# Learning Interpretable Characteristic Kernels via Decision Forests

Sambit Panda

Neurodata Lab - Dr. Joshua Vogelstein

# Independence Testing Problem

- Testing whether there is dependence between random variables
- Data are often very high dimensional and highly nonlinear, making testing difficult

# Independence Testing Problem

- Testing whether there is dependence between random variables
- Data are often very high dimensional and highly nonlinear, making testing difficult

| $X$ | $Y$ |
|---|---|
| Brain Shape | Health |
| Brain Connectvity | Mental State |
| Gene Expression | Cancer Stage |

# General Independence Testing

Suppose we have $n$ samples of $(x_i, y_i) \overset{iid}{\sim} F_{XY}$, i.e., $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}^q$. $X$ and $Y$ have distributions $F_X$ and $F_Y$ and joint distribution $F_{XY}$. We are testing:

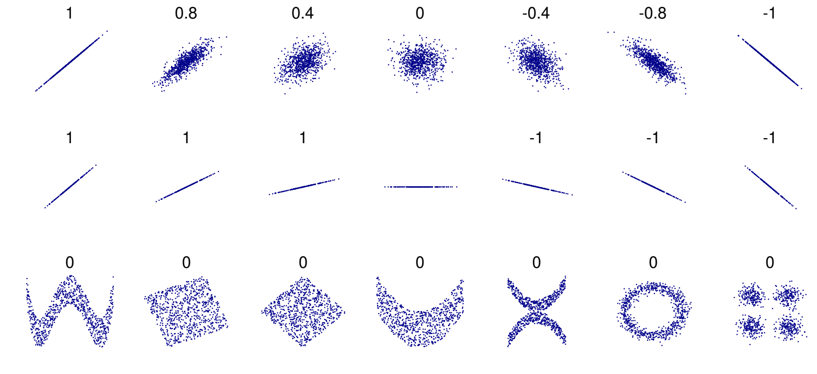
$$H_0 : F_{XY} = F_X F_Y,$$
$$H_A : F_{XY} \neq F_X F_Y.$$

# Desired Testing Properties

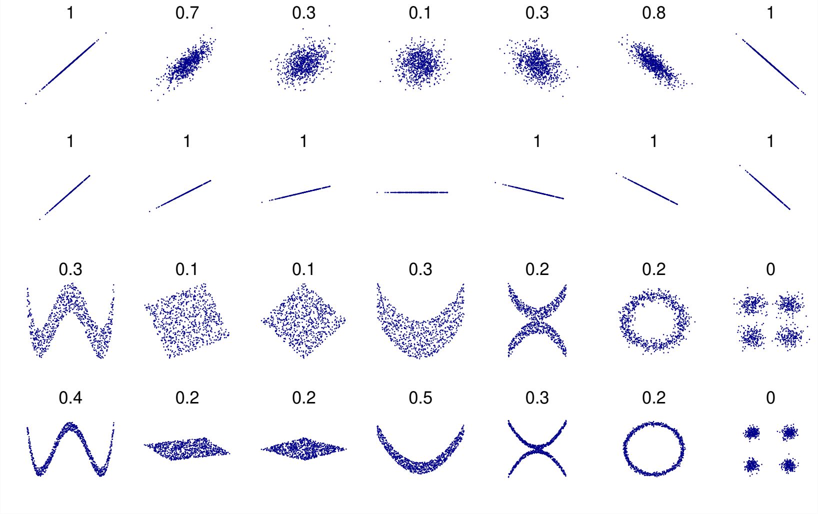
- Universally consistent for any distribution with finite second moments
- Valid
- Strong empirical performance on a range of linear and nonlinear relationships in finite sample

# Intuition

# Pearson’s Correlation Can Only Detect Linear Relationships



# Distance Correlation (Dcorr) Picks Up Both Linear and Nonlinear



# Distance Correlation (Dcorr) [Szekely and Rizzo, 2014]

1. Compute pairwise distance matrices $D^x$, $D^y$

# Distance Correlation (Dcorr) [Szekely and Rizzo, 2014]

1. Compute pairwise distance matrices $D^x$, $D^y$
2. Center (biased) or doubly center (unbiased) $D^x$ and $D^y$
3. Compute distance covariance statistic
4. Normalize to get distance correlation statistic $\mathrm{Dcorr}_n(x, y)$
5. Compute p-values via a permutation test or chi-square approximation [Shen et al., 2022]

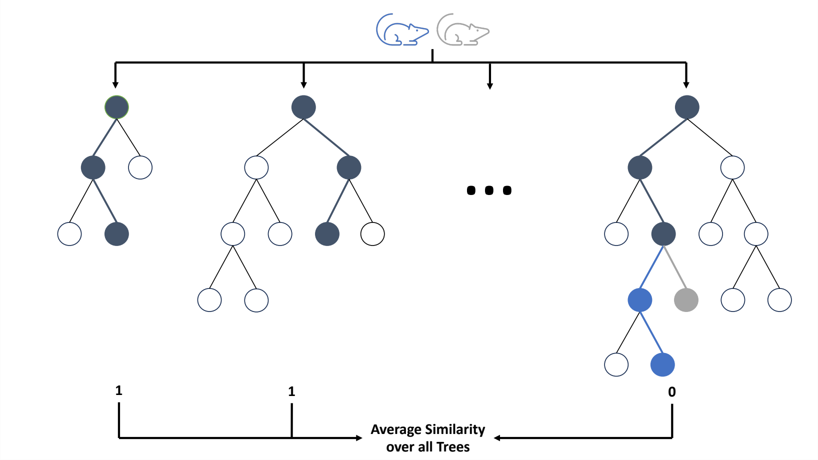
# Distance Correlation (Dcorr) [Szekely and Rizzo, 2014]

1. Compute pairwise distance matrices $D^x$, $D^y$
2. Center (biased) or doubly center (unbiased) $D^x$ and $D^y$
3. Compute distance covariance statistic
4. Normalize to get distance correlation statistic $\mathrm{Dcorr}_n(x, y)$
5. Compute p-values via a permutation test or chi-square approximation [Shen et al., 2022]

# Random Forest Proximity Kernel [Breiman, 2002]

- Random forest is an ensemble of decision trees
- Induces a proximity kernel which is how often that two observations lie in the same leaf node across all trees.

# Random Forest Proximity Kernel [Breiman, 2002]



# Why Care About This Kernel?

- We can prove the random forest proximity kernel is characteristic, which allows Dcorr to be **universally consistent** for any distribution with finite second moments

# Why Care About This Kernel?

- We can prove the random forest proximity kernel is characteristic, which allows Dcorr to be **universally consistent** for any distribution with finite second moments
- Dcorr may have lower power when **sample size is low** and when data has strong nonlinear dependencies, excessive noise, or high-dimensional [Ramdas et al., 2015]

# Why Care About This Kernel?

- We can prove the random forest proximity kernel is characteristic, which allows Dcorr to be **universally consistent** for any distribution with finite second moments
- Dcorr may have lower power when **sample size is low** and when data has strong nonlinear dependencies, excessive noise, or high-dimensional [Ramdas et al., 2015]
- Literature has shown that better power can be achieved with data-adaptive kernels [Gretton et al., 2012]

# Kernel Mean Embedding Random Forest (KMERF)

1. Compute the random forest proximity kernel for $x$, $K^{\Phi(x)}$
2. Transform similarities to distances for $x$ [Shen and Vogelstein, 2021]

$$D^x = 1 - \frac{K^{\Phi(x)}}{\max(K^{\Phi(x)})}$$

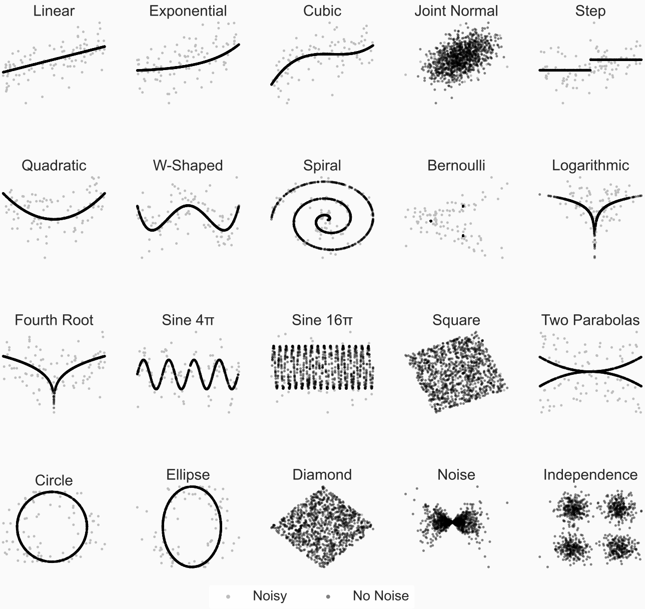
3. Compute pairwise distances for $y$ using a distance metric, $D^y$
4. Compute Dcorr test statistic and p-value

# Desired Testing Properties

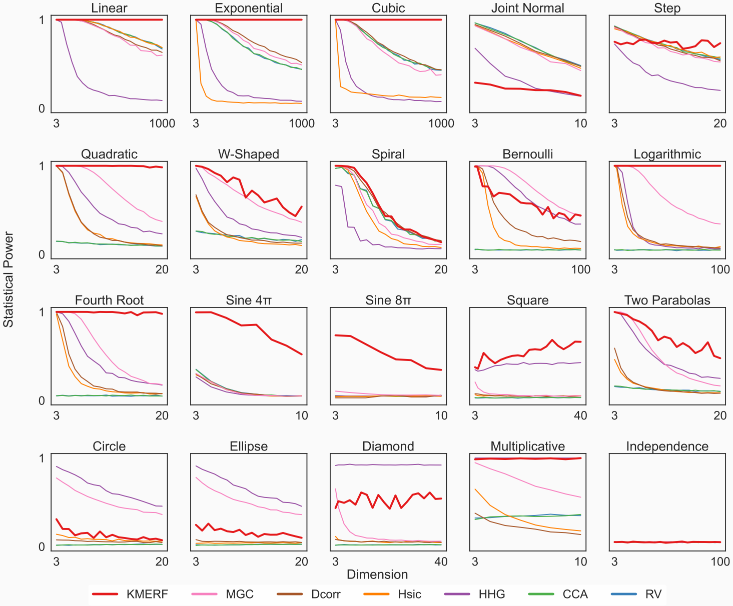
- Universally consistent for any distribution with finite second moments
- Valid
- Strong empirical performance on a range of linear and nonlinear relationships in finite sample

# Simulations

# 20 Independence Testing Simulation Settings (1D)



# HD Independence Testing Power (KMERF Nearly Dominates)

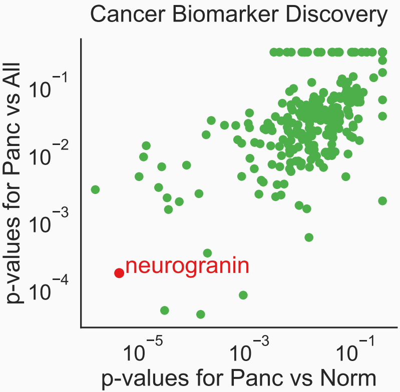


# Real Data

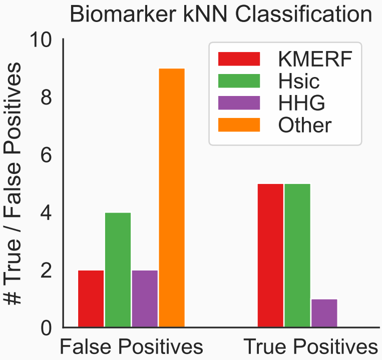
# Cancer Biomarker Discovery

1. 318 peptides were identified from 33 normal, 10 pancreatic cancer, 28 colorectal cancer, and 24 ovarian cancer samples [Wang et al., 2017].
2. Created a binary label vector, where 1 indicated the presence of pancreatic cancer in the patients, and 0 indicated its absence
3. Applied the Benjamini-Hochberg procedure [Benjamini and Hochberg, 1995] to control the false discovery rate

# KMERF Identifies a Unique Biomarker for Pancreatic Cancer



Cancer Biomarker Discovery



Biomarker kNN Classification

# Conclusion

- KMERF is universally consistent for distributions with finite second moments due to the kernel being characteristic
- Empirically demonstrated KMERF is valid
- Demonstrated strong empirical performance for KMERF on a range of multivariate linear and nonlinear relationships

# Acknowledgements

Paper: [Panda et al., 2023]

## NeuroData

- Joshua Vogelstein
- Cencheng Shen
- Jayanta Dey
- Tommy Athey
- Alice Wang

## Faculty

- Eliza O'Reilly
- Minh Tang
- Shangsi Wang

## Funding Agencies

- NIH
- Child Mind Institute
- NSF
- NSSEFF
- DARPA

Questions?

# Characteristic Kernel

**Definition ([Fukumizu et al., 2007])**

Let $(\mathcal{X}, \mathcal{B})$, $X$ is a random variable on $\mathcal{X}$ and $(\mathcal{H}, k)$ is a RKHS on $\mathcal{X}$. The mean element of $X$ in $\mathcal{H}$ is a unique element $m_X \in \mathcal{H}$ such that $\langle m_X, f \rangle_{\mathcal{H}} = E[f(X)]$ for all $f \in \mathcal{H}$. If the distribution of $X$ is $F_X$, and $\mathcal{P}$ is the family of all probabilities on $\mathcal{X}, \mathcal{B}$, we define a map $\mathcal{M}_k$ by

$$\mathcal{M}_k : \mathcal{P} \to \mathcal{H}, \quad F_X \mapsto m_X.$$

The kernel $k$ is characteristic if the map $\mathcal{M}_k$ is injective, or equivalently, if $E_{X \sim F_{X_1}}[f(X)] = E_{X \sim F_{X_2}}[f(X)]$ for all $f \in \mathcal{H}$ implies that $F_{X_1} = F_{X_2}$ and vice versa.

# Two Sample Testing Problem

We are testing differences in distributions between groups (*i.e.* control vs. cancer).

# Two Sample Testing Problem

We are testing differences in distributions between groups (*i.e.* control vs. cancer).

Let $u_i \in \mathbb{R}^p$ be the realization of random variable $U$ with distribution $F_U$ for $i = 1, \ldots, n_u$. Let $v_j \in \mathbb{R}^p$ be the realization of random variable $V$ with distribution $F_V$ for $i = 1, \ldots, n_v$. Then,

$$H_0 : F_U = F_V,$$
$$H_A : F_U \neq F_V.$$

# Two Sample Testing Problem

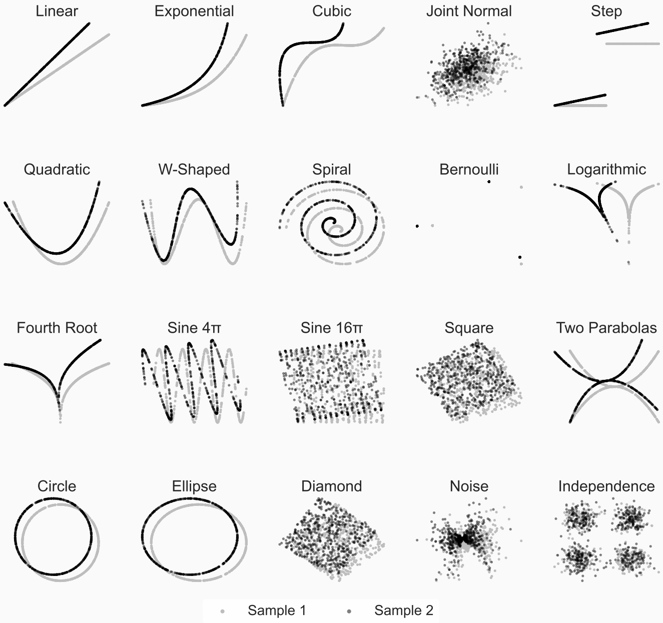
We are testing differences in distributions between groups (*i.e.* control vs. cancer).

Let $u_i \in \mathbb{R}^p$ be the realization of random variable $U$ with distribution $F_U$ for $i = 1, \ldots, n_u$. Let $v_j \in \mathbb{R}^p$ be the realization of random variable $V$ with distribution $F_V$ for $i = 1, \ldots, n_v$. Then,

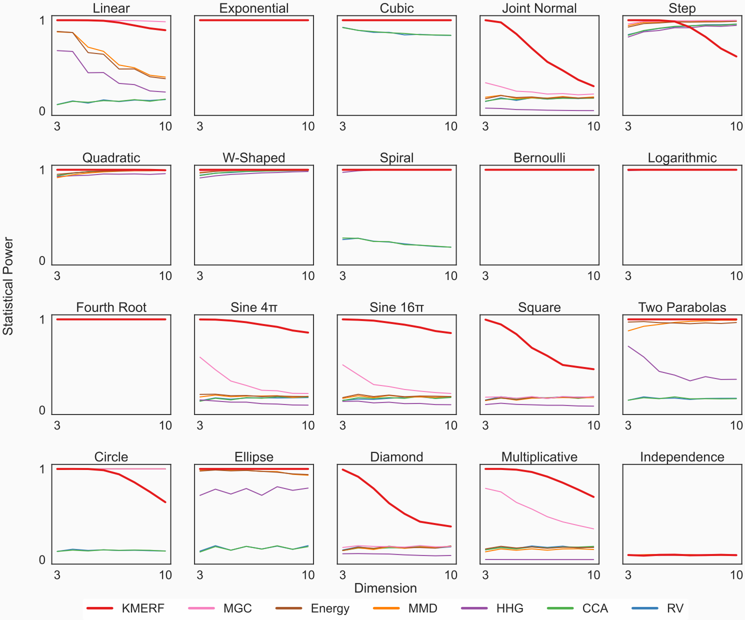
$$H_0 : F_U = F_V,$$
$$H_A : F_U \neq F_V.$$

This can be easily extended to $k$ samples. This problem can be reduced to the independence testing problem [Panda et al., 2021].

# 20 Two-Sample Simulation Settings (1D)



# HD Two-Sample Testing Power (KMERF Nearly Dominates)



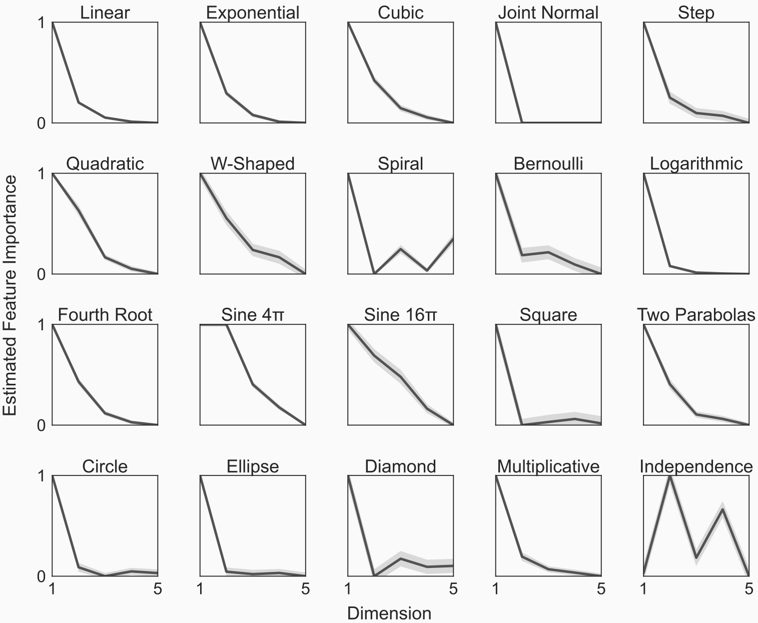
# Simulation Settings

1. Compared KMERF at 500 trees to other multivariate independence tests (MGC, Dcorr, Hsic, HHG, CCA, and RV)
2. $n = 100$ samples for $x$ and $y$ are sampled from each simulation, p-values were computed, and repeat 1000 times
3. Empirical power was estimated at $\alpha = 0.05$
4. Dimension for each simulation was varied and the process was repeated and repeat

# Empirical Feature Importances

Gini Importance calculates each feature importance as the sum over the number of splits (across all tress) that include the feature, proportional to the number of samples it splits.

# 5D Sims Estimated Feature Importance vs. Dimension



# References i

Benjamini, Y. and Hochberg, Y. (1995).
**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.

Breiman, L. (2002).
**Some infinity theory for predictor ensembles.**
*Journal of Combinatorial Theory, Series A*, 98:175–191.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007).
**Kernel measures of conditional dependence.**
In *Advances in neural information processing systems*.

# References ii

- Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., M. Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012).
  **Optimal kernel choice for large-scale two-sample tests.**
  In *Advances in neural information processing systems 25*, pages 1205–1213.

- Panda, S., Shen, C., Perry, R., Zorn, J., Lutz, A., Priebe, C. E., and Vogelstein, J. T. (2021).
  **Nonpar manova via independence testing.**

- Panda, S., Shen, C., and Vogelstein, J. T. (2023).
  **Learning interpretable characteristic kernels via decision forests.**
  *arXiv preprint arXiv:1812.00029*.

# References iii

- Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. (2015).
  **On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions.**
  In *29th AAAI Conference on Artificial Intelligence*.

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

# References iv

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

Szekely, G. and Rizzo, M. (2014).
Partial distance correlation with methods for dissimilarities.
*Annals of Statistics*, 42(6):2382–2412.

# References v

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