Kernelized Cumulants: Beyond Kernel Mean Embeddings

Patric Bonnier[1] Harald Oberhauser [1] Zoltán Szabó[2]

[1]University of Oxford, [2]London School of Economics

Abstract

In $\mathbb{R}^d$, it is well-known that cumulants provide an alternative to moments that can achieve the same goals with numerous benefits such as lower variance estimators. In this paper we extend cumulants to reproducing kernel Hilbert spaces (RKHS) using tools from tensor algebras and show that they are computationally tractable by a kernel trick. These kernelized cumulants provide a new set of all-purpose statistics; the classical maximum mean discrepancy and Hilbert-Schmidt independence criterion arise as the degree one objects in our general construction. We argue both theoretically and empirically (on synthetic, environmental, and traffic data analysis) that going beyond degree one has several advantages and can be achieved with the same computational complexity and minimal overhead in our experiments.

Cumulants

$$\sum_{i\in\mathbb{N}} \kappa^{(i)}(\gamma)\frac{\theta^i}{i!} = \log\left(\sum_{i\in\mathbb{N}} \mathbb{E}_\gamma\left(X^i\right)\frac{\theta^i}{i!}\right).$$

$\kappa^{(1)}(\gamma) = \mathbb{E}_\gamma(X)$,
$\kappa^{(2)}(\gamma) = \mathbb{E}_\gamma\left(X^2\right) - \mathbb{E}_\gamma(X)^2$
$\kappa^{(3)}(\gamma) = \mathbb{E}_\gamma\left(X^3\right) - 3\mathbb{E}_\gamma\left(X^2\right)\mathbb{E}_\gamma(X) + 2\mathbb{E}_\gamma(X)^2$

Theorem:[1] Let γ be a probability measure on a bounded subset of $\mathbb{R}^d$ with cumulants $\kappa(\gamma)$ and let $(X_1,\dots,X_d)\sim\gamma$. Then
1. $\gamma \mapsto \kappa(\gamma)$ is injective.
2. $X_1,\dots,X_d$ are independent $\Leftrightarrow \kappa^{\mathbf{i}}(\gamma) = 0$ for all $\mathbf{i}\in\mathbb{N}_+^d$.

Kernelised cumulants

- Repetition (diagonal measure): $\mathbf{i} = (i_1,\dots,i_d)\in\mathbb{N}^d$,
$$\gamma^{\mathbf{i}} := \mathrm{Law}(\underbrace{X_1,\dots,X_1}_{i_1 \text{ times}},\underbrace{X_2,\dots,X_2}_{i_2 \text{ times}},\dots,\underbrace{X_d,\dots,X_d}_{i_d \text{ times}}).$$
- Partitioning (partition measure): $\pi\in P(d)$, $b = |\pi|$,
$$\gamma_\pi := \gamma|_{\mathcal{X}_{\pi_1}}\otimes\cdots\otimes\gamma|_{\mathcal{X}_{\pi_b}}.$$
- Kernelized cumulants: $m = \deg(\mathbf{i}) := \sum_{j=1}^d i_j$, $\gamma_\pi^{\mathbf{i}} = \left(\gamma^{\mathbf{i}}\right)_\pi$,
$$\kappa_{k_1,\dots,k_d}(\gamma) := \left(\kappa^{\mathbf{i}}_{k_1,\dots,k_d}(\gamma)\right)_{\mathbf{i}\in\mathbb{N}^d},$$
$$\kappa^{\mathbf{i}}_{k_1,\dots,k_d}(\gamma) := \sum_{\pi\in P(m)} c_\pi \mathbb{E}_{\gamma_\pi^{\mathbf{i}}} k^{\otimes\mathbf{i}}(\cdot,(X_1,\dots,X_m)).$$

Characterisation of measures

Assume:
- γ, η: probability measures on $\times_{j=1}^d\mathcal{X}_j$,
- $(\mathcal{X}_j)_{j=1}^d$ are Polish spaces,
- k_j: bounded, continuous, point-separating kernel ($j\in[d]$).

Then,
$$\gamma = \eta \Leftrightarrow \kappa_{k_1,\dots,k_d}(\gamma) = \kappa_{k_1,\dots,k_d}(\eta),$$
$$\begin{aligned} d^{\mathbf{i}}(\gamma,\eta) &:= \|\kappa^{\mathbf{i}}_{k_1,\dots,k_d}(\gamma) - \kappa^{\mathbf{i}}_{k_1,\dots,k_d}(\eta)\|^2_{\mathcal{H}^{\otimes\mathbf{i}}} \\ &= \sum_{\pi,\tau\in P(m)} c_\pi c_\tau\Big[\mathbb{E}_{\gamma_\pi^{\mathbf{i}}\otimes\gamma_\tau^{\mathbf{i}}} k^{\otimes\mathbf{i}}((X_1,\dots,X_m),(Y_1,\dots,Y_m)) \\ &\quad + \mathbb{E}_{\eta_\pi^{\mathbf{i}}\otimes\eta_\tau^{\mathbf{i}}} k^{\otimes\mathbf{i}}((X_1,\dots,X_m),(Y_1,\dots,Y_m)) \\ &\quad - 2\mathbb{E}_{\gamma_\pi^{\mathbf{i}}\otimes\eta_\tau^{\mathbf{i}}} k^{\otimes\mathbf{i}}((X_1,\dots,X_m),(Y_1,\dots,Y_m))\Big]. \end{aligned}$$

Characterisation of independence

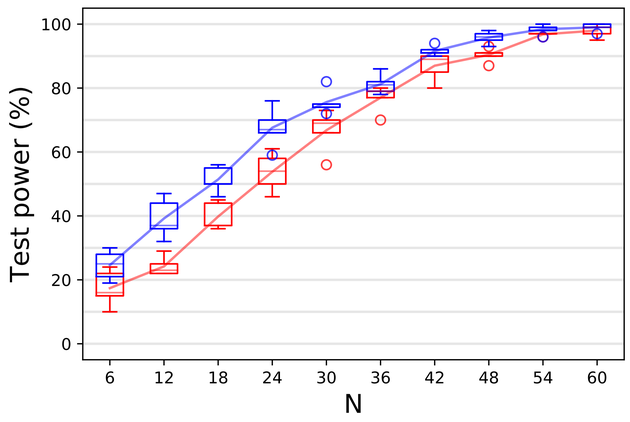
Assume:
- γ: probability measure on $\times_{j=1}^d\mathcal{X}_j$,
- $(\mathcal{X}_j)_{j=1}^d$ are Polish spaces,
- k_j: bounded, continuous, point-separating kernel ($j\in[d]$).

Then,
$$\gamma = \gamma|_{\mathcal{X}_1}\otimes\cdots\otimes\gamma|_{\mathcal{X}_d} \Leftrightarrow \kappa^{\mathbf{i}}_{k_1,\dots,k_d}(\gamma) = 0$$
for every $\mathbf{i}\in\mathbb{N}_+^d$, and
$$\|\kappa^{\mathbf{i}}_{k_1,\dots,k_d}(\gamma)\|^2_{\mathcal{H}^{\otimes\mathbf{i}}} = \sum_{\pi,\tau\in P(m)} c_\pi c_\tau \mathbb{E}_{\gamma_\pi^{\mathbf{i}}\otimes\gamma_\tau^{\mathbf{i}}} k^{\otimes\mathbf{i}}((X_j)_{j=1}^m,(Y_j)_{j=1}^m),$$
where $m = \deg(\mathbf{i})$.

Estimators

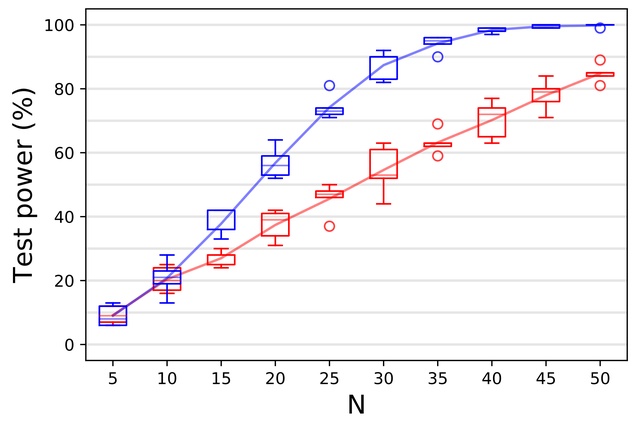
Theorem: V-statistic estimator of $d^{(2)}(\gamma,\eta)$:
$$\frac{1}{N^2}\mathrm{Tr}\left[(\mathbf{K}_x\mathbf{J}_N)^2\right] + \frac{1}{M^2}\mathrm{Tr}\left[(\mathbf{K}_y\mathbf{J}_M)^2\right] - \frac{2}{NM}\mathrm{Tr}\left[\mathbf{K}_{xy}\mathbf{J}_M\mathbf{K}_{xy}^\top\mathbf{J}_N\right]$$
with
$$(x_n)_{n=1}^N \overset{\text{i.i.d.}}{\sim} \gamma, \quad (y_m)_{m=1}^M \overset{\text{i.i.d.}}{\sim} \eta,$$
$$\mathbf{K}_x = [k(x_i,x_j)]_{i,j=1}^N, \quad \mathbf{K}_y = [k(y_i,y_j)]_{i,j=1}^M,$$
$$\mathbf{K}_{x,y} = [k(x_i,y_j)]_{i,j=1}^{N,M}, \quad \mathbf{J}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top.$$
See [2] for other estimators.

HSIC vs. CSIC



Independence testing using HSIC (red) and CSIC statistics (blue).

MMD vs. $d^{(3)}$



Two-sample testing using MMD (red) and $d^{(3)}$ (blue) on the Sao Paulo traffic dataset.

References

References

[1] S. R. Jammalamadaka, T. S. Rao, and G. Terdik, "Higher order cumulants of random vectors and applications to statistical inference and time series," *Indian Journal of Statistics*, vol. 68, no. 2, pp. 326–356, 2006.

[2] P. Bonnier, Z. Szabó, and H. Oberhauser, "Kernelized cumulants: Beyond kernel mean embeddings," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.