Kernel Integrated R^2: A Measure of Dependence

Pouya Roudaki[1] Shakeel Gavioli-Akilagun[1,2] Florian Kalinke[3] Mona Azadkia[1] Zoltán Szabó[1]

[1]London School of Economics [2]City University of Hong Kong [3]Karlsruhe Institute of Technology

Quick Summary

We introduce a new measure of dependence $D(Y,X)$, kernel integrated R^2, s.t.

1. $D(Y,X) \in [0,1]$,
2. $D(Y,X) = 0$ iff $Y \perp X$ and
3. $D(Y,X) = 1$ iff $\exists f : \mathcal{X} \to \mathcal{Y}$ s.t. $Y = f(X)$ holds $\mathbb{P}_{XY}$-a.s.

We also propose two estimators of $D(Y,X)$, one based on nearest-neighbor graphs and the other based on reproducing kernel Hilbert space (RKHS) methods.

Existing: Integrated R^2

The kernel integrated R^2 that we propose extends the recently introduced measure of dependence, *integrated* R^2, denoted by ν [2].

- **Assumptions:** $\boxed{\textit{(Space)}\ (X,Y) \in \mathbb{R}^d \times \mathbb{R}}$.
- **Definition:**

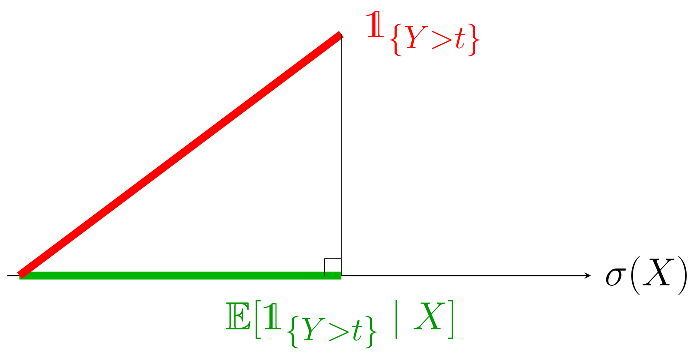
$$\nu(Y,X) = \int_{\mathbb{R}} \frac{\mathbb{V}_X\left(\mathbb{E}_{Y|X}[\mathbb{1}_{\{Y>t\}}]\right)}{\mathbb{V}_Y\left(\mathbb{1}_{\{Y>t\}}\right)} \mathrm{d}\mathbb{P}_Y(t) \overset{\text{total var}}{=} 1 - \int_{\mathbb{R}} \frac{\mathbb{E}_X\left[\mathbb{V}_{Y|X}\left(\mathbb{1}_{\{Y>t\}}\right)\right]}{\mathbb{V}_Y\left(\mathbb{1}_{\{Y>t\}}\right)} \mathrm{d}\mathbb{P}_Y(t).$$

- **Limitations:** Restricted to univariate responses Y and finite-dimensional Euclidean covariates X.

• Intuition

- Randomness from X and Y as a 2D plane; the horizontal axis is $\sigma(X)$: the sigma algebra of X.
- **Conditioning on** X **is a projection** onto axis $\sigma(X)$.
- Variance is the **length** (**L^2 norm**) of these vectors.
- So

$$\nu(Y,X) = \int_{\mathbb{R}} \frac{\mathbb{V}_X\left(\mathbb{E}_{Y|X}[\mathbb{1}_{\{Y>t\}}]\right)}{\mathbb{V}_Y\left(\mathbb{1}_{\{Y>t\}}\right)} \mathrm{d}\mathbb{P}_Y(t) = \text{average}\left(\frac{\text{projection length}}{\text{ambient-space vector length}}\right).$$



- if $Y = f(X) \Rightarrow \mathbb{1}_{\{Y>t\}} \in \sigma(X) \Rightarrow$ projection length = ambient-space vector length $\Rightarrow \nu(Y,X) = 1$.
- if $Y \perp X \Rightarrow \mathbb{1}_{\{Y>t\}} \perp \sigma(X) \Rightarrow$ projection length $= 0 \Rightarrow \nu(Y,X) = 0$.

• R^2 Interpretation

Coefficient of determination R_t^2 from the linear regression of $\mathbb{1}_{\{Y>t\}}$ on X:

$$R_t^2 = \frac{\mathbb{V}_X(\mathbb{E}_{Y|X}[\mathbb{1}_{\{Y>t\}}])}{\mathbb{V}_Y(\mathbb{1}_{\{Y>t\}})} = 1 - \frac{\mathbb{E}_X[\mathbb{V}_{Y|X}(\mathbb{1}_{\{Y>t\}})]}{\mathbb{V}_Y(\mathbb{1}_{\{Y>t\}})},$$

that is, $\nu(Y,X)$ is the integrated R_t^2 over the law of Y.

$$\nu(Y,X) = \int_{\mathbb{R}} R_t^2 \, \mathrm{d}\mathbb{P}_Y(t),$$

an average of explained-variance by coefficients over all thresholds t: the overall proportion of variation in Y explainable by X.

Proposed: Kernel Integrated R^2

- **Assumptions:**
  - $\boxed{\textit{(Space)}\ (X,Y) \in \mathcal{X} \times \mathcal{Y}, \text{ where } \mathcal{X} \text{ is a } \textit{topological space} \text{ and } \mathcal{Y} \text{ is a } \textit{Polish space}}$.
  - *(RKHS)* Let $k_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a characteristic kernel with associated RKHS $\mathcal{H}_{\mathcal{Y}}$.
- **Definition:**

$$D(Y,X) = 1 - \int_{\mathcal{Y}} \frac{\mathbb{E}_X\left[\mathbb{V}_{Y|X}\left(k_{\mathcal{Y}}(Y,y)\right)\right]}{\mathbb{V}_Y\left(k_{\mathcal{Y}}(Y,y)\right)} \mathrm{d}\mathbb{P}_Y(y).$$

Choosing $k_{\mathcal{Y}}(u,y) = \mathbb{1}_{\{u>y\}}$—which is not a kernel—in D formally gives back ν.

Comparison with Related Measures

- Chatterjee's correlation coefficient [3, 1]:

$$\xi(X,Y) = \int_{\mathbb{R}} \frac{\mathbb{V}_X\left(\mathbb{E}_{Y|X}[\mathbb{1}_{\{Y>t\}}]\right)}{\int_{\mathbb{R}} \mathbb{V}_Y\left(\mathbb{1}_{\{Y>u\}}\right) \mathrm{d}\mathbb{P}_Y(u)} \mathrm{d}\mathbb{P}_Y(t).$$

- Key difference between ξ and ν: Global vs. local normalization. **The mean of ratios preserves dependence signals better than the ratio of means**, especially when the variance of $\mathbb{1}_{\{Y>t\}}$ is small for some t.
- Measure of association on topological spaces [4] extends Chatterjee's correlation coefficient to topological spaces:

$$\eta_{k_{\mathcal{Y}}}(Y,X) = 1 - \frac{\mathbb{E}_X\left[\mathbb{E}_{Y,Y'|X}\left[\left\|k_{\mathcal{Y}}(\cdot,Y) - k_{\mathcal{Y}}(\cdot,Y')\right\|_{\mathcal{H}_{\mathcal{Y}}}^2\right]\right]}{\mathbb{E}_{Y_1,Y_2}\left[\left\|k_{\mathcal{Y}}(\cdot,Y_1) - k_{\mathcal{Y}}(\cdot,Y_2)\right\|_{\mathcal{H}_{\mathcal{Y}}}^2\right]},$$

where (i) $Y|X$, $Y'|X \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{Y|X}$, (ii) Y_1, $Y_2 \overset{\text{i.i.d.}}{\sim} \mathbb{P}_Y$, and (iii) $k_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a kernel on the Polish space $\mathcal{Y}$ with associated RKHS $\mathcal{H}_{\mathcal{Y}}$.

Property	ξ	ν	$\eta_{k_{\mathcal{Y}}}$	D
Value in $[0,1]$	✓	✓	✓	✓
0-independent	✓	✓	✓	✓
1-full dependence	✓	✓	✓	✓
Bijective invariance	✓	✓	×	×
Multivariate Y	×	×	✓	✓
Kernel-endowed domain $\mathcal{X}, \mathcal{Y}$	×	×	✓	✓
Local normalization	×	✓	×	✓

Estimation

- **Goal:** Estimate $D(Y,X)$ for $(X_i,Y_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{XY}$.
- Approximate outer integral by empirical mean,

$$\int_{\mathcal{Y}} \frac{\mathbb{E}_X\left[\mathbb{V}_{Y|X}\left(k_{\mathcal{Y}}(Y,y)\right)\right]}{\mathbb{V}_Y\left(k_{\mathcal{Y}}(Y,y)\right)} \mathrm{d}\mathbb{P}_Y(y) \approx \frac{1}{n}\sum_{i=1}^n \frac{\mathbb{E}_X\left[\mathbb{V}_{Y|X}\left(k_{\mathcal{Y}}(Y,Y_i)\right)\right]}{\mathbb{V}_Y\left(k_{\mathcal{Y}}(Y,Y_i)\right)},$$

  and estimate the **numerator** and **denominator** separately.
- Estimation of the denominator: *easy*.
- Estimation of $\mathbb{V}_{Y|X}\left(k_{\mathcal{Y}}(Y,Y_i)\right)$ in the numerator: *tricky*. Two approaches:

1. ***Graph-based estimator:***

$$\mathbb{V}_{Y|X}\left(k_{\mathcal{Y}}(Y,Y_i)\right) = \frac{1}{2}\mathbb{E}_{Y,Y'|X}\left[\left(k_{\mathcal{Y}}(Y,Y_i) - k_{\mathcal{Y}}(Y',Y_i)\right)^2\right],$$

   use the distribution of Y-values among the K nearest neighbors of X_j in $\mathcal{X}$ as a proxy for $Y \mid X = X_j$, i.e., $Y \mid X = X_{\mathcal{N}_j} \approx Y \mid X = X_j$, where $\mathcal{N}_j$ is the set of indices of the K nearest neighbors of X_j in $\mathcal{X}$.

2. ***RKHS method-based estimator:***

$$\mathbb{V}_{Y|X}\left(k_{\mathcal{Y}}(Y,Y_i)\right) = \mathbb{E}_{Y|X}\left[k_{\mathcal{Y}}^2(Y,Y_i)\right] - \mathbb{E}_{Y|X}\left[k_{\mathcal{Y}}(Y,Y_i)\right]^2,$$

   equip $\mathcal{X}$ with a characteristic kernel $k_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and use the notion of conditional mean embedding [5] and its ridge estimator to estimate $\mathbb{E}_{Y|X}\left[k_{\mathcal{Y}}(Y,Y_i)\right]$ and $\mathbb{E}_{Y|X}\left[k_{\mathcal{Y}}^2(Y,Y_i)\right]$.

Graph-based Estimator

- **Assumptions:** $\boxed{\mathcal{X} \text{ is a metric space}}$.
- **Notations:** $\mathcal{N}_j^{\setminus i}$ is the K nearest neighbors of X_j among $\{X_l\}_{l \neq i,j}$.
- Let

$$E_{n,i}^{\text{K-NN}} = \frac{1}{2K(n-1)} \sum_{j \neq i} \sum_{l \in \mathcal{N}_j^{\setminus i}} \left(k_{\mathcal{Y}}(Y_j,Y_i) - k_{\mathcal{Y}}(Y_l,Y_i)\right)^2,$$

$$V_{n,i}^{\text{K-NN}} = \frac{1}{n-1}\sum_{j \neq i} k_{\mathcal{Y}}^2(Y_j,Y_i) - \left[\frac{1}{n-1}\sum_{j \neq i} k_{\mathcal{Y}}(Y_j,Y_i)\right]^2,$$

$$\hat{D}^{\text{K-NN}}(Y,X) = 1 - \frac{1}{n}\sum_{i=1}^n \frac{E_{n,i}^{\text{K-NN}}}{V_{n,i}^{\text{K-NN}}}.$$

- **Runtime:** $\mathcal{O}(n^2 \log n)$.
- Convergence rate:

$$|\hat{D}^{\text{K-NN}}(Y,X) - D(Y,X)| = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} + \left(\frac{K}{n}\right)^{\beta/d} + \frac{(\log n)^{\beta/\alpha}}{n^2}\right),$$

  with $\alpha \approx$ rate of tail decay of X, $\beta \approx$ smoothness of cond. moments of $k_{\mathcal{Y}}(\cdot,Y)$, $d \approx$ intrinsic dim.
- Problem is *easier* when $d \downarrow$, $\beta \uparrow$, and *just a bit easier* when $\alpha \uparrow$.

RKHS Method-based Estimator

- **Assumptions:** $\boxed{\mathcal{X} \text{ is separable}}$; $k_{\mathcal{X}}$, $k_{\mathcal{Y}}^2$: characteristic, associated with RKHSs $\mathcal{H}_{\mathcal{X}}$, $\mathcal{H}_{\mathcal{Y}} \otimes \mathcal{H}_{\mathcal{Y}}$, resp.
- **Notations:** Let $\epsilon_n > 0$, j-th canonical basis vector: $\mathbf{e}_j \in \mathbb{R}^n$. Centering matrix: $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \in \mathbb{R}^{n \times n}$,

$$\mathbf{K}_X = [k_{\mathcal{X}}(X_i,X_j)]_{i,j=1}^n, \quad \tilde{\mathbf{K}}_X = \mathbf{H}\mathbf{K}_X\mathbf{H}, \quad \mathbf{K}_Y = [k_{\mathcal{Y}}(Y_i,Y_j)]_{i,j=1}^n, \quad \mathbf{M} = \mathbf{K}_Y\tilde{\mathbf{K}}_X(\tilde{\mathbf{K}}_X + n\epsilon_n\mathbf{I}_n)^{-1}.$$

- Let

$$E_{n,i}^{\text{RKHS}} = \frac{1}{n}\Big\langle \mathbf{e}_i, (\mathbf{K}_Y \circ \mathbf{K}_Y)\mathbf{1}_n + (\mathbf{K}_Y \circ \mathbf{K}_Y)\tilde{\mathbf{K}}_X(\tilde{\mathbf{K}}_X + n\epsilon_n\mathbf{I}_n)^{-1}\mathbf{1}_n - \frac{1}{n}\mathbf{K}_Y\mathbf{1}_n\mathbf{1}_n^\top\mathbf{K}_Y\mathbf{e}_i - \frac{2}{n}\mathbf{K}_Y\mathbf{1}_n\mathbf{1}_n^\top\mathbf{M}^\top\mathbf{e}_i - \mathbf{M}\mathbf{M}^\top\mathbf{e}_i\Big\rangle_{\mathbb{R}^n},$$

$$V_{n,i}^{\text{RKHS}} = \frac{1}{n}\mathbf{1}_n^\top(\mathbf{K}_Y \circ \mathbf{K}_Y)\mathbf{e}_i - \left(\frac{1}{n}\mathbf{1}_n^\top\mathbf{K}_Y\mathbf{e}_i\right)^2,$$

$$\hat{D}^{\text{RKHS}}(Y,X) = 1 - \frac{1}{n}\sum_{i=1}^n \frac{E_{n,i}^{\text{RKHS}}}{V_{n,i}^{\text{RKHS}}}.$$

- **Runtime:** $\mathcal{O}(n^3)$.

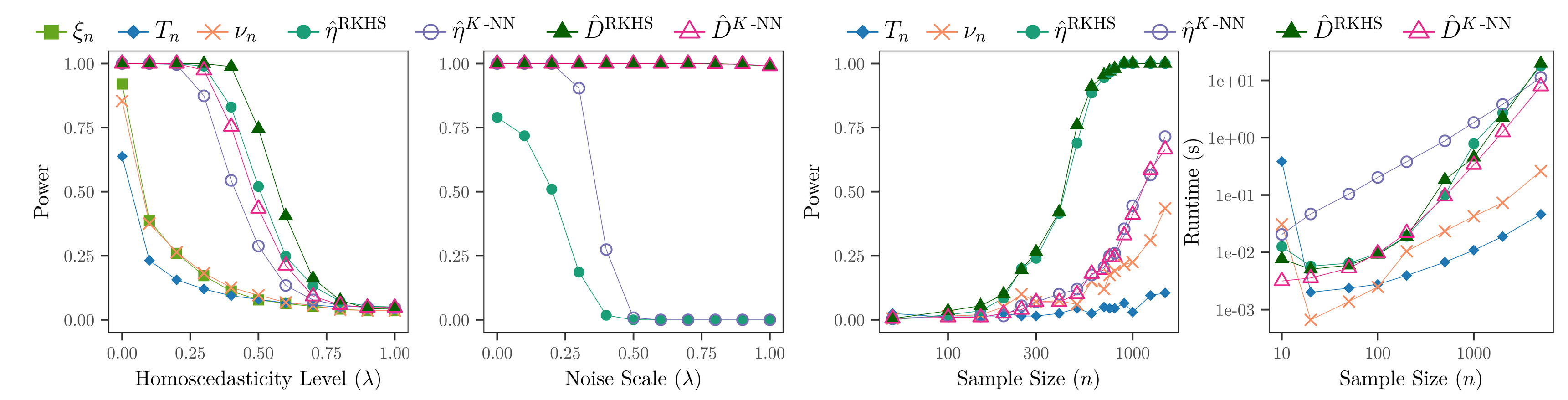
Comparison of Estimators

Property	$\hat{D}^{\text{K-NN}}$	$\hat{D}^{\text{RKHS}}$
Space $\mathcal{X}$	Metric	Separable topological
Time complexity	$\mathcal{O}(n^2 \log n)$	$\mathcal{O}(n^3)$

Numerical Experiments

We assess the performance of independence tests based on $\hat{D}^{\text{RKHS}}$ and $\hat{D}^{\text{K-NN}}$. Code, replicating experiments, is available at `https://github.com/PouyaRoudaki/KernelIR.git`

- **Test setting:** $n = 100$, permutation-test repetitions: 100, $\alpha = 0.05$. Easy: easy to detect $Y = f(X)$.
- **Euclidean (heterosced.):** $(X,Y) \in (\mathbb{R},\mathbb{R})$. Coeff. of homosced.: λ. Easy: $\lambda \to 0$. Difficult: $\lambda \to 1$.
- **Non-Euclidean (rotations):** $(X,Y) \in (\mathbb{R}^3, \mathrm{SO}(3))$. Coeff. of noise: λ. Easy: $\lambda \to 0$. Difficult: $\lambda \to 1$.
- **Real-world, Million Song Data (MSD):** Song features: X, year of release: Y, $(X,Y) \in (\mathbb{R}^{90}, \mathbb{R})$.



(a) Power. Left: Euclidean. Right: Non-Euclidean. (b) Left: Power on MSD. Right: Runtimes.

References

[1] M. Azadkia and S. Chatterjee. A simple measure of conditional dependence. *The Annals of Statistics*, 49(6):3070–3102, 2021.

[2] M. Azadkia and P. Roudaki. A new measure of dependence: Integrated R^2. Technical report, 2025. `https://arxiv.org/abs/2505.18146`.

[3] S. Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022, 2021.

[4] N. Deb, P. Ghosal, and B. Sen. Measuring association on topological spaces using kernels and geometric graphs. Technical report, 2020. `https://arxiv.org/abs/2010.01768`.

[5] I. Klebanov, I. Schuster, and T. J. Sullivan. A rigorous theory of conditional mean embeddings. *SIAM Journal on Mathematics of Data Science*, 2 (3):583–606, 2020.