Nyström M-HSIC

Zoltán Szabó

Joint work with:

- Florian Kalinke @ Karlsruhe Institute of Technology (KIT), Germany.

Economic Data Analysis and Statistical Inference to Unfold Uncertainty session,
CMStatistics
Dec. 18, 2023

In a nutshell

- Hilbert-Schmidt independence criterion (HSIC):
  - popular dependency measure, various applications.
- Bottleneck:
  1. quadratic runtime: $\mathcal{O}\left(n^2\right)$, $n =$ sample size,
  2. existing accelerations [Zhang et al., 2018]: $M = 2$, no guarantees.

In a nutshell

- Hilbert-Schmidt independence criterion (HSIC):
  - popular dependency measure, various applications.
- Bottleneck:
  1. quadratic runtime: $\mathcal{O}\left(n^2\right)$, $n =$ sample size,
  2. existing accelerations [Zhang et al., 2018]: $M = 2$, no guarantees.

Contributions

1. $M \geq 2$, computational gain even for $M = 2$,
2. improved runtime: $\mathcal{O}\left(n^{\frac{3}{2}}\right)$ instead of $\mathcal{O}\left(n^2\right)$,
3. convergence rate: $\mathcal{O}_P\left(\frac{1}{\sqrt{n}}\right)$ – optimal in minimax sense.

In a nutshell

- Hilbert-Schmidt independence criterion (HSIC):
  - popular dependency measure, various applications.
- Bottleneck:
  1. quadratic runtime: $\mathcal{O}\left(n^2\right)$, $n =$ sample size,
  2. existing accelerations [Zhang et al., 2018]: $M = 2$, no guarantees.

Contributions

1. $M \geq 2$, computational gain even for $M = 2$,
2. improved runtime: $\mathcal{O}\left(n^{\frac{3}{2}}\right)$ instead of $\mathcal{O}\left(n^2\right)$,
3. convergence rate: $\mathcal{O}_P\left(\frac{1}{\sqrt{n}}\right)$ – optimal in minimax sense.

This is what we unfold in the sequel.

Kernel (generalization of $\mathbf{a}^\top \mathbf{b}$), RKHS

[Aronszajn, 1950, Steinwart and Christmann, 2008]

- Def-1 (feature space):

$$k(a, b) = \langle \Phi(a), \Phi(b) \rangle_{\mathcal{H}}.$$

- Def-2 (reproducing kernel):

$$k(\cdot, b) \in \mathcal{H}, \qquad f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}.$$

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n \times n} \succeq 0$.

Kernel (generalization of $\mathbf{a}^\top \mathbf{b}$), RKHS

[Aronszajn, 1950, Steinwart and Christmann, 2008]

- Def-1 (feature space):

$$k(a,b) = \langle \Phi(a), \Phi(b) \rangle_{\mathcal{H}}.$$

- Def-2 (reproducing kernel):

$$k(\cdot, b) \in \mathcal{H}, \qquad f(b) = \langle f, k(\cdot, b) \rangle_{\mathcal{H}}.$$

- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n\times n} \succeq 0$.

Notes

- $k \overset{1:1}{\leftrightarrow} \mathcal{H}_k = \overline{\text{Span}}(k(\cdot, x) : x \in \mathcal{X})$: Fourier analysis, polynomials, splines, ...
- Examples: $k_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + \gamma)^p$, $k_G(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|_2^2}$.
- Kernels exist on various domains!

Some kernel-enriched domains: $(\mathcal{X}, k)$

- Strings
  [Watkins, 1999, Lodhi et al., 2002, Leslie et al., 2002, Kuang et al., 2004, Leslie and Kuang, 2004, Saigo et al., 2004, Cuturi and Vert, 2005],
- time series [Rüping, 2001, Cuturi et al., 2007, Cuturi, 2011, Király and Oberhauser, 2019],
- trees [Collins and Duffy, 2001, Kashima and Koyanagi, 2002],
- groups and specifically rankings [Cuturi et al., 2005, Jiao and Vert, 2016],
- sets [Haussler, 1999, Gärtner et al., 2002, Balanca and Herbin, 2012, Fellmann et al., 2023], probability distributions [Berlinet and Thomas-Agnan, 2004, Hein and Bousquet, 2005, Smola et al., 2007, Sriperumbudur et al., 2010],
- various generative models [Jaakkola and Haussler, 1999, Tsuda et al., 2002, Seeger, 2002, Jebara et al., 2004],
- fuzzy domains [Guevara et al., 2017], or
- graphs [Kondor and Lafferty, 2002, Gärtner et al., 2003, Kashima et al., 2003, Borgwardt and Kriegel, 2005, Shervashidze et al., 2009, Vishwanathan et al., 2010, Kondor and Pan, 2016, Bai et al., 2020, Borgwardt et al., 2020, Schulz et al., 2022, Nikolentzos and Vazirgiannis, 2023].

Mean embedding

- Mean embedding [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]:

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x) \in \mathcal{H}_k} \, \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

Mean embedding, MMD

- Mean embedding [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]:

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x) \in \mathcal{H}_k} \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy [Smola et al., 2007, Gretton et al., 2012]:

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) := \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}.$$

Mean embedding, MMD, HSIC

- Mean embedding [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]:

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x) \in \mathcal{H}_k} \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy [Smola et al., 2007, Gretton et al., 2012]:

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) := \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}.$$

- HSIC [Gretton et al., 2005] (M=2), [Quadrianto et al., 2009, Sejdinovic et al., 2013a, Pfister et al., 2018, Szabó and Sriperumbudur, 2018] ($M \geq 3$), $k := \otimes_{m=1}^{M} k_m$:

$$\mathrm{HSIC}_k(\mathbb{P}) := \mathrm{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^{M} \mathbb{P}_m\right)$$

Mean embedding, MMD, HSIC

- Mean embedding [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]:

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x)\in\mathcal{H}_k} \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy [Smola et al., 2007, Gretton et al., 2012]:

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) := \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}.$$

- HSIC [Gretton et al., 2005] (M=2), [Quadrianto et al., 2009, Sejdinovic et al., 2013a, Pfister et al., 2018, Szabó and Sriperumbudur, 2018] ($M \ge 3$), $k := \otimes_{m=1}^M k_m$:

$$\mathrm{HSIC}_k(\mathbb{P}) := \mathrm{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^M \mathbb{P}_m\right)$$
$$= \Big\| \underbrace{\mu_{\otimes_{m=1}^M k_m}(\mathbb{P}) - \otimes_{m=1}^M \mu_{k_m}(\mathbb{P}_m)}_{\text{cross-covariance operator}} \Big\|_{\mathcal{H}_k}.$$

Mean embedding, MMD, HSIC

- Mean embedding [Berlinet and Thomas-Agnan, 2004, Smola et al., 2007]:

$$\mu_k(\mathbb{P}) := \int_{\mathcal{X}} \underbrace{k(\cdot, x)}_{\Phi(x) \in \mathcal{H}_k} \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k.$$

- Maximum mean discrepancy [Smola et al., 2007, Gretton et al., 2012]:

$$\mathrm{MMD}_k(\mathbb{P}, \mathbb{Q}) := \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}.$$

- HSIC [Gretton et al., 2005] (M=2), [Quadrianto et al., 2009, Sejdinovic et al., 2013a, Pfister et al., 2018, Szabó and Sriperumbudur, 2018] ($M \geq 3$), $k := \otimes_{m=1}^M k_m$:

$$\mathrm{HSIC}_k(\mathbb{P}) := \mathrm{MMD}_k\left(\mathbb{P}, \otimes_{m=1}^M \mathbb{P}_m\right) = \Big\| \underbrace{\mu_{\otimes_{m=1}^M k_m}(\mathbb{P}) - \otimes_{m=1}^M \mu_{k_m}(\mathbb{P}_m)}_{\text{cross-covariance operator}} \Big\|_{\mathcal{H}_k}.$$

Notes before clarification of what $\otimes_{m=1}^M k_m$ and $\otimes_{m=1}^M \mu_{k_m}(\mathbb{P}_m)$ are.

MMD, HSIC: statistical relations, validness of HSIC

- M MD:

$$\text{MMD}_k(\mathbb{P},\mathbb{Q}) = \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k} = \sup_{f\in B_k} \underbrace{\langle f, \mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k}}_{\mathbb{E}_{x\sim\mathbb{P}} f(x) - \mathbb{E}_{x\sim\mathbb{Q}} f(x)}$$

MMD, HSIC: statistical relations, validness of HSIC

- M MD:

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k} = \sup_{f\in B_k} \underbrace{\langle f, \mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k}}_{\mathbb{E}_{x\sim\mathbb{P}} f(x) - \mathbb{E}_{x\sim\mathbb{Q}} f(x)}$$

  - $\in$ IPMs [Zolotarev, 1983, Müller, 1997],

MMD, HSIC: statistical relations, validness of HSIC

- M MD:

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k} = \sup_{f \in B_k} \underbrace{\langle f, \mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k}}_{\mathbb{E}_{x\sim\mathbb{P}} f(x) - \mathbb{E}_{x\sim\mathbb{Q}} f(x)}$$

  - $\in$ IPMs [Zolotarev, 1983, Müller, 1997],
  - $\overset{\dagger}{\Leftrightarrow}$ energy distance [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005], a.k.a. N-distance [Zinger et al., 1992, Klebanov, 2005].

MMD, HSIC: statistical relations, validness of HSIC

- M MD:

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k} = \sup_{f\in B_k} \underbrace{\langle f, \mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k}}_{\mathbb{E}_{x\sim\mathbb{P}} f(x) - \mathbb{E}_{x\sim\mathbb{Q}} f(x)}$$

  - $\in$ IPMs [Zolotarev, 1983, Müller, 1997],
  - $\overset{\dagger}{\Leftrightarrow}$ energy distance [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005], a.k.a. N-distance [Zinger et al., 1992, Klebanov, 2005].
- HSIC ($M = 2$) $\xLeftrightarrow{\text{[Sejdinovic et al., 2013b]}}$ distance covariance [Székely et al., 2007, Székely and Rizzo, 2009, Lyons, 2013].

MMD, HSIC: statistical relations, validness of HSIC

- M MD:

$$\text{MMD}_k(\mathbb{P},\mathbb{Q}) = \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k} = \sup_{f\in B_k} \underbrace{\langle f, \mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\rangle_{\mathcal{H}_k}}_{\mathbb{E}_{x\sim\mathbb{P}} f(x) - \mathbb{E}_{x\sim\mathbb{Q}} f(x)}$$

  - $\in$ IPMs [Zolotarev, 1983, Müller, 1997],
  - $\overset{\dagger}{\Leftrightarrow}$ energy distance [Baringhaus and Franz, 2004, Székely and Rizzo, 2004, Székely and Rizzo, 2005], a.k.a. N-distance [Zinger et al., 1992, Klebanov, 2005].
- HSIC ($M = 2$) $\overset{\text{[Sejdinovic et al., 2013b]}}{\Longleftrightarrow}$ distance covariance [Székely et al., 2007, Székely and Rizzo, 2009, Lyons, 2013].
- Validness of HSIC $\overset{\text{[Szabó and Sriperumbudur, 2018]}}{\Longleftarrow}$ k_m-s are universal [Steinwart, 2001, Micchelli et al., 2006, Sriperumbudur et al., 2011].

Tensor product: $a_1 \otimes a_2$

- If $\mathbf{a} \in \mathbb{R}^{n_1}$, $\mathbf{b} \in \mathbb{R}^{n_2}$:

$$\mathbb{R} \ni \mathbf{v}^\top \left(\mathbf{a}\mathbf{b}^\top\right) \mathbf{w} = \left(\mathbf{v}^\top \mathbf{a}\right) \left(\mathbf{b}^\top \mathbf{w}\right) = \langle \mathbf{a}, \mathbf{v} \rangle_{\mathbb{R}^{n_1}} \langle \mathbf{b}, \mathbf{w} \rangle_{\mathbb{R}^{n_2}},$$

  $\mathbf{a} \otimes \mathbf{b} := \mathbf{a}\mathbf{b}^\top$ is an $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ bilinear form.

Tensor product: $a_1 \otimes a_2$

- If $\mathbf{a} \in \mathbb{R}^{n_1}$, $\mathbf{b} \in \mathbb{R}^{n_2}$:

$$\mathbb{R} \ni \mathbf{v}^\top \left(\mathbf{a}\mathbf{b}^\top\right) \mathbf{w} = \left(\mathbf{v}^\top \mathbf{a}\right) \left(\mathbf{b}^\top \mathbf{w}\right) = \langle \mathbf{a}, \mathbf{v} \rangle_{\mathbb{R}^{n_1}} \langle \mathbf{b}, \mathbf{w} \rangle_{\mathbb{R}^{n_2}},$$

  $\mathbf{a} \otimes \mathbf{b} := \mathbf{a}\mathbf{b}^\top$ is an $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ bilinear form.
- For $a \in \mathcal{H}_1$, $b \in \mathcal{H}_2$ Hilbert spaces, i.e. for $M = 2$:

$$(a \otimes b)(v, w) := \langle a, v \rangle_{\mathcal{H}_1} \langle b, w \rangle_{\mathcal{H}_2}.$$

Tensor product: $\otimes_{m=1}^{M} a_m$

- If $\mathbf{a} \in \mathbb{R}^{n_1}$, $\mathbf{b} \in \mathbb{R}^{n_2}$:

$$\mathbb{R} \ni \mathbf{v}^\top \left(\mathbf{a}\mathbf{b}^\top\right) \mathbf{w} = \left(\mathbf{v}^\top \mathbf{a}\right) \left(\mathbf{b}^\top \mathbf{w}\right) = \langle \mathbf{a}, \mathbf{v} \rangle_{\mathbb{R}^{n_1}} \langle \mathbf{b}, \mathbf{w} \rangle_{\mathbb{R}^{n_2}},$$

  $\mathbf{a} \otimes \mathbf{b} := \mathbf{a}\mathbf{b}^\top$ is an $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}$ bilinear form.
- For $a \in \mathcal{H}_1$, $b \in \mathcal{H}_2$ Hilbert spaces, i.e. for $M = 2$:

$$(a \otimes b)(v, w) := \langle a, v \rangle_{\mathcal{H}_1} \langle b, w \rangle_{\mathcal{H}_2}.$$

- For $M \geq 2$ and $a_m \in \mathcal{H}_m$,

$$\left(\otimes_{m=1}^{M} a_m\right)(b_1, \ldots, b_M) := \prod_{m=1}^{M} \langle a_m, b_m \rangle_{\mathcal{H}_m}.$$

Tensor product: $\otimes_{m=1}^{M}\mathcal{H}_m$

$$\otimes_{m=1}^{M}\mathcal{H}_m := \overline{\mathrm{Span}}(\otimes_{m=1}^{M} a_m \,:\, a_m \in \mathcal{H}_m).$$

Tensor product: $\otimes_{m=1}^M \mathcal{H}_m$

$$\otimes_{m=1}^M \mathcal{H}_m := \overline{\mathrm{Span}}(\otimes_{m=1}^M a_m \,:\, a_m \in \mathcal{H}_m).$$

$\xrightarrow{\text{spec.}}$ The tensor product of RKHSs is an RKHS

$$\mathcal{H}_k = \otimes_{m=1}^M \mathcal{H}_{k_m},$$

$$k(x, x') := \left(\otimes_{m=1}^M k_m\right)(x, x') := \prod_{m=1}^M \underbrace{k_m(x_m, x'_m)}_{\text{coordinate-wise similarity}}.$$

A few HSIC applications

- independence testing [Gretton et al., 2008, Bilodeau and Nangue, 2017, Górecki et al., 2018, Pfister et al., 2018, Albert et al., 2022],
- feature selection [Camps-Valls et al., 2010, Song et al., 2012, Yamada et al., 2014, Wang et al., 2022], with apps in biomarker detection [Climente-González et al., 2019] & wind power prediction [Bouche et al., 2023],
- clustering [Song et al., 2007, Climente-González et al., 2019],
- causal discovery [Mooij et al., 2016, Pfister et al., 2018, Chakraborty and Zhang, 2019, Schölkopf et al., 2021],
- sensitivity analysis [Veiga, 2015, Fellmann et al., 2023],
- uncertainty quantification [Stenger et al., 2020],
- analysis of data augmentation methods for brain tumor detection [Anaya-Isaza and Mera-Jiménez, 2022].

HSIC estimators: classical vs. proposed

- Samples: $\hat{\mathbb{P}}_n := \left\{\left(x_1^1, \ldots, x_M^1\right), \ldots, \left(x_1^n, \ldots, x_M^n\right)\right\} \subset \mathcal{X}$.

HSIC estimators: classical vs. proposed

- Samples: $\hat{\mathbb{P}}_n := \left\{\left(x_1^1, \ldots, x_M^1\right), \ldots, \left(x_1^n, \ldots, x_M^n\right)\right\} \subset \mathcal{X}$.
- Classical:

$$\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) = \frac{1}{n^2}\,\mathbf{1}_n^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m}\right)\mathbf{1}_n + \frac{1}{n^{2M}}\prod_{m\in[M]}\mathbf{1}_n^\mathsf{T}\mathbf{K}_{k_m}\mathbf{1}_n - \frac{2}{n^{M+1}}\mathbf{1}_n^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m}\mathbf{1}_n\right),$$

$$\mathbf{K}_{k_m} = \left[k_m\left(x_m^i, x_m^j\right)\right]_{i,j\in[n]} \in \mathbb{R}^{n\times n}, \quad m \in [M].$$

HSIC estimators: classical vs. proposed

- Samples: $\hat{\mathbb{P}}_n := \left\{\left(x_1^1, \ldots, x_M^1\right), \ldots, \left(x_1^n, \ldots, x_M^n\right)\right\} \subset \mathcal{X}$.
- Classical:

$$\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) = \frac{1}{n^2}\,\mathbf{1}_n^{\mathsf{T}}\left(\circ_{m\in[M]}\mathbf{K}_{k_m}\right)\mathbf{1}_n + \frac{1}{n^{2M}}\prod_{m\in[M]}\mathbf{1}_n^{\mathsf{T}}\mathbf{K}_{k_m}\mathbf{1}_n$$

$$-\frac{2}{n^{M+1}}\mathbf{1}_n^{\mathsf{T}}\left(\circ_{m\in[M]}\mathbf{K}_{k_m}\mathbf{1}_n\right),$$

$$\mathbf{K}_{k_m} = \left[k_m\left(x_m^i, x_m^j\right)\right]_{i,j\in[n]} \in \mathbb{R}^{n\times n}, \quad m\in[M].$$

- Proposed: $\tilde{\mathbb{P}}_{n'}$ subsample of $\hat{\mathbb{P}}_n$ (with replacement),

$$\mathbf{K}_{k_m,n',n'} = \left[k_m\left(\tilde{x}_m^i, \tilde{x}_m^j\right)\right]_{i,j\in[n']} \in \mathbb{R}^{n'\times n'}, \quad m\in[M],$$

HSIC estimators: classical vs. proposed

- Samples: $\hat{\mathbb{P}}_n := \left\{\left(x_1^1, \ldots, x_M^1\right), \ldots, \left(x_1^n, \ldots, x_M^n\right)\right\} \subset \mathcal{X}$.
- Classical:

$$\mathrm{HSIC}_k^2\left(\hat{\mathbb{P}}_n\right) = \frac{1}{n^2}\,\mathbf{1}_n^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m}\right)\mathbf{1}_n + \frac{1}{n^{2M}}\prod_{m\in[M]}\mathbf{1}_n^\mathsf{T}\mathbf{K}_{k_m}\mathbf{1}_n$$

$$-\frac{2}{n^{M+1}}\mathbf{1}_n^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m}\mathbf{1}_n\right),$$

$$\mathbf{K}_{k_m} = \left[k_m\left(x_m^i, x_m^j\right)\right]_{i,j\in[n]} \in \mathbb{R}^{n\times n}, \quad m \in [M].$$

- Proposed: $\tilde{\mathbb{P}}_{n'}$ subsample of $\hat{\mathbb{P}}_n$ (with replacement),

$$\mathrm{HSIC}_{k,\mathrm{N}}^2\left(\hat{\mathbb{P}}_n\right) = \boldsymbol{\alpha}_k^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m,n',n'}\right)\boldsymbol{\alpha}_k + \prod_{m\in[M]}\boldsymbol{\alpha}_{k_m}^\mathsf{T}\mathbf{K}_{k_m,n',n'}\boldsymbol{\alpha}_{k_m}$$

$$-2\boldsymbol{\alpha}_k^\mathsf{T}\left(\circ_{m\in[M]}\mathbf{K}_{k_m,n',n'}\boldsymbol{\alpha}_{k_m}\right),$$

$$\mathbf{K}_{k_m,n',n'} = \left[k_m\left(\tilde{x}_m^i, \tilde{x}_m^j\right)\right]_{i,j\in[n']} \in \mathbb{R}^{n'\times n'}, \quad m \in [M],$$

with appropriately chosen $\boldsymbol{\alpha}_k, \boldsymbol{\alpha}_{k_m} \in \mathbb{R}^{n'}$ $(m \in [M])$.

Towards the appropriately chosen α_k, α_{k_m}-s

- Recall: $k = \otimes_{m=1}^{M} k_m$,

$$\text{HSIC}_k(\mathbb{P}) = \|C_X\|_{\mathcal{H}_k}, \quad C_X = \underbrace{\mu_{\otimes_{m=1}^{M} k_m}(\mathbb{P})}_{1 \text{ mean}} - \underbrace{\otimes_{m=1}^{M} \mu_{k_m}(\mathbb{P}_m)}_{M \text{ means}}.$$

Towards the appropriately chosen α_k, α_{k_m}-s

- Recall: $k = \otimes_{m=1}^M k_m$,

$$\mathrm{HSIC}_k(\mathbb{P}) = \|C_X\|_{\mathcal{H}_k}, \quad C_X = \underbrace{\mu_{\otimes_{m=1}^M k_m}(\mathbb{P})}_{1 \text{ mean}} - \underbrace{\otimes_{m=1}^M \mu_{k_m}(\mathbb{P}_m)}_{M \text{ means}}.$$

Idea

Approximate the $M + 1$ means, with the Nyström method [Chatalic et al., 2022], and analyze the error propagation.

The classical Nyström approach: mean approximation

- We will choose

$$(\mathcal{Y},\ell,\mathbb{Q}) = (\mathcal{X},k,\mathbb{P}), \quad (\mathcal{Y},\ell,\mathbb{Q}) = (\mathcal{X}_m,k_m,\mathbb{P}_m), \quad m\in[M],$$

- $\tilde{\mathbb{Q}}_{n'} = \left\{\tilde{y}^1,\dots,\tilde{y}^{n'}\right\}$: subsample of $\hat{\mathbb{Q}}_n = \{y^1,\dots,y^n\} \stackrel{\text{i.i.d.}}{\sim} \mathbb{Q}$.

The classical Nyström approach: mean approximation

- We will choose

$$(\mathcal{Y},\ell,\mathbb{Q}) = (\mathcal{X},k,\mathbb{P}), \quad (\mathcal{Y},\ell,\mathbb{Q}) = (\mathcal{X}_m,k_m,\mathbb{P}_m), \quad m\in[M],$$

- $\tilde{\mathbb{Q}}_{n'} = \left\{\tilde{y}^1,\ldots,\tilde{y}^{n'}\right\}$: subsample of $\hat{\mathbb{Q}}_n = \left\{y^1,\ldots,y^n\right\} \stackrel{\text{i.i.d.}}{\sim} \mathbb{Q}$.
- Idea:

$$\mu_\ell(\mathbb{Q}) \approx \mu_\ell(\hat{\mathbb{Q}}_n) \approx \sum_{i\in[n']} \alpha_i \phi_\ell(\tilde{y}^i) =: \mu_\ell\left(\tilde{\mathbb{Q}}_{n'}\right) \in \mathcal{H}_\ell^{\text{Nys}},$$

$$\mathcal{H}_\ell^{\text{Nys}} = \text{Span}\left(\phi_\ell(\tilde{y}^i)\ :\ i\in[n']\right) \subset \mathcal{H}_\ell.$$

Nyström approximation: computational task

- Find the minimum norm solution of

$$\min_{\boldsymbol{\alpha}_\ell \in \mathbb{R}^{n'}} \left\| \mu_\ell\left(\hat{\mathbb{Q}}_n\right) - \sum_{i \in [n']} \alpha_i \phi_\ell\left(\tilde{y}^i\right) \right\|_{\mathcal{H}_\ell}^2.$$

Nyström approximation: computational task

- Find the minimum norm solution of

$$\min_{\boldsymbol{\alpha}_\ell \in \mathbb{R}^{n'}} \left\| \mu_\ell\left(\hat{\mathbb{Q}}_n\right) - \sum_{i\in[n']} \alpha_i \phi_\ell\left(\tilde{y}^i\right) \right\|_{\mathcal{H}_\ell}^2 .$$

- Solution [Laub, 2004, Chatalic et al., 2022]:

$$\mu_\ell\left(\tilde{\mathbb{Q}}_{n'}\right) = \sum_{i\in[n']} \alpha_\ell^i \phi_\ell\left(\tilde{y}^i\right), \quad \boldsymbol{\alpha}_\ell = \frac{1}{n}\left(\mathbf{K}_{\ell,n',n'}\right)^{-} \mathbf{K}_{\ell,n',n} \mathbf{1}_n,$$

with Gram matrices

$$\mathbf{K}_{\ell,n',n'} = \left[\ell\left(\tilde{x}^i, \tilde{x}^j\right)\right]_{i,j\in[n']} \in \mathbb{R}^{n'\times n'},$$

$$\mathbf{K}_{\ell,n',n} = \left[\ell\left(\tilde{x}^i, x^j\right)\right]_{i\in[n'],j\in[n]} \in \mathbb{R}^{n'\times n}.$$

Results for the proposed HSIC estimator [Kalinke and Szabó, 2023]

- Runtime: $\mathcal{O}\left(Mn'^3 + Mn'n\right)$ vs. $\mathcal{O}\left(Mn^2\right) \Rightarrow$
  - Saving if $n' = o\left(n^{2/3}\right)$.
- Finite-sample guarantee $\Rightarrow \sqrt{n}$-consistency if the effective dimension decays
  - polynomially ($\leq c\lambda^{-\gamma}$, $c > 0, \gamma \in (0,1]$) and $n' = \tilde{\mathcal{O}}\left(n^{1/(2-\gamma)}\right)$, or
  - exponentially ($\leq \log(1 + c/\gamma)/\beta$, $c, \beta > 0$) and $n' = \tilde{\mathcal{O}}\left(\sqrt{n}\right)$. $\Rightarrow$

Runtime can be $\mathcal{O}\left(n^{\frac{3}{2}}\right)$.

- This matches the rate of the quadratic-time estimator.

Key lemma: error propagation on tensor products

- Let
  - $X = (X_m)_{m=1}^{M} \in \mathcal{X} = \times_{m=1}^{M} \mathcal{X}_m$, $X_m \sim \mathbb{P}_m$,
  - $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ bounded ($\sup_{x_m \in \mathcal{X}_m} \sqrt{k_m(x_m, x_m)} \le a_{k_m}$), $k = \otimes_{m=1}^{M} k_m$,
  - $\tilde{\mathbb{P}}_{m,n'}$: Nyström sample of the m-th marginal.

Key lemma: error propagation on tensor products

- Let
  - $X = (X_m)_{m=1}^M \in \mathcal{X} = \times_{m=1}^M \mathcal{X}_m$, $X_m \sim \mathbb{P}_m$,
  - $k_m : \mathcal{X}_m \times \mathcal{X}_m \to \mathbb{R}$ bounded ($\sup_{x_m \in \mathcal{X}_m} \sqrt{k_m(x_m, x_m)} \le a_{k_m}$), $k = \otimes_{m=1}^M k_m$,
  - $\tilde{\mathbb{P}}_{m,n'}$: Nyström sample of the m-th marginal.
- Then

$$\left\| \otimes_{m=1}^M \mu_{k_m}(\mathbb{P}_m) - \otimes_{m=1}^M \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right) \right\|_{\mathcal{H}_k} \le \prod_{m \in [M]} \left(a_{k_m} + d_{k_m}\right) - \prod_{m \in [M]} a_{k_m},$$

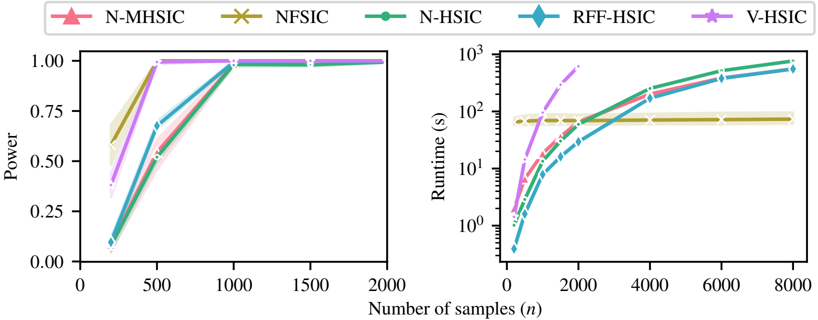
where $d_{k_m} = \left\| \mu_{k_m}(\mathbb{P}_m) - \mu_{k_m}\left(\tilde{\mathbb{P}}_{m,n'}\right) \right\|_{\mathcal{H}_{k_m}}$.

Demo-1: million song data – media annotations

- Task: test the dependence of (X, Y) [$M = 2$, H_1 holds],
  - X: 90 acoustic features; Y: year of release.
- $M = 2$: allows comparing to existing methods.

Demo-1: million song data – media annotations

- Task: test the dependence of (X, Y) [$M = 2$, H_1 holds],
  - X: 90 acoustic features; Y: year of release.
- $M = 2$: allows comparing to existing methods.



1. runtime & power like N-HSIC and RFF-HSIC, but $M \geq 2$✓
2. lower complexity than V-HSIC,
3. NFSIC: restricted to $\mathbb{R}^d$, $M = 2$, and analytic kernels.

Demo-2: weather causal discovery [Mooij et al., 2016]

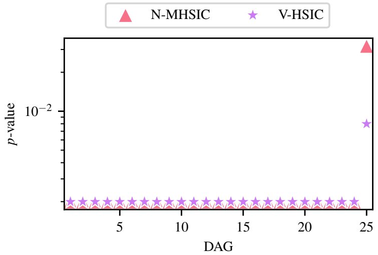
- Observation: (altitude, temperature, sunshine) triplets, $M = 3$, $n = 349$.
- Task: infer the most plausible DAG on the 3 nodes.
- Approach: for each DAG candidate†
  1. additive model regression: $X_m = \sum_{j \in \mathrm{PA}_m} f_{m,j}(X_j) + e_m$, $m \in [M]$,
  2. independence testing of $(\hat{e}_m)_{m=1}^{M}$.

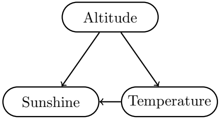
† $3^3 - 2 = 25$, 2 graphs contain cycles.

Demo-2: weather causal discovery [Mooij et al., 2016]

- Observation: (altitude, temperature, sunshine) triplets, $M = 3$, $n = 349$.
- Task: infer the most plausible DAG on the 3 nodes.
- Approach: for each DAG candidate[†]
  1. additive model regression: $X_m = \sum_{j \in \mathrm{PA}_m} f_{m,j}(X_j) + e_m$, $m \in [M]$,
  2. independence testing of $(\hat{e}_m)_{m=1}^M$.

[†] $3^3 - 2 = 25$, 2 graphs contain cycles.





Compared to V-HSIC

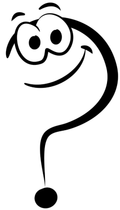
Both methods find the most plausible DAG.

Summary

- Focus: HSIC acceleration with the Nyström method.
- Results:
  1. $M \geq 2$, computational gain even for $M = 2$,
  2. $\sqrt{n}$-consistency (upon appropriate effective dimension decay),
  3. improved runtime: $O\left(n^{\frac{3}{2}}\right)$ instead of $O\left(n^2\right)$,
  4. numerical demo: dependency testing of media annotations, causal discovery.
- Details @ UAI [Kalinke and Szabó, 2023], code.

Summary

- Focus: HSIC acceleration with the Nyström method.
- Results:
  1. $M \geq 2$, computational gain even for $M = 2$,
  2. $\sqrt{n}$-consistency (upon appropriate effective dimension decay),
  3. improved runtime: $O\left(n^{\frac{3}{2}}\right)$ instead of $O\left(n^2\right)$,
  4. numerical demo: dependency testing of media annotations, causal discovery.
- Details @ UAI [Kalinke and Szabó, 2023], code.



Albert, M., Laurent, B., Marrel, A., and Meynaoui, A. (2022).
Adaptive test of independence based on HSIC measures.
The Annals of Statistics, 50(2):858–879.

Anaya-Isaza, A. and Mera-Jiménez, L. (2022).
Data augmentation and transfer learning for brain tumor detection in magnetic resonance imaging.
IEEE Access, 10:23217–23233.

Aronszajn, N. (1950).
Theory of reproducing kernels.
Transactions of the American Mathematical Society, 68:337–404.

Bai, L., Cui, L., Rossi, L., Xu, L., Bai, X., and Hancock, E. (2020).
Local-global nested graph kernels using nested complexity traces.
Pattern Recognition Letters, 134:87–95.

Balanca, P. and Herbin, E. (2012).

A set-indexed Ornstein-Uhlenbeck process.
Electronic Communications in Probability, 17:1–14.

Baringhaus, L. and Franz, C. (2004).
On a new multivariate two-sample test.
Journal of Multivariate Analysis, 88:190–206.

Berlinet, A. and Thomas-Agnan, C. (2004).
Reproducing Kernel Hilbert Spaces in Probability and Statistics.
Kluwer.

Bilodeau, M. and Nangue, A. G. (2017).
Tests of mutual or serial independence of random vectors with applications.
Journal of Machine Learning Research, 18:1–40.

Borgwardt, K., Ghisu, E., Llinares-López, F., O'Bray, L., and Riec, B. (2020).
Graph kernels: State-of-the-art and future challenges.

Foundations and Trends in Machine Learning, 13(5-6):531–712.

Borgwardt, K. M. and Kriegel, H.-P. (2005).
Shortest-path kernels on graphs.
In *International Conference on Data Mining (ICDM)*, pages 74–81.

Bouche, D., Flamary, R., d'Alché Buc, F., Plougonven, R., Clausel, M., Badosa, J., and Drobinski, P. (2023).
Wind power predictions from nowcasts to 4-hour forecasts: a learning approach with variable selection.
Renewable Energy, 211:938–947.

Camps-Valls, G., Mooij, J. M., and Schölkopf, B. (2010).
Remote sensing feature selection by kernel dependence measures.
IEEE Geoscience and Remote Sensing Letters, 7(3):587–591.

Chakraborty, S. and Zhang, X. (2019).

Distance metrics for measuring joint dependence with application to causal inference.
Journal of the American Statistical Association, 114(528):1638–1650.

Chatalic, A., Schreuder, N., Rudi, A., and Rosasco, L. (2022).
Nyström kernel mean embeddings.
In *International Conference on Machine Learning (ICML)*, pages 3006–3024.

Climente-González, H., Azencott, C.-A., Kaski, S., and Yamada, M. (2019).
Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data.
Bioinformatics, 35(14):i427–i435.

Collins, M. and Duffy, N. (2001).
Convolution kernels for natural language.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 625–632.

Cuturi, M. (2011).
Fast global alignment kernels.
In *International Conference on Machine Learning (ICML)*, pages 929–936.

Cuturi, M., Fukumizu, K., and Vert, J.-P. (2005).
Semigroup kernels on measures.
Journal of Machine Learning Research, 6:1169–1198.

Cuturi, M. and Vert, J.-P. (2005).
The context-tree kernel for strings.
Neural Networks, 18(8):1111–1123.

Cuturi, M., Vert, J.-P., Birkenes, O., and Matsui, T. (2007).
A kernel for time series based on global alignments.
In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 413–416.

Fellmann, N., Blanchet-Scalliet, C., Helbert, C., Spagnol, A., and Sinoquet, D. (2023).

Kernel-based sensivity analysis for (excursion) sets.
Technical report.
(https://arxiv.org/abs/2305.09268).

Gärtner, T., Flach, P., Kowalczyk, A., and Smola, A. (2002).
Multi-instance kernels.
In *International Conference on Machine Learning (ICML)*, pages 179–186.

Gärtner, T., Flach, P., and Wrobel, S. (2003).
On graph kernels: Hardness results and efficient alternatives.
Learning Theory and Kernel Machines, pages 129–143.

Górecki, T., Krzyśko, M., and Wolyński, W. (2018).
Independence test and canonical correlation analysis based on the alignment between kernel matrices for multivariate functional data.
Artificial Intelligence Review, pages 1–25.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2012).

A kernel two-sample test.
Journal of Machine Learning Research, 13(25):723–773.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005).
Measuring statistical dependence with Hilbert-Schmidt norms.
In *Algorithmic Learning Theory (ALT)*, pages 63–78.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. (2008).
A kernel statistical test of independence.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 585–592.

Guevara, J., Hirata, R., and Canu, S. (2017).
Cross product kernels for fuzzy set similarity.
In *International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6.

Haussler, D. (1999).
Convolution kernels on discrete structures.

Technical report, University of California at Santa Cruz.
(http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf).

Hein, M. and Bousquet, O. (2005).
Hilbertian metrics and positive definite kernels on probability measures.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 136–143.

Jaakkola, T. S. and Haussler, D. (1999).
Exploiting generative models in discriminative classifiers.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 487–493.

Jebara, T., Kondor, R., and Howard, A. (2004).
Probability product kernels.
Journal of Machine Learning Research, 5:819–844.

Jiao, Y. and Vert, J.-P. (2016).
The Kendall and Mallows kernels for permutations.

In *International Conference on Machine Learning (ICML)*, volume 37, pages 2982–2990.

Kalinke, F. and Szabó, Z. (2023).
Nyström M-Hilbert-Schmidt independence criterion.
In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1005–1015.

Kashima, H. and Koyanagi, T. (2002).
Kernels for semi-structured data.
In *International Conference on Machine Learning (ICML)*, pages 291–298.

Kashima, H., Tsuda, K., and Inokuchi, A. (2003).
Marginalized kernels between labeled graphs.
In *International Conference on Machine Learning (ICML)*, pages 321–328.

Király, F. J. and Oberhauser, H. (2019).
Kernels for sequentially ordered data.
Journal of Machine Learning Research, 20:1–45.

Klebanov, L. (2005).
N-Distances and Their Applications.
Charles University, Prague.

Kondor, R. and Pan, H. (2016).
The multiscale Laplacian graph kernel.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2982–2990.

Kondor, R. I. and Lafferty, J. (2002).
Diffusion kernels on graphs and other discrete input.
In *International Conference on Machine Learning (ICML)*, pages 315–322.

Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., and Leslie, C. (2004).
Profile-based string kernels for remote homology detection and motif extraction.
Journal of Bioinformatics and Computational Biology, 13(4):527–550.

Laub, A. J. (2004).
Matrix analysis for scientists and engineers.
SIAM.

Leslie, C., Eskin, E., and Noble, W. S. (2002).
The spectrum kernel: A string kernel for SVM protein classification.
Biocomputing, pages 564–575.

Leslie, C. and Kuang, R. (2004).
Fast string kernels using inexact matching for protein sequences.
Journal of Machine Learning Research, 5:1435–1455.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2002).
Text classification using string kernels.
Journal of Machine Learning Research, 2:419–444.

Lyons, R. (2013).
Distance covariance in metric spaces.

The Annals of Probability, 41:3284–3305.

Micchelli, C., Xu, Y., and Zhang, H. (2006).
Universal kernels.
Journal of Machine Learning Research, 7:2651–2667.

Mooij, J., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016).
Distinguishing cause from effect using observational data: Methods and benchmarks.
Journal of Machine Learning Research, 17:1–102.

Müller, A. (1997).
Integral probability metrics and their generating classes of functions.
Advances in Applied Probability, 29:429–443.

Nikolentzos, G. and Vazirgiannis, M. (2023).
Graph alignment kernels using Weisfeiler and Leman hierarchies.

In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2019–2034.

Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2018).
Kernel-based tests for joint independence.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(1):5–31.

Quadrianto, N., Song, L., and Smola, A. (2009).
Kernelized sorting.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 1289–1296.

Rüping, S. (2001).
SVM kernels for time series analysis.
Technical report, University of Dortmund.
(http://www.stefan-rueping.de/publications/rueping-2001-a.pdf).

Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T. (2004).

Protein homology detection using string alignment kernels.
Bioinformatics, 20(11):1682–1689.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021).
Toward causal representation learning.
Proceedings of the IEEE, 109(5):612–634.

Schulz, T. H., Welke, P., and Wrobel, S. (2022).
Graph filtration kernels.
In *AAAI Conference on Artifical Intelligence (AAAI)*, pages 8196–8203.

Seeger, M. (2002).
Covariance kernels from Bayesian generative models.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 905–912.

Sejdinovic, D., Gretton, A., and Bergsma, W. (2013a).
A kernel test for three-variable interactions.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 1124–1132.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013b).
Equivalence of distance-based and RKHS-based statistics in hypothesis testing.
Annals of Statistics, 41:2263–2291.

Shervashidze, N., Vishwanathan, S. V. N., Petri, T., Mehlhorn, K., and Borgwardt, K. M. (2009).
Efficient graphlet kernels for large graph comparison.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 488–495.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007).
A Hilbert space embedding for distributions.
In *Algorithmic Learning Theory (ALT)*, pages 13–31.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. (2012).

Feature selection via dependence maximization.
Journal of Machine Learning Research, 13(1):1393–1434.

Song, L., Smola, A. J., Gretton, A., and Borgwardt, K. M. (2007).
A dependence maximization view of clustering.
In *International Conference on Machine Learning (ICML)*, pages 815–822.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. (2011).
Universality, characteristic kernels and RKHS embedding of measures.
Journal of Machine Learning Research, 12:2389–2410.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. (2010).
Hilbert space embeddings and metrics on probability measures.
Journal of Machine Learning Research, 11:1517–1561.

Steinwart, I. (2001).

On the influence of the kernel on the consistency of support vector machines.
Journal of Machine Learning Research, 6(3):67–93.

Steinwart, I. and Christmann, A. (2008).
Support Vector Machines.
Springer.

Stenger, J., Gamboa, F., Keller, M., and Iooss, B. (2020).
Optimal uncertainty quantification of a risk measurement from a thermal-hydraulic code using canonical moments.
International Journal for Uncertainty Quantification, 10(1).

Szabó, Z. and Sriperumbudur, B. K. (2018).
Characteristic and universal tensor product kernels.
Journal of Machine Learning Research, 18(233):1–29.

Székely, G. and Rizzo, M. (2004).
Testing for equal distributions in high dimension.
InterStat, 5:1249–1272.

Székely, G. and Rizzo, M. (2005).
A new test for multivariate normality.
Journal of Multivariate Analysis, 93:58–80.

Székely, G. J. and Rizzo, M. L. (2009).
Brownian distance covariance.
The Annals of Applied Statistics, 3:1236–1265.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007).
Measuring and testing dependence by correlation of distances.
The Annals of Statistics, 35:2769–2794.

Tsuda, K., Kin, T., and Asai, K. (2002).
Marginalized kernels for biological sequences.
Bioinformatics, 18:268–275.

Veiga, S. D. (2015).
Global sensitivity analysis with dependence measures.
Journal of Statistical Computation and Simulation, 85(7):1283–1305.

Vishwanathan, S. N., Schraudolph, N., Kondor, R., and Borgwardt, K. (2010).
Graph kernels.
Journal of Machine Learning Research, 11:1201–1242.

Wang, A., Du, J., Zhang, X., and Shi, J. (2022).
Ranking features to promote diversity: An approach based on sparse distance correlation.
Technometrics, 64(3):384–395.

Watkins, C. (1999).
Dynamic alignment kernels.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 39–50.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., and Sugiyama, M. (2014).
High-dimensional feature selection by feature-wise kernelized lasso.
Neural Computation, 26(1):185–207.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. (2018).
Large-scale kernel methods for independence testing.
Statistics and Computing, 28(1):1–18.

Zinger, A., Kakosyan, A., and Klebanov, L. (1992).
A characterization of distributions by mean values of statistics and certain probabilistic metrics.
Journal of Soviet Mathematics.

Zolotarev, V. (1983).
Probability metrics.
Theory of Probability and its Applications, 28:278–302.