An Alternative Kernel Method for the Two-Sample Problem

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present an alternative kernel method for the two-sample problem that is based on Friedman's approach of using any binary classification learning machine to score the data. When the learning machine is chosen to be a support vector machine, we show that this approach is a generalization of the permutation t-test. Previous work has yielded a normal rate of convergence bound using Stein's Method in the simple setting of univariate data and a linear kernel with simulations, suggesting that this proof technique may be extended to address a more general setting. Despite a lack of tuning of the SVM parameters, this method is shown to be competitive with the Maximum Mean Discrepancy (MMD) test.

1 Introduction

The two-sample problem addresses the issue of comparing samples from two possibly different probability distributions. They range from simple parametric, location alternative tests on univariate data such as the t-test to more general non-parametric, asymptotically consistent tests, which have power against all alternatives. Many options exist for vectorial data, and kernels provide an enticing avenue for extensions to more general data types.

The two-sample problem is also widely prevalent: ensuring cross-platform compability of microarray data allows for the merging samples to achieve larger sample sizes. Biologists would like to know whether gene expression levels on a set of genes differ between cancer and control groups. Further uses for two-sample testing include authorship validation: Given two sets of documents, is the hypothesis of a single author consistent with the data?

The two-sample problem is generally posed in the following fashion: $\{\mathbf{x}_i\}_1^n$ are drawn from $p(\mathbf{x})$ and $\{\mathbf{y}_i\}_1^m$ are drawn from $q(\mathbf{y})$, where $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^p$. The goal is to test $H_0 : p(\mathbf{x}) = q(\mathbf{y})$ against $H_A : p(\mathbf{x}) \neq q(\mathbf{y})$. An ideal test should have power against all alternatives. That is, as $n, m \to \infty$, the test will always reject when $p \neq q$ for any non-zero significance level α.

2 The Friedman Two-Sample Test

Friedman proposed the following approach to the two-sample problem [1]:

For $\{\mathbf{x}_i\}_1^N$ drawn from $p(\mathbf{x})$ and $\{\mathbf{z}_i\}_1^M$ drawn from $q(\mathbf{x})$, we would like to test $\mathcal{H}_A : p \neq q$ against $\mathcal{H}_0 : p = q$.

1. Pool the two samples $\{\mathbf{u}_i\}_1^{N+M} = \{\mathbf{x}_i\}_1^N \cup \{\mathbf{z}_i\}_1^M$ to create a predictor variable training set.

2. Assign a response value $y_i = 1$ to the observations from the first sample ($1 \leq i \leq N$) and $y_i = -1$ to the observations from the second sample ($N + 1 \leq i \leq N + M$).
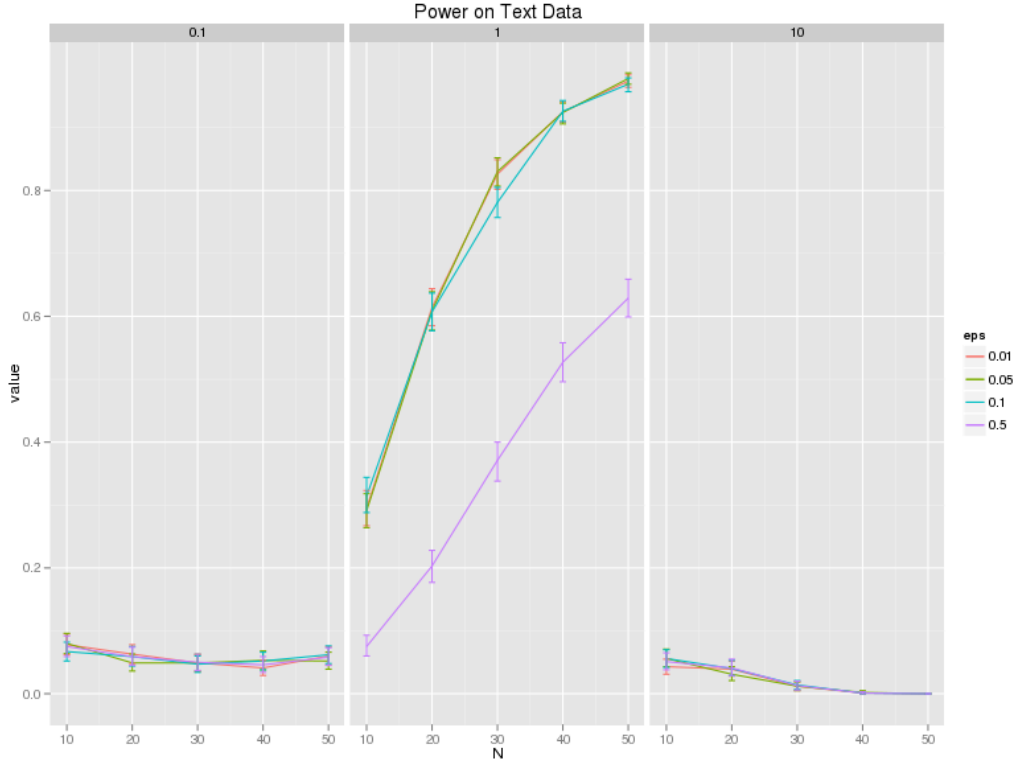
1

Figure 1: Friedman Test (SVM with 3-spectrum kernel) for string data of subsection 6.2 demonstrating power for columns $C \in \{.1, 1, 10\}$ and colors $\epsilon \in \{.01, .05, .1, .5\}$. Error bars indicate 95% bootstrap confidence intervals. Tuning parameter choice is *critical*. We fix $C = 1$ and $\epsilon = .1$ for computational considerations, but cross-validation is recommended.

3. Apply a binary classification learning machine to the training data to produce a scoring function $f(\mathbf{u})$ to score each of the observations $\{s_i = f(\mathbf{u}_i)\}_1^{N+M}$.

4. Calculate a univariate two-sample test statistic $\hat{t} = T(\{s_i\}_1^N, \{s_i\}_{N+1}^{N+M})$.

5. Determine the permutation null distribution of the above statistic to yield a p-value.

6. The test rejects $\mathcal{H}_0$ at significance level α if $p < \alpha$.

The Friedman Test (FT) is a simple, elegant idea that leverages the many advancements made over the past several decades in the fields of prediction and classification and applies them to the problem of two-sample testing. In short, as long as there exists a learning machine for the problem at hand, the Friedman Test provides a recipe for turning that learning machine into a two-sample test. This immediately yields two-sample tests for many kinds of data, including all types for which kernels have been defined. But there still remains some choice in the scoring function $F(\mathbf{u})$. It must be flexible enough to discriminate between the potential distributional differences of the problem at hand. The operating characteristics of the new two-sample test is *solely* a function of the paired learning algorithm.

By virtue of its permutation construction, the test has level α—the probability that we reject the null hypothesis given that the null hypothesis is true, also known as type I error. Given a threshold α, we wish to minimize the type II error, accepting the null hypothesis given that the alternative hypothesis is true. Equivalently, we wish to maximize the power, one minus the type II error [2]. The downside of the permutation design is, of course, that any computational cost is naïvely multiplied by the number of permutations. However, there are many situations for which the cost is sublinear in the number of permutations. For instance, caching the computation of the kernel matrix yields substantial savings when re-using it for permutation based inference. This is especially true when
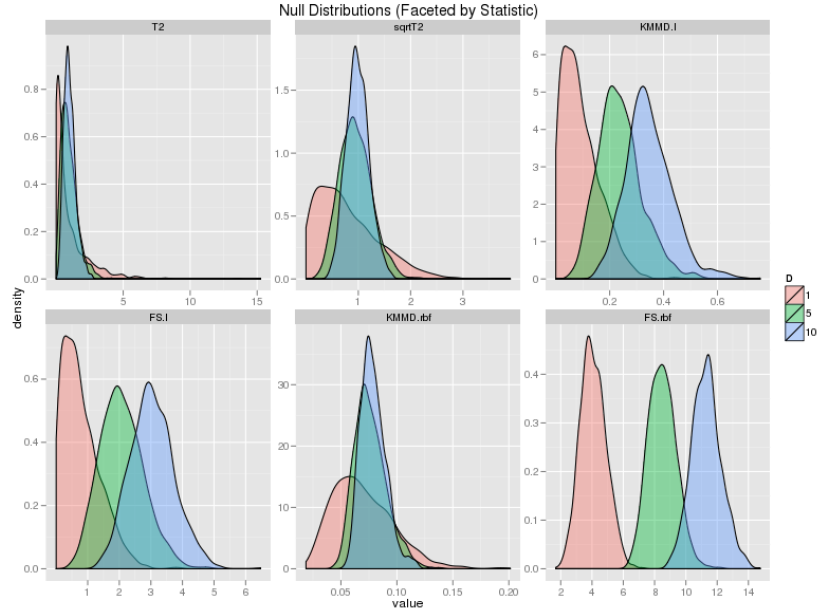
2

108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161

Figure 2: T2: Hotelling's T^2-statistic; sqrtT2: $|T|$; KMMD.l: kernel MMD with a linear kernel; FS.l: FS with a linear kernel; KMMD.rbf: kernel MMD with a radial basis function (RBF) kernel; FS.rbf: FS with RBF kernel
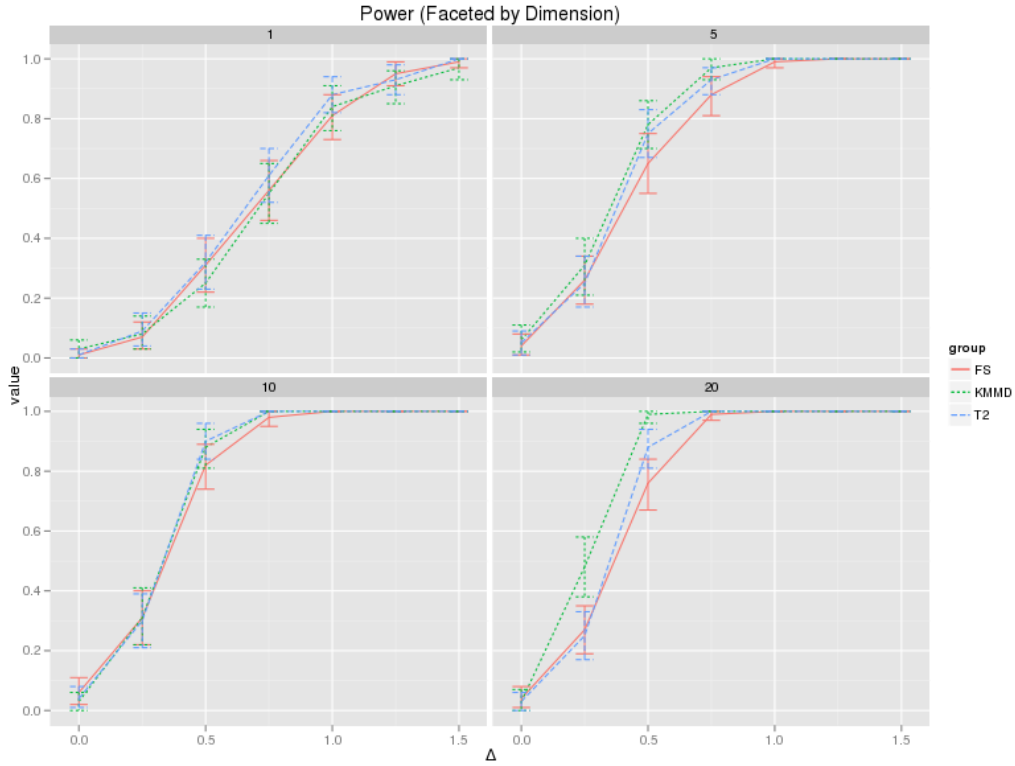

Figure 3: FS: Friedman statistic; KMMD: kernel Maximum Mean Discrepancy; T2: Hotelling's T^2-statistic; Error bars indicate 95% bootstrap confidence intervals. The tests perform similarly, and the kernel-based tests use a linear kernel.
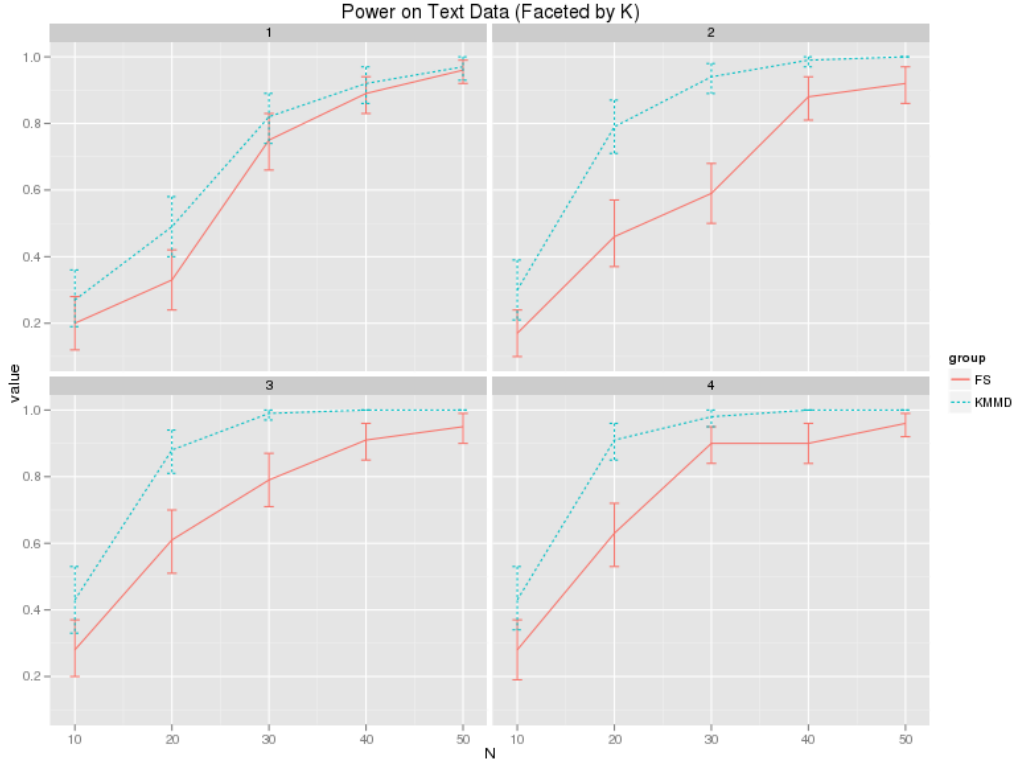
Figure 4: FS: Friedman statistic; KMMD: kernel Maximum Mean Discrepancy; Error bars indicate 95% bootstrap confidence intervals. The MMD test is more powerful.

Figure 5: Images of roosters and pigeons for use in discrimination test.

computation of the kernel matrix is expensive relative to finding the SVM parameters via quadratic programming.

3 SVM

We experience better computational results with Support Vector Machine (SVM) regression rather than classification as implemented in the **ksvm** function of the **R** [3] package **kernlab** [4].

Recall that SVM regression solves the following problem [5]:

$$
\begin{aligned}
& \underset{\mathbf{w}\in\mathcal{H},\xi^{(*)}\in\mathbb{R}^m,b\in\mathbb{R}}{\text{minimize}} && \tau(\mathbf{w},\xi^{(*)}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{M}(\xi_i + \xi_i^*) \\
& \text{subject to} && f(\mathbf{x}_i) - y_i \leq \epsilon + \xi_i \\
& && y_i - f(\mathbf{x}_i) \leq \epsilon + \xi_i^* \\
& && \xi_i, \xi_i^* \geq 0 \qquad \text{for all } i = 1,\ldots,m.
\end{aligned}
$$

4

with solution is given by

$$f(x) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) k(x_i, x) + b.$$

3.1 Tuning Parameters

The cost parameter C controls the complexity of the prediction function, and ϵ controls the leniency of the loss function. These parameters are typically chosen via cross-validation over a grid of choices. However, due to computational considerations, we mostly fix these values at $C = 1$ and $\epsilon = .1$. In subsection 6.2, we describe a sample of string data from Twitter. In figure 1 we demonstrate the statistical power of the test for the Twitter data over a grid of SVM parameters. It is clear that these parameters play a *crucial* role in the operating characteristics of the resultant test.

We emphasize that the proper strategy is to conduct the search anew for each statistic calculation in each permutation. That is, use cross-validation to find the best performing pair (C_0, ϵ_0) in terms of the Friedman Statistic. For each permutation i, use cross-validation over the same grid to find the ith pair (C_i, ϵ_i). This ensures symmetry of protocol and enforces that the test have level α. The grid search likely maximizes the power over the set of tuning parameters: it is hoped that the search benefits the actual labeling of values by at least as much as it does permuted labels.

3.2 Equivalence to Permutation t-test

Theorem 3.1. *The Friedman Test paired with support vector regression generalizes the two-sample permutation t-test. Namely, the two procedures are equivalent with univariate data and a linear kernel.*

Proof.

$$f(x) = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) k(x_i, x) + b = \sum_{i=1}^{m} (\alpha_i^* - \alpha_i) x_i x + b = wx + b$$

since we have univariate data and a linear kernel. Therefore, the SVM score is simply a linear transformation of the data. Welch's t-statistic is given by

$$T(\{x_i\}_1^N, \{z_i\}_1^M) = \frac{\bar{x} - \bar{z}}{\sqrt{\frac{s_X^2}{N} + \frac{s_z^2}{M}}}$$

where

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \text{ and } s_X^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2.$$

Let $z = f(x) = wx + b$ and note that

$$\bar{z} = \frac{1}{N} \sum_{i=1}^{N} z_i = \frac{w}{N} \sum_{i=1}^{N} x_i + b = w\bar{x} + b$$

and

$$s_Z^2 = \frac{1}{N-1} \sum_{i=1}^{N} (z_i - \bar{z})^2 = \frac{1}{N-1} \sum_{i=1}^{N} (wx_i + b - w\bar{x} + b)^2 = w^2 s_X^2.$$

Therefore,

$$T(\{f(x_i)\}_1^N, \{f(z_i)\}_1^M) = \frac{w\bar{x} + b - w\bar{z} + b}{|w|\sqrt{\frac{s_X^2}{N} + \frac{s_z^2}{M}}} = \text{sign}(w) T(\{x_i\}_1^N, \{z_i\}_1^M).$$

Since we are interested in two-sided testing, we consider

$$|T(\{f(x_i)\}_1^N, \{f(z_i)\}_1^M)| = |T(\{x_i\}_1^N, \{z_i\}_1^M)|.$$

Thus, the t-statistics are identical, and since the permutation procedure is the same, the tests are equivalent. $\square$
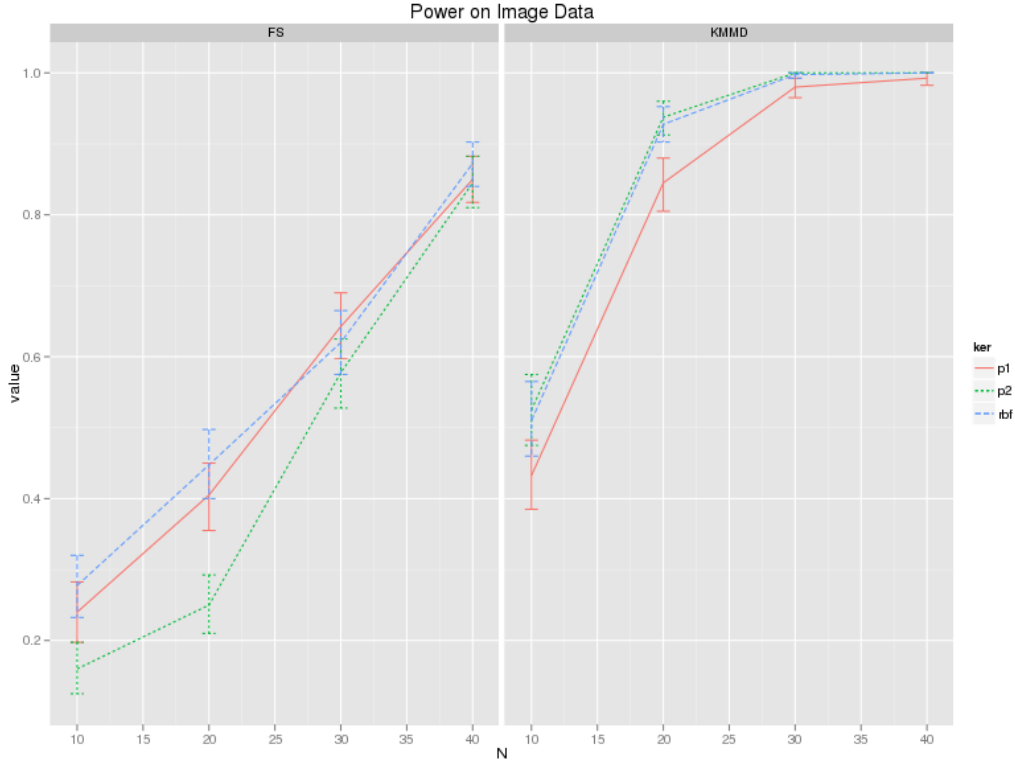
Figure 6: p1: linear kernel; p2: inhomogeneous degree 2 polynomial kernel; rbf: radial basis function kernel; Error bars indicate 95% bootstrap confidence intervals.

Despite the slight dependence, the randomization distribution of the t-statistic converges weakly to the normal distribution [6]. Anonymous and Anonymous [7] use Stein's Method of exchangeable pairs [8, 9] to prove a conservative $\mathcal{O}(N^{-1/4})$ rate of convergence in Kolmogorov-Smirnov distance between the two distributions. The problem is not as straightforward as in the i.i.d. case because the permutation structure induces a global—though mild and diminishing in sample size—negative dependence in the data. This dependence thwarts traditional Fourier-analytic techniques yet can be managed via Stein's eponymous method of proof.

4   Maximum Mean Discrepancy

Gretton et al. [10, 11, 12, 13] introduce a kernel based approach for the two-sample problem based on the Maximum Mean Discrepancy (MMD) statistic, an integral probability metric. MMD provides good performance in practice, strong theoretical guarantees, and is the first two-sample test for comparing distributions over graphs.

Definition 1. With $\mathfrak{F}$ a class of functions $f : \mathcal{X} \to \mathbb{R}$, p and q probability distributions, and $X \sim p$ and $Z \sim q$ random variables, the maximum mean discrepancy (MMD) and its empirical estimate are defined as

$$\text{MMD}[\mathfrak{F}, p, q] := \sup_{f \in \mathfrak{F}} (\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{z \sim q}[f(z)]),$$

$$\text{MMD}[\mathfrak{F}, X, Z] := \sup_{f \in \mathfrak{F}} \left(\frac{1}{N} \sum_{i=1}^{N} f(x_i) - \frac{1}{M} \sum_{i=1}^{M} f(z_i) \right).$$

The function class $\mathfrak{F}$ is typically taken to be the unit ball in a reproducing kernel Hilbert space (RKHS), however, well-known metrics can be obtained over other function classes. Although Gretton et al. provide several distribution-free tests based on MMD theory, we instead compare the

6

Friedman Test (FT) against the permutation-based MMD so as to compare statistic with statistic. In this way, the theory is dissociated from the comparison. We feel that this is the most fair comparison of the two tests because many of the theoretical results are inexact. We also do not have big enough sample sizes in our real datasets to ensure low error in theoretical approximations. Even if we did, the power for the tests would be very nearly one, making comparisons on non-simulated data difficult.

5 Null Distributions

The null distribution plays a fundamental role in frequentist statistical inference. Hotelling's T^2-statistic has null distribution that corresponds to a scaled central $F_{(p,n+m-1-p)}$ distribution, where p is the dimensionality of the data and n, m are the sample sizes of the two groups. As its name suggests, the T^2-test is a generalization of Student's t-test, and for $T \sim t(n + m - 2)$, we have that $T^2 \sim F_{(1,n+m-2)}$. As a consequence of Theorem 3.1, the Friedman Statistic in the univariate data, linear kernel setting is equal to the $|T|$. In figure 2 we simulate 200 standard multivariate normal draws from each class with dimension $D \in \{1, 5, 10\}$.

For the FS, the SVM cost parameter C is fixed at 1, with $\epsilon = .1$. We choose the RBF kernel hyperparameter via estimation techniques such as those implemented in the **sigest** function of **kernlab** [4]. Due to the different scales, it is not easy to see that FS and $|T|$ in fact have the same distribution. The T^2 densities correspond to a parametrized family of F-distributions. It is not surprising that the MMD linear kernel null distributions shift rightward as a function of dimension: the higher dimensionality affords the function in the RKHS to better find discrepancies between the two empirical distributions. The same rationale holds true for the FS when thinking of separating hyperplanes. Interestingly, there are marked differences between the MMD and FS for the RBF kernel.

6 Experiments

6.1 Vectorial Data

We consider $\{x_i\}_{i=1}^{20} \sim \text{MVN}_d(\mathbf{0}, \mathbf{I})$ and $\{y_i\}_{i=1}^{20} \sim \text{MVN}_d(\Delta\mathbf{1}, \mathbf{I})$ where our dimensionality $d \in \{1, 5, 10, 20\}$ and mean difference $\Delta \in \{0, .25, \ldots, 1.5\}$ in figure 3.

For FS and MMD, we used the the RBF kernel with the same method of hyperparameter estimation. In this simple setting, all three methods perform similarly with perhaps a small edge to MMD.

6.2 String Data

For a string data comparison, we consider Twitter data and look at the latest 1,000 tweets from Barack Obama (@BarackObama) and Sarah Palin (@SarahPalinUSA) obtained from the **R** package **twitteR** [14]. We pre-process each tweet by removing all hyperlinks and anything that is neither a letter nor a space. Finally, we convert all letters to lowercase. For simplicity, we choose the k-spectrum kernel [15] with $k = 4$ as our kernel for both the FT and MMD. Thus, each string is mapped to a 27^k dimensional feature vector of counts of the number of k letter and space combinations. We draw samples of various sizes from both the Barack Obama tweets and Sarah Palin tweets in order to empirically determine the power, with results detailed in figure 4.

The MMD test outperforms the Friedman test on this task. Power increases as a function of k for both tests, and it is somewhat surprising to see the strong performance from considering only frequencies of unigrams.

6.3 Image Data

We consider the task of discriminating between images of roosters and pigeons from the Caltech 101 Object Categories dataset [16]. Samples of the birds are in figure 5. We resize images to a common resolution of 300×297 and convert to a vector of monochrome bitmap values. To correct for global differences in illumination and ensure that only local patterns would be used for discrimination, we center and scale each vector. Power comparisons can be seen in figure 6.

Again, MMD performs better. However, it appears that the linear kernel performs significantly worse for the MMD than for the FS. This could reflect a difference in the function classes over which each technique operates.

7 Extensions

7.1 Heterogeneous Data

This procedure extends naturally to the heterogeneous data setting via multiple kernel learning (MKL) [17, 18]. Qiu et al. [19] develop MKL for support vector regression. Given j different data modalities, it suffices to match a kernel K_i to each—or perhaps more than one kernel for each data source, so as to better target specific features. The semidefinite programming approach (SDP) to MKL finds the best linear combination $K = \sum_{i=1}^{j} \mu_i K_i$ for some relevant objective function. For computational reasons, the best non-negative linear combination is frequently sought, as this yields a simpler quadratically constrained quadratic program (QCQP).

7.2 Missing Data

If we further consider entire missing modalities (e.g. one sample is missing some biometric reading), Poh et al. [20] develop the *neutral point substitution* technique to allow substitution of the missing modality with a new kernel that is *unbiased* with regard to the classification at hand. This allows for full use of both modalities that are present for all samples as well as those that are present only for a subset of the samples and effective utilization of all the data in the training set. Panov et al. [21] modify the NPS method to allow for missing modalities in the test set.

7.3 Theoretical Guarantees

Having proved a bound in the univariate data, linear kernel case by constructing an exchangeable pair, Anonymous and Anonymous [7] use simulations to suggest that the same pair is likely to yield success in more general settings: the key *approximate regression condition* holds more universally for multivariate data, a non-linear kernel, and a combination of the two settings. Further simulations demonstrate that the $\mathcal{O}(N^{-1/4})$ rate of convergence does not appear to be tight and a more typical $\mathcal{O}(N^{-1/2})$ is within reach.

A rate of convergence result with known constant allows for a single calculation of the Friedman statistic—rather than the N_{perm} required for randomization-based inference. Theoretical inference could be done on the limiting distribution, with error characterized by the proven bound. This large savings in computation comes only at the known cost of the limiting distribution approximation, which falls rapidly in sample size.

8 Discussion

We have tested a two-sample testing method of Friedman's [1] with a particular choice of learning algorithm—support vector regression. This Friedman Test can be seen as a generalization of the celebrated permutation t-test, or randomization test. Without tuning, performance is competitive in some settings with the MMD test. Simulations suggest that more powerful tests may be achieved with the added complexity of tuning—at some computational cost. Further work is required to determine a good set of heuristic choices for the SVM tuning parameters.

Modern data sources often consist of different modalities. Wireless sensor networks (including cellular phones) are deployed to collect large quantities of *diverse* data. These networks may be heterogeneous, with newer and upgraded hardware logging novel sources of data. Because Friedman's idea leverages *any* learning algorithm, we can at present easily incorporate extensions such as both the treatment of heterogeneous data *and* an allowance for missing data modalities. Future developments in regression and classification can be incorporated to advance the state-of-the-art in two-sample testing.

References

[1] J. Friedman, "On Multivariate Goodness–of–Fit and Two–Sample Testing," *Proceedings of Phystat2003, http://www.slac.stanford.edu/econf/C*, vol. 30908, 2003.

[2] E. Lehmann and J. Romano, *Testing Statistical Hypotheses*. Springer Verlag, 2005.

[3] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[4] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab – an S4 package for kernel methods in R," *Journal of Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004.

[5] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. the MIT Press, 2002.

[6] E. Lehmann, *Elements of large-sample theory*. Springer Verlag, 1999.

[7] A. Anonymous and B. Anonymous, "A rate of convergence bound for the randomization t-distribution using stein's method," *Unpublished Technical Report*.

[8] L. Chen, L. Goldstein, and Q. Shao, *Normal Approximation by Stein's Method*. Springer Verlag, 2010.

[9] C. Stein, "Approximate computation of expectations," *Lecture Notes-Monograph Series*, vol. 7, 1986.

[10] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel method for the two-sample-problem," *Advances in Neural Information Processing Systems*, vol. 19, pp. 513–520, 2007.

[11] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. Sriperumbudur, "A fast, consistent kernel two-sample test," *Advances in Neural Information Processing Systems*, vol. 22, pp. 673–681, 2010.

[12] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, pp. 723–773, 2012.

[13] K. Borgwardt, A. Gretton, M. Rasch, H. Kriegel, B. Schölkopf, and A. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[14] J. Gentry, *twitteR: R based Twitter client*, 2011. R package version 0.99.6.

[15] C. Leslie, E. Eskin, and W. Noble, "The spectrum kernel: A string kernel for SVM protein classification," in *Proceedings of the Pacific Symposium on Biocomputing*, vol. 7, pp. 566–575, Hawaii, USA., 2002.

[16] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

[17] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan, "Learning the kernel matrix with semidefinite programming," *The Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

[18] M. Gönen and E. Alpaydın, "Multiple kernel learning algorithms," *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[19] S. Qiu and T. Lane, "Multiple kernel learning for support vector regression," *Computer Science Department, The University of New Mexico, Albuquerque, NM, USA, Tech. Rep*, 2005.

[20] N. Poh, D. Windridge, V. Mottl, A. Tatarchuk, and A. Eliseyev, "Addressing missing values in kernel-based multimodal biometric fusion using neutral point substitution," *Information Forensics and Security, IEEE Transactions on*, vol. 5, no. 3, pp. 461–469, 2010.

[21] M. Panov, A. Tatarchuk, V. Mottl, and D. Windridge, "A modified neutral point method for kernel-based fusion of pattern-recognition modalities with incomplete data sets," *Multiple Classifier Systems*, pp. 126–136, 2011.