hat{P}_X) = \sqrt{\frac{1}{m}} \left[\frac{\hat{\mathbb{E}}\xi_{\omega_1}(X)}{\left\|\hat{\mathbb{E}}\xi_{\omega_1}(X)\right\|}, \ldots, \frac{\hat{\mathbb{E}}\xi_{\omega_m}(X)}{\left\|\hat{\mathbb{E}}\xi_{\omega_m}(X)\right\|}\right] \tag{3}$$

where we have replaced the expectations by their empirical estimates. Because $\left\|\Psi(\hat{P}_X)\right\| = 1$, we can construct

$$\begin{aligned} \widehat{\mathrm{PhD}}(\hat{P}_X, \hat{P}_Y) &= \left\|\Psi(\hat{P}_X) - \Psi(\hat{P}_Y)\right\|^2 \\ &= 2 - 2\Psi(\hat{P}_X)^\top \Psi(\hat{P}_Y), \end{aligned}$$

Algorithm 1 Phase/Fourier Neural Network

```
Input: Batch of bag of samples $X \in \mathbb{R}^{b \times N \times p}$, where $b$ is the batch size, $N$ is the bag size and $p$ is the dimension

Output: Classification or Regression Output
1. Compute $f(X) = XW$ where $W \in \mathbb{R}^{p \times m}$
2. Apply a sin and cos activation function
       $l_1(X) = [\sin(f(X)) \cos(f(X))]$
3. Apply mean pooling operation over $N$, effectively computing $\hat{\mathbb{E}}\xi_{\omega_i}(X)$ for each $\omega_i \in \mathbb{R}^p$
       $l_2(X) = \left[\hat{\mathbb{E}}\xi_{\omega_1}(X), \dots, \hat{\mathbb{E}}\xi_{\omega_m}(X)\right] \in \mathbb{R}^{2m}$
4. For Phase Neural Network, compute $\left\|\hat{\mathbb{E}}\xi_{\omega_1}(X)\right\|$ for each frequency and normalise to obtain:
       $l_3(X) = \left[\frac{\hat{\mathbb{E}}\xi_{\omega_1}(X)}{\left\|\hat{\mathbb{E}}\xi_{\omega_1}(X)\right\|}, \dots, \frac{\hat{\mathbb{E}}\xi_{\omega_m}(X)}{\left\|\hat{\mathbb{E}}\xi_{\omega_m}(X)\right\|}\right]$
5. Batch Normalisation Layer
6. Output layer
```

which is an unbiased estimator of $\text{PhD}(\hat{P}_X, \hat{P}_Y)$. In summary, $\Psi(\hat{P}) \in \mathbb{R}^{2m}$ is an explicit feature vector constructed to be invariant to additive SPD noise components. It can now be directly applied to (1) two-sample testing up to SPD components, where the distance between the phase features, i.e. an estimate of the phase discrepancy, can be used as a summary statistic, and (2) learning on distributions where it is believed that only the indecomposable part of a distribution contains the signal about the response. Broadly speaking, even if this assumption is invalid, given that the underlying distribution is irregular, it may still be useful to encode invariance as long as the benefit of removing the noise components outweighs the signal in the SPD part of the distribution.

4. Learning Discriminative Features

To construct the approximate mean embeddings, we compute an explicit feature map by taking averages of the Fourier features. They are given by

$$\Phi(\hat{P}_X) = \sqrt{\frac{1}{m}} \left[\hat{\mathbb{E}}\xi_{\omega_1}(X), \dots, \hat{\mathbb{E}}\xi_{\omega_m}(X)\right] \in \mathbb{R}^{2m}.$$

The phase function approach similarly explicitly computes a feature map for each bag of samples, but it has an additional normalisation term over each frequency as in (3). In both these cases, given a supervised signal, we can optimise a set of frequencies $\{w_i\}_{i=1}^m$ that will give us a useful representation and good discriminative performance. In other words, we are no longer focusing on a specific shift-invariant kernel k (equivalently, a specific spectral measure Λ), but are learning discriminative Fourier/phase features.

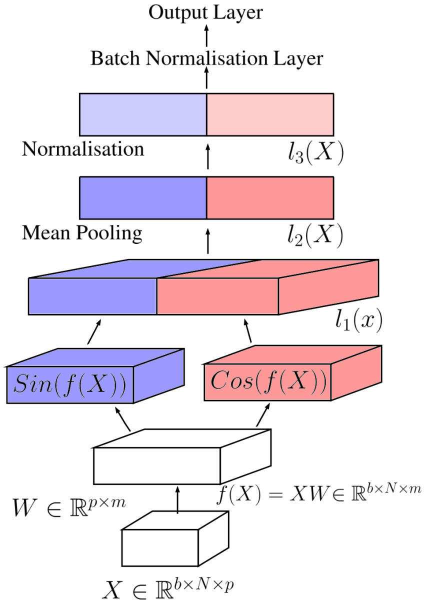
Output Layer
Batch Normalisation Layer
Normalisation
$l_3(X)$
Mean Pooling
$l_2(X)$
$l_1(x)$
$Sin(f(X))$
$Cos(f(X))$
$W \in \mathbb{R}^{p \times m}$
$f(X) = XW \in \mathbb{R}^{b \times N \times m}$
$X \in \mathbb{R}^{b \times N \times p}$

Figure 1. Main structure of the phase neural network

To learn these discriminative Fourier/phase features, we construct a neural network with special activation functions, pooling layers and batch normalisation layers as shown in Algorithm 1 and Figure 1. Although the batch normalisation is not required, it is highly recommended for faster training of the network (Ioffe & Szegedy, 2015), due to the normalisation for the phase neural network in step 5 of the algorithm. Because of the neural network structure, we can take advantage of the rich literature, as well as alter the network in order to target a variety of different problems. For example, setting now the loss function as the squared loss, cross entropy or pinball loss, we can solve tasks in regression, classification or quantile regression on distributional inputs with discriminative frequencies. The Fourier neural network can also be extended to inputs in $\mathbb{R}^p$ for normal regression and classification problems by removing the mean pooling operation in step 3 of the algorithm.

5. Asymmetry in Paired Differences

We now consider a separate approach to nonparametric two-sample test, where we wish to test the null hypothesis that $H_0 : P = Q$ vs. the general alternative, but we only have iid samples arising from $X \sim P \star \mathcal{E}_1$ and $Y \sim Q \star \mathcal{E}_2$, i.e.

$$X = X_0 + U \qquad Y = Y_0 + V$$

where $X_0 \sim P$, $Y_0 \sim Q$ lie in the space of $\mathcal{P}(\mathbb{R}^d)$ of indecomposable distributions uniquely determined by phase functions and U and V are SPD noise components. Under the null hypothesis, since X_0 has the same distribution as Y_0, then so do $X - Y = X_0 - Y_0 + U - V$ and $Y - X = Y_0 - X_0 + V - U$ as $U - V$ is symmetric. Moreover, $\varphi_{X-Y} = |\varphi_{X_0}|^2 \varphi_U \varphi_V > 0$, so $X - Y$ is SPD. Conversely, if we assume that $X - Y$ is SPD, i.e. $\varphi_X \overline{\varphi_Y} > 0$, then $\rho_{X_0} \overline{\rho_{Y_0}} > 0$. Since $|\rho_{X_0}| = |\rho_{Y_0}| = 1$, this implies that $\rho_{X_0} = \rho_{Y_0}$, and hence $X_0 \overset{d}{=} Y_0$, since we assumed that X_0 and Y_0 belong to $\mathcal{P}(\mathbb{R}^d)$. Hence, we have that

$$X - Y \text{ is SPD} \iff X_0 \overset{d}{=} Y_0.$$

This motivates us to simply perform a two-sample test on $X - Y$ and $Y - X$ since its rejection would imply rejection of $X_0 \overset{d}{=} Y_0$. However, note that this is a test for symmetry only and that for consistency against all alternatives, positivity of characteristic function would need to be checked separately. Now, given two i.i.d. samples $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$ with n even, we split the two samples into two halves and compute $W_i = X_i - Y_i$ on one half and $Z_i = Y_i - X_i$ on the other half, and perform a nonparametric two sample test on W and Z (which are, by construction, independent of each other). The advantage of this regime is that we can use any two-sample test – in particular in this paper, we will focus on the linear time mean embedding (ME) test (Jitkrittum et al., 2016), which was found to have performance similar to or better than the original MMD two-sample test (Gretton et al., 2012), and explicitly formulates a criterion which maximises the test power. We will refer to the resulting test on paired differences as the Symmetric Mean Embedding (SME). Following the numerical evaluation in section 6.1 where we highlight the properties of each test on simulated and real datasets, we recommend using the ME test and the SME test together, as this allows us to understand whether the difference in the two sets of samples can be explained by SPD noise components, or whether there are also differences in the indecomposable parts of the underlying distributions.

It is tempting to also consider learning on distributions with invariances using this formalism. However note that the MMD on paired differences is *not invariant to the additive SPD noise components* under the alternative, i.e. in general

$$\mathrm{MMD}(X - Y, Y - X) \neq \mathrm{MMD}(X_0 - Y_0, Y_0 - X_0).$$

This means that the paired differences approach to learning is sensitive to the actual type and scale of the additive SPD noise components, as we demonstrate in section 6.2.

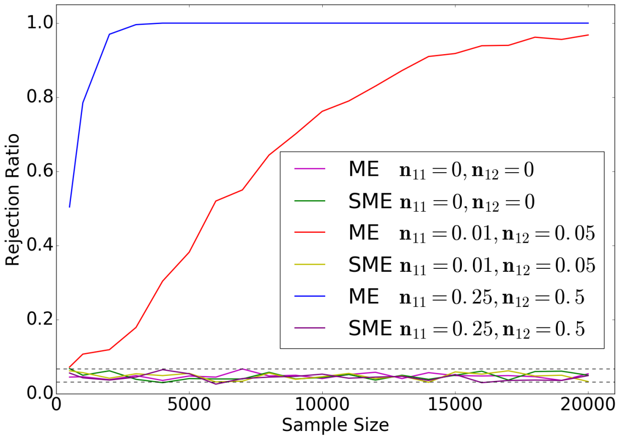

Figure 2. Type I error up to various additional symmetric noise for samples in the synthetic χ^2 dataset. n_{11} denotes the noise to signal ratio for the first set of samples and n_{12} for the second set.

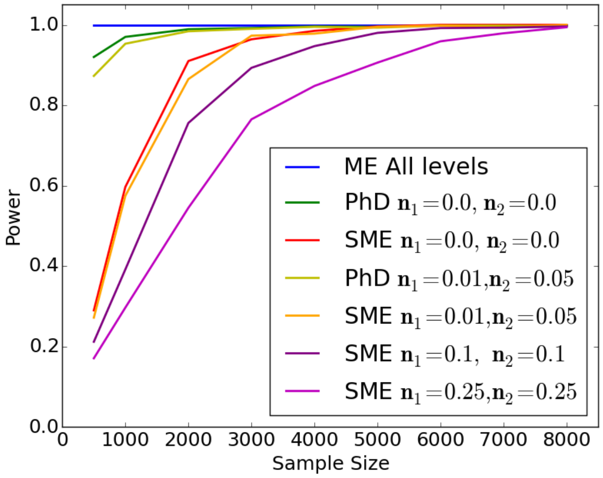

Figure 3. Power under different additional symmetric noise and samples in the synthetic χ^2 dataset. n_1 denotes the noise to signal ratio for the X set of samples and n_2 denotes the noise to signal ratio for the Y set of samples. Note here the PhD test for the small noise levels where the type I error is controlled is also included.

6. Experimental Results

6.1. Two-Sample Tests with Invariances

In this section, we demonstrate the performance of the SME test and the PhD test on both artificial and real-world data for testing the hypothesis $H_0 : X_0 \overset{d}{=} Y_0$ based on samples $\{X_i\}_{i=1}^N$ from $X_0 + U$ and $\{Y_i\}_{i=1}^N$ from $Y_0 + V$, where U and V are arbitrary SPD noise components (we assume the same number of samples for simplicity). SME test follows the setup in Jitkrittum et al. (2016) but applied to $\{X_i - Y_i\}_{i=1}^{N/2}$ and $\{Y_i - X_i\}_{i=N/2+1}^{N}$. For the PhD test, we use as the test statistic the Monte Carlo estimate $\widehat{\mathrm{PhD}}(\hat{P}_X, \hat{P}_Y)$ of the phase discrepancy, as it is invariant to additive SPD components. It is unclear what the exact form of the null distribution is, so we use a permutation test, by recomputing the statistic on the samples which are first merged and then split in the original proportions. While we are combining samples with different distributions, the permutation test is still correct since, under the null hypothesis $X_0 \overset{d}{=} Y_0$, the resulting characteristic function φ_{null} of the mixture can be written as

$$\varphi_{null} = \frac{1}{2}\varphi_{X_0}\varphi_U + \frac{1}{2}\varphi_{X_0}\varphi_V = \varphi_{X_0}(\frac{1}{2}\varphi_U + \frac{1}{2}\varphi_V),$$

and since the mixture of the noise terms is also SPD, we have that $\rho_{null} = \rho_{X_0} = \rho_{Y_0}$. For our experiments, we denote by N the sample size, d the dimension of the samples, and we take $\alpha = 0.05$ to be the significance level. In the SME test, we take the number of test locations J to be 10, and use 20% of the samples to optimise the test locations. All experimental results are averaged over 1000 runs, where each run repeats the simulation or randomly samples without replacement from the real dataset.

Synthetic example: Noisy χ^2. We start by demonstrating our tests with invariances on a simulated dataset where X_0 and Y_0 are random vectors with dimension $d = 5$, each dimension is the same in distribution and follows $\chi^2(4)/4$ and $\chi^2(8)/8$ respectively. Note that these distributions have the same mean (1) but different variances ($1/2$ and $1/4$ respectively). An illustration of the true and empirical phase and characteristic function with noise for these two distributions can be found in the Appendix. We construct samples $\{X_{n_1,i}\}_{i=1}^N$ and $\{Y_{n_2,i}\}_{i=1}^N$ such that $X_{n_1} \sim X_0 + U$, where $U \sim \mathcal{N}(0, \sigma_1^2 I)$ and similarly $Y_{n_2} \sim Y_0 + V$, where $V \sim \mathcal{N}(0, \sigma_2^2 I)$, n_i denotes the noise-to-signal ratio given by the ratio of variances in each dimension, i.e. $n_1 = 2\sigma_1^2$ and $n_2 = 4\sigma_2^2$. We first verify that Type I error is indeed controlled at our design level of $\alpha = 0.05$ *up to various additive SPD noise components*. This is shown in Figure 2 for the SME test when $X_0 \overset{d}{=} Y_0$, both constructed using $\chi^2(4)/4$. It is noted here that the ME test rejects the null hypothesis for even a small difference in noise levels, hence it is unable to let us *target the underlying distributions* we are concerned with. This is unlike the SME test which controls the Type I error even for large differences in noise levels. The test based on empirical phase discrepancy, on the other hand, was found to have inflated Type I error rates for large noise levels, as shown in the Appendix. This is likely due to the problems arising when sampling high frequencies, leading to a poor estimation of the empirical PhD features at large noise levels and a biased null distribution estimate. Next, we investigate the power of SME tests, as shown in Figure 3 for varying sample sizes and noise-to-signal ratios. Even at large noise levels, the SME test discovers the difference between the underlying χ^2-distributions given a sufficient sample size. For a comparison, we have also included the PhD test power for small noise levels, in which the Type I error is controlled at the design level, as shown in the Appendix. The PhD test has better power than the SME test. This is not surprising, as for the SME we have to halve the sample size in order to construct a valid test. However, recall that PhD test has an inflated Type I and should be used with caution. ME test rejects at all levels at all sample sizes as it picks up all possible differences. SME is by construction a more conservative test and SME rejection provides a much stronger statement: two samples differ even when all arbitrary additive SPD components have been stripped off. In practice, we recommend using both the SME and ME test together, as this can provide insights about the data generating processes at hand. We demonstrate an example of this next.

Higgs Dataset. The UCI Higgs dataset (Lichman, 2013) described in Baldi et al. (2014) is a dataset with 11 million observations, where the problem is to distinguish between the signal process where Higgs bosons are found, versus the background process that do not produce Higgs bosons. In particular, we will consider a two-sample test with the ME and SME test on the high level features derived by physicists, as well as a two-sample test on four extremely low level features (azimuthal angular momentum ϕ measured by four particle jets in the detector). The high level features here (in $\mathbb{R}^7$) are derived from the set of low-level features, and have been shown to have good discriminative properties in Baldi et al. (2014). Thus, we expect them to have different distributions across two processes. Denoting by X the high level features of the process without Higgs Boson, and Y as the corresponding distribution for the processes where Higgs bosons are produced, we test the null hypothesis that the indecomposable parts of X and Y agree. The results can be found in Table 1 in the Appendix, which shows that the high level features are discriminative even up to additive SPD components, with a high power for the SME and ME test even at small sample sizes (rejection rate of 0.94 at sample size $N = 500$). Now we perform the same experiment, but with the low level features, jet ϕ-momentum distribution $\in \mathbb{R}^4$, following the same experimental setup from Chwialkowski et al. (2015). The results for the ME and SME test can be found in Figure 4, where sample sizes vary between 1000 to 36000. Here we observe that while ME test clearly rejects and finds the difference between the two distributions, there is no evidence that the indecomposable parts of the joint distributions of the angular momentum actually differ. In fact, the test rejection rate remains around the chosen design level of $\alpha = 0.05$ for all sample sizes. This highlights the significance in using the SME test, suggesting that the nature of the difference between the two processes can be explained by some additive SPD noise components, potentially irrelevant for discrimination, providing an insight into the dataset. Furthermore, this also highlights the argument that given two samples from complex data collection and generation processes, a nonparametric two sample test like ME will likely reject

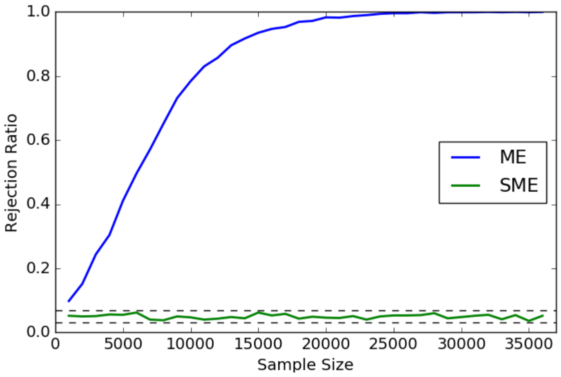

Figure 4. Rejection ratio vs. sample size for extremely low level features for Higgs dataset. Dashed line is the 99% Wald interval for 1000 repetitions for $\alpha = 0.05$.

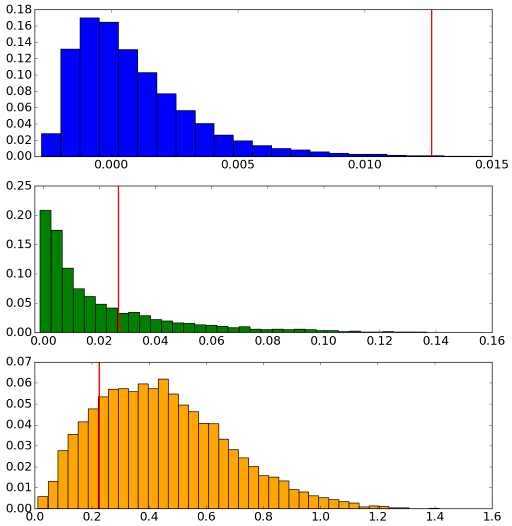

Figure 5. Histograms on various estimates for all pairs of bags with varying additive noise, red line denotes the noiseless case. **Top:** Estimated MMD on paired differences for all pair of bags, the red line given by the mean of the estimated MMD on paired differences for bags without noise. **Middle:** the squared distance between Fourier features (an estimate of MMD). **Bottom:** the squared distance between phase features (an estimate of PhD).

given sufficient sample sizes, even if the discovered difference are between the inherent noise in the two processes. With the SME test however, we can ask a much more subtle question about the differences between the assumed true underlying processes. Figures showing that the Type I error is controlled at the design level of $\alpha = 0.05$ for low and high level features can be found in Appendix.

6.2. Learning with Phase Features

While it performed well when testing the null hypothesis, the MMD on paired differences is not invariant to the additive SPD noise components under the alternative hypothesis. Using the synthetic experimental setup as before, we simulate 100 noiseless bags from the two scaled χ^2-distributions $X_0 \sim \chi^2(4)/4$ and $Y_0 \sim \chi^2(8)/8$, where each bag contains 1000 samples. We add varying levels of Gaussian noise to each bag, i.e. the bags are of the form $X_i = X_0 + \mathcal{N}(0, Z_i)$ and $Y_i = Y_0 + \mathcal{N}(0, W_i)$, where $Z_i, W_i \sim U[0, 1]$. We compute the estimate of the MMD on paired differences, the squared distance between Fourier features (an estimate of MMD) and the squared distance between phase features (an estimate of PhD) for all pairs of bags. In all computations, we used the same set of frequencies (sampled from a Gaussian distribution). We do the same for the noiseless samples (or use analytic expressions where available). The results are shown in figure 5. We see that the MMD on paired differences is not invariant to SPD noise components (clearly, the noiseless case indicated by the red line has a much higher level of asymmetry than the noisy case where due to the presence of high levels of symmetric noise, differences often do appear symmetric). This is unlike the phase features, which maintain some level of invariance even for this low signal-to-noise ratio: the estimates stay away from 0 – preserving the signal about the difference of indecomposable χ^2 components – and the mode is nearer the true value, even though a bias is clearly present (which is to be expected due to the normalisation of empirical means inside the estimator (3)). This suggests that phase features are more suitable for invariant learning on distributions than MMD on paired differences. The Fourier features are also given for comparison, but these are not expected to be invariant.

Synthetic example: regression. We now demonstrate the use of phase features in the dataset generated by the following model:

$$\begin{aligned} \theta \sim \Gamma(\alpha, \beta), \quad Z \quad &\sim \quad U[0, \sigma], \quad \epsilon \sim \mathcal{N}(0, Z), \\ X \quad &\sim \quad \frac{\Gamma\big(\theta/2, 1/2\big)}{\sqrt{2\theta}} + \epsilon, \end{aligned} \tag{4}$$

where we take $\alpha = 7.0$, $\beta = 1.0$ and θ to be the parameter we are interested in predicting, given a bag of samples from $X|\theta$. Note in the model, by normalising, our underlying signal has variance 1, this enables us to better control the signal-to-noise ratio. For the experiment, we generate 500 bags of samples from the model, where each bag contains 1000 observations as training data for the Fourier and phase neural networks. We use a mean squared error (MSE) loss, using L_2 weight regularisation with coefficient λ and perform a 3-fold cross-validation, optimising the learning rate, the number of frequencies and λ. For the test data, we generate 500 bags, and check the MSE, we repeat this process 100 times and results are shown in Figure 6. Figure 6 shows that the phase features is more stable under increasing noise, due to invariance to the additive SPD noise components, as demonstrated by the slower rate of

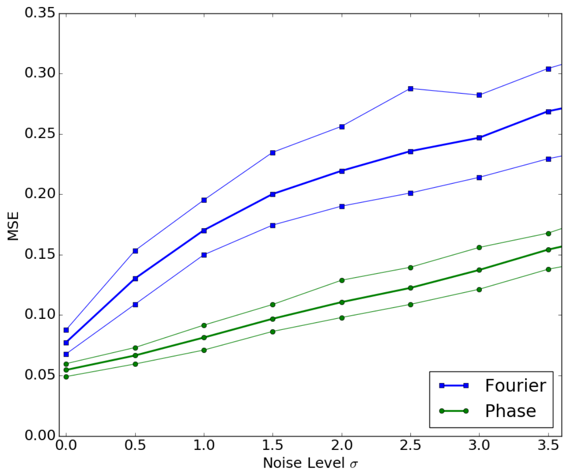

Figure 6. MSE of θ, using the Fourier and phase neural network averaged over 100 runs. Here noise σ is varied between 0 and 3.5, and the 5^{th} and the 95^{th} percentile is shown.

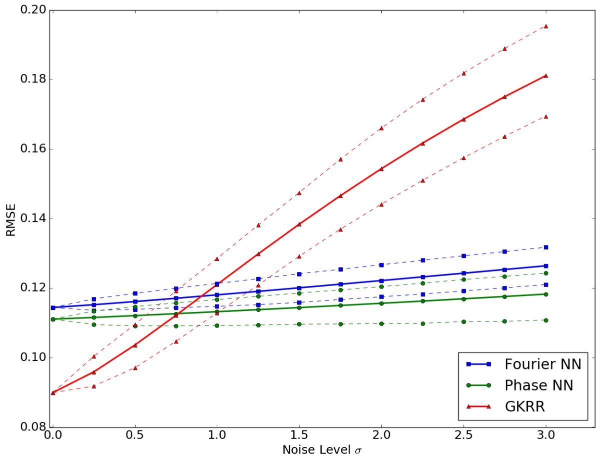

Figure 7. RMSE on the test set, corrupted by various levels of noise, using the Fourier and phase neural network and GKKR averaged over 100 runs. Here noise-to-signal ratio σ is varied between 0 and 3.0, and the 5^{th} and the 95^{th} percentile is shown.

increase of MSE relative to that of the Fourier features. It is of interest to note that under no noise, the phase features actually outperform the Fourier features slightly, possibly due to how the normalisation of Fourier features interacts with the network structure.

Covariate Shift: Aerosol Dataset. We now demonstrate the phase and Fourier neural networks on Aerosol MISR1 dataset also studied by Wang et al. (2012) and Szabó et al. (2015). We consider a situation with *covariate shift* on distribution inputs: the testing data is impaired by additive SPD components different from those in the bags used for training. Here, we have an aerosol optical depth (AOD) multi-instance learning problem with 800 bags, where each bag contains 100 randomly selected multispectral pixels within 20km radius around an AOD sensor. The label y_i for each bag is given by the AOD sensor measurements and each sample x_i is 16 dimensional. This can be understood as a distribution regression problem where each bag is treated as a set of samples from some distribution. The experimental setup is as follows: we use 640 bags for training and 160 bags for testing. Here in the bags for testing *only*, we add varying levels of Gaussian noise $\epsilon \sim \mathcal{N}(0, Z)$ to each bag, where Z is a diagonal matrix with diagonal components $z_i \sim U[0, \sigma v_i]$ with v_i being the empirical variance in dimension i across all samples, this is in order to account for different scales across different dimensions, and σ is the noise-to-signal ratio. For tuning we perform a 3-fold cross validation to select the number of frequencies, learning rate, as well as individual weight decay parameter for the two layers in the Fourier and phase neural networks. For a comparison, we use the Gaussian Bag kernel Ridge Regression (GKRR) as described in section 2.2 (i.e. we use a Gaussian kernel at the bag level, before using a linear kernel for ridge regression), tuning the bandwidth of the Gaussian kernel and the regularisation parameter in ridge regression. With the same trained model, we now measure Root Mean Square Error (RMSE) on the test sets corrupted with noise. We repeat testing 100 times with various noise corrupted test set for different σ, the results are shown in figure 7. We notice that although GKRR clearly outperforms both the Fourier NN and Phase NN in this case for the noiseless dataset (likely due to the small size of the dataset), we clearly see that the Phase neural network implementation is almost insensitive to the covariate shift in the test sets, even at low signal-to-noise ratios, whereas the performance of GKRR degrades significantly. This highlights the fact that the phase features are stable under additive SPD noise, and are applicable to the scenario where the test data are impaired by different measurement noises to that in the training set. It is also noted that the Fourier NN performs similarly to that of the Phase NN on this example. Interestingly, discriminative frequencies learnt on the training data correspond to Fourier features that are nearly normalised (i.e. they are close to unit norm like phase features - the figures showing this are in the Appendix). This means that even the Fourier NN has *learned to be approximately invariant* based on training data, indicating that even the original Aerosol data potentially has irrelevant SPD noise components.

7. Conclusion

No dataset is immune from measurement noise and often this noise differs across different data generation and collection processes. When measuring distances between distributions, can we disentangle the differences in noise from the differences in the signal? We considered two different ways to encode invariances to additive symmetric noise in those distances, each with different strengths: a nonparametric measure of asymmetry in paired sample differences and a weighted distance between the empirical phase functions. The former was used to construct a hypothesis test on whether the difference between the two generating processes can be explained away by the difference in postulated noise, whereas the latter allowed us to introduce a flexible framework for invariant feature construction and learning algorithms on distribution inputs which are robust to measurement noise and covariate shift.

References

Baldi, Pierre, Sadowski, Peter, and Whiteson, Daniel. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5, 2014.

Chwialkowski, Kacper P, Ramdas, Aaditya, Sejdinovic, Dino, and Gretton, Arthur. Fast two-sample testing with analytic representations of probability measures. In *Advances in Neural Information Processing Systems*, pp. 1981–1989, 2015.

Delaigle, Aurore and Hall, Peter. Methodology for nonparametric deconvolution when the error distribution is unknown. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):231–252, 2016.

Fearnhead, Paul and Prangle, Dennis. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012.

Gretton, Arthur, Borgwardt, Karsten M, Rasch, Malte J, Schölkopf, Bernhard, and Smola, Alexander. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, pp. 448–456, 2015.

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

Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Linnik, Yu V and Ostrovskii, IV. *Decomposition of random variables and vectors*. 1977.

Mitrovic, J., Sejdinovic, D., and Teh, Y.W. DR-ABC: Approximate Bayesian Computation with Kernel-Based Distribution Regression. In *International Conference on Machine Learning (ICML)*, pp. 1482–1491, 2016.

Muandet, Krikamol, Fukumizu, Kenji, Dinuzzo, Francesco, and Schölkopf, Bernhard. Learning from distributions via support measure machines. In *Advances in Neural Information Processing Systems 25*, pp. 10–18. 2012.

Muandet, Krikamol, Fukumizu, Kenji, Sriperumbudur, Bharath, and Schölkopf, Bernhard. Kernel mean embedding of distributions: A review and beyonds. *arXiv preprint arXiv:1605.09522*, 2016.

Quinonero-Candela, Joaquin, Sugiyama, Masashi, Schwaighofer, Anton, and Lawrence, Neil D. *Dataset Shift in Machine Learning*. The MIT Press, 2009.

Rahimi, Ali and Recht, Benjamin. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pp. 1177–1184, 2007.

Rossberg, H-J. Positive definite probability densities and probability distributions. *Journal of Mathematical Sciences*, 76(1):2181–2197, 1995.

Sejdinovic, Dino, Sriperumbudur, Bharath, Gretton, Arthur, and Fukumizu, Kenji. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Statist.*, 41(5):2263–2291, October 2013.

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

Sriperumbudur, Bharath K., Gretton, Arthur, Fukumizu, Kenji, Schölkopf, Bernhard, and Lanckriet, Gert R.G. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, August 2010.

Sutherland, Dougal J., Oliva, Junier B., Póczos, Barnabás, and Schneider, Jeff G. Linear-time learning on distributions with approximate kernel embeddings. In *Proc. AAAI Conference on Artificial Intelligence*, pp. 2073–2079, 2016.

Szabó, Zoltán, Gretton, Arthur, Póczos, Barnabás, and Sriperumbudur, Bharath K. Two-stage sampled learning theory on distributions. In *Proc. International Conference on Artificial Intelligence and Statistics, AISTATS 2015*, 2015.

Wang, Z., Lan, L., and Vucetic, S. Mixture model for multiple instance regression and applications in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2226–2237, June 2012.

Wendland, H. *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK, 2004.

A. Phase Discrepancy

In this section, we will provide further details of the definitions, calculations and proofs in section 3.
Phase discrepancy is defined as the weighted L_2-distances between the phase functions, i.e.

$$\mathrm{PhD}(X,Y) = \int \left|\rho_X\left(\omega\right) - \rho_Y\left(\omega\right)\right|^2 d\Lambda\left(\omega\right),$$

for some positive measure Λ (w.l.o.g. a probability measure). Phase discrepancy measures how much X and Y differ up to an independent SPD noise component. We first have the following proposition:

Proposition 3.

$$\mathit{PhD}(X,Y) = 2 - 2\int \frac{\mathbb{E}\cos\left(\omega^\top\left(X-Y\right)\right)}{\sqrt{\mathbb{E}\cos\left(\omega^\top\left(X-X'\right)\right)\mathbb{E}\cos\left(\omega^\top\left(Y-Y'\right)\right)}} d\Lambda(\omega).$$

Proof.

$$\begin{aligned} \mathrm{PhD}(X,Y) &= \int \left|\rho_X\left(\omega\right) - \rho_Y\left(\omega\right)\right|^2 d\Lambda\left(\omega\right) \\ &= \int \left|\rho_X\left(\omega\right)\right|^2 d\Lambda\left(\omega\right) + \int \left|\rho_Y\left(\omega\right)\right|^2 d\Lambda\left(\omega\right) \\ &\quad - \int \left(\rho_X\overline{\rho_Y} + \overline{\rho_X}\rho_Y\right) d\Lambda \\ &= 2 - \int \frac{\varphi_X\overline{\varphi_Y} + \overline{\varphi_X}\varphi_Y}{\left|\varphi_X\right|\left|\varphi_Y\right|} d\Lambda \\ &= 2 - 2\int \frac{\varphi_Z}{\sqrt{\varphi_{X-X'}\varphi_{Y-Y'}}} d\Lambda, \end{aligned}$$

where X and X' are iid, Y and Y' are iid and Z is an equal mixture of $X-Y$ and $Y-X$. Indeed,

$$\varphi_X\overline{\varphi_Y} + \overline{\varphi_X}\varphi_Y = \varphi_{X-Y} + \varphi_{Y-X} = 2\varphi_Z,$$

and

$$\varphi_{X-X'} = \varphi_X\overline{\varphi_X} = \left|\varphi_X\right|^2.$$

Note that $X-X'$,$Y-Y'$ and Z are all symmetric. Thus,

$$\varphi_Z(\omega) = \mathbb{E}\left[\cos\left(\omega^\top Z\right)\right] = \frac{1}{2}\mathbb{E}\left[\cos\left(\omega^\top\left(X-Y\right)\right)\right] + \frac{1}{2}\mathbb{E}\left[\cos\left(\omega^\top\left(Y-X\right)\right)\right] = \mathbb{E}\left[\cos\left(\omega^\top\left(X-Y\right)\right)\right].$$

Substituting provides us the result. □

Proposition 4. *$K_\omega\left(\mathsf{P}_X,\mathsf{P}_Y\right) = \left(\frac{\mathbb{E}\xi_\omega(X)}{\|\mathbb{E}\xi_\omega(X)\|}\right)^\top \left(\frac{\mathbb{E}\xi_\omega(Y)}{\|\mathbb{E}\xi_\omega(Y)\|}\right)$ is a positive definite kernel on probability measures $\forall\omega$, where here $\xi_\omega\left(x\right) = \left[\cos\left(\omega^\top x\right), \sin\left(\omega^\top x\right)\right]$, and so is $K\left(\mathsf{P}_X,\mathsf{P}_Y\right) = \int K_\omega\left(\mathsf{P}_X,\mathsf{P}_Y\right) d\Lambda(\omega)$ for any positive measure Λ.*

Proof. Define a feature map $\xi_\omega : \mathcal{X} \to \mathbb{R}^2$ with $\xi_\omega\left(x\right) = \left[\cos\left(\omega^\top x\right), \sin\left(\omega^\top x\right)\right]$, which induces a kernel on $\mathcal{X}$ given by $k_\omega(x,y) = \cos\left(\omega^\top\left(x-y\right)\right)$. Then $\kappa_\omega\left(\mathsf{P}_X,\mathsf{P}_Y\right) = \mathbb{E}\cos\left(\omega^\top\left(X-Y\right)\right) = \mathbb{E}k_\omega(X,Y) = \left(\mathbb{E}\xi_\omega(X)\right)^\top\mathbb{E}\xi_\omega(Y)$ is a valid kernel on probability measures and so is the normalised kernel

$$K_\omega\left(\mathsf{P}_X,\mathsf{P}_Y\right) = \frac{\kappa_\omega\left(\mathsf{P}_X,\mathsf{P}_Y\right)}{\sqrt{\kappa_\omega\left(\mathsf{P}_X,\mathsf{P}_X\right)\kappa_\omega\left(\mathsf{P}_Y,\mathsf{P}_Y\right)}} = \left(\frac{\mathbb{E}\xi_\omega(X)}{\|\mathbb{E}\xi_\omega(X)\|}\right)^\top \left(\frac{\mathbb{E}\xi_\omega(Y)}{\|\mathbb{E}\xi_\omega(Y)\|}\right),$$

where we used that $\mathbb{E}\cos\left(\omega^\top\left(X-X'\right)\right) = \left(\mathbb{E}\xi_\omega(X)\right)^\top\mathbb{E}\xi_\omega(X') = \|\mathbb{E}\xi_\omega(X)\|^2$. For the last claim, simply note that integrating through the positive measure preserves positive semidefinitess, i.e. $\sum\alpha_i\alpha_j K(\mathsf{P}_i,\mathsf{P}_j) = \int\left(\sum\alpha_i\alpha_j K_\omega(\mathsf{P}_i,\mathsf{P}_j)\right) d\Lambda\left(\omega\right) \geq 0$. □

As a direct corollary,

Proposition 5. $\mathrm{PhD}(X,Y) = 2 - 2K\left(\mathsf{P}_X,\mathsf{P}_Y\right) = 2\int\left(1 - \left(\frac{\mathbb{E}\xi_\omega(X)}{\|\mathbb{E}\xi_\omega(X)\|}\right)^\top \left(\frac{\mathbb{E}\xi_\omega(Y)}{\|\mathbb{E}\xi_\omega(Y)\|}\right)\right) d\Lambda(\omega).$

B. Characteristic and Phase Function Plots

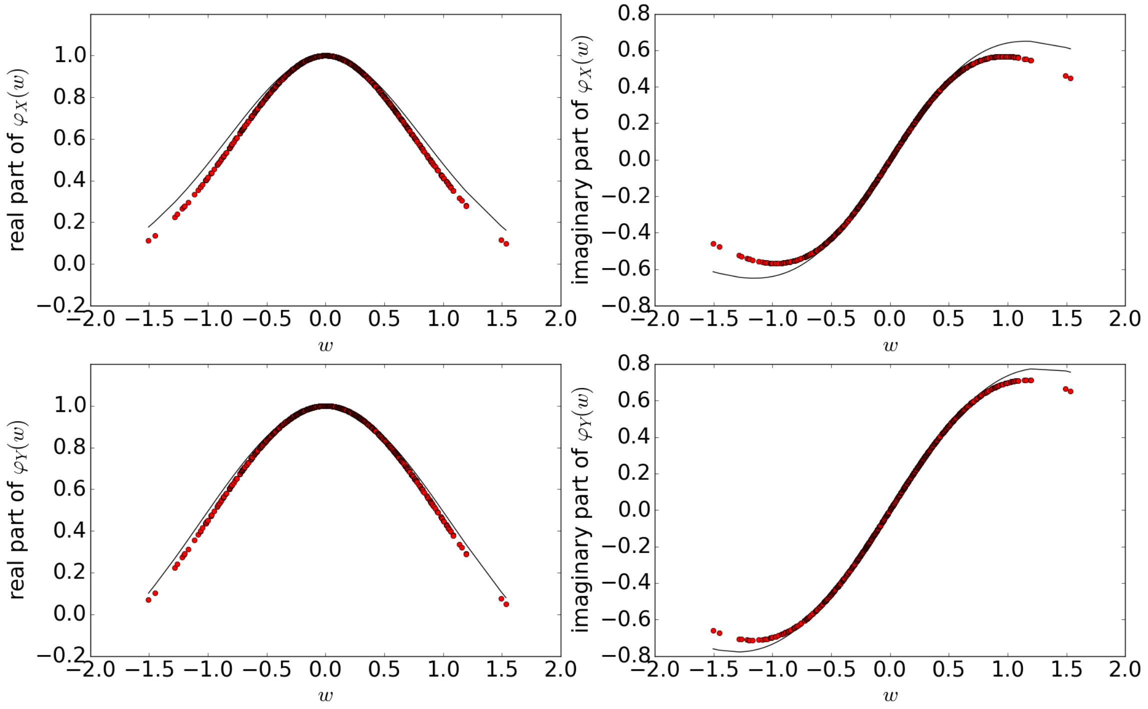

Figure 8. The black line here correspond to the real and imaginary part of the true characteristic function of the $\chi^2(4)/4$ and $\chi^2(8)/8$ distribution, denoted X, Y on the top and bottom graphs respectively. The red points denote the empirical characteristic function constructed with 750 frequencies sampled from a Gaussian kernel with $\sigma = 2$ using a bag size of 1000 observations, with some additional Gaussian noise.

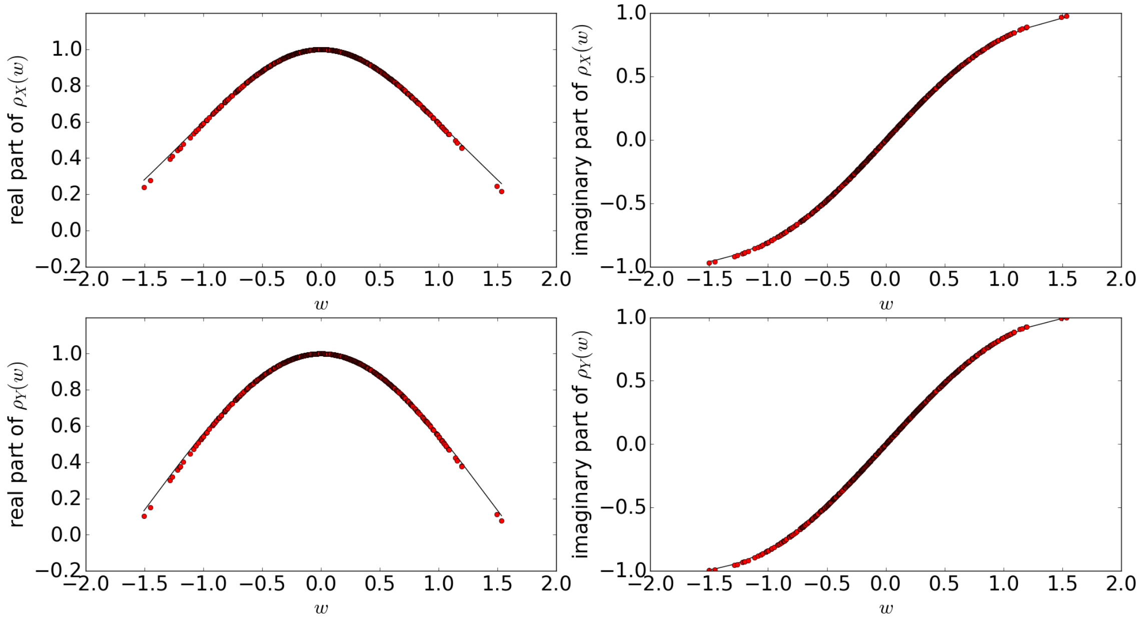

Figure 9. The black line here correspond to the real and imaginary part of the true phase function of the $\chi^2(4)/4$ and $\chi^2(8)/8$ distribution, denoted X, Y on the top and bottom graphs respectively. The red points denote the empirical phase function constructed with 750 frequencies from a Gaussian kernel with $\sigma = 2$ using a bag size of 1000 observations, with some additional Gaussian noise.

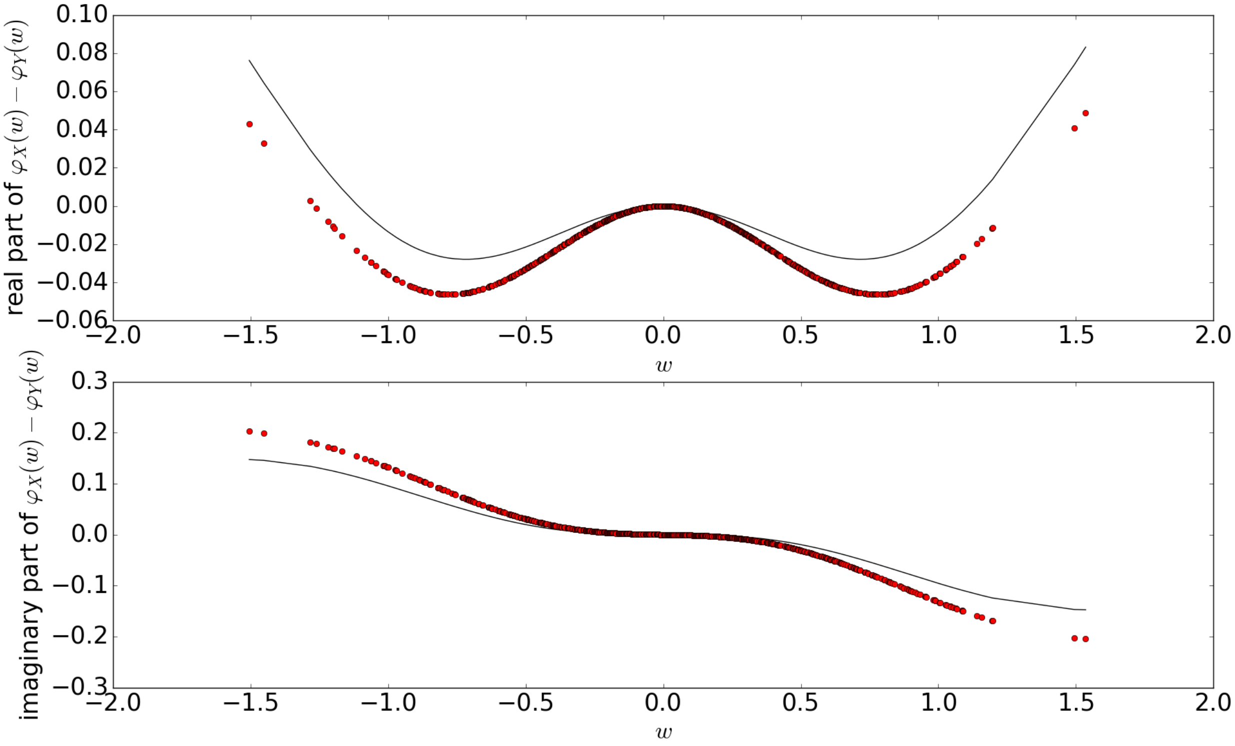

Figure 10. The top and bottom graph denotes the difference in the real and imaginary part of the characteristic function for the $\chi^2(4)/4$ and $\chi^2(8)/8$ as in figure 8.

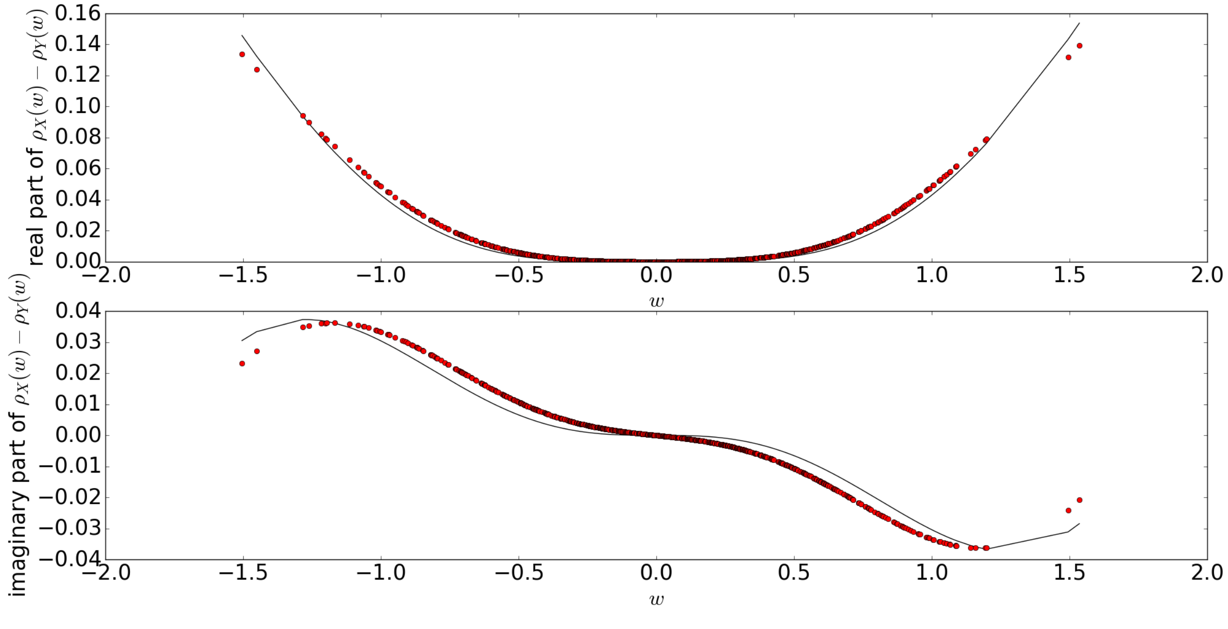

Figure 11. The top and bottom graph denotes the difference in the real and imaginary part of the phase function for the $\chi^2(4)/4$ and $\chi^2(8)/8$ as in figure 9.

C. Different Indecomposable Distributions Can Coincide in Phase

Let X and Y be (univariate) random variables with densities

$$f_X(x) = \frac{1}{\sqrt{2\pi}} x^2 \exp(-x^2/2), \quad f_Y(x) = \frac{1}{2}|x| \exp(-|x|).$$

Then it can be directly checked that their characteristic functions are given by

$$\varphi_X(\omega) = (1-\omega^2)\exp(-\omega^2/2), \quad \varphi_Y(\omega) = \frac{1-\omega^2}{(1+\omega^2)^2}.$$

Thus, the phase functions coincide and are equal to

$$\rho_X(\omega) = \rho_Y(\omega) = \begin{cases} +1, |\omega| < 1, \\ -1, |\omega| > 1, \\ \text{undefined}, \omega \in \{-1, 1\}. \end{cases}$$

However, it is can also checked that even though they are symmetric, X and Y are indecomposable, cf. e.g. Linnik & Ostrovskii (1977), which use a related but distinct notion of indecomposability of random variables. The plots of the densities and characteristic functions of X and Y are given in Fig. 12.

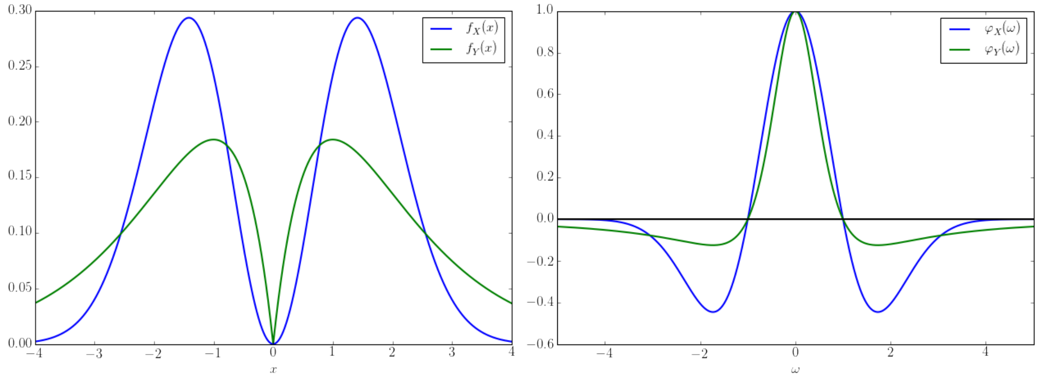

Figure 12. Example of two indecomposable distributions which have the same phase function. **Left**: densities. **Right**: charactersitic functions.

D. Paired Differences

Another way to measure asymmetry of the difference between random vectors X and Y is to use $\text{MMD}(X-Y, Y-X)$ instead of $\text{PhD}(X,Y)$. However, this quantity is not invariant, i.e., $\text{MMD}(X-Y, Y-X) \neq \text{MMD}(X_0 - Y_0, Y_0 - X_0)$, and in fact the values will heavily depend on the distributions of U and V. We note that

$$\varphi_{X-Y}(\omega) - \varphi_{Y-X}(\omega) \quad = \quad 2i\mathbb{E}\sin\left(\omega^\top (X-Y)\right),$$

so that we are effectively measuring the size of the imaginary part of the characteristic function of $X - Y$ (which should not be there if it is symmetric). There are several different ways in which we can write this quantity:

$$\begin{aligned} \text{MMD}(X-Y, Y-X) &= \|\mathbb{E}k(\cdot, X-Y) - \mathbb{E}k(\cdot, Y-X)\|^2_{\mathcal{H}_k} \\ &= \int \left|\varphi_X \overline{\varphi_Y} - \overline{\varphi_X}\varphi_Y\right|^2 d\Lambda \\ &= 4\int \left[\mathbb{E}\sin\left(\omega^\top (X-Y)\right)\right]^2 d\Lambda(\omega) \\ &= \int |\varphi_X|^2 |\varphi_Y|^2 \left(2 - \frac{\varphi_X \overline{\varphi_Y}}{\overline{\varphi_X}\varphi_Y} - \frac{\overline{\varphi_X}\varphi_Y}{\varphi_X \overline{\varphi_Y}}\right) d\Lambda. \end{aligned}$$

The last expression indicates that this quantity is affected by the amplitude of the individual characteristic functions. Moreover, the quantity does not appear to lend itself to the *feature on distributions* formalism, i.e. we were unable to derive some Hilbert space features $\Upsilon(\mathsf{P}) \in \mathcal{H}$ such that $\mathrm{MMD}(X - Y, Y - X) = \|\Upsilon(\mathsf{P}_X) - \Upsilon(\mathsf{P}_Y)\|^2_{\mathcal{H}}$, and it is thus unclear whether this approach can be used to define a valid kernel on distributions.

E. Distribution Regression with Invariance for ABC

We have designed an explicit feature map for a bag of samples that can be used for any distribution regression problem. We now present its potential application to Approximate Bayesian Computation (ABC). Motivated by the approach of Fearnhead & Prangle (2012) and Mitrovic et al. (2016), we propose to use the phase features to construct an optimal summary statistic (under some loss function) for ABC. ABC is a Bayesian framework that allows us to approximate the posterior distribution of some parameter θ by approximating the likelihood function through simulations. To capture this approximation of the likelihood function, simulated datasets from the model are compared with the observed data using some lower dimensional summary statistics. If the summary statistic is sufficient, then there is no loss of information when projecting the data onto lower dimensional space. In practice however, sufficient statistics are not available for complex models of interest and instead using the strategy of Fearnhead & Prangle (2012), one can construct summary statistics that provide inference of θ which is optimal with respect to a given loss function.

In particular, we will focus on the squared loss function as given by $L(\theta, \theta') = (\theta - \theta')^2$. Fearnhead & Prangle (2012) showed that under this loss, the posterior mean of the θ given observations $\mathbf{X}$ is in fact the optimal summary statistic of $\mathbf{X}$ for the ABC procedure. However, since this quantity can not be analytically computed, one approach is to estimate it by fitting a regression model from simulated data, some examples of this include the semi-automatic ABC (Fearnhead & Prangle, 2012) and DR-ABC (Mitrovic et al., 2016). Here we focus on ideas from DR-ABC, which uses a kernel distribution regression approach, treating each simulated dataset (given θ simulated from the prior) as a bag of samples and taking its label to be θ. After training the regression model, it proceeds to using it as a summary statistic as given in algorithm 3. The DR-ABC paper further proposed the conditional DR-ABC (CDR-ABC), which makes the assumption that only certain aspects of the data have an influence on θ. By conditioning on such nuisance variables and then using conditional distribution regression (by embedding conditional distributions (Song et al., 2013)), it can better account for the functional relationship inside the model. However, one problem with this approach is that the nuisance variables have to be observed directly, even for the true dataset, which may often not be the case. For example, consider the hierarchical model we used to illustrate the utility of phase features for regression in section 6.2.

$$\begin{aligned} \theta \sim \Gamma(\alpha, \beta), \quad Z &\sim U[0, \sigma], \quad \epsilon \sim \mathcal{N}(0, Z), \\ X &\sim \frac{\Gamma\big(\theta/2, 1/2\big)}{\sqrt{2\theta}} + \epsilon, \end{aligned} \tag{5}$$

for some fixed values of α, β and σ. Here θ is the parameter we are interested in, ϵ is a latent noise variable (unobserved) and X is the observation. Since neither ϵ nor Z are observed on the true dataset, we can only use DR-ABC, not CDR-ABC. But DR-ABC then does not take into account the model structure which tells us that ϵ is irrelevant for inferring θ, and it is thus likely to give poor performance for large values of σ. Hence, we propose to use phase features inside such regression model, which will be invariant to the noise variable ϵ which is an SPD component in observations. By using phase features for distribution regression, we should be able to better capture the functional relationship between θ and its corresponding dataset, a bag from $X|\theta$ and hence build better summary statistics for ABC. In some sense, this approach can be thought of as implicitly conditioning out the latent nuisance variable ϵ, similarly as CDR-ABC does when it is observed. Furthermore, although we have chosen this example as an illustration, the phase features could be applied to many complex models with nuisance latent variables, even when we cannot write their contribution explicitly as here. The algorithms 2 and 3 shows the approach as in DR-ABC, but now replaced by our phase or Fourier regression approaches to compute summary statistics, and we denote these as Phase-ABC and Fourier-ABC.

F. Implementation Details

F.1. Toy Example

We implement the phase and fourier neural network in TensorFlow. For the network, we use a squared loss function with an additional L_2 weight decay for regularisation. For optimisation, we use ADAM (Kingma & Ba, 2014) with fixed

Algorithm 2 Phase Regression, Fourier Regression

```
Input: prior $\pi$ for $\theta$, data-generating process $P$, phase or fourier features
Output: Phase or Fourier Regression Neural Network
for $i = 1, \ldots, n$ do
    Sample $\theta_i \sim \pi$
    Sample dataset $B_i = \{x_{ij}\}_{j=1}^{N}$ from $P(\cdot|\theta_i)$
end for
Train Phase or Fourier neural network with $\{B_i, y_i\}_{i=1}^{n}$
```

Algorithm 3
Phase-ABC or Fourier-ABC

```
Input: prior $\pi$ for $\theta$, data-generating process $P$, observed data $B^* = \{x_j^*\}_{j=1}^{N^*}$, $\epsilon$, number of particles $K$
Output: Weighted Posterior sample $\sum_k w_k \delta_{\theta_k}$
1. Perform Phase or Fourier Regression, obtain $m(\cdot)$
2. ABC
for $k = 1, \ldots, K$ do
    Sample $\theta_k \sim \pi$
    Sample dataset $B_k = \{x_{kj}\}_j$ from $P(\cdot|\theta_k)$
    Compute $\widetilde{w}_k = \exp\left(-\frac{||m(B_k) - m(B^*)||_2^2}{\epsilon}\right)$
end for
$w_k = \widetilde{w}_k / \sum_k \widetilde{w}_k$
```

learning rate decay and 120 epochs, with a batch size of 10. To tune this network, we perform 3-fold cross validation over, where we initialise the network 3 times, and the average error is computed on the test fold. We tune with the following set of parameters for the neural network (after some preliminary tuning to explore the parameter space):

For Phase Network:

- learning rate: $0.01, 0.05, 0.1, 0.25$
- number of frequencies (i.e. width of the layer): $50, 70, 90, 110, 130$
- λ for regularisation $0, 1.0 \times 10^{-5}, 1.0 \times 10^{-4}, 1.0 \times 10^{-3}, 5.0 \times 10^{-2}, 1.0 \times 10^{-1}, 1.0$

For Fourier Network:

- learning rate: $0.01, 0.05, 0.1, 0.25$
- number of frequencies (i.e. width of the layer): $50, 70, 90, 110, 130$
- λ for regularisation $0, 1.0 \times 10^{-13}, 1.0 \times 10^{-12}, 1.0 \times 10^{-11}, 5.0 \times 10^{-10}, 1.0 \times 10^{-09}, 1.0 \times 10^{-08}$

Furthermore, we initialise the network with the optimally tuned parameters 6 times and test its performance on an independent validation set, before choosing the best performing model. We also keep a history of the mean and variance of the batches (just before the batch normalisation layer) from the last training epochs, and we take the mean of those to be used during testing.

F.2. Aerosol Dataset

For the network, we use a squared loss function with an additional L_2 weight decay for regularisation, with a separate regularisation parameter for the two individual layers. For optimisation, we again use ADAM (Kingma & Ba, 2014) with fixed learning rate decay and 120 epochs, with a batch size of 10. We perform a 3-fold cross validation, and compute the MSE. We tune with the following set of parameters for the neural network (after some preliminary tuning to explore the parameter space), here we initialise the first layer with Gaussian distribution with standard deviation = $1/\gamma_0$, where γ_0 denote the median heuristic for kernel bandwidth. In testing, we initialise the network only once (since we do have a independent validation set).

For Phase and Fourier Network:

- learning rate: $0.2, 0.4, 0.6, 0.8$
- number of frequencies (i.e. width of the layer): $60, 90, 120, 150$
- λ_1 for first layer weights regularisation $0.0, 0.1, 1.0, 5.0, 10.0, 20.0, 50.0, 100.0$
- λ_2 for second layer weights regularisation $0.0, 1\times 10^{-6}, 1\times 10^{-5}, 1\times 10^{-4}, 1\times 10^{-3}, 1\times 10^{-2}, 1\times 10^{-1}, 1.0$

For Gaussian Bag Kernel Ridge Regression:
We use random Fourier features for the Gaussian bag kernel (with parameter γ) with 250 frequencies. We denote γ_0 by the median heuristic for kernel bandwidth and perform a 3-fold cross validation over the following parameters:

- γ: $0.25\,\gamma_0, 0.5\,\gamma_0, 0.75\,\gamma_0, \gamma_0, 1.25\,\gamma_0, 1.5\,\gamma_0, 1.75\,\gamma_0$
- λ regularisation for ridge regression: $0, 0.0001, 0.001, 0.01, 0.1, 1.0, 10.0$

F.3. PhD two sample test

For the PhD two sample test for the toy dataset, for each of the 1000 runs, we use a permutation size of 400, with the number of frequencies sampled set at 50. Here the frequencies are sampled using the radial frequency distribution, where Σ is chosen to be $\sigma^2\mathbf{I}$, with σ^2 being the empirical variance of the two set of samples. The Radial Frequency Distribution is defined as follows:

$$\mathbf{w} = R\Sigma^{-\frac{1}{2}}\boldsymbol{\psi}$$

where $\boldsymbol{\psi} \in \mathbb{R}^n$ is uniformly distributed on the L_2 unit sphere $\mathcal{S}_{n-1}$, and $R \in \mathbb{R}_+$ is a radius drawn independently from a folded Gaussian $\mathcal{N}^+(0,1)$. The radial frequency distribution is useful in high dimensions, as unlike the normal distributions, which 'under samples' the low or middle frequencies, it is able to sample a broader range of frequencies due to its form. By covering a broader range of frequencies, we may be able to 'better encode' information of the distribution represented by the bags, leading to a feature map that is more informative.

G. Additional Results

G.1. Two-Sample Tests with Invariances

In figure above, we see that the PhD statistic controls Type I error for no added Gaussian noise, and also control Type I error for small differences in additive Gaussian components, unlike the ME test. However, we see that the type I error for a larger noise to signal ratio on the two set of samples indeed does alleviate the Type I error. This is not surprising, as the null distribution was constructed by using a permutation test, using:

$$\varphi_{null} = \frac{1}{2}\varphi_{X_0}\varphi_U + \frac{1}{2}\varphi_{X_0}\varphi_V = \varphi_{X_0}(\frac{1}{2}\varphi_U + \frac{1}{2}\varphi_V),$$

and if the estimated phase features are biased, in the regime with large additive Gaussian noise, then the following may not be true approximately: $\hat{\rho}_{null} = \hat{\rho}_{X_0} = \hat{\rho}_{Y_0}$, leading a to a biased null distribution.

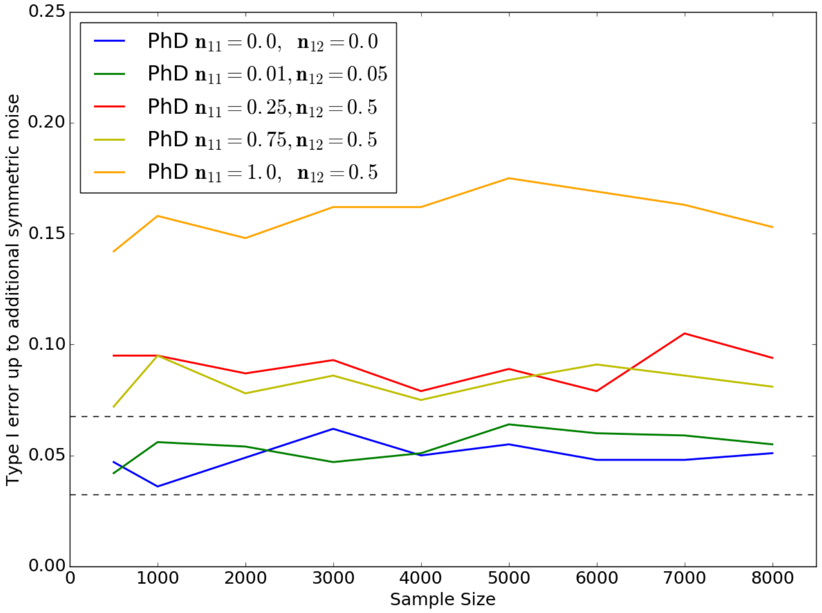

Figure 13. Type I error for the synthetic example with χ^2 for the PhD statistic for various additive Gaussian components, our base distribution without addition of noise is $\chi^2(4)/4$. Here n_{11} refers to the noise to signal ratio for the first set of samples and n_{12} refers to the second set of samples

Table 1. Power for various sample size for high level features of the Higgs dataset

SAMPLE SIZE N	SME POWER	ME POWER
500	0.94	1.0
600	0.969	0.999
700	0.987	1.0
800	0.989	1.0
900	0.994	1.0
1000	0.995	1.0

The table here refers to the high level features of the Higgs dataset, which have been shown to be discriminative in (Baldi et al., 2014). In this case, clearly both the ME and SME achieve good power, note here the SME has slightly less power, due to using only half of the samples to keep independence.

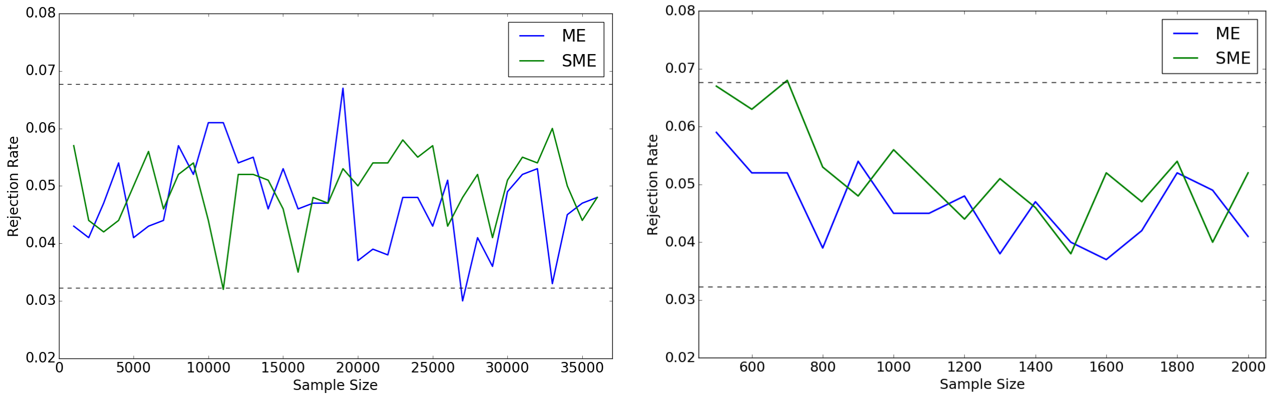

Figure 14. Type I error for the Higgs Dataset. **Left:** Extremely low level features **Right:** High level features. The black dashed line is the 99% Wald interval $\alpha \pm 2.57\sqrt{\alpha(1-\alpha)/1000}$, where here $\alpha = 0.05$ is the significance level and 1000 is the number of repetitions.

The two figures here show that the Type I error is controlled for the ME and SME test, when we have $X_0 \overset{d}{=} Y_0$, where we only consider samples drawn from Y, corresponding to the distribution of the processes where the Higgs Boson are produced. Note that on the right graph, the Type I error at first may be slightly alleviated due to small set of samples.

G.2. Aerosol Dataset

We here provide some additional results for the Aerosol Dataset. We first describe the Paired Difference MMD Bag Kernel (PDMMD). We start by constructing a bag 'kernel' with the paired difference MMD (it is unclear that this kernel is actually positive definite). Assuming that MMD$(X-Y, Y-X)$ is a metric of negative type (Sejdinovic et al., 2013), we write

$$K(X,Y) = \frac{1}{2}(\text{MMD}(X,-X) + \text{MMD}(Y,-Y) - \text{MMD}(X-Y, Y-X))$$

To estimate $K(X,Y)$, we need to again form independent samples in each MMD term. We now use this bag kernel to perform ridge regression with a linear kernel at the second level and we call this method PDMMD ridge regression, PDMMDRR in short. With the same setup as in section 6.2, we also investigate the case where there is no covariate shift, i.e. the training data and the test data have the same σ level of Gaussian noise added. To reduce variance in the results, we use 10 different train and test splits of the data, where each time cross validation is done on the train set to choose the optimal parameters, in order to keep independence from the test set. For the PDMMDRR, we use a Gaussian kernel for the MMD, and again use random Fourier feature approximation to approximate each MMD, using the same setup as in the GKRR, with cross validation over Gaussian kernel bandwidth and also regularisation parameter for ridge regression. We show the results below for the original dataset, and also for the case where Gaussian noise with noise-to-signal ratio of $\sigma = 10.0$ is added to the dataset.

Table 2. Average RMSE for the Aerosol Dataset and the Noisy ($\sigma = 10.0$) Aerosol Dataset across 10 runs, for different train and test splits.

	Fourier NN	Phase NN	GKRR	PDMMDRR
No noise	0.100	0.100	0.0799	0.188
Noise added	0.124	0.125	0.127	0.168

We can see that the GKRR outperforms all other method on the original dataset, and reaches the result as in Szabó et al. (2015). It is clear the PDMMDRR performs poorly compared to the rest of the methods and in fact the errors have very high variance, suggesting possible instability in this method. We can see that although phase NN and Fourier NN have a lower performance on the original dataset, with additive Gaussian noise it is indeed competitive with the GKRR approach. This again highlights that the phase neural network is stable under additive noise. Here, we again see that the Fourier NN performs similarly well with the phase NN – we explain this below.

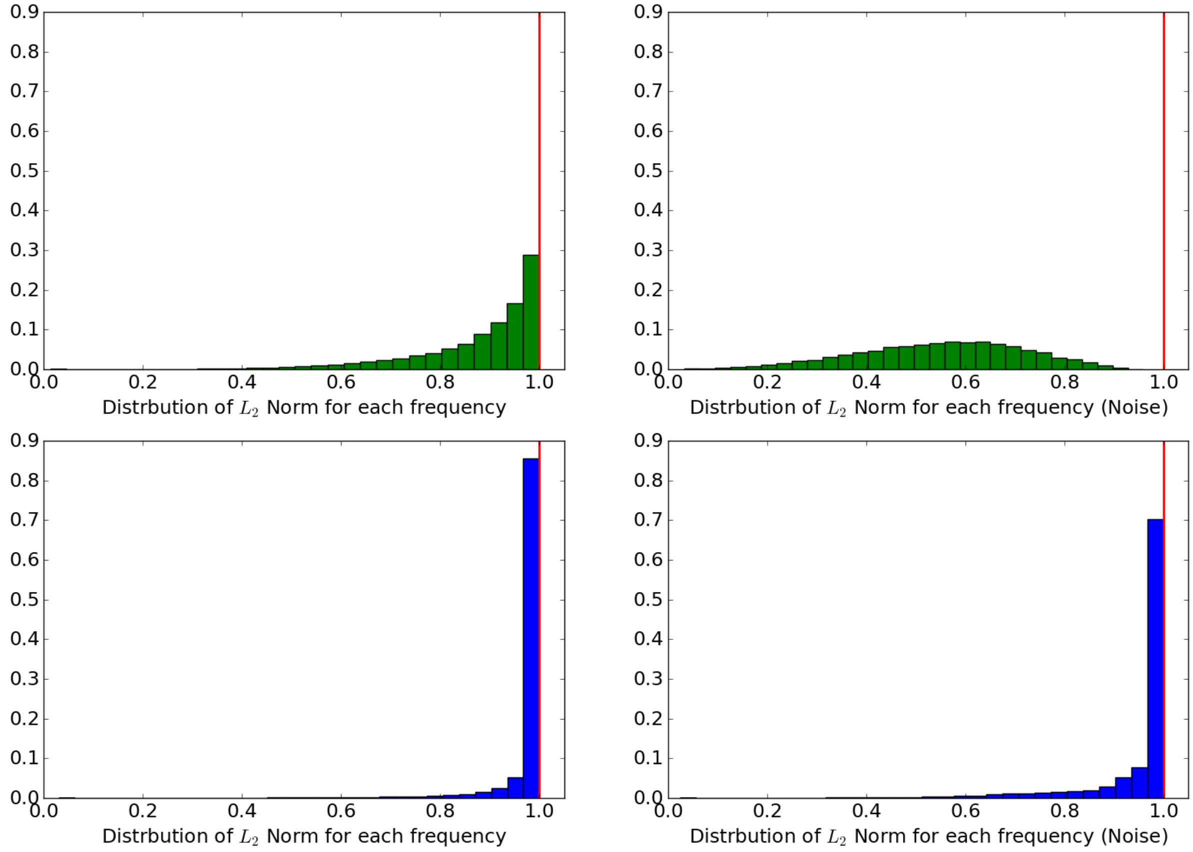

Figure 15. Histograms for the distribution of the L_2 norm of the averages of fourier features over each frequency w for the original aerosol test set and the aerosol test set with added noise ($\sigma = 3$), here red line denotes the unit norm representing the phase features
Top Green: Random Fourier Features w (with the optimised kernel bandwidth)
Bottom Blue: Learnt Fourier features w from the Fourier Neural Network

In the experiments for the Aerosol covariate shift and above, we have seen that the Fourier NN performs similarly to the Phase NN, even under the addition of Gaussian noise, here we provide some possible insights. From the trained Fourier NN on the original dataset, we extract the frequencies w learnt and compute $\left\|\hat{\mathbb{E}}\xi_\omega(X)\right\|$ for each frequency over the original and noisy test set, similarly we do this for the frequencies generated from the Gaussian kernel (with the optimised bandwidth on the aerosol dataset). We show the empirical distribution of both of these in the figure above, we see that the discriminative frequencies learnt on the training data correspond to the Fourier features which are nearly normalised (i.e. they are close to unit norm like phase features, shown by the red line), this may suggest that the learnt frequencies have captured a notion of invariance to additive SPD components on just the training data. This provides insight into good performance of Fourier NN even under the covariate shift. It also indicates that the original Aerosol data potentially has irrelevant SPD noise components that the Fourier NN has learnt to ignore.