Distinguishing Distributions with Interpretable Features

Wittawat Jitkrittum WITTAWAT@GATSBY.UCL.AC.UK
Zoltán Szabó Z.SZABO@UCL.AC.UK
Kacper Chwialkowski KACPER.CHWIALKOWSKI@GMAIL.COM
Arthur Gretton ARTHUR.GRETTON@GMAIL.COM
Gatsby Computational Neuroscience Unit, University College London.

Abstract

Two semimetrics on probability distributions are proposed, based on a difference between features chosen from each, where these features can be in either the spatial or Fourier domains. The features are chosen so as to maximize the distinguishability of the distributions, by optimizing a lower bound of power for a statistical test using these features. The result is a parsimonious and interpretable indication of how and where two distributions differ, which can be used even in high dimensions, and when the difference is localized in the Fourier domain. A real-world benchmark image data demonstrates that the returned features provide a meaningful and informative indication as to how the distributions differ.

1. Introduction

We address the problem of discovering features of distinct probability distributions P and Q, such that they can most easily be distinguished. The distributions may be in high dimensions, can differ in non-trivial ways (i.e., not simply in their means), and are observed only through i.i.d. samples. We take a two-sample hypothesis testing approach to discovering features which best distinguish P and Q. Our approach builds on the analytic representations of probability distributions of Chwialkowski et al. (2015), where differences in expectations of analytic functions at particular spatial (ME test) or frequency locations (SCF test) are used to construct a two-sample test statistic, which can be computed in linear time. Despite the differences in these analytic functions being evaluated at a finite set of locations, the analytic tests have greater power than linear time tests based on subsampled estimates of the MMD (Gretton et al., 2012b; Zaremba et al., 2013).

Given two samples $\mathsf{X} := \{\mathbf{x}_i\}_{i=1}^n, \mathsf{Y} := \{\mathbf{y}_i\}_{i=1}^n \subset \mathbb{R}^d$ independently and identically distributed (i.i.d.) according to P and Q, respectively, the goal of a two-sample test is to decide whether P is different from Q on the basis of the two samples. The task is formulated as a statistical hypothesis test proposing a null hypothesis $H_0 : P = Q$ (samples are drawn from the same distribution) against an alternative hypothesis $H_1 : P \neq Q$ (the sample generating distributions are different). A test calculates a test statistic $\hat{\lambda}_n$ from X and Y, and rejects H_0 if $\hat{\lambda}_n$ exceeds a predetermined test threshold (critical value). The threshold T_α is given by the $(1-\alpha)$-quantile of the distribution of $\hat{\lambda}_n$ under H_0 i.e., null distribution, and α is the significance level of the test.

Mean Embedding Test (ME Test) The ME test uses as its test statistic $\hat{\lambda}_n$, a form of Hotelling's T-squared statistic, defined as $\hat{\lambda}_n := n\overline{\mathbf{z}}_n^\top \mathbf{S}_n^{-1}\overline{\mathbf{z}}_n$, where $\overline{\mathbf{z}}_n := \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i$, $\mathbf{S}_n := \frac{1}{n-1}\sum_{i=1}^n (\mathbf{z}_i - \overline{\mathbf{z}}_n)(\mathbf{z}_i - \overline{\mathbf{z}}_n)^\top$, and $\mathbf{z}_i := (k(\mathbf{x}_i, \mathbf{v}_j) - k(\mathbf{y}_i, \mathbf{v}_j))_{j=1}^J \in \mathbb{R}^J$. The statistic depends on a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (with $\mathcal{X} \subseteq \mathbb{R}^d$), and a set of J test locations $\mathcal{V} = \{\mathbf{v}_j\}_{j=1}^J \subset \mathbb{R}^d$. Under H_0, asymptotically $\hat{\lambda}_n$ follows $\chi^2(J)$, a chi-squared distribution with J degrees of freedom. The ME test rejects H_0 if $\hat{\lambda}_n > T_\alpha$, where the test threshold T_α is given by the $(1-\alpha)$-quantile of the asymptotic null distribution $\chi^2(J)$. Although the distribution of $\hat{\lambda}_n$ under H_1 was not derived, Chwialkowski et al. (2015) showed that if k is analytic, integrable and characteristic (in the sense of Sriperumbudur et al. (2011)), under H_1 $\hat{\lambda}_n$ can be arbitrarily large as $n \to \infty$, allowing the test to correctly reject H_0.

Smooth Characteristic Function Test (SCF Test) The SCF uses the test statistic which has the same form as $\hat{\lambda}_n$ in the ME test with a modified

$$\mathbf{z}_i := [\hat{l}(\mathbf{x}_i)\sin(\mathbf{x}_i^\top \mathbf{v}_j) - \hat{l}(\mathbf{y}_i)\sin(\mathbf{y}_i^\top \mathbf{v}_j), \\ \hat{l}(\mathbf{x}_i)\cos(\mathbf{x}_i^\top \mathbf{v}_j) - \hat{l}(\mathbf{y}_i)\cos(\mathbf{y}_i^\top \mathbf{v}_j)]_{j=1}^J \in \mathbb{R}^{2J},$$

where $\hat{l}(\mathbf{x}) = \int_{\mathbb{R}^d} \exp(-i\mathbf{u}^\top \mathbf{x}) l(\mathbf{u})\,\mathrm{d}\mathbf{u}$ is the Fourier transform of $l(\mathbf{x})$, and $l : \mathbb{R}^d \to \mathbb{R}$ is an analytic smoothing kernel. In contrast to the ME test defining the statistic in terms of spatial locations, the locations $\mathcal{V} = \{\mathbf{v}_j\}_{j=1}^J$ in the SCF test are in the frequency domain.

2. Main Contributions

The statistic $\hat{\lambda}_n$ for both ME and SCF tests depends on a set of test locations $\mathcal{V}$ and a kernel k. For simplicity, assume a Gaussian kernel $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)$. A well chosen $\theta := \{\mathcal{V}, \sigma\}$ will increase the probability of correctly rejecting H_0 when H_1 holds i.e., $\mathbb{P}(\hat{\lambda}_n \geq T_\alpha | H_1)$ or test power. We propose to optimize θ by maximizing a test power proxy, defined as a lower bound on the test power. The optimization of θ brings two benefits: first, it significantly increases the probability of rejecting H_0 when H_1 holds; second, the learned test locations act as discriminative features allowing an interpretation of how the two distributions differ.

Let $\lambda_n := n\boldsymbol{\mu}^\top \boldsymbol{\Sigma}\boldsymbol{\mu}$, $\boldsymbol{\mu} := \mathbb{E}[\mathbf{z}_1]$, and $\boldsymbol{\Sigma} := \mathbb{E}[(\mathbf{z} - \boldsymbol{\mu})(\mathbf{z} - \boldsymbol{\mu})^\top]$. We have $\mathbb{P}(\hat{\lambda}_n > T_\alpha | H_1) \geq 1 - 2\exp\left(\frac{[(n-1)(\lambda_n - T_\alpha) - 24Jcn]^2}{72J^4(2n-1)^2c^2n^2}\right) - 4\exp\left(-\frac{(\lambda_n - T_\alpha)^2}{72c^2nJ^4}\right)$ as a lower bound on the test power, where c is a global constant bounding $\|\mathbf{S}_n^{-1}\|_F$ and $\|\boldsymbol{\Sigma}^{-1}\|_F$ for all $\mathcal{V}$ and for all Gaussian kernels. This lower bound can be derived by applying Hoeffding's inequality to bound $\|\overline{\mathbf{z}}_n - \boldsymbol{\mu}\|_2$ and $\|\mathbf{S}_n - \boldsymbol{\Sigma}\|_F$, and combining the results with a union bound. It can be seen that, for large n, to maximize the lower bound on the power, it is sufficient to maximize λ_n. In practice, since $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown, in place of λ_n we use $\hat{\lambda}^{tr}_{n/2} \propto \overline{\mathbf{z}}_n^\top \mathbf{S}_n^{-1} \overline{\mathbf{z}}_n$, an empirical quantity computed on a held-out training set of size $n/2$. The actual test statistic is denoted by $\hat{\lambda}^{te}_{n/2}$ which is computed on a test sample of size $n/2$.

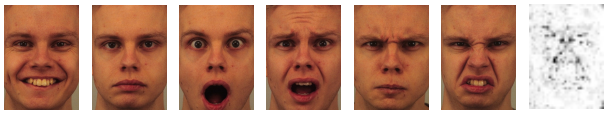
We also derive a finite-sample bound to $|\sup_{\mathcal{V},\sigma} \overline{\mathbf{z}}_n^\top \mathbf{S}_n^{-1} \overline{\mathbf{z}}_n - \sup_{\mathcal{V},\sigma} \boldsymbol{\mu}^\top \boldsymbol{\Sigma}\boldsymbol{\mu}|$. The result implies that the optimization objective converges almost surely to its population quantity uniformly over the class of Gaussian kernels, and all distinct test locations $\mathcal{V}$. We omit the technical details due to the lack of space. We note that optimizing parameters by maximizing a test power proxy (Gretton et al., 2012b) is valid under both H_0 and H_1 as long as the data used for parameter tuning and for testing are disjoint.

3. Distinguishing Pos. and Neg. Emotions

We study empirically how well the ME and SCF tests can distinguish two samples of photos of people showing positive and negative facial expressions. We use Karolinska Directed Emotional Faces (KDEF) dataset (Lundqvist et al., 1998) containing face images of 70 amateur actors, 35 females and 35 males. Each actor poses six expressions: happy (HA), neutral (NE), surprised (SU), afraid (AF), angry (AN), and disgusted (DI). We assign HA, NE, and SU faces into the positive emotion group (i.e., samples from P), and AF, AN and DI faces into the negative emotion group (samples from Q). We denote this problem as "+ vs. −". Examples of six facial expressions from one actor are shown in Fig. 1. Each image is cropped to exclude the background, resized to $48 \times 34 = 1632$ pixels (d dimensions), and converted to grayscale.

Table 1. Type-I errors and powers in the problem of distinguishing positive (+) and negative (-) facial expressions. $\alpha = 0.01$. $J = 1$.

Problem	n^{te}	ME-full	SCF-full	MMD-lin
$\pm$ vs. $\pm$	201	.010	.014	.008
$+$ vs. $-$	201	.998	1.00	.618



(a) HA (b) NE (c) SU (d) AF (e) AN (f) DI (g) $\mathbf{v}_1$

Figure 1. (a)-(f): Six facial expressions of actor AM05 in the KDEF data. (g): Average across trials of the learned test locations $\mathbf{v}_1$.

For the SCF test, we set $\hat{l}(\mathbf{x}) = k(\mathbf{x}, 0)$. Denote by ME-full and SCF-full the ME and SCF tests whose test locations and the Gaussian kernel width σ are fully optimized using gradient ascent on a separate training sample of the same size as the test set. MMD-lin refers to the nonparameteric test based on maximum mean discrepancy of Gretton et al. (2012a), where we use a linear-time estimator for the MMD (see Gretton et al. (2012a, Section 6)). We run the tests 500 times with $J = 1$ and $\alpha = 0.01$. Samples are partitioned randomly into training and test sets in each trial. We report an empirical estimate of $\mathbb{P}(\hat{\lambda}^{te}_{n/2} > T_\alpha)$ which is the proportion of the number of times the statistic $\hat{\lambda}^{te}_{n/2}$ is above T_α. The quantity $\mathbb{P}(\hat{\lambda}^{te}_{n/2} > T_\alpha)$ is type-I error (false positive) under H_0, and corresponds to test power when H_1 is true.

The type-I errors and test powers are shown in Table 1. In the table, "$\pm$ vs. $\pm$" is a problem in which all faces expressing the six emotions are randomly split into two samples of equal sizes i.e., H_0 is true. Evidently, both ME-full and SCF-full achieve high test powers while maintaining the right type-I errors. As a way to interpret how positive and negative emotions differ, we take an average across trials of the learned test locations of ME-full in the "+ vs. −" problem. This average is shown in Fig. 1g. Indeed, we see that the test locations faithfully capture the difference of positive and negative emotions by giving more weights to the regions of nose, upper lip, and nasolabial folds (smile lines), confirming the interpretability of the test in a high-dimensional problem.

References

Chwialkowski, Kacper, Ramdas, Aaditya, Sejdinovic, Dino, and Gretton, Arthur. Fast two-sample testing with analytic representations of probability measures. In *NIPS*, pp. 1972–1980, 2015.

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

Gretton, Arthur, Sejdinovic, Dino, Strathmann, Heiko, Balakrishnan, Sivaraman, Pontil, Massimiliano, Fukumizu, Kenji, and Sriperumbudur, Bharath K. Optimal kernel choice for large-scale two-sample tests. In *NIPS*, pp. 1205–1213, 2012b.

Lundqvist, Daniel, Flykt, Anders, and Öhman, Arne. The Karolinska directed emotional faces-KDEF. Technical report, ISBN 91-630-7164-9, 1998.

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

Zaremba, Wojciech, Gretton, Arthur, and Blaschko, Matthew. B-test: A non-parametric, low variance kernel two-sample test. In *NIPS*, pp. 755–763, 2013.