Appendix for Learning Interpretable Characteristic Kernels via Decision Forests

Anonymous Author(s)
Affiliation
Address
email

A Proofs

Theorem 1. *The random forest induced kernel* $\mathbf{K}^{\mathbf{x}}$ *is always positive definite.*

Proof. The forest-induced kernel is a summation of permuted block diagonal matrix, with ones in each block and zeros elsewhere [5], i.e.,

$$\mathbf{K}^{\mathbf{x}} = \frac{1}{m}\sum_{w=1}^{m} Q_w B_w Q_w^T,$$

where Q is a permutation matrix, and B is a block diagonal matrix with each block representing a leaf node, and the sum is over all m trees in the forest. For example, when each leaf node only contains one observation, B becomes the identity matrix.

Each block matrix is always positive definite and still positive definite after permutation, because permutation does not change eigenvalues. As summation of positive definite matrices is still positive definite, $\mathbf{K}^{\mathbf{X}}$ is always positive definite. □

Next we show the kernel can be characteristic, when the tree partition area converges to zero. A similar property is also used for proving classification consistency in k-nearest-neighbor [6], and we shall denote $N(\phi_w) \in \mathbb{R}_{\geq 0}^{L_w}$ as the maximum area of each part.

Theorem 2. *Suppose as* $n, m \to \infty$, $N(\phi_w) \to 0$ *for each tree* $\phi_w \in \mathbf{P}$ *and each observation* x_i*. Then the random forest induced kernel* $\mathbf{K}^{\mathbf{x}}$ *is asymptotically characteristic.*

Proof. since the kernel is positive semidefinite, it suffices to prove

$$E[k(\cdot, X_1)] = E[k(\cdot, X_2)] \text{ if and only if } F_{X_1} = F_{X_2}. \tag{1}$$

The forward implication is trivial. To prove the converse, it suffices to investigate when $E[\mathbf{K}^{\mathbf{x}}(\cdot, X_1)] = E[\mathbf{K}^{\mathbf{x}}(\cdot, X_2)]$, or equivalently

$$E_{X_1}\left(\frac{1}{m}\sum_{w=1}^{m}[\mathbf{I}(\phi_w(X_1) = \phi_w(z))]\right) = E_{X_2}\left(\frac{1}{m}\sum_{w=1}^{m}[\mathbf{I}(\phi_w(X_2) = \phi_w(z))]\right)$$

for any observation z.

We first show the above equality occurs if and only if $\phi_w(X_1) \overset{D}{=} \phi_w(X_2)$. Once again, the forward implication is trivial. The converse can be shown by contradiction: without loss of generality, suppose there exists a leaf node region U such that $\phi_w(X_1) \in U$ with probability p_1 while $\phi_w(X_2) \in U$ with probability p_2. Then for any point-mass observation z always in U, $E[\mathbf{K}^{\mathbf{x}}(z, X_1)] = p_1 \neq E[\mathbf{K}^{\mathbf{x}}(z, X_2)] = p_2$, which is a contradiction.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Next we show $\phi_w(X_1) \overset{D}{=} \phi_w(X_2)$ if and only if $F_{X_1} = F_{X_2}$. The forward implication is again trivial. The converse is shown by contradiction. Suppose $F_{X_1} \neq F_{X_2}$. Without loss of generality, there always exists a neighborhood $N(x)$ such that $Prob(X_1 \in N(x)) = p_3 \neq Prob(X_2 \in N(x)) = p_4$. Now, because each tree partition area converges to 0, we can always make $N(x)$ small enough so that $\phi_w(N(x)) = U$ almost surely for some leaf node region U. Then $Prob(\phi_w(X_1) = U) = p_3 \neq Prob(\phi_w(X_2) = R) = p_4$. Thus $\phi_w(X_1) \overset{D}{\neq} \phi_w(X_2)$, contradiction. Therefore, Equation 1 is proved, and the kernel is characteristic. □

Corollary 2.1. *KMERF satisfies*

$$\lim_{n \to \infty} c_k^n(\mathbf{x}, \mathbf{y}) = c \geq 0,$$

with equality to 0 *if and only if* $F_{XY} = F_X F_Y$. *Moreover, for sufficiently large* n *and sufficiently small type 1 error level* α, *this method is valid and consistent for independence and k-sample testing.*

Proof. As $n \to \infty$, $\mathbf{K}$ is asymptotically a characteristic kernel by Theorem 2. By Shen and Vogelstein [9], the Euclidean distance induced kernel is also characteristic. Therefore by Gretton et al. [7], Lyons [8], $c_k^n(\mathbf{x}, \mathbf{y})$ is asymptotically 0 if and only if independence.

By Shen et al. [10], for sufficiently large n, the chi-square distribution $\frac{\chi_1^2 - 1}{n}$ dominates the true null distribution of the unbiased correlation in upper tail. Therefore, when X and Y are actually independent, the testing power is no more than the type 1 error level α, making it a valid test. When X and Y are dependent, the distribution $\frac{\chi_1^2 - 1}{n}$ converges to 0 in probability, such that the p-value converges to 0 and the testing power converges to 1, making it a consistent test. □

B Limitations

There are a few limitations to this approach. In the problem setting that we are considering (composite null vs. composite alternative), there is no uniformly most powerful test [3]. So, while this paper presents its argument with simulated data, it is not yet known how this statistical method will perform against other statistics with real data. This is difficult to determine as distributions of data are oftentimes unknown and so may not fall cleanly in one of the 20 distributions that were tested. Given the performance of KMERF, it is likely safer to use KMERF over others as it appears to perform better than alternatives in most cases.

In addition, we have currently have not explored the performance of our algorithm with respect to other decision forests types [4, 1, 11], and hyper-parameter tuning. It would be interesting the extend this approach using these decision forests to answer additional hypothesis testing problems, such as paired k-sample testing, etc.

C Simulations

C.1 Independence Simulations

For the independence simulation, we test independence between X and Y. For the random variable $X \in \mathbb{R}^p$, we denote $X_{|d|}, d = 1, \ldots, p$ as the d^{th} dimension of X. $w \in \mathbb{R}^p$ is a decaying vector with $w_{|d|} = 1/d$ for each d, such that $w^\mathsf{T} X$ is a weighted summation of all dimensions of X. Furthermore, $\mathcal{U}(a, b)$ denotes the uniform distribution on the interval (a, b), $\mathcal{B}(p)$ denotes the Bernoulli distribution with probability p, $\mathcal{N}(\mu, \Sigma)$ denotes the normal distribution with mean μ and covariance Σ, U and V represent some auxiliary random variables, κ is a scalar constant to control the noise level, and ϵ is sampled from an independent standard normal distribution unless mentioned otherwise.

1. `Linear`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$:

$$X \sim \mathcal{U}(-1, 1)^p,$$

$$Y = w^\mathsf{T} X + \kappa\epsilon.$$

2. `Exponential`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$:

$$X \sim \mathcal{U}\left(0, 3\right)^p,$$

$$Y = \exp\left(w^\mathsf{T} X\right) + 10\kappa\epsilon.$$

3. `Cubic`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$:

$$X \sim \mathcal{U}\left(-1, 1\right)^p,$$

$$\begin{aligned} Y =& 128\left(w^\mathsf{T} X - \frac{1}{3}\right)^3 + 48\left(w^\mathsf{T} X - \frac{1}{3}\right)^2 \\ &- 12\left(w^\mathsf{T} X - \frac{1}{3}\right) + 80\kappa\epsilon. \end{aligned}$$

4. `Joint Normal`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: Let $\rho = 1/2p$, I_p be the identity matrix of size $p \times p$, J_p be the matrix of ones of size $p \times p$, and $\Sigma = \begin{bmatrix} I_p & \rho J_p \\ \rho J_p & \left(1 + 0.5\kappa\right) I_p \end{bmatrix}$. Then,

$$\left(X, Y\right) \sim \mathcal{N}\left(0, \Sigma\right).$$

5. `Step Function`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$:

$$X \sim \mathcal{U}\left(-1, 1\right)^p,$$

$$Y = \mathbb{I}\left(w^\mathsf{T} X > 0\right) + \epsilon,$$

where $\mathbb{I}$ is the indicator function; that is, $\mathbb{I}\left(z\right)$ is unity whenever z is true, and 0 otherwise.

6. `Quadratic`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$:

$$X \sim \mathcal{U}\left(-1, 1\right)^p,$$

$$Y = \left(w^\mathsf{T} X\right)^2 + 0.5\kappa\epsilon.$$

7. `W-Shape`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: For $U \sim \mathcal{U}\left(-1, 1\right)^p$,

$$X \sim \mathcal{U}\left(-1, 1\right)^p,$$

$$Y = 4\left[\left(\left(w^\mathsf{T} X\right)^2 - \frac{1}{2}\right)^2 + \frac{w^\mathsf{T} U}{500}\right] + 0.5\kappa\epsilon.$$

8. `Spiral`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: For $U \sim \mathcal{U}\left(0, 5\right)$, $\epsilon \sim \mathcal{N}\left(0, 1\right)$,

$$X_{|d|} = U \sin\left(\pi U\right) \cos^d\left(\pi U\right) \text{ for } d = 1, ..., p-1,$$

$$X_{|p|} = U \cos^p\left(\pi U\right),$$

$$Y = U \sin\left(\pi U\right) + 0.4p\epsilon.$$

9. `Uncorrelated Bernoulli`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: For $U \sim \mathcal{B}\left(0.5\right)$, $\epsilon_1 \sim \mathcal{N}\left(0, I_p\right)$, $\epsilon_2 \sim \mathcal{N}\left(0, 1\right)$,

$$X \sim \mathcal{B}\left(0.5\right)^p + 0.5\epsilon_1,$$

$$Y = \left(2U - 1\right) w^\mathsf{T} X + 0.5\epsilon_2.$$

10. `Logarithmic`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: For $\epsilon \sim \mathcal{N}\left(0, I_p\right)$,

$$X \sim \mathcal{N}\left(0, I_p\right),$$

$$Y_{|d|} = 2\log_2\left(\left|X_{|d|}\right|\right) + 3\kappa\epsilon_{|d|} \text{ for } d = 1, ..., p.$$

11. `Fourth Root`$(X, Y) \in \mathbb{R}^p \times \mathbb{R}$:

$$X \sim \mathcal{U}\left(-1, 1\right)^p,$$

$$Y = \left|w^\mathsf{T} X\right|^{1/4} + \frac{\kappa}{4}\epsilon.$$

12. $\texttt{Sine Period } 4\pi (X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: For $U \sim \mathcal{U}\left(-1, 1\right)$, $V \sim \mathcal{N}\left(0, 1\right)^p$, $\theta = 4\pi$,
$$\begin{aligned} X_{|d|} &= U + 0.02 p V_{|d|} \text{ for } d = 1, ..., p, \\ Y &= \sin(\theta X) + \kappa \epsilon. \end{aligned}$$
13. $\texttt{Sine Period } 16\pi (X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: Same as above except $\theta = 16\pi$ and the noise on Y is changed to $0.5 \kappa \epsilon$.
14. $\texttt{Square}(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: For $U \sim \mathcal{U}\left(-1, 1\right)$, $V \sim \mathcal{U}\left(-1, 1\right)$, $\epsilon \sim \mathcal{N}\left(0, 1\right)^p$, $\theta = -\frac{\pi}{8}$,
$$\begin{aligned} X_{|d|} &= U \cos\left(\theta\right) + V \sin\left(\theta\right) + 0.05 p \epsilon_{|d|}, \\ Y_{|d|} &= -U \sin\left(\theta\right) + V \cos\left(\theta\right). \end{aligned}$$
15. $\texttt{Diamond}(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: Same as above except $\theta = \pi / 4$.
16. $\texttt{Two Parabolas}(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: For $\epsilon \sim \mathcal{U}\left(0, 1\right)$, $U \sim \mathcal{B}\left(0.5\right)$,
$$\begin{aligned} X &\sim \mathcal{U}\left(-1, 1\right)^p, \\ Y &= \left(\left(w^{\mathsf{T}} X\right)^2 + 2 \kappa \epsilon\right) \cdot \left(U - \frac{1}{2}\right). \end{aligned}$$
17. $\texttt{Circle}(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: For $U \sim \mathcal{U}\left(-1, 1\right)^p$, $\epsilon \sim \mathcal{N}\left(0, I_p\right)$, $r = 1$,
$$\begin{aligned} X_{|d|} &= r \left(\sin\left(\pi U_{|d+1|}\right) \prod_{j=1}^{d} \cos\left(\pi U_{|j|}\right) \right) \\ &\text{for } d = 1, ..., p-1, \\ X_{|p|} &= r \left(\prod_{j=1}^{p} \cos\left(\pi U_{|j|}\right) \right), \\ Y &= \sin\left(\pi U_{|1|}\right) + 0.4 \epsilon_{|p|}. \end{aligned}$$
18. $\texttt{Ellipse}(X, Y) \in \mathbb{R}^p \times \mathbb{R}^p$: Same as above except $r = 5$.
19. $\texttt{Multiplicative Noise}(x, y) \in \mathbb{R}^p \times \mathbb{R}^p$: $u \sim \mathcal{N}\left(0, I_p\right)$,
$$\begin{aligned} x &\sim \mathcal{N}\left(0, I_p\right), \\ y_{|d|} &= u_{|d|} x_{|d|} \text{ for } d = 1, ..., p. \end{aligned}$$
20. $\texttt{Multimodal Independence}(X, Y) \in \mathbb{R}^p \times \mathbb{R}$: For $U \sim \mathcal{N}\left(0, I_p\right)$, $V \sim \mathcal{N}\left(0, I_p\right)$, $U' \sim \mathcal{B}\left(0.5\right)^p$, $V' \sim \mathcal{B}\left(0.5\right)^p$,
$$\begin{aligned} X &= U/3 + 2U' - 1, \\ Y &= V/3 + 2V' - 1. \end{aligned}$$

Figure E1 visualizes these equations. The light grey points in the figure are each simulation with noise added and the dark grey points are each simulation without noise added. Note that the last two simulations don't have any noise parameters.

C.2 Two-Sample Simulations

We do two-sample testing between Z and Z', generated as follows: let $Z = [X|Y]$ be the respective random variables from the independence simulation setup. Then define Q_θ as a rotation matrix for a given angle θ, i.e.,
$$Q_\theta = \begin{bmatrix} \cos\theta & 0 & \dots & -\sin\theta \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \sin\theta & 0 & \dots & \cos\theta \end{bmatrix}$$
Then we let
$$Z' = Q_\theta Z^{\mathsf{T}}$$
be the rotated versions of Z.

Figure E2 visualizes the above simulations. The simulations light grey points is a simulated data set and the dark grey points are the same dataset rotated by 10 degrees counter-clockwise. Simulations were plotted with min-max normalization for visualization purposes.

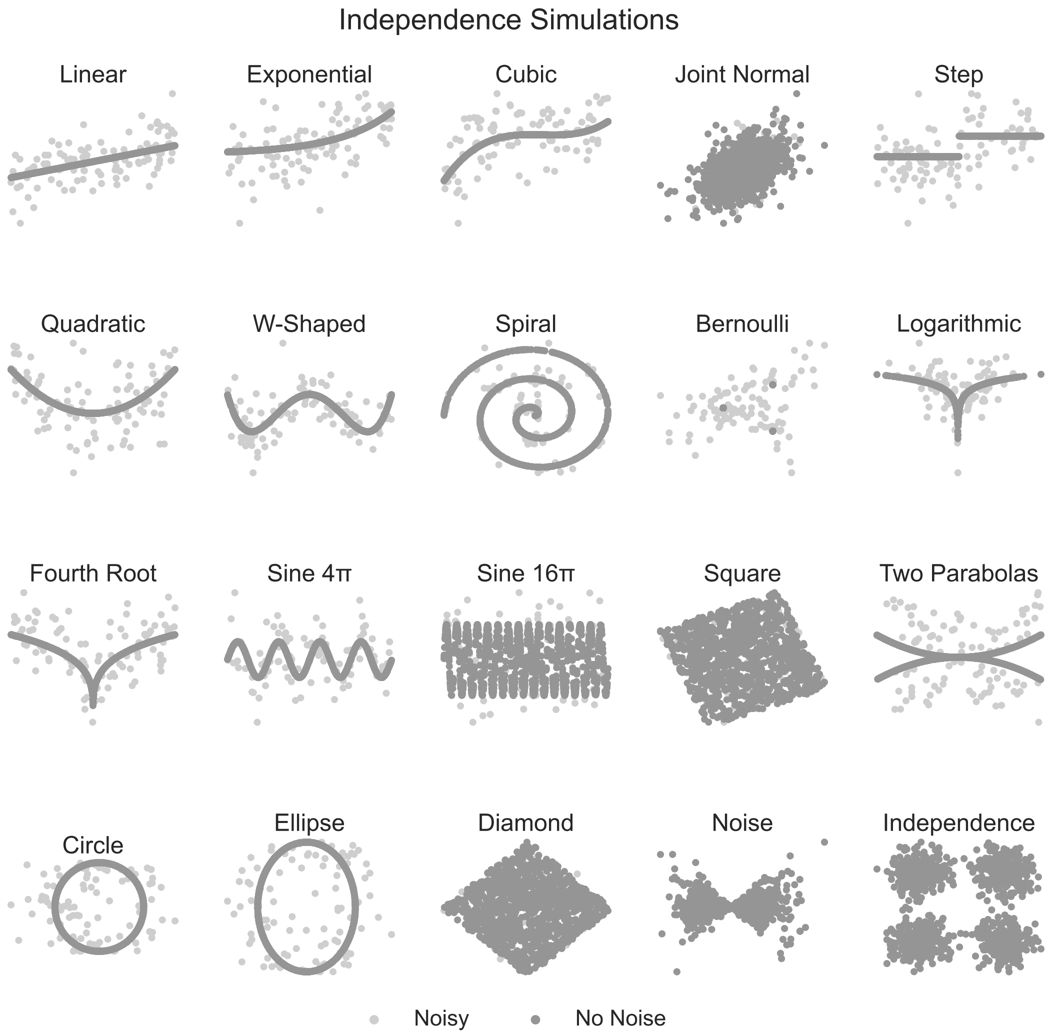

Figure E1: Simulations used for Figures 1 and 3. 100 points from noisy simulations (light grey points) on 1000 points from simulations without noise (dark grey points) for each of the 20 dimensional simulations shown above.

D Real Data

Previous studies have shown the utility of selection reaction monitoring when measuring protein and peptide abundance [12], and one was used to identify 318 peptides from 33 normal, 10 pancreatic cancer, 28 colorectal cancer, and 24 ovarian cancer samples [13]. For all tests, we created a binary label vector, where 1 indicated presence of pancreatic cancer in the patients and 0 otherwise. We then evaluated at a type 1 error level of $\alpha = 0.05$, and used the Benjamini-Hochberg procedure to control the false discovery rate [2] for our 318 p-values. All data used in this experiment are provided in the supplmental.

E Hardware and Software Configurations

- Operating System: Linux (ubuntu 20.04)
- VM Size: Azure Standard D96as v4 (96 vcpus, 384 GiB memory), Azure Standard B20ms (20 vcpus, 80 GiB memory)
- Software: Python 3.8, hyppo v0.4.0

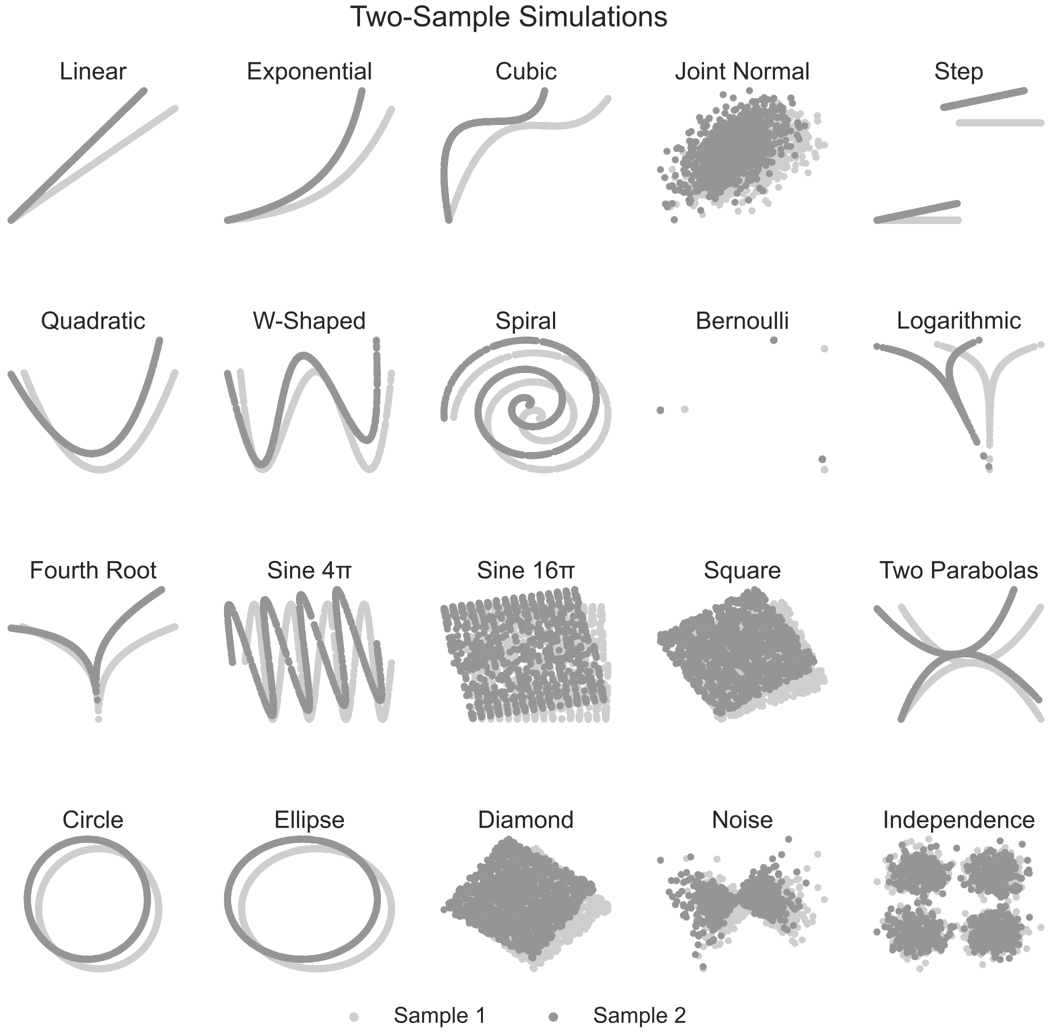

Figure E2: Simulations used for Figure 2. The first dataset (black dots) is 1000 samples from each of the 20 two-dimensional, no-noise simulation settings. The second dataset is the first dataset rotated by 10 degrees counter-clockwise. Simulations were normalized using min-max normalization for visualization purposes.

References

[1] S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *Annals of Statistics*, 47(2): 1148–1178, 2018.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[3] P. J. Bickel and K. A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics, Vol I.* Prentice Hall, 2nd edition, 2006.

[4] R. Blaser and P. Fryzlewicz. Random rotation ensembles. *Journal of Machine Learning Research*, 17(4):1–26, 2016.

[5] A. Davies and Z. Ghahramani. The random forest kernel and creating other kernels for big data from random partitions. *arXiv:1402.4293v1*, 2014.

[6] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[7] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Scholkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6:2075–2129, 2005.

[8] R. Lyons. Distance covariance in metric spaces. *Annals of Probability*, 41(5):3284–3305, 2013.

[9] C. Shen and J. T. Vogelstein. The exact equivalence of distance and kernel methods in hypothesis testing. *AStA Advances in Statistical Analysis*, 105(3):385–403, 2021.

[10] C. Shen, S. Panda, and J. T. Vogelstein. The chi-square test of distance correlation. *Journal of Computational and Graphical Statistics*, 31(1):254–262, 2022.

[11] T. Tomita, J. Browne, C. Shen, J. Chung, J. Patsolic, B. Falk, J. Yim, C. E. Priebe, R. Burns, M. Maggioni, and J. T. Vogelstein. Sparse projection oblique randomer forests. *Journal of Machine Learning Research*, 21(104):1–39, 2020.

[12] Q. Wang, R. Chaerkady, J. Wu, H. J. Hwang, N. Papadopoulos, L. Kopelovich, A. Maitra, H. Matthaei, J. R. Eshleman, R. H. Hruban, et al. Mutant proteins as cancer-specific biomarkers. *Proceedings of the National Academy of Sciences*, 108(6):2444–2449, 2011.

[13] Q. Wang, M. Zhang, T. Tomita, J. T. Vogelstein, S. Zhou, N. Papadopoulos, K. W. Kinzler, and B. Vogelstein. Selected reaction monitoring approach for validating peptide biomarkers. *Proceedings of the National Academy of Sciences*, 114(51):13519–13524, 2017.