A Linear-Time Kernel Goodness-of-Fit Test

Authored by:

Kenji Fukumizu
Arthur Gretton
Wittawat Jitkrittum
Zoltan Szabo
Wenkai Xu

Abstract

We propose a novel adaptive test of goodness-of-fit, with computational cost linear in the number of samples. We learn the test features that best indicate the differences between observed samples and a reference model, by minimizing the false negative rate. These features are constructed via Stein's method, meaning that it is not necessary to compute the normalising constant of the model. We analyse the asymptotic Bahadur efficiency of the new test, and prove that under a mean-shift alternative, our test always has greater relative efficiency than a previous linear-time kernel test, regardless of the choice of parameters for that test. In experiments, the performance of our method exceeds that of the earlier linear-time test, and matches or exceeds the power of a quadratic-time kernel test. In high dimensions and where model structure may be exploited, our goodness of fit test performs far better than a quadratic-time two-sample test based on the Maximum Mean Discrepancy, with samples drawn from the model.

1 Paper Body

The goal of goodness of fit testing is to determine how well a model density p(x) fits an observed sample D = {xi }ni=1 ? X ? Rd from an unknown distribution q(x). This goal may be achieved via a hypothesis test, where the null hypothesis H0 : p = q is tested against H1 : p 6= q. The problem of testing goodness of fit has a long history in statistics [11], with a number of tests proposed for particular parametric models. Such tests can require space partitioning [18, 3], which works poorly in high dimensions; or closed-form integrals under the model, which may be difficult to obtain, besides in certain special cases [2, 5, 30, 26]. An alternative is to conduct a two-sample test using samples drawn from both p and q. This approach was taken by [23], using a test based on the (quadratic-time) Maximum Mean Discrepancy [16], however this does not take advantage of the known structure of p (quite apart from the increased computational cost of dealing with samples from p). More recently, measures of discrepancy with respect to a model have been proposed based on Stein?s method [21]. A Stein operator for p may be applied to a class of test functions, yielding functions that have zero expectation under p. Classes of test functions can include the W 2,? Sobolev space [14], and reproducing kernel Hilbert spaces (RKHS) [25]. Statistical tests have been proposed by [9, 22] based on classes of Stein transformed RKHS functions, where the test statistic is the norm of the smoothness-constrained function with largest expectation under q . We will refer to this statistic as the Kernel Stein Discrepancy (KSD). For consistent tests, it is sufficient to use C0 -universal kernels [6, Definition 4.1], as shown by [9, Theorem 2.2], although inverse multiquadric kernels may be preferred if uniform tightness is required [15].2 ?

Zolt?n Szab??s ORCID ID: 0000-0001-6183-7603. Arthur Gretton?s ORCID ID: 0000-0003-3169-7624. Briefly, [15] show that when an exponentiated quadratic kernel is used, a sequence of sets D may be constructed that does not correspond to any q, but for which the KSD nonetheless approaches zero. In a statistical testing setting, however, we assume identically distributed samples from q, and the issue does not arise. 2

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

The minimum variance unbiased estimate of the KSD is a U-statistic, with computational cost quadratic in the number n of samples from q. It is desirable to reduce the cost of testing, however, so that larger sample sizes may be addressed. A first approach is to replace the U-statistic with a running average with linear cost, as proposed by [22] for the KSD, but this results in an increase in variance and corresponding decrease in test power. An alternative approach is to construct explicit features of the distributions, whose empirical expectations may be computed in linear time. In the two-sample and independence settings, these features were initially chosen at random by [10, 8, 32]. More recently, features have been constructed explicitly to maximize test power in the two-sample [19] and independence testing [20] settings, resulting in tests that are not only more interpretable, but which can yield performance matching quadratic-time tests. We propose to construct explicit linear-time features for testing goodness of fit, chosen so as to maximize test power. These features further reveal where the model and data differ, in a readily interpretable way. Our first theoretical contribution is a derivation of the null and alternative distributions for tests based on such features, and a corresponding power optimization criterion. Note that the goodness-of-fit test requires somewhat different strategies to those employed for two-sample and independence testing [19, 20], which become computationally prohibitive in high dimensions for the Stein discrepancy (specifically, the normalization used in prior work to simplify the asymptotics would incur a cost cubic in the dimension d and the number of features in the optimization). Details may be found in Section 3. Our second theoretical contribution, given in Section 4, is an analysis of the relative Bahadur efficiency of our test vs the linear time test of [22]: this represents the relative rate at which the pvalue decreases under H1 as we observe more samples. We prove that our test has greater asymptotic Bahadur efficiency relative to the test of [22], for Gaussian distributions under the mean-shift alternative. This is shown to hold regardless of the bandwidth of the exponentiated quadratic kernel used for the earlier test. The proof techniques developed are of independent interest, and we anticipate that they may provide a foundation for the analysis of relative efficiency of linear-time tests in the two-sample and independence testing domains. In experiments (Section 5), our new linear-time test is able to detect subtle local differences between the density p(x), and the unknown q(x) as observed through samples. We show that our linear-time test constructed based on optimized features has comparable performance to the quadratic-time test of [9, 22], while uniquely providing an explicit visual indication of where the model fails to fit the data.

2

Kernel Stein Discrepancy (KSD) Test

We begin by introducing the Kernel Stein Discrepancy (KSD) and associated statistical test, as proposed independently by [9] and [22]. Assume that the data domain is a connected open set X ? Rd . Consider a Stein operator Tp that takes in a multivariate function f (x) = (f1 (x), . . . , fd (x))¿ ? Rd and constructs a function (Tp f) (x) : Rd ? R. The constructed function has the key property that for all f in an appropriate function class, Ex?q [(Tp f)(x)] = 0 if and only if q = p. Thus, one can use this expectation as a statistic for testing goodness of fit. The function class F d for the function f is chosen to be a unit-norm ball in a reproducing kernel Hilbert space (RKHS) in [9, 22]. More precisely, let F be an RKHS associated with a positive definite kernel k : X ? X ? R. Let ?(x) = k(x, ?) denote a feature map of k so that k(x, x0) = h?(x), ?(x0)iF . Assume that fi ? F for all i = 1, . . . , d so that f ? F ? ? ? ? ? F := F d where F d is equipped with Pd the standard inner product hf , giF d := i=1 hfi , gi iF . The kernelized Stein operator Tp studied

Pd p(x) ?fi (x) (a)

in [9] is (Tp f) (x) := i=1 ? log f (x) + = f , ? p (x, ?) F d , where at (a) we use the i ?xi ?xi reproducing property of F, i.e., fi (x) = hfi , k(x, ?)iF , and that hence ? p (x, ?) := ? log?xp(x) k(x, ?)+ ?k(x,?) is in F d . ?x is defined such that (Tp f) (x) ? Rd . This distinction

?k(x,?) ?xi

? F [28, Lemma 4.34],

We note that the Stein operator presented in [22] is not crucial and leads to the same goodness-offit test. Under appropriate conditions, e.g. that limkxk?? p(x)fi (x) = 0 for all i = 1, . . . , d, it can be shown using integration by parts that Ex?p (Tp f)(x) = 0 for any f ? F d [9, Lemma 5.1]. Based on the Stein operator, [9, 22] define the kernelized Stein discrepancy as Sp (q) :=

sup kf kF d ?1

(a) Ex?q f , ? p (x, ?) F d =

sup kf kF d ?1

2

f , Ex?q ? p (x, ?) F d = kg(?)kF d ,

(1)

where at (a), ? p (x, ?) is Bochner integrable [28, Definition A.5.20] as long as Ex?q k? p (x, ?)kF d ¡ ?, and g(y) := Ex?q ? p (x, y) is what we refer to as the Stein witness function. The Stein witness function will play a crucial role in our new test statistic in Section 3. When a C0 -universal kernel is used [6, Definition 4.1], and as long as Ex?q k?x log p(x) ? ?x log q(x)k2 ¡ ?, it can be shown that Sp (q) = 0 if and only if p = q [9, Theorem 2.2]. Ex?q Ex0 ?q hp (x, x0), where hp (x, y) := Pd ? 2 k(x,y) ¿ ¿ s¿ p (x)sp (y)k(x, y) + sp (y)?x k(x, y) + sp (x)?y k(x, y) + i=1 ?xi ?yi , and sp (x) := c2 = ?x log p(x) is a column vector. An unbiased empirical estimator of Sp2 (q), denoted by S P 2 i¡j hp (xi , xj) [22, Eq. 14], is a degenerate U-statistic under H0 . For the goodness-of-fit n(n?1) c2 test, the rejection threshold can be computed by a bootstrap procedure. All these properties make S a very flexible criterion to detect the discrepancy of p and q: in particular, it can be computed even if p is known only up to a normalization constant. Further studies on nonparametric Stein operators can be found in [25, 14].

The KSD Sp (q) can be written as Sp2 (q)

=

c2 costs O(n2). To reduce this cost, a Linear-Time Kernel Stein (LKS) Test Computation of S linear-time (i.e., O(n)) estimator based on an incomplete U-statistic is proposed in [22, Eq. 17], c2 := 2 Pn/2 h (x given by S i=1 p 2i?1 , x2i), where we assume n is even for simplicity. Empirically l n [22] observed that the linear-time estimator performs much worse (in terms of test power) than the quadratic-time U-statistic estimator, agreeing with our findings presented in Section 5.

3

New Statistic: The Finite Set Stein Discrepancy (FSSD)

Although shown to be powerful, the main drawback of the KSD test is its high computational cost of O(n2). The LKS test is one order of magnitude faster. Unfortunately, the decrease in the test power outweighs the computational gain [22]. We therefore seek a variant of the KSD statistic that can be computed in linear time, and whose test power is comparable to the KSD test. Key Idea The fact that Sp (q) = 0 if and only if p = q implies that g(v) = 0 for all v ? X if and only if p = q, where g is the Stein witness function in (1). One can see g as a function witnessing the differences of p, q, in such a way that —gi (v)— is large when there is a discrepancy in the region around v, as indicated by the ith output of g. The test statistic of [22, 9] is essentially given by the degree of ?flatness? of g as measured by the RKHS norm k ? kF d . The core of our proposal is to use a different measure of flatness of g which can be computed in linear time. The idea is to use a real analytic kernel k which makes g1 , . . . , gd real analytic. If gi 6= 0 is an analytic function, then the Lebesgue measure of the set of roots {x — gi (x) = 0} is zero [24]. This property suggests that one can evaluate gi at a finite set of locations V = {v1 , . . . , vJ }, drawn from a distribution with a density (w.r.t. the Lebesgue measure). If gi 6= 0, then almost surely gi (v1), . . . , gi (vJ) will not be zero. This idea was successfully exploited in recently proposed linear-time tests of [8] and [19, 20]. Our new test statistic based on this idea is called the Finite Set Stein Discrepancy (FSSD) and is given in Theorem 1. All proofs are given in the appendix. Theorem 1 (The Finite Set Stein Discrepancy (FSSD)). Let V = {v1 , . . . , vJ } ? Rd be random vectors drawn i.i.d. from a distribution ? which has a density. Let X be a connected open set Pd PJ 1 2 in Rd . Define FSSD2p (q) := dJ j=1 gi (vj). Assume that 1) k : X ? X ? R is C0 i=1 universal [6, Definition 4.1] and real analytic i.e., for all v ? X , f (x) := k(x, v) is a real analytic function on X . 2) Ex?q Ex0 ?q hp (x, x0) ¡ ?. 3) Ex?q k?x log p(x) ? ?x log q(x)k2 ¡ ?. 4) limkxk?? p(x)g(x) = 0. Then, for any J ? 1, ?-almost surely FSSD2p (q) = 0 if and only if p = q. This measure depends on a set of J test locations (or features) {vi }Ji=1 used to evaluate the Stein witness function, where J is fixed and is typically small.2 A kernel which is C0 -universal and real kx?yk analytic is the Gaussian kernel k(x, y) = exp ? 2?2 2 (see [20, Proposition 3] for the result k on analyticity). Throughout this work, we will assume all the conditions stated in Theorem 1, and consider only the Gaussian kernel. Besides the requirement that the kernel be real and analytic, the remaining conditions in Theorem 1 are the same as given in [9, Theorem 2.2]. Note that if the 3

FSSD is to be employed in a setting otherwise than testing, for instance to obtain pseudo-samples converging to p, then stronger conditions may be needed [15]. 3.1

Goodness-of-Fit Test with the FSSD Statistic

Given a significance level ? for the goodness-of-fit test, the test can be constructed so that H0 is 2 is 2 ¿ T? , where T? is the rejection threshold (critical value), and FSSD rejected when nFSSD an empirical estimate of FSSD2p (q). The threshold which guarantees that the type-I error (i.e., the probability of rejecting H0 when it is true) is bounded above by ? is given by the (1 ? ?)-quantile of 2 under H0 . In the following, we start by giving the null distribution i.e., the distribution of nFSSD 2 , and summarize its asymptotic distributions in Proposition 2. the expression for FSSD ? Let ?(x) ? Rd?J such that [?(x)]i,j = ?p,i (x, vj)/ dJ. Define ? (x) := vec(?(x)) ? RdJ where vec(M) concatenates columns of the matrix M into a column vector. We note that ? (x) depends on the test locations V = {vj }Jj=1 . Let ?(x, y) := ? (x)¿ ? (y) = tr(?(x)¿ ?(y)). Given an i.i.d. sample {xi }ni=1 ? q, a consistent, unbiased estimator of FSSD2p (q) is d X n X J X X X 2 1 2 = 1 FSSD ?(xi , xj), ?p,l (xi , vm)?p,l (xj , vm) = dJ n(n ? 1) i=1 n(n ? 1) i¡j m=1 l=1

(2)

j6=i

which is a one-sample second-order U-statistic with ? as its U-statistic kernel [27, Section 5.1.1]. d Being a U-statistic, its asymptotic distribution can easily be derived. We use ? to denote convergence in distribution. 2). Let Z1 , . . . , ZdJ i.i.d. ? N (0, 1). Let ? := Proposition 2 (Asymptotic distributions of FSSD dJ?dJ Ex?q [? (x)], ?r := covx?r [? (x)] ? R for r ? {p, q}, and {?i }dJ i=1 be the eigenvalues of ?p = Ex?p [? (x)? ¿ (x)]. Assume that Ex?q Ey?q ?2 (x, y) ¡ ?. Then, for any realization of V = {vj }Jj=1 , the following statements hold. d

2 ? 1. Under H0 : p = q, nFSSD

PdJ
2 i=1 (Zi
? 1)?i .
2 2. Under H1 : p 6= q, if ?H := 4?¿ ?q ? ¿ 0, then 1
?
d 2 2 ? FSSD2) ? n(FSSD N (0, ?H). 1

Proof. Recognizing that (2) is a degenerate U-statistic, the results follow directly from [27, Section 5.5.1, 5.5.2]. Claims 1 and 2 of Proposition 2 imply that under H1 , the test power (i.e., the probability of correctly rejecting H1) goes to 1 asymptotically, if the threshold T? is defined as above. In practice, simulating from the asymptotic null distribution in Claim 1 can be challenging, since the plug-in estimator of ?p requires a sample from p, which is not available. A straightforward solution is to draw sample from p, either by assuming that p can be sampled easily or by using a Markov chain Monte Carlo (MCMC) method, although this adds an additional computational burden to the test procedure. A more subtle issue is that when dependent samples from p are used in obtaining the test threshold, the test may become more conservative than required for i.i.d. data [7]. An alternative approach is to use ? q instead of ?p . The covariance matrix ? ? q can be directly computed from the the plug-in estimate ? data. This is the approach we take. Theorem 3 guarantees that the replacement of the covariance in the computation of the asymptotic null distribution still yields a consistent test. We write PH1 for the 2 under H1 . distribution of nFSSD ? q := 1 Pn ? (xi)? ¿ (xi) ? [1 Pn ? (xi)][1 Pn ? (xj)]¿ with {xi }n ? Theorem 3. Let ? i=1 i=1 i=1 j=1 n n n PdJ q. Suppose that the test threshold T? is set to the (1??)-quantile of the distribution of i=1 (Zi2 ?1)??i i.i.d. ? q . Then, under H0 , asymptotically where {Zi }dJ ?1 , . . . , ??dJ are eigenvalues of ? i=1 ? N (0, 1), and ? J the false positive rate is ?. Under H1 , for {vj }j=1 drawn from a distribution with a density, the test 2 ¿ T?) ? 1 as n ? ?. power PH (nFSSD 1

?q = ? ? p i.e., the plug-in Remark 1. The proof of Theorem 3 relies on two facts. First, under H0 , ? ? estimate of ?p . Thus, under H0 , the null distribution approximated with ?q is asymptotically 4

? p to ?p . Second, the rejection threshold obtained from the correct, following the convergence of ? approximated null distribution is asymptotically constant. Hence, under H1 , claim 2 of Proposition 2 d 2 ? 2 ¿ T?) ? 1. implies that nFSSD ? as n ? ?, and consequently PH1 (nFSSD 3.2

Optimizing the Test Parameters

Theorem 1 guarantees that the population quantity FSSD2 = 0 if and only if p = q for any choice of {vi }Ji=1 drawn from a distribution with a density. In practice, we are forced to rely on the empirical 2 , and some test locations will give a higher detection rate (i.e., test power) than others for FSSD J finite n. Following the approaches of [17, 20, 19, 29], we choose the test locations V = {vj }j=1 and kernel bandwidth ?k2 so as to maximize the test power i.e., the probability of rejecting H0 when it is false. We first give an approximate expression for the test power when n is large. 2). Under H1 , for large n and fixed r, the Proposition 4 (Approximate test power of nFSSD

? 2 r 2 ¿ r) ? 1 ? ? ? ? n FSSD test power PH1 (nFSSD , where ? denotes the cumulative ?H1 n?H1 distribution function of the standard normal distribution, and ?H1 is defined in Proposition 2.

? r/n?FSSD2 2 ?FSSD2 2 ¿ r) = PH (FSSD 2 ¿ r/n) = PH ?n FSSD Proof. PH1 (nFSSD n . ¿ 1 1 ?H1 ?H1 For sufficiently large n, the alternative distribution is approximatelynormal as given in Proposition 2.

2 ¿ r) ? 1 ? ? ? r ? ?n FSSD2 . It follows that PH1 (nFSSD ?H n?H 1

1

Let ? := {V, ?k2 } be the collection of all tuning parameters. Assume that n is sufficiently large. ? 2 r Following the same argument as in [29], in ?n? ? n FSSD ?H1 , we observe that the first term H1 ? 2 ? r = O(n1/2), dominating = O(n?1/2) going to 0 as n ? ?, while the second term n FSSD ?H n?H 1

1

the first for large n. Thus, the best parameters that maximize the test power are given by ? ? = 2 ¿ T?) ? arg max? FSSD2 . Since FSSD2 and ?H are unknown, we divide arg max? PH1 (nFSSD 1 ?H 1

2

the sample {xi }ni=1 into two disjoint training and test sets, and use the training set to compute ??FSSD , H1 +? where a small regularization parameter ? ¿ 0 is added for numerical stability. The goodness-of-fit test is performed on the test set to avoid overfitting. The idea of splitting the data into training and test sets to learn good features for hypothesis testing was successfully used in [29, 20, 19, 17]. To find a local maximum of {vi }Ji=1

2 FSSD ? ?H1 +? ,

we use gradient ascent for its simplicity. The initial points of

are set to random draws from a normal distribution fitted to the training data, a heuristic we found to perform well in practice. The objective is non-convex in general, reflecting many possible ways to capture the differences of p and q. The regularization parameter ? is not tuned, and is 2 fixed to a small constant. Assume that ?x log p(x) costs O(d2) to evaluate. Computing ?? ??FSSD H +? 1

2 2 and ? costs O(d2 J 2 n). The computational complexity of nFSSD ?H is O(d2 Jn). Thus, finding 1 a local optimum via gradient ascent is still linear-time, for a fixed maximum number of iterations. ? q costs O(d2 J 2 n), and obtaining all the eigenvalues of ? ? q costs O(d3 J 3) (required Computing ? only once). If the eigenvalues decay to zero sufficiently rapidly, one can approximate the asymptotic null distribution with only a few eigenvalues. The cost to obtain the largest few eigenvalues alone can be much smaller. P ? := n1 ni=1 ? (xi). It is possible to normalize the FSSD statistic to get a new Remark 2. Let ? ? n := n? ? q + ?I)?1 ? ? ¿ (? ? where ? ? 0 is a regularization parameter that goes to 0 statistic ? as n ? ?. This was done in the case of the ME (mean embeddings) statistic of [8, 19]. The asymptotic null distribution of this statistic takes the convenient form of ?2 (dJ) (independent of ? q . It turns out that the test power p and q), eliminating the need to obtain the eigenvalues of ? ? criterion for tuning the parameters in this case is the statistic ?n itself. However, the optimization ? q + ?I)?1 (costing O(d3 J 3)) needs to be reevaluated in each is

computationally expensive as (? gradient ascent iteration. This is not needed in our proposed FSSD statistic.

5

4

Relative Efficiency and Bahadur Slope

Both the linear-time kernel Stein (LKS) and FSSD tests have the same computational cost of O(d2 n), and are consistent, achieving maximum power of 1 as n ? ? under H1 . It is thus of theoretical interest to understand which test is more sensitive in detecting the differences of p and q. This can be quantified by the Bahadur slope of the test [1]. Two given tests can then be compared by computing the Bahadur efficiency (Theorem 7) which is given by the ratio of the slopes of the two tests. We note that the constructions and techniques in this section may be of independent interest, and can be generalised to other statistical testing settings. We start by introducing the concept of Bahadur slope for a general test, following the presentation of [12, 13]. Consider a hypothesis testing problem on a parameter ?. The test proposes a null hypothesis H0 : ? ? ?0 against the alternative hypothesis H1 : ? ? ??0 , where ?, ?0 are arbitrary sets. Let Tn be a test statistic computed from a sample of size n, such that large values of Tn provide an evidence to reject H0 . We use plim to denote convergence in probability, and write Er for Ex?r Ex0 ?r . Approximate Bahadur Slope (ABS) For ?0 ? ?0 , let the asymptotic null distribution of Tn be F (t) = limn?? P?0 (Tn ¡ t), where we assume that the CDF (F) is continuous and common to all ?0 ? ?0 . The continuity of F will be important later when Theorem 9 and 10 are used to compute the slopes of LKS and FSSD tests. Assume that there exists a continuous strictly increasing function (Tn)) ? : (0, ?) ? (0, ?) such that limn?? ?(n) = ?, and that ?2 plimn?? log(1?F = c(?) ?(n) where Tn ? P? , for some function c such that 0 ¡ c(?A) ¡ ? for ?A ? ??0 , and c(?0) = 0 when ?0 ? ?0 . The function c(?) is known as the approximate Bahadur slope (ABS) of the sequence Tn . The quantifier ?approximate? comes from the use of the asymptotic null distribution instead of the exact one [1]. Intuitively the slope c(?A), for ?A ? ??0 , is the rate of convergence of p-values (i.e., 1 ? F (Tn)) to 0, as n increases. The higher the slope, the faster the p-value vanishes, and thus the lower the sample size required to reject H0 under ?A . (1)

(2)

Approximate Bahadur Efficiency Given two sequences of test statistics, Tn and Tn having the (1) (2) same ?(n) (see Theorem 10), the approximate Bahadur efficiency of Tn relative to Tn is defined (1) as E(?A) := c(1) (?A)/c(2) (?A) for ?A ? ??0 . If E(?A) ¿ 1, then Tn is asymptotically more (2) efficient than Tn in the sense of Bahadur, for the particular problem specified by ?A ? ??0 . We now give approximate Bahadur slopes for two sequences of linear time test statistics: the proposed c2 discussed in Section 2. 2 , and the LKS test statistic ?nS nFSSD l 2 is c(FSSD) := FSSD2 /?1 , where ?1 is the Theorem 5. The approximate Bahadur slope of nFSSD ¿ maximum eigenvalue of ?p := Ex?p [?(x)? (x)] and ?(n) = n. ? c2 Theorem 6. The approximate Bahadur slope of the linear-time kernel Stein (LKS) test statistic nS l 2 0 1 [Eq hp (x,x)] (LKS) is c

= 2 E h2 (x,x0) , where hp is the U-statistic kernel of the KSD statistic, and ?(n) = n.] p[p To make these results concrete, we consider the setting where p = N (0, 1) andq = N (?q , 1). We assume that both tests use the Gaussian kernel k(x, y) = exp ?(x ? y)2 /2?k2 , possibly with different bandwidths. We write ?k2 and ?2 for the FSSD and LKS bandwidths, respectively. Under these assumptions, the slopes given in Theorem 5 and Theorem 6 can be derived explicitly. The full expressions of the slopes are given in Proposition 12 and Proposition 13 (in the appendix). By [12, 13] (recalled as Theorem 10 in the supplement), the approximate Bahadur efficiency can be computed by taking the ratio of the two slopes. The efficiency is given in Theorem 7. Theorem 7 (Efficiency in the Gaussian mean shift problem). Let E1 (?q , v, ?k2 , ?2) be the approxic2 for the case where p = N (0, 1), q = N (? , 1), 2 relative to ?nS mate Bahadur efficiency of nFSSD l

q

2). Fix ? 2 = 1 for nFSSD 2 . Then, for any ?q 6= 0, and J = 1 (i.e., one test location v for nFSSD k for some v ? R, and for any ?2 ¿ 0, we have E1 (?q , v, ?k2 , ?2) ¿ 2.

When p = N (0, 1) and q = N (?q , 1) for ?q 6= 0, Theorem 7 guarantees that our FSSD test is asymptotically at least twice as efficient as the LKS test in the Bahadur sense. We note that the 6

efficiency is conservative in the sense that ?k2 = 1 regardless of ?q . Choosing ?k2 dependent on ?q will likely improve the efficiency further.

5

Experiments

In this section, we demonstrate the performance of the proposed test on a number of problems. The primary goal is to understand the conditions under which the test can perform well. p Sensitivity to Local Differences We start by demonstrating that q 2 the test power objective FSSD /?H1 captures local differences FSSD ? of p and q, and that interpretable features v are found. Consider a one-dimensional problem in which p = N (0, 1) and ? ?4 ?2 v ? 0 v ? 2 4 q = Laplace(0, 1/ 2), a zero-mean Laplace distribution with scale ? parameter 1/ 2. These parameters are chosen so that p and q have Figure 1: The power criterion the same mean and variance. Figure 1 plots the (rescaled) objective FSSD2 /?H1 as a function of as a function of v. The objective illustrates that the best features test location v. (indicated by v ?) are at the most discriminative locations. 2

H1

Test Power We next investigate the power of different tests on two problems: ? Qd 1. Gaussian vs. Laplace: p(x) = N (x—0, Id) and q(x) = i=1 Laplace(xi —0, 1/ 2) where the dimension d will be varied. The two distributions have the same mean and variance. The main characteristic of this problem is local differences of p and q (see Figure 1). Set n = 1000. 2. Restricted Boltzmann Machine (RBM): p(x) is the marginal distribution of p(x, h) =

1 1 ¿ ¿ ¿ 2 exp x Bh + b x + c x ? kxk , where x ? Rd , h ? {?1}dh is a random vector of Z 2 hidden variables, and Z is the normalization constant. The exact marginal density p(x) = P dh terms. h?{?1,1}dh p(x, h) is intractable when

dh is large, since it involves summing over 2 Recall that the proposed test only requires the score function ?x log p(x) (not the normalization constant), which can be computed in closed form in this case. In this problem, q is another RBM where entries of the matrix B are corrupted by Gaussian noise. This was the problem considered in [22]. We set d = 50 and dh = 40, and generate samples by n independent chains (i.e., n independent samples) of blocked Gibbs sampling with 2000 burn-in iterations. We evaluate the following six kernel-based nonparametric tests with ? = 0.05, all using the Gaussian kernel. 1. FSSD-rand: the proposed FSSD test where the test locations set to random draws from a multivariate normal distribution fitted to the data. The kernel bandwidth is set by the commonly used median heuristic i.e., ?k = median({kxi ? xj k, i ¡ j}). 2. FSSD-opt: the proposed FSSD test where both the test locations and the Gaussian bandwidth are optimized (Section 3.2). 3. KSD: the quadratic-time Kernel Stein Discrepancy test with the median heuristic. 4. LKS: the linear-time version of KSD with the median heuristic. 5. MMD-opt: the quadratic-time MMD two-sample test of [16] where the kernel bandwidth is optimized by grid search to maximize a power criterion as described in [29]. 6. ME-opt: the linear-time mean embeddings (ME) two-sample test of [19] where parameters are optimized. We draw n samples from p to run the two-sample tests (MMD-opt, ME-opt). For FSSD tests, we use J = 5 (see Section A for an investigation of test power as J varies). All tests with optimization use 20% of the sample size n for parameter tuning. Code is available at https://github.com/wittawatj/kernel-gof. Figure 2 shows the rejection rates of the six tests for the two problems, where each problem is repeated for 200 trials, resampling n points from q every time. In Figure 2a (Gaussian vs. Laplace), high performance of FSSD-opt indicates that the test performs well when there are local differences between p and q. Low performance of FSSD-rand emphasizes the importance of the optimization of FSSD-opt to pinpoint regions where p and q differ. The power of KSD quickly drops as the dimension increases, which can be understood since KSD is the RKHS norm of a function witnessing differences in p and q across the entire domain, including where these differences are small. We next consider the case of RBMs. Following [22], b, c are independently drawn from the standard multivariate normal distribution, and entries of B ? R50?40 are drawn with equal probability from {?1}, in each trial. The density q represents another RBM having the same b, c as in p, and with all entries of B corrupted by independent zero-mean Gaussian noise with standard deviation ?per . Figure 7

0.0

1

5 10 dimension d

15

1.0 0.5 0.0 0.00

ME-opt

300

0.75 0.50 0.25 0.00

0.02 0.04 0.06 Perturbation SD ?per

MMD-opt

Time (s)
0.5
LKS Rejection rate
1.0
KSD
2000 Sample size n
Rejection rate
FSSD-rand Rejection rate
Rejection rate
FSSD-opt
1.0
4000
200 100 0 1000
2000 3000 Sample size n
4000
0.5 (d) Runtime (RBM) (a) Gaussian vs. Laplace. (b) RBM. n = 1000. Per- (c) RBM. ?per = 0.1. Pern = 1000. turb all entries of B. turb B1,1 . 0.0 Figure 2: Rejection rates of the six tests. The proposed linear-time FSSD-opt has0.02 a comparable or 0.00 0.04 0.06 higher test power in some cases than the quadratic-time KSD test. Perturbation SD ?per

2b shows the test powers as ?per increases, for a fixed sample size n = 1000. We observe that all the tests have correct false positive rates (type-I errors) at roughly ? = 0.05 when there is no perturbation noise. In particular, the optimization in FSSD-opt does not increase false positive rate when H0 holds. We see that the performance of the proposed FSSD-opt matches that of the quadratic-time KSD at all noise levels. MMD-opt and ME-opt perform far worse than the goodness-of-fit tests when the difference in p and q is small (?per is low), since these tests simply represent p using samples, and do not take advantage of its structure. The advantage of having O(n) runtime can be clearly seen when the problem is much harder, requiring larger sample sizes to tackle. Consider a similar problem on RBMs in which the parameter B ? R50?40 in q is given by that of p, where only the first entry B1,1 is perturbed by random N (0, 0.12) noise. The results are shown in Figure 2c where the sample size n is varied. We observe that the two two-sample tests fail to detect this subtle difference even with large sample size. The test powers of KSD and FSSD-opt are comparable when n is relatively small. It appears that KSD has higher test power than FSSD-opt in this case for large n. However, this moderate gain in the test power comes with an order of magnitude more computation. As shown in Figure 2d, the runtime of the KSD is much larger than that of FSSD-opt, especially at large n. In these problems, the performance of the new test (even without optimization) far exceeds that of the LKS test. Further simulation results can be found in Section B. 0.20 Interpretable Features In the 0.16 final simulation, we demonstrate 0.12 that the learned test locations are 0.08 informative in visualising where 0.04 the model does not fit the data 0.00 well. We consider crime data ?0.04 from the Chicago Police Depart?0.08 ment, recording n = 11957 locations (latitude-longitude co- (a) p = 2-component GMM. (b) p =

10-component GMM ordinates) of robbery events in optimization objective as a function of Chicago in 2016.3 We address Figure 3: Plots of the 2 the situation in which a model p test location v ? R in the Gaussian mixture model (GMM) for the robbery location density is evaluation task. given, and we wish to visualise where it fails to match the data. We fit a Gaussian mixture model (GMM) with the expectationmaximization algorithm to a subsample of 5500 points. We then test the model on a held-out test set of the same size to obtain proposed locations of relevant features v. Figure 3a shows the test robbery locations in purple, the model with two Gaussian components in wireframe, and the optimization objective for v as a grayscale contour plot (a red star indicates the maximum). We observe that the 2-component model is a poor fit to the data, particularly in the right tail areas of the data, as indicated in dark gray (i.e., the objective is high). Figure 3b shows a similar plot with a 10-component GMM. The additional components appear to have eliminated some mismatch in the right tail, however a discrepancy still exists in the left region. Here, the data have a sharp boundary on the right side following the geography of Chicago, and do not exhibit exponentially decaying Gaussian-like tails. We note that tests based on a learned feature located at the maximum both correctly reject H0 . 3

Data can be found at https://data.cityofchicago.org.

8

Acknowledgement WJ, WX, and AG thank the Gatsby Charitable Foundation for the financial support. ZSz was financially supported by the Data Science Initiative. KF has been supported by KAKENHI Innovative Areas 25120012.

2 References

[1] R. R. Bahadur. Stochastic comparison of tests. The Annals of Mathematical Statistics, 31(2): 276?295, 1960. [2] L. Baringhaus and N. Henze. A consistent test for multivariate normality based on the empirical characteristic function. Metrika, 35:339?348, 1988. [3] J. Beirlant, L. Gy?rfi, and G. Lugosi. On the asymptotic normality of the l1 - and l2 -errors in histogram density estimation. Canadian Journal of Statistics, 22:309?318, 1994. [4] R. Bhatia. Matrix analysis, volume 169. Springer Science & Business Media, 2013. [5] A. Bowman and P. Foster. Adaptive smoothing and density based tests of multivariate normality. Journal of the American Statistical Association, 88:529?537, 1993. [6] C. Carmeli, E. De Vito, A. Toigo, and V. Umanit?. Vector valued reproducing kernel Hilbert spaces and universality. Analysis and Applications, 08(01):19?61, Jan. 2010. [7] K. Chwialkowski, D. Sejdinovic, and A. Gretton. A wild bootstrap for degenerate kernel tests. In NIPS, pages 3608?3616, 2014. [8] K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In NIPS, pages 1981?1989, 2015. [9] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In ICML, pages 2606?2615, 2016. [10] T. Epps and K. Singleton. An omnibus test for the two-sample problem using

the empirical characteristic function. Journal of Statistical Computation and Simulation, 26(3?4):177?203, 1986. [11] J. Frank J. Massey. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association, 46(253):68?78, 1951. [12] L. J. Gleser. On a measure of test efficiency proposed by R. R. Bahadur. 35(4):1537?1544, 1964. [13] L. J. Gleser. The comparison of multivariate tests of hypothesis by means of Bahadur efficiency. 28(2):157?174, 1966. [14] J. Gorham and L. Mackey. Measuring sample quality with Stein?s method. In NIPS, pages 226?234, 2015. [15] J. Gorham and L. Mackey. Measuring sample quality with kernels. In ICML, pages 1292?1301. PMLR, 06?11 Aug 2017. [16] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Sch?lkopf, and A. Smola. A kernel two-sample test. JMLR, 13:723?773, 2012. [17] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In NIPS, pages 1205?1213. 2012. [18] L. Gy?rfi and E. C. van der Meulen. A consistent goodness of fit test based on the total variation distance. In G. Roussas, editor, Nonparametric Functional Estimation and Related Topics, pages 631?645, 1990. [19] W. Jitkrittum, Z. Szab?, K. P. Chwialkowski, and A. Gretton. Interpretable Distribution Features with Maximum Testing Power. In NIPS, pages 181?189. 2016. [20] W. Jitkrittum, Z. Szab?, and A. Gretton. An adaptive test of independence with analytic kernel embeddings. In ICML, pages 1742?1751. PMLR, 2017. [21] C. Ley, G. Reinert, and Y. Swan. Stein?s method for comparison of univariate distributions. Probability Surveys, 14:1?52, 2017. 9

[22] Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In ICML, pages 276?284, 2016. [23] J. Lloyd and Z. Ghahramani. Statistical model criticism using kernel two sample tests. In NIPS, pages 829?837, 2015. [24] B. Mityagin. The Zero Set of a Real Analytic Function. Dec. 2015. arXiv: 1512.07276. [25] C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3):695?718, 2017. [26] M. L. Rizzo. New goodness-of-fit tests for Pareto distributions. ASTIN Bulletin: Journal of the International Association of Actuaries, 39(2):691?715, 2009. [27] R. J. Serfling. Approximation Theorems of Mathematical Statistics. John Wiley & Sons, 2009. [28] I. Steinwart and A. Christmann. Support Vector Machines. Springer, New York, 2008. [29] D. J. Sutherland, H.-Y. Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized Maximum Mean Discrepancy. In ICLR, 2016. [30] G. J. Sz?kely and M. L. Rizzo. A new test for multivariate normality. Journal of Multivariate Analysis, 93(1):58?80, 2005. [31] A. W. van der Vaart. Asymptotic Statistics. Cambridge University Press, 2000. [32] Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic. Large-scale kernel methods for independence testing. Statistics and Computing, pages 1?18, 2017.

10