Testing and Learning on Distributions with Symmetric Noise Invariance

Ho Chung Leon Law
Department of Statistics
University Of Oxford
ho.law@spc.ox.ac.uk

Christopher Yau
Centre for Computational Biology
University of Birmingham
c.yau@bham.ac.uk

Dino Sejdinovic
Department of Statistics
University Of Oxford
dino.sejdinovic@stats.ox.ac.uk

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.

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 [2, 5, 7]. 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), 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* [14, 23, 24], 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) [5], can be sensitive to irrelevant sources of variation and may lead to suboptimal or even misleading results, in which case 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, exploit the irregular and complicated nature of real life distributions and *encode invariances* to sources

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

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 also require that it 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.

The paper is outlined as follows. In section 2, we provide an overview of the background. In section 3, we provide details of the construction and implementation of phase features. In section 4, we discuss the approach based on asymmetry in paired differences for two sample testing with invariances. Section 5 provides experiments on synthetic and real data, before concluding in section 6.

2 Background and Setup

We will say that a random vector E on $\mathbb{R}^d$ is a *symmetric positive definite (SPD) 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 [20]. 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 components.

Following the terminology similar to that of [3], 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 mostly the indecomposable components of distributions are of interest and 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. [15] 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 [26, 6.2], the squared MMD can be written as a weighted L_2-distance between characteristic functions [22, Corollary 4]

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

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 (RFF) [19] 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] $\tilde{k}(x, y) = \frac{1}{m}\sum_{j=1}^m \left[\cos(\omega_j^T x)\cos(\omega_j^T y) + \sin(\omega_j^T x)\sin(\omega_j^T y)\right] = \langle\phi(x), \phi(y)\rangle_{\mathbb{R}^{2m}}$ where

[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.

$\omega_1, \ldots, \omega_m \sim \Lambda$, giving an explicit 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 [14] or regression [24]. Approximate kernel embeddings have also been applied in this context [23].

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 [19], the relevant notion in our context is the *phase function* of a probability measure, recently used for nonparametric deconvolution by [3]. 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 [3], the goal is to estimate the density function f_0 of a univariate r.v. 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 is an SPD noise component, and that X_0 is indecomposable, i.e. X_0 itself does not contain any SPD noise components, [3] 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 $\rho_X(\omega) = \frac{\varphi_X(\omega)}{|\varphi_X(\omega)|}$, where $\varphi_X(\omega)$ denotes the characteristic function of X. 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 and similarly to the motivation of [3], we argue that real-world distributions of interest often exhibit certain amount of irregularity and it is exactly this irregularity which is exploited in our methodology. In analogy to the MMD, we first define the phase discrepancy (PhD) as a weighted L_2-distances between the phase functions:

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

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 $\text{PhD}(X, Y) = \text{PhD}(X_0, Y_0)$. It is clear then that the PhD is not affected by additive SPD noise components, so it captures desired invariance. However PhD for Λ supported everywhere is in fact not a proper metric on the indecomposable probability measures $\mathcal{I}(\mathbb{R}^d)$, as one can find indecomposable random variables X and Y s.t. $\rho_X = \rho_Y$ and thus $\text{PhD}(X, Y) = 0$. An example is given in Appendix A.

While such cases appear contrived, we hence 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 Appendix B).

Proposition 1.

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

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.*

Proposition 2.

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

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:

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

where we have replaced the expectations by their empirical estimates. Because $\left\|\Psi(\hat{P}_X)\right\| = 1$, we can construct $\widehat{\text{PhD}}(\hat{P}_X, \hat{P}_Y) = \left\|\Psi(\hat{P}_X) - \Psi(\hat{P}_Y)\right\|^2 = 2 - 2\Psi(\hat{P}_X)^\top\Psi(\hat{P}_Y)$, 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 of the empirical distribution which encodes invariance to additive SPD noise components present in P [2], as demonstrated in Figure F.1 in the Appendix. It can now be directly applied to (1) two-sample testing up to SPD components, where the distance between the phase features, i.e. a Monte Carlo estimate of the PhD in equation 2, can be used as a test statistic, with details given in section 5.1 and (2) learning on distributions, where we use phase features as the explicit feature map for a bag of samples.

Although we have assumed an indecomposable underlying distribution so far, this assumption is not strict. For distribution regression, if the indecomposable 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 SPD components irrelevant for learning outweighs the signal in the SPD part of the distribution, i.e. there is a trade off between SPD noise and SPD signal. In practice, the phase features we propose can be used to encode such invariance where appropriate or in conjunction with other features which do not encode invariance, thus letting the data decide which features are discriminative for the problem at hand.

In order to construct the approximate mean embeddings for learning, we first compute an explicit feature map by taking averages of the Fourier features, as given by $\Phi(\hat{P}_X) = \sqrt{\frac{1}{m}}\left[\hat{\mathbb{E}}\xi_{\omega_1}(X), \ldots, \hat{\mathbb{E}}\xi_{\omega_m}(X)\right]$. For phase features, we need to compute an additional normalisation term over each frequency as in (3). To obtain the set of frequencies $\{w_i\}_{i=1}^m$, we can draw samples from a probability measure Λ corresponding to an inverse Fourier transform of a shift-invariant kernel, e.g. Gaussian Kernel. However, given a supervised signal, we can also 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 no longer focus on a specific shift-invariant kernel k, but are *learning discriminative Fourier/phase features*. To do this, we can construct a neural network (NN) with special activation functions, pooling layers as shown in Algorithm D.1 and Figure D.1 in the Appendix.

4 Asymmetry in Paired Differences

We now consider a separate approach to nonparametric two-sample test, where we wish to test the null hypothesis that $H_0 : P \overset{d}{=} Q$ vs. the general alternative, but we only have iid samples arising from

[2] Note that, unlike the population expression $\Psi(P)$, the empirical estimator $\Psi(\hat{P})$ will in general have a distribution affected by the noise components and is thus only approximately invariant, but we observe that it captures invariance very well as long as the signal-to-noise regime remains relatively high (Section 5.1).

$X \sim P \star \mathcal{E}_1$ and $Y \sim Q \star \mathcal{E}_2$. i.e.

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

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. With this setting (proof in Appendix B):

Proposition 3. *Under the null hypothesis H_0, $X - Y$ 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 $X_0 \overset{d}{=} Y_0$, as it tests for symmetry. 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 [7], which was found to have performance similar to or better than the original MMD two-sample test [5], 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). Note although we have assumed X_0, Y_0 lie in the space of $\mathcal{P}(\mathbb{R}^d)$ of indecomposable distributions, in practice the test would only not reject if the underlying distribution *only differs* in the symmetric components for the SME test (or SPD components for the PhD test). We argue this is unlikely due to real life distributions being complex in nature. In practice, we recommend the use of the ME and SME or PhD test together to provide an exploration tool to understand the underlying difference, as demonstrated in the Higgs Data experiment in section 5.1.

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 $\text{MMD}(X - Y, Y - X) \neq \text{MMD}(X_0 - Y_0, Y_0 - X_0)$. This means that the paired differences approach to learning is sensitive to the actual type and scale of the additive SPD noise components, hence not suitable for learning. The mathematical details and empirical experiments to show this are presented in Appendix C and F.1.

5 Experimental Results

5.1 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 [7] 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 estimate $\widehat{\text{PhD}}(\hat{P}_X, \hat{P}_Y)$ of (2). It is unclear what the exact form of the null distribution is, so we use a permutation test, by recomputing this statistic on the samples which are first merged and then randomly split in the original proportions. While we are combining samples with different distributions, the permutation test is still justified 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 SPD 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 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

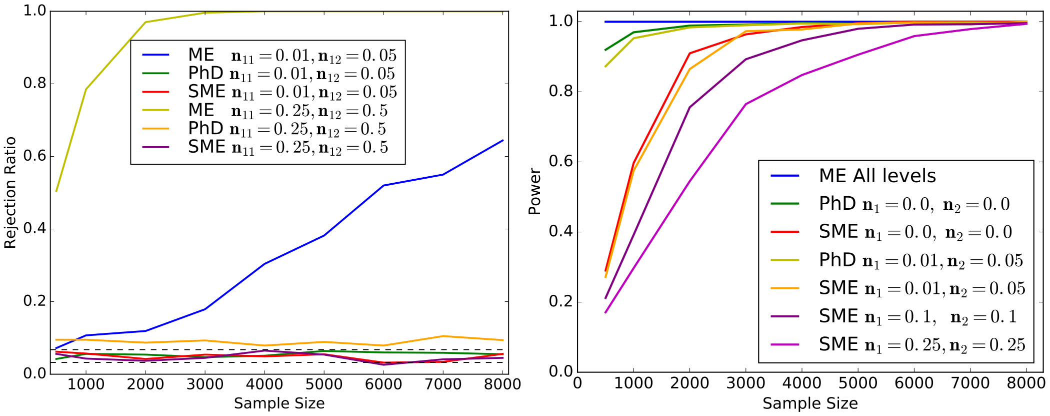

Figure 1: Type I error and Power under various additional symmetric noise in the synthetic χ^2 dataset. Dashed line is the 99% Wald interval here. **Left:** Type I error, n_{11} denotes the noise to signal ratio for the first set of samples and n_{12} for the second set. **Right:** Power, n_1 denotes 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.

empirical phase and characteristic function with noise for these two distributions can be found in Appendix F.2. 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 1 (left), where $X_0 \overset{d}{=} Y_0$, both constructed using $\chi^2(4)/4$, with the noiseless case found in Figure F.6 in the Appendix. 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 PhD test, on the other hand, while correctly controlling Type I at small noise levels, was found to have inflated Type I error rates for large noise, with more results provided in Figure F.6 in the Appendix. This is due to the sensitivity to noise in the permutation test. Namely, the test uses invariance to SPD of the population expression of PhD, but the estimator of the null distribution of the test statistic will in general be affected by the differing noise levels.

Next, we investigate the power, shown in Figure 1 (right). For a fair comparison, we have included the PhD test power only for small noise levels, in which the Type I error is controlled at the design level. In these cases, 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 the PhD test has an inflated Type I error for large noises, which means that its results should be considered with caution in practice. ME test rejects at all levels at all sample sizes as it picks up all possible differences. SME and PhD are by construction more conservative tests whose rejection provides a much stronger statement: two samples differ even when *all arbitrary additive SPD components* have been stripped off.

Higgs Dataset The UCI Higgs dataset [1, 11] 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 $\mathbb{R}^7$) have been shown to have good discriminative properties in [1]. 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 F.1 in the Appendix, which shows that the high level features differ 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 $\in \mathbb{R}^4$,

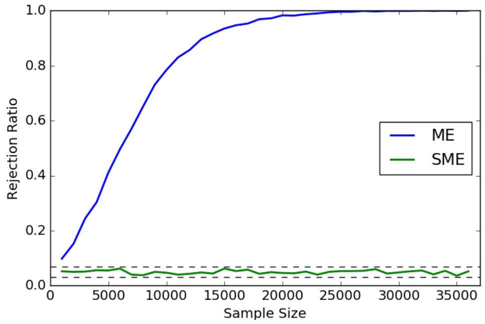

Figure 2: Rejection ratio vs. sample size for extremely low level features for Higgs dataset. Dashed line is the 99% Wald interval for 1000 repetitions for $\alpha = 0.05$. Note PhD is not used here, due to its expensive computational cost.

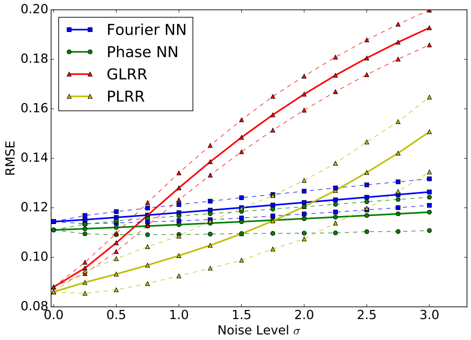

Figure 3: RMSE on the test set, corrupted by various levels of noise averaged over 100 runs, with the 5^{th} and the 95^{th} percentile. The noiseless case is shown with one run. RMSE from mean is 0.206.

commented in [1] to carry very little discriminating information, following the same experimental setup from [2].

The results for the ME and SME test can be found in Figure 2. 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 potentially be explained by some additive symmetric noise components which may be 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 given sufficient sample sizes, even if the discovered difference may not be of interest. 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 both low and high level features can be found in Figure F.7 in the Appendix.

5.2 Learning with Phase Features

Aerosol Dataset To demonstrate the phase features invariance to SPD noise component, we use the Aerosol MISR1 dataset also studied by [24] and [25] and consider a situation with *covariate shift* [18] on distribution inputs: the testing data is impaired by additive SPD components different to that in the training data. Here, we have an aerosol optical depth (AOD) multi-instance learning problem with 800 bags, where each bag contains 100 randomly selected multispectral (potentially cloudy) 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.

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, accounting for different scales across dimensions. For comparisons, we consider linear ridge regression on embeddings with respect to a Gaussian kernel, approximated with RFF (GLRR) as described in section 2.2 (i.e. a linear kernel is applied on approximate embeddings), linear ridge regression on phase features (PLRR) (i.e. normalisation step is applied to obtain (3)), and also the phase and Fourier neural networks (NN), described in Appendix D, tuning all hyperparameters with 3-fold cross validation. 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 sets and results are shown in figure 3. It is also noted that a second level non-linear kernel K does not improve performance on this problem [24].

We see that GLRR and PLRR are competitive (see Appendix Table F.2 for average across different splits) in the noiseless case, and these clearly outperform both the Fourier NN and Phase NN (likely

Table 1: Mean Square Error (MSE) on dark matter dataset for 500 runs with 5^{th} and 95^{th} percentile.

Algorithm	MSE
Mean	0.16
PLRR	**0.021** (0.018, 0.024)
GLRR	0.033 (0.030, 0.037)
LGRR	0.032 (0.028, 0.036)
PGRR	0.021 (0.017, 0.024)
GGRR	**0.018** (0.015, 0.019)

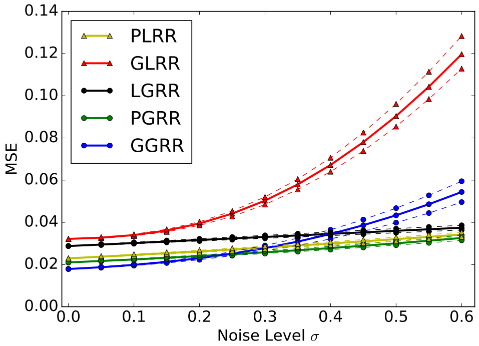

Figure 4: MSE with various levels of noise added on test set, with 5^{th} and 95^{th} percentile.

due to the small size of the dataset). For increasing noise, the performance of GLRR degrades significantly, and while the performance of PLRR degrades also, the model is much more robust under additional SPD noise. In comparison, the Phase NN implementation is almost insensitive to covariate shift in the test sets, unlike the performance of PLRR, highlighting the importance of learning discriminative frequencies w in a very low signal-to-noise setting. It is 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 - as shown in Figure F.9 in the Appendix). This means that even the Fourier NN has *learned to be approximately invariant* based on training data, indicating that the original Aerosol data potentially has irrelevant SPD noise components. This is reinforced by the nature of the dataset (each bag contains 100 randomly selected potentially cloudy pixels, known to be noisy [25]) and no loss of performance from going from GLRR to PLRR. The results highlights that phase features are stable under additive SPD noise, even under such a difficult setting.

Dark Matter Dataset We now study the use of phase features on the dark matter dataset, composing of a catalog of galaxy clusters. In this setting, we would like to predict the total mass of galaxy clusters, using the dispersion of velocities in the direction along our line of sight. In particular, we will use the 'ML1' dataset, as obtained from the authors of [16, 17], who constructed a catalog of massive halos from the MultiDark `mdpl` simulation [9]. The dataset contains 5028 bags, with each sample consisting of its sub-object velocity and its mass label. By viewing each galaxy cluster at multiple lines of sights, we obtain 15 000 bags, using the same experimental setup as in [10]. For experiments, we use approximately 9000 bags for training, and 3000 bags each for validation and testing, keeping those of multiple lines of sight in the same set. As before, we use GLRR and PLRR and we also include in comparisons methods with a second level Gaussian kernel (with RFF) applied to phase features (PGRR) and to approximate embeddings (GGRR). For a baseline, we also include a first level linear kernel (equivalent to representing each bag with its mean), before applying a second level gaussian kernel (LGRR). We use the same set of randomly sampled frequencies across the methods, tuning for the scale of the frequencies and for regularisation parameters.

Table 1 shows the results of the methods across 10 different data splits, with 50 sets of randomised frequencies for each data split. We see that PLRR is significantly better than GLRR. This suggests that under this model structure, by removing SPD components from each bag, we can target the underlying signal and obtain superior performance, highlighting the applicability of phase features. Considering a second level gaussian kernel, we see that the GGRR has a slight advantage over PGRR, with PGRR performing similar to PLRR. This suggests that the SPD components of the distribution of sub-object velocity may be useful for predicting the mass of a galaxy cluster if an additional nonlinearity is applied to embeddings – whereas the benefits of removing them outweigh the signal present in them without this additional nonlinearity. To show that indeed the phase features are robust to SPD components, we perform the same covariate shift experiment as in the aerosol dataset, with the results given in Figure 4. Note that LGRR is also robust to noise, as each bag is simply represented by its mean.

6 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 target underlying signal distributions.

References

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

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

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

[4] Paul Fearnhead and Dennis Prangle. 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.

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

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

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

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

[9] Anatoly Klypin, Gustavo Yepes, Stefan Gottlober, Francisco Prada, and Steffen Hess. MultiDark simulations: the story of dark matter halo concentrations and density profiles. 2014. arXiv:1411.4001.

[10] H. C. L. Law, Dougal J. Sutherland, Dino Sejdinovic, and Seth Flaxman. Bayesian Distribution Regression. 2017. arXiv:1705.04293.

[11] M. Lichman. UCI machine learning repository, 2013.

[12] Yu V Linnik and IV Ostrovskii. *Decomposition of random variables and vectors*. 1977.

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

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

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

[16] Michelle Ntampaka, Hy Trac, Dougal J. Sutherland, Nicholas Battaglia, Barnabás Póczos, and Jeff Schneider. A machine learning approach for dynamical mass measurements of galaxy clusters. *The Astrophysical Journal*, 803(2):50, 2015. arXiv:1410.0686.

[17] Michelle Ntampaka, Hy Trac, Dougal J. Sutherland, S. Fromenteau, B. Poczos, and Jeff Schneider. Dynamical mass measurements of contaminated galaxy clusters using machine learning. *The Astrophysical Journal*, 831(2):135, 2016. arXiv:1509.05409.

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

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

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

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

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

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

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

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

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

A 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. [12], 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. A.1.

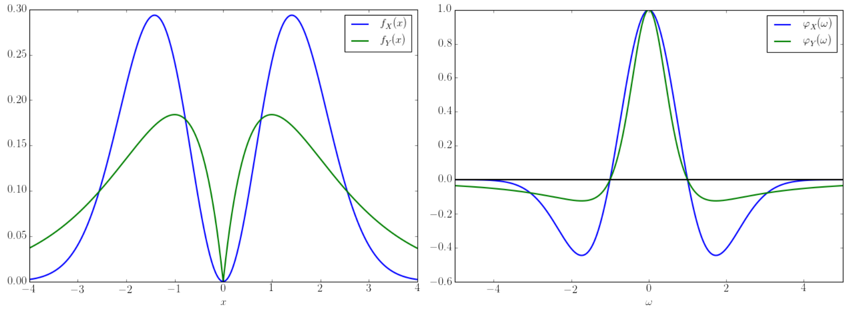

Figure A.1: Example of two indecomposable distributions which have the same phase function. **Left**: densities. **Right**: charactersitic functions.

B Phase Discrepancy and Asymmetry in Paired Differences Proofs

In this section, we will provide further details of the definitions, calculations and proofs in section 3 and 4. Phase discrepancy is defined as the weighted L_2-distances between the phase functions, i.e.

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

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 4.

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

Proof.

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

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,

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

Substituting provides us the result. □

Proposition 5. $K_\omega\left(\mathsf{P}_X, \mathsf{P}_Y\right) = \left(\frac{\mathbb{E}\xi_\omega(X)}{\|\mathbb{E}\xi_\omega(X)\|}\right)^\top\left(\frac{\mathbb{E}\xi_\omega(Y)}{\|\mathbb{E}\xi_\omega(Y)\|}\right)$ *is a positive definite kernel on probability measures* $\forall\omega$, *where here* $\xi_\omega(x) = \left[\cos\left(\omega^\top x\right), \sin\left(\omega^\top x\right)\right]$, *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 $\xi_\omega(x) = \left[\cos\left(\omega^\top x\right), \sin\left(\omega^\top x\right)\right]$, 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 6. $\text{PhD}(X,Y) = 2 - 2K\left(\mathsf{P}_X, \mathsf{P}_Y\right) = 2\int\left(1 - \left(\frac{\mathbb{E}\xi_\omega(X)}{\|\mathbb{E}\xi_\omega(X)\|}\right)^\top\left(\frac{\mathbb{E}\xi_\omega(Y)}{\|\mathbb{E}\xi_\omega(Y)\|}\right)\right)d\Lambda(\omega)$.

Proposition 7. *Under the null hypothesis,* $X-Y$ *is SPD* $\iff X_0 \overset{d}{=} Y_0$.

Proof. Under H_0, 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 $X-Y$ is SPD $\iff X_0 \overset{d}{=} Y_0$. □

C Paired Differences

Another way to measure asymmetry of the difference between random vectors X and Y is to use MMD$(X-Y, Y-X)$ instead of PhD(X,Y). However, this quantity is not invariant, i.e., MMD$(X-Y, Y-X) \neq$ 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

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

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:

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

The last expression indicates that this quantity is affected by the amplitude of the individual characteristic functions, experimental details to show this empirically can be found in F.1. 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.

D Learning Discriminative Features

Algorithm D.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:
      $l_3(X) = \left[\frac{\hat{\mathbb{E}}\xi_{\omega_1}(X)}{\|\hat{\mathbb{E}}\xi_{\omega_1}(X)\|},\dots,\frac{\hat{\mathbb{E}}\xi_{\omega_m}(X)}{\|\hat{\mathbb{E}}\xi_{\omega_m}(X)\|}\right]$
5. Batch Normalisation Layer
6. Output layer
```

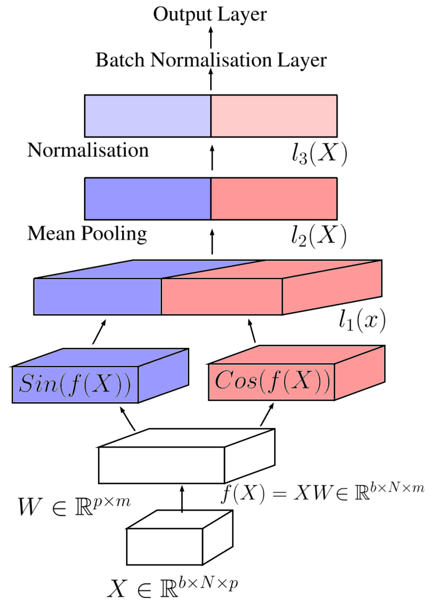
Output Layer
Batch Normalisation Layer
Normalisation
$l_3(X)$
Mean Pooling
$l_2(X)$
$l_1(x)$
$Sin(f(X))$
$Cos(f(X))$
$W \in \mathbb{R}^{p\times m}$
$f(X) = XW \in \mathbb{R}^{b\times N\times m}$
$X \in \mathbb{R}^{b\times N\times p}$

Figure D.1: Main structure of the phase neural network

Algorithm D.1 shows the phase Neural Network (phase NN) and the Fourier Neural Network (Fourier NN), where the latter can be obtained by simply removing step 4 in the algorithm. Although the batch normalisation is not required, it is highly recommended for faster training of the network [6], 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.

E Distribution Regression with Invariance for ABC

Algorithm E.1 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** $i = 1, \dots, n$ **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 E.2 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** $k = 1, \dots, K$ **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**
$w_k = \widetilde{w}_k / \sum_k \widetilde{w}_k$
```

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 [4] and [13], 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 [4], 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 $L(\theta, \theta') = (\theta - \theta')^2$. [4] 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 [4] and DR-ABC [13]. 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 E.2. 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 [21]), 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 below.

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

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 E.1 and E.2 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. Some experimental results can be found in F.4.1.

F Additional Results

F.1 Asymmetry in Paired Differences Experiment

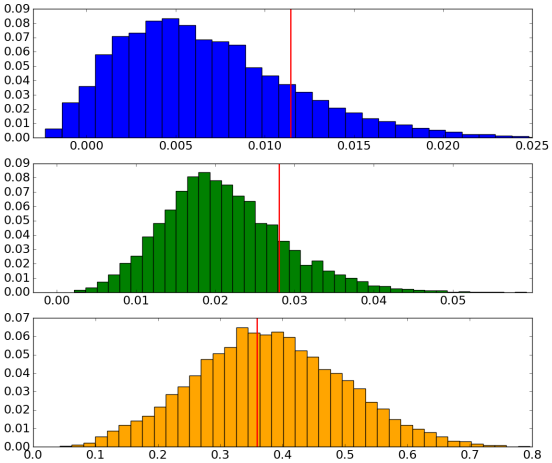

Figure F.1: Histograms on various estimates for all pairs of bags with varying additive noise, red line denotes the noiseless case. **Top:** Estimated MMD on paired differences for all pair of bags, the red line given by the mean of the estimated MMD on paired differences for bags without noise. **Middle:** the squared distance between Fourier features (an estimate of MMD). **Bottom:** the squared distance between phase features (an estimate of PhD).

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 $Z_i, W_i \sim U[0, 0.1]$. 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 $\{w_i\}_{i=1}^{100}$ (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 F.1. 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, 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 there is clearly some variance, however this is expected as its PhD population expression is invariant, but not its estimator, furthermore the frequencies are sampled (with the median heuristic bandwidth) and not learnt. 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, as shown.

F.2 Characteristic and Phase Function Plots

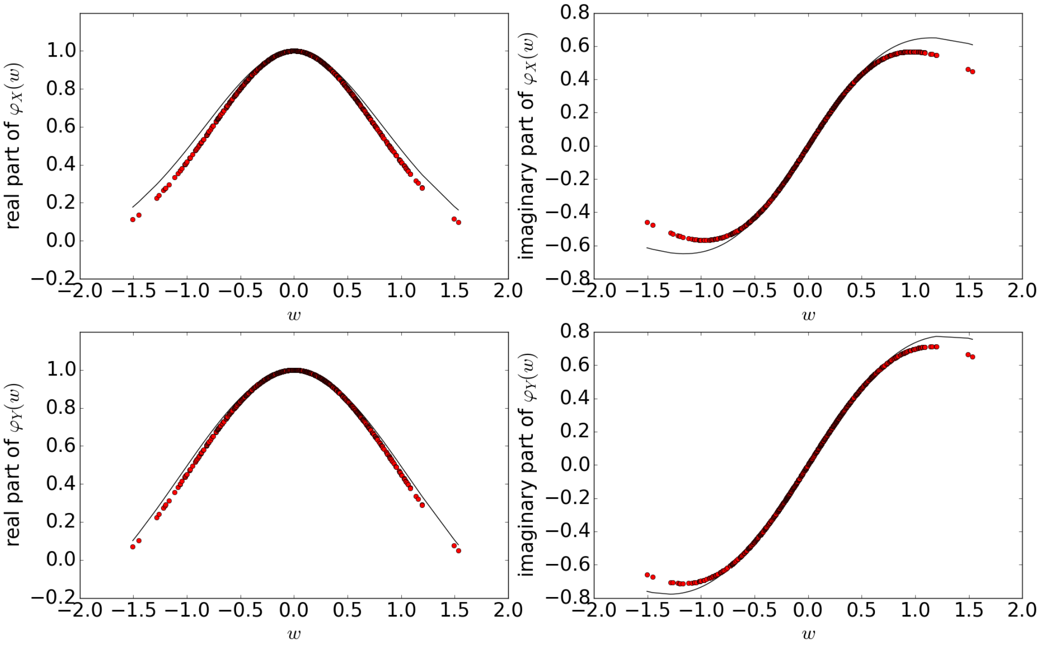

Figure F.2: The black line here correspond to the real and imaginary part of the true characteristic function of the $\chi^2(4)/4$ and $\chi^2(8)/8$ distribution, denoted X, Y on the top and bottom graphs respectively. The red points denote the empirical characteristic function constructed with 750 frequencies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

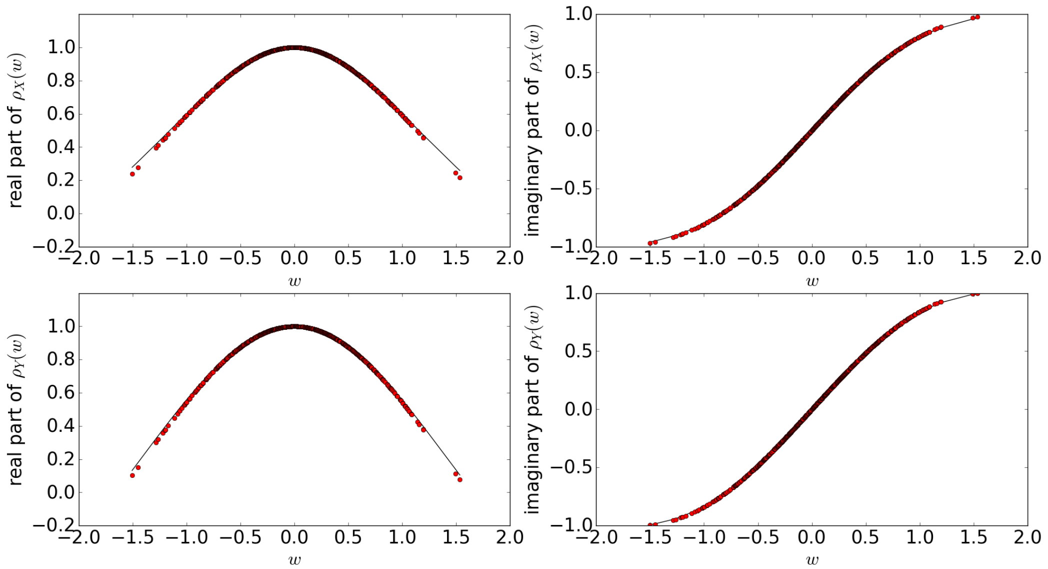

Figure F.3: The black line here correspond to the real and imaginary part of the true phase function of the $\chi^2(4)/4$ and $\chi^2(8)/8$ distribution, denoted X, Y on the top and bottom graphs respectively. The red points denote the empirical phase function constructed with 750 frequencies from a Gaussian kernel with $\sigma = 2$ using a bag size of 1000 observations, with some additional Gaussian noise.

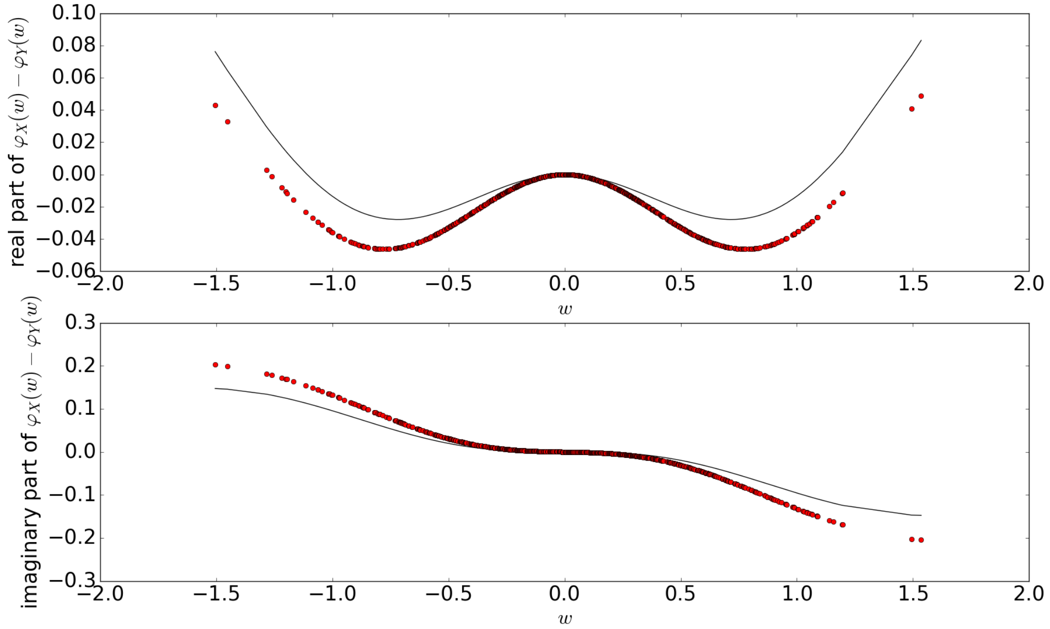

Figure F.4: The top and bottom graph denotes the difference in the real and imaginary part of the characteristic function for the $\chi^2(4)/4$ and $\chi^2(8)/8$ as in figure F.2.

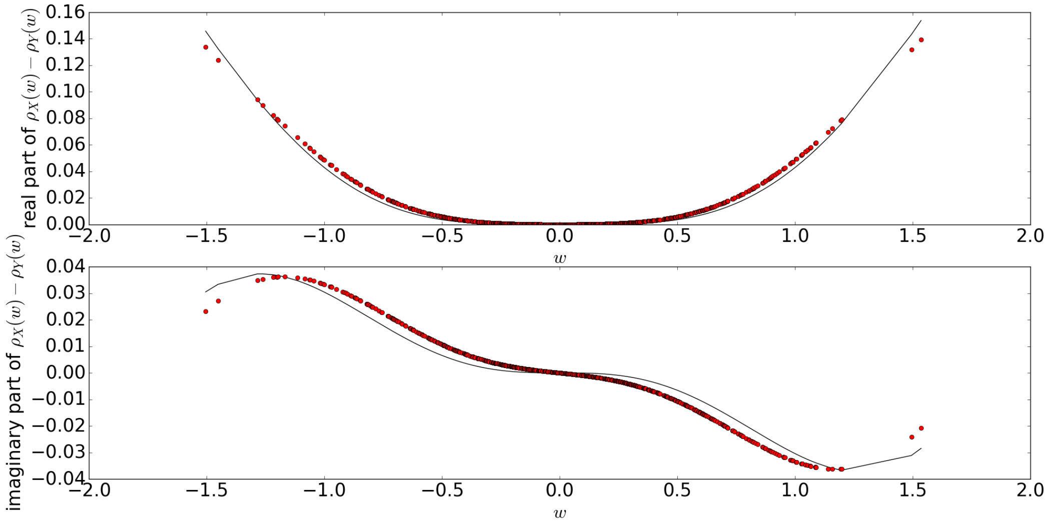

Figure F.5: The top and bottom graph denotes the difference in the real and imaginary part of the phase function for the $\chi^2(4)/4$ and $\chi^2(8)/8$ as in figure F.3.

F.3 Two-Sample Tests with Invariances

F.3.1 Synthetic χ^2 Dataset

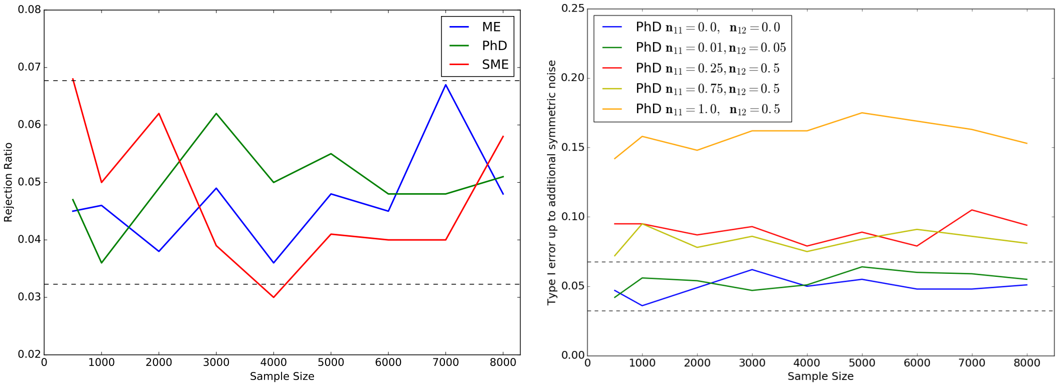

Figure F.6: Extra Type I error results for the synthetic example with χ^2 **Left:** With no noise added for the ME, PhD and SME test. **Right:** Various additive Gaussian components, our base distribution without addition of noise is $\chi^2(4)/4$. Here n_{11} refers to the noise to signal ratio for the first set of samples and n_{12} refers to the second set of samples.

In figure above, 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.

On the left figure, we see that indeed all three test considered in this paper indeed controls the Type I error, when the underlying distribution between the two sets of sample is the same, note here no additional noise is added.

On the right figure, 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.

F.3.2 Higgs Dataset

Table F.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 [1]. 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.

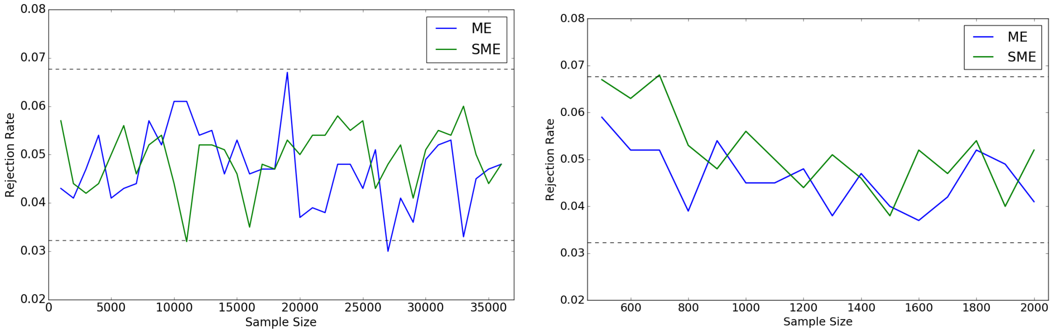

Figure F.7: Type I error for the Higgs Dataset. **Left:** Extremely low level features **Right:** High level features. The black dashed line is the 99% Wald interval $\alpha \pm 2.57\sqrt{\alpha(1-\alpha)/1000}$, where here $\alpha = 0.05$ is the significance level and 1000 is the number of repetitions.

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.

F.4 Learning with Phase Features

F.4.1 Toy Dataset

We demonstrate the use of phase features in the synthetic dataset generated by the following model (note this is the ABC hierarchical model discussed above):

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

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 $X|\theta$, sampled iid across dimensions. 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

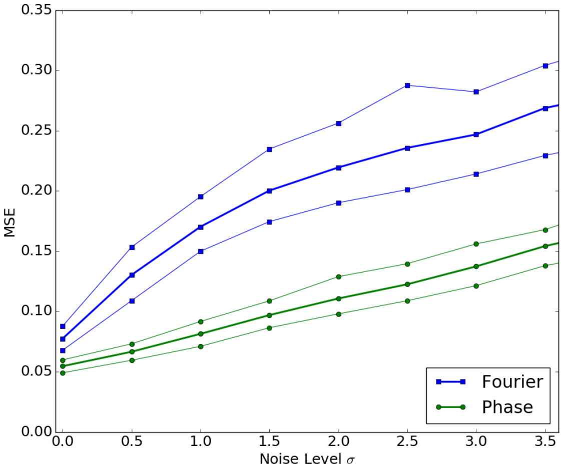

Figure F.8: MSE of θ, using the Fourier and phase neural network averaged over 100 runs. Here noise σ is varied between 0 and 3.5, and the 5^{th} and the 95^{th} percentile is shown.

(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 F.8. Figure F.8 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 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.

F.4.2 Aerosol Dataset

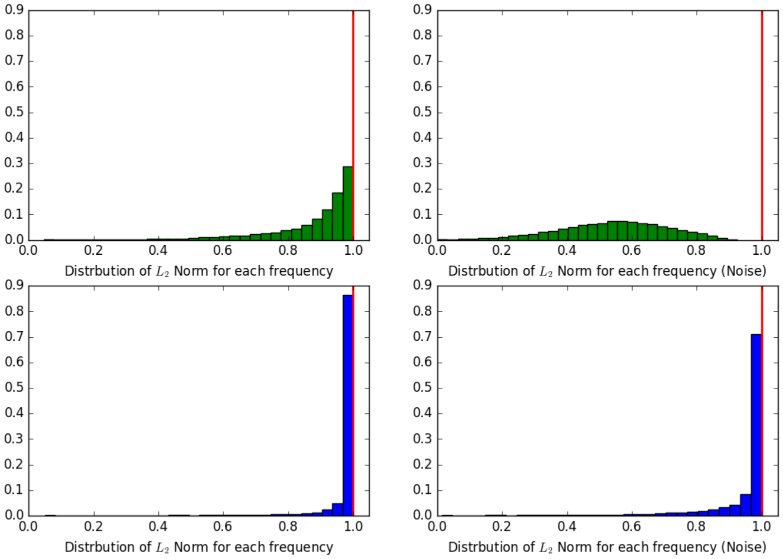

Figure F.9: Histograms for the distribution of the L_2 norm of the averages of Fourier features over each frequency w for the original aerosol test set and the aerosol test set with added noise ($\sigma = 3$), here red line denotes the unit norm representing the phase features **Top Green:** Random Fourier Features w (with the optimised kernel bandwidth) **Bottom Blue:** Learnt Fourier features w from the Fourier Neural Network.

We here provide some additional results for the Aerosol Dataset. First, we provide the average RMSE on the aerosol dataset (without noise on test set), based on 10 runs, for different train and test splits in Table F.2.

Table F.2: Average RMSE for the Aerosol Dataset across 10 runs, for different train and test splits, with standard deviation in brackets

	Fourier NN	Phase NN	GLRR	PLRR
No noise	0.101 (0.011)	0.101 (0.008)	0.079 (0.010)	0.085 (0.009)

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 original 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.

G Implementation Details

G.1 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.2 Toy Example in Appendix

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 [8] with fixed 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 the learning rate, the number of frequencies and the regularisation parameter λ for the neural network. 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.

G.3 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 [8] 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 the learning rate, regularisation parameters and also number of frequencies for the neural network, 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.

G.4 Dark Matter Dataset

For all methods we sample frequencies from the normal distribution (with standard deviation = $1/\gamma_0$, where γ_0 denote the median heuristic for kernel bandwidth.). After sampling a set of frequencies, we tune the scale of the set of frequencies and also the ridge regularisation parameter using the validation set. In particular we use 75 frequencies on the first and second level of the kernel whenever they are used. Note we use the same set of frequencies (at each individual kernel level) across all the methods in a single run to allow for easier comparison, with potentially different scale tuned on the validation set.