Post-selection inference with HSIC-Lasso
Supplementary Materials

A. Technical Appendix on HSIC and HSIC-Lasso

A.1. Measuring Dependence with HSIC

The main incentive to develop advanced techniques to describe dependence relations between two random variables X and Y arises from the fact that the covariance

$$\operatorname{cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y],$$

is designed for linear relationships only. If the dependence structure, however, is of non-linear nature, the covariance can only partly capture the relationship between X and Y or completely fails to do so. Nevertheless, general, or rather model-free, independence can be expressed in terms of the covariance as follows, cf. (Gretton et al., 2005b).

Proposition 11. *The random variables X and Y are independent if and only if* $\operatorname{cov}(f(X), g(Y)) = 0$ *for each pair (f, g) of bounded, continuous functions.*

There are two lines of thought leading to the Hilbert-Schmidt independence criterion: one presented by Gretton et al. (2005a) regarding HSIC as the Hilbert-Schmidt norm of a cross-covariance operator and one thinking of HSIC as maximum mean discrepancy on a product space according to Zhang et al. (2018). In this work, we follow the latter derivation and link it with the first approach and Proposition 11 at the end.
First, we introduce the concept of reproducing kernel Hilbert spaces.

Definition 12. Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on D with scalar product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. A function $k\colon D \times D \to \mathbb{R}$ is called a *reproducing kernel* of $\mathcal{H}$ if

1. $k(\cdot, x) \in \mathcal{H} \quad \forall x \in D$,
2. $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x) \quad \forall x \in D\ \forall f \in \mathcal{H}$.

If $\mathcal{H}$ has a reproducing kernel, it is called a *reproducing kernel Hilbert space (RKHS)*.

Remark 1. As an immediate consequence of the upper definition, we get

$$k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}} \quad \forall x, y \in D.$$

The following theorem, proved by Aronszajn (1950), provides sufficient conditions for a function k to be a reproducing kernel.

Theorem 13 (Moore-Aronszajn)**.** *Let $k\colon D \times D \to \mathbb{R}$, be symmetric and positive definite, that is*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0, \quad \forall n \geq 1\ \forall a \in \mathbb{R}^n\ \forall x \in D^n.$$

Then there is a unique RKHS $\mathcal{H}_k$ with reproducing kernel k.

Against this backdrop, we may ask how properties of the kernel k translate into characteristics of $\mathcal{H}_k$. The notion of a universal kernel, introduced by Steinwart (2002), helps to shed light on this issue.

Definition 14. A continuous kernel k on a compact metric space (D, d) is called *universal* if $\mathcal{H}_k$ is dense in $C(D)$, the space of continuous functions on D, with respect to $\|\cdot\|_\infty$.

It is shown that both the Gaussian and exponential kernel, defined by

$$k(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right),\ \sigma^2 > 0, \qquad k(x,y) = \exp\left(-\frac{\|x-y\|_2}{2\sigma}\right),\ \sigma > 0,$$

respectively, are universal.
Second, we introduce the particularly useful framework of embedding distributions into Hilbert spaces according to Smola et al. (2007).

Definition 15. Let k be a bounded kernel on D and $\mathbb{P}$ a probability measure on D. The *kernel embedding* of $\mathbb{P}$ into the RKHS $\mathcal{H}_k$ is $\mu_k(\mathbb{P}) \in \mathcal{H}_k$ such that

$$\mathrm{E}\left[f(X)\right] = \int_D f(x)\,\mathrm{d}\mathbb{P}(x) = \langle f, \mu_k(\mathbb{P})\rangle_{\mathcal{H}_k}, \quad X \sim \mathbb{P},\ \forall f \in \mathcal{H}_k.$$

Remark 2. Alternatively, $\mu_k(\mathbb{P})$ can be defined by

$$\mu_k(\mathbb{P}) = \int_D k(\cdot, x)\,\mathrm{d}\mathbb{P}(x).$$

Definition 15 allows us to use Hilbert space theory on distributions which gives rise to the definition of maximum mean discrepancy (MMD), see for example (Borgwardt et al., 2006) and (Gretton et al., 2012), which measures the distance between probability measures.

Definition 16. Let k be a bounded kernel and $\mathbb{P}$ and $\mathbb{Q}$ probability measures on D. The *maximum mean discrepancy (MMD)* between $\mathbb{P}$ and $\mathbb{Q}$ with respect to k is defined as

$$\mathrm{MMD}_k(\mathbb{P},\mathbb{Q}) = \|\mu_k(\mathbb{P}) - \mu_k(\mathbb{Q})\|_{\mathcal{H}_k}^2.$$

Lemma 17. *In the setting of Definition 16,* MMD_k *is a metric on probability measures if* k *is a universal kernel.*

Proof. Theorem 1 of (Smola et al., 2007) states that $\mathbb{P} \mapsto \mu_k(\mathbb{P})$ is injective for universal k. Hence, any two different measures have two distinct embeddings. The statement directly follows from the norm properties of $\|\cdot\|_{\mathcal{H}_k}$. □

The maximum mean discrepancy can be used to test whether two given data samples stem from the same distribution. Since our goal is to find a measure for the dependence between two random variables X and Y, we use MMD to compare the joint distribution $\mathbb{P}_{X,Y}$ and the product of the marginals $\mathbb{P}_X\mathbb{P}_Y$.
To this end, consider any two kernels k and l on the domains D_X and D_Y. It is easy to verify that $K = k \otimes l$, given by

$$K\big((x,y),(x',y')\big) = k(x,x')\,l(y,y'), \quad x,x' \in D_X,\ y,y' \in D_Y,$$

is a valid kernel on the product space $D_X \times D_Y$. Employing Remark 2, we can define a dependence measure between X and Y based on RKHSs.

Definition 18. Let X and Y be random variables and k and l be bounded kernels on the domains D_X and D_Y, respectively. The *Hilbert-Schmidt independence criterion* $\mathrm{HSIC}_{k,l}(X,Y)$ for X and Y based on the kernels k and l is given by

$$\begin{aligned}\mathrm{HSIC}_{k,l}(X,Y) &= \mathrm{MMD}_{k\otimes l}(\mathbb{P}_{X,Y}, \mathbb{P}_X\mathbb{P}_Y)\\ &= \big\|\mathrm{E}_{XY}[k(\cdot,X)\otimes l(\cdot,Y)] - \mathrm{E}_X[k(\cdot,X)]\ \mathrm{E}_Y[l(\cdot,Y)]\ \big\|_{\mathcal{H}_{k\otimes l}}^2. \end{aligned} \tag{1}$$

The name of HSIC stems from the point of view, held by Gretton et al. (2005a). The term within the norm in (1) can be identified with the cross-covariance operator $C_{XY}\colon \mathcal{H}_k \to \mathcal{H}_l$ for which

$$\langle f, C_{XY}\, g\rangle_{\mathcal{H}_k} = \mathrm{cov}\left(f(X), g(Y)\right) \quad \forall f \in \mathcal{H}_k\, \forall g \in \mathcal{H}_l \tag{2}$$

holds. Consequently, HSIC is the squared Hilbert-Schmidt norm $\|C_{XY}\|_{\mathrm{HS}}^2$.
Coming full circle, we see that using universal kernels k and l, which causes $k \otimes l$ to be universal as well, has two important implications. First, Lemma 17 states that HSIC is indeed a valid metric to measure dependence between random variables. Second, Definition 14 yields that $\mathcal{H}_k$ and $\mathcal{H}_l$ are dense in $C(D_X)$ and $C(D_Y)$, respectively. Hence, (2) directly reflects the characterisation of independence given in Proposition 11.
Moreover, it can be shown that Definition 4 and Definition 18 are equivalent; yet, the former is more convenient to develop estimators.

A.2. HSIC-Estimation

Since the introduction of the Hilbert-Schmidt independence criterion several estimators have been proposed. We assume that a data sample $\{(x_j, y_j)\}_{j=1}^n$ is given and that the kernels k and l are universal and w.l.o.g. bounded by 1. Gretton et al. (2005a) proposed a simple estimator, which, however, exhibits a bias of order $\mathcal{O}(n^{-1})$, whereas Song et al. (2012) corrected this unfavourable trait putting forward an unbiased estimator.

Definition 19. Let K and L be defined by $K_{ij} = k(x_i, x_j)$ and $L_{ij} = l(x_i, x_j)$ for $1 \leq i, j \leq n$ and set $\tilde{K} = K - \mathrm{diag}(K)$, $\tilde{L} = L - \mathrm{diag}(L)$ and $\Gamma = \mathrm{Id} - \frac{1}{n} 1 1^T$, where $1 \in \mathbb{R}^n$ has one at every entry. The *biased* and *unbiased HSIC-estimators* $\widehat{\mathrm{HSIC}}_{\mathrm{b}}(X, Y)$ and $\widehat{\mathrm{HSIC}}_{\mathrm{u}}(X, Y)$ are defined as

$$\widehat{\mathrm{HSIC}}_{\mathrm{b}}(X, Y) = \frac{1}{(n-1)^2} \operatorname{tr}(K \Gamma L \Gamma),$$
$$\widehat{\mathrm{HSIC}}_{\mathrm{u}}(X, Y) = \frac{1}{n(n-3)} \left(\operatorname{tr}(\tilde{K}\tilde{L}) + \frac{1^T \tilde{K} 1 \, 1^T \tilde{L} 1}{(n-1)(n-2)} - \frac{2}{n-2} 1^T \tilde{K} \tilde{L} 1 \right).$$

In order to develop and establish properties of estimators, it proves advantageous to use the theory of U-statistics (1948). This broad class of estimators was pioneered by Hoeffding and provides a versatile framework to establish useful properties for a multitude of estimators. We use the definition of (Lee, 1990).

Definition 20. Let $X_1, \ldots, X_n$ be i.i.d. random variables, which take values in a measurable space $(A, \mathcal{A})$ and share the same distribution, and let $h \colon A^k \to \mathbb{R}$ be a symmetric function. We denote $\mathcal{S}_{n,k}$ as the set of all k-subsets of $\{1, \ldots, n\}$. For $n \geq k$,

$$U_n = \binom{n}{k}^{-1} \sum_{(i_1, \ldots, i_k) \in \mathcal{S}_{n,k}} h(X_{i_1}, \ldots, X_{i_k})$$

is a *U-statistic of degree k with kernel h.*

Song et al. (2012) proved that $\widehat{\mathrm{HSIC}}_{\mathrm{u}}$ indeed has an according representation.

Theorem 21. *Using the notation of Definition 19, $\widehat{\mathrm{HSIC}}_{\mathrm{u}}$ is a U-statistic of degree 4 with kernel*

$$h(i, j, q, r) = \frac{1}{24} \sum_{(s,t,u,v)}^{(i,j,q,r)} K_{st}(L_{st} + L_{uv} - 2L_{su}).$$

The sum is taken over all 24 quadruples (s, t, u, v) that can be selected without replacement from (i, j, q, r) and the notation of h was reduced to only contain the indices.

A.3. Lasso Formulation of Normal Weighted HSIC-Lasso

We consider a normal weighted HSIC-Lasso selection with associated estimate

$$\hat{\beta} = \underset{\beta \in \mathbb{R}_+^p}{\operatorname{argmin}} -\beta^T H + \frac{1}{2} \beta^T M \beta + \lambda \beta^T w, \tag{3}$$

according to Definition 8.
Assuming that M is positive definite, we can reformulate (3) in terms of a Lasso-problem as follows

$$\hat{\beta} = \underset{\beta \in \mathbb{R}_+^p}{\operatorname{argmin}} \frac{1}{2} \|Y - U\beta\|_2^2 + \lambda \beta^T w.$$

U is determined by the Cholesky decomposition $M = U^T U$ and Y is the solution to $H = U^T Y$. This formulation facilitates the computation of the estimate as there is a variety of efficient algorithms and software packages for Lasso problems available. These are tailored to optimise expressions with a regularisation term and, therefore, yield sparse solutions.

B. Hypothesis Testing

We assume that the response Y follows a normal distribution $\mathcal{N}(\mu, \Sigma)$, where μ is unknown and Σ is given, and that the selection event can be represented as a polyhedron, i.e. $\{\hat{S} = S\} = \{AY \leq b\}$. Furthermore, the quantity of interest can be expressed as $\eta_S^T \mu$, but we drop the dependence on S in the following. Tibshirani et al. (2016) described how one- and two-sided hypothesis testing and confidence interval calculation can be done in this setting. Suppose we want to test

$$\mathrm{H}_0 : \eta_S^T \mu = 0 \quad \text{against} \quad \mathrm{H}_1 : \eta_S^T \mu > 0.$$

Then the statistic

$$T_1 = 1 - F_{0, \eta^T \Sigma \eta}^{[\mathcal{V}^-(Z), \mathcal{V}^+(Z)]}(\eta^T Y)$$

is a valid p-value for H_0 conditional on $\{AY \leq b\}$. Further, defining δ_α for $0 \leq \alpha \leq 1$ such that

$$1 - F_{\delta_\alpha, \eta^T \Sigma \eta}^{[\mathcal{V}^-(Z), \mathcal{V}^+(Z)]}(\eta^T Y) = \alpha$$

yields a valid one-sided confidence interval $[\delta_\alpha, \infty)$ conditional on $\{AY \leq b\}$.
Likewise, we consider the two-sided hypothesis testing problem

$$\mathrm{H}_0 : \eta_S^T \mu = 0 \quad \text{against} \quad \mathrm{H}_1 : \eta_S^T \mu \neq 0$$

and use the statistic

$$T_2 = 2 \min \left\{ F_{0, \eta^T \Sigma \eta}^{[\mathcal{V}^-(Z), \mathcal{V}^+(Z)]}(\eta^T Y), 1 - F_{0, \eta^T \Sigma \eta}^{[\mathcal{V}^-(Z), \mathcal{V}^+(Z)]}(\eta^T Y) \right\}.$$

Again, T_2 is a valid conditional p-value and defining $\delta_{\alpha/2}$ and $\delta_{1-\alpha/2}$ such that

$$1 - F_{\delta_{\alpha/2}, \eta^T \Sigma \eta}^{[\mathcal{V}^-(Z), \mathcal{V}^+(Z)]}(\eta^T Y) = \alpha/2,$$
$$1 - F_{\delta_{1-\alpha/2}, \eta^T \Sigma \eta}^{[\mathcal{V}^-(Z), \mathcal{V}^+(Z)]}(\eta^T Y) = 1 - \alpha/2$$

yields a valid confidence interval $[\delta_{\alpha/2}, \delta_{1-\alpha/2}]$ conditional on $\{AY \leq b\}$.

C. Proofs

C.1. Intermediary Results

This subsection collects technical results that will be used in the proofs of the following subsections. First, we state an auxiliary lemma that corresponds to Lemma A in Section 4.3.3 of (Lee, 1990).

Lemma 22. *Let the random variables $Z_1, \ldots, Z_N$ have a multinomial distribution* $\mathrm{Mult}(m; N^{-1}, \ldots, N^{-1})$ *and let $(a_i)_{i \in \mathbb{N}}$ be a sequence having the properties $\lim_{N \to \infty} N^{-1} \sum_{i=1}^N a_i = 0$ and $\lim_{N \to \infty} N^{-1} \sum_{i=1}^N a_i^2 = \sigma^2$. Then,*

$$m^{-\frac{1}{2}} \sum_{i=1}^N a_i (Z_i - m/N) \xrightarrow{D} \mathcal{N}(0, \sigma^2), \quad \text{as } m, N \to \infty.$$

Korolyuk & Borovskikh presented a multidimensional version of the central limit theorem for U-statistics (1994).

Theorem 23. *Let $U_n^{(1)}, \ldots, U_n^{(m)}$ be U-statistics according to Definition 20 and let $X_1^{(i)}, \ldots, X_n^{(i)}, i \in \{1, \ldots, m\}$, be the corresponding i.i.d. random variables. The respective kernels, degrees and expectations are denoted by $h^{(i)}$, $k^{(i)}$ and $\theta^{(i)}$. We introduce the definitions*

$$\psi^{(i)}(x) := \mathrm{E}\left[h^{(i)}(x, X_2^{(i)}, \ldots, X_{k^{(i)}}^{(i)}) - \theta^{(i)} \right], \quad \sigma^{(i,j)} := \mathrm{E}\left[\psi^{(i)}(X_1^{(i)}) \, \psi^{(j)}(X_1^{(j)}) \right], \quad i, j \in \{1, \ldots, m\}.$$

If $\sigma^{(i,i)} > 0$ and $\mathrm{E}\left[\left(h^{(i)}(X_1^{(i)}, \ldots, X_{k^{(i)}}^{(i)}) \right)^2 \right] < \infty$ hold for all $i \in \{1, \ldots, m\}$, then

$$\sqrt{n} \begin{pmatrix} (U_n^{(1)} - \theta^{(1)})/k^{(1)} \\ \vdots \\ (U_n^{(m)} - \theta^{(m)})/k^{(m)} \end{pmatrix} \xrightarrow{D} \mathcal{N}(0, \Sigma), \quad \text{as } n \to \infty,$$

where the elements of Σ are given by $\Sigma_{ij} = \sigma^{(i,j)}, i, j \in \{1, \ldots, m\}$.

C.2. Proof of Theorem 6

The statement for H_{block} is a direct consequence of the multidimensional central limit theorem. The expression $\sqrt{n/B}\big(H_{\text{block}} - H_0\big)$ can be written as

$$\sqrt{n/B}\left(\frac{1}{n/B}\sum_{b=1}^{n/B}\begin{pmatrix}\widehat{\text{HSIC}}_{\text{u}}(X^{b,(1)},Y^b)\\ \vdots \\ \widehat{\text{HSIC}}_{\text{u}}(X^{b,(p)},Y^b)\end{pmatrix} - \begin{pmatrix}\text{HSIC}(X^{(1)},Y)\\ \vdots \\ \text{HSIC}(X^{(p)},Y)\end{pmatrix}\right).$$

The n/B random variables in the sum are independent and identically distributed due to the i.i.d. assumption and data subdivision. Moreover, the involved estimators are unbiased and $n/B \to \infty$.

In order to prove the second statement of Theorem 6, we use an adaptation of the one-dimensional proof of asymptotic normality for an incomplete U-statistics estimator using random subset selection, cf. Theorem 1 in Section 4.3.3 of (Lee, 1990). We prove multidimensional convergence with the Cramér-Wold device, see e.g. Theorem 11.2.3 of (Lehmann & Romano, 2005). That is it suffices to show that $\sqrt{m}\,\nu^T\big(H_{\text{inc}} - H\big)$ converges to a one-dimensional Gaussian distribution as $m \to \infty$ for any $\nu \in \mathbb{R}^p$.
We introduce the independent random vectors $Z^{(j)}, j \in \{1,\dots,p\}$ and index their entries with $\mathcal{S}_{n,4}$; hence, their elements are $\{Z_S^{(j)} \colon S \in \mathcal{S}_{n,4}\}$. All of them follow a multinomial distribution $\text{Mult}(m; N^{-1},\dots,N^{-1})$, where $N = \binom{n}{4}$. Hence, we can write

$$m^{\frac{1}{2}}\,\nu^T\big(H_{\text{inc}} - H\big) = m^{-\frac{1}{2}}\,\nu^T \sum_{S\in\mathcal{S}_{n,4}} Z_S\big(h(S) - H\big), \tag{4}$$

where the sum as well as the product within is to be understood componentwise, and $Z = (Z^{(1)},\dots,Z^{(p)})$ as well as h are used in a vectorised way, slightly abusing notation. In order to derive the asymptotic distribution of (4), we consider its characteristic function ϕ_n. In the following manipulations we drop the indices for the summation $\sum_{S\in\mathcal{S}_{n,4}}$, introduce the notation $X^{(j)} = (X_1^{(j)},\dots,X_n^{(j)}), j \in \{1,\dots,p\}$, and Y accordingly, and denote the p-dimensional vector of (complete) U-statistics by U_n, that is the vector of unbiased HSIC-estimators. Then:

$$\begin{aligned}
\phi_n(t) &= \text{E}\left[\exp\left(it\,m^{-\frac{1}{2}}\,\nu^T\sum Z_S\big(h(S)-H\big)\right)\right]\\
&= \text{E}\left[\text{E}\left[\exp\left(it\,m^{-\frac{1}{2}}\,\nu^T\sum Z_S\big(h(S)-H\big)\right)\middle| X^{(1)},\dots,X^{(p)},Y\right]\right]\\
&= \text{E}\left[\exp\left(it\,m^{\frac{1}{2}}\sum_{j=1}^{p}\nu_j(U_n^{(j)}-H_j)\right)\text{E}\left[\exp\left(it\,m^{-\frac{1}{2}}\sum_{j=1}^{p}\nu_j\sum\left(Z_S^{(j)}-\frac{m}{N}\right)\big(h_j(S)-H_j\big)\right)\middle| X^{(1)},\dots,X^{(p)},Y\right]\right]\\
&= \text{E}\left[\exp\left(it\,m^{\frac{1}{2}}\sum_{j=1}^{p}\nu_j(U_n^{(j)}-H_j)\right)\text{E}\left[\exp\left(it\,m^{-\frac{1}{2}}\nu_j\sum\left(Z_S^{(j)}-\frac{m}{N}\right)\big(h_j(S)-H_j\big)\right)\middle| X^{(1)},\dots,X^{(p)},Y\right]\right].
\end{aligned}$$

In the manipulations above we used the tower law of conditional expectation and the independence of the $Z_S^{(j)}, j \in \{1,\dots,p\}$. Moreover, we inserted $\pm\, m(U_n - H) = m/N\sum(h(S) - H)$.
Having separated the randomness coming from the data and the subset selection, we treat the second factor in the product above. Standard U-statistics theory implies that

$$\lim_{N\to\infty} N^{-1}\sum_{S\in\mathcal{S}_{n,4}}(h_j(S)-H_j) = 0 \quad\text{and}\quad \lim_{N\to\infty} N^{-1}\sum_{S\in\mathcal{S}_{n,4}}(h_j(S)-H_j)^2 = \sigma_j^2$$

almost surely, where Song et al. (2012) stated a formula for σ_j^2. Ergo, the requirements of Lemma 22 are fulfilled and applying it together with the dominated convergence theorem yields

$$\begin{aligned}
\lim_{n\to\infty}\phi_n(t) &= \lim_{n\to\infty}\text{E}\left[\exp\left(it\,m^{\frac{1}{2}}\textstyle\sum_{j=1}^{p}\nu_j(U_n^{(j)}-H_j)\right)\right]\prod_{j=1}^{p}\exp\big(-(\sigma_j\nu_j)^2t^2/2\big)\\
&= \lim_{n\to\infty}\text{E}\left[\exp\left(it\sqrt{m/n}\,\nu^T\sqrt{n}\,(U_n-H)\right)\right]\prod_{j=1}^{p}\exp\big(-(\sigma_j\nu_j)^2t^2/2\big).
\end{aligned}$$

Lim et al. (2020) pointed out that $\sigma^{(i,i)} = 0$, according Theorem 23, holds for $Y \perp\!\!\!\perp X^{(i)}$, whereas $\sigma^{(i,i)} > 0$ is true if the response and the i-th covariate are dependent. Therefore, we define the index set $I := \{i\colon \mathrm{HSIC}(X^{(i)}, Y) > 0\}$ and the positive definite matrix Ξ by $\Xi_{ij} = 16\,\sigma^{(i,j)}, i, j \in I$. Using Theorem 23 and Slutsky's theorem, we arrive at

$$\lim_{n\to\infty} \phi_n(t) = \exp\big(-(\sqrt{l}\,\nu_I^T\,\Xi\,\nu_I)\,t^2/2\big) \prod_{j=1}^{p} \exp\big(-(\sigma_j\nu_j)^2 t^2/2\big).$$

The limit of ϕ_n is clearly a Gaussian characteristic function which proves asymptotic normality. □

C.3. Proof of Theorem 9

We prove Theorem 9 in two steps: First, we establish $(H_n, M_n, \widehat{\Sigma}_n)|\{A_nH_n \le b_n\} \to (H, M, \Sigma)|\{AH \le b\}$ in distribution; second, we apply the continuous mapping theorem (CMT), see for example Theorem 2.3 (i) in (van der Vaart, 1998), to the cdf of a truncated Gaussian. The ideas of this proof are heavily influenced by (Tibshirani et al., 2018).

Let S be a set of selected covariates and $\tilde{S} \subseteq S$. In the following, we use the abbreviations $\eta_n = \eta_{\tilde{S}}(M_n)$, $A_n = A_{\tilde{S}}(M_n)$ and $b_n = b_{\tilde{S}}(M_n)$. Applying Theorem 2.7 (v) of (van der Vaart, 1998), we get $(H_n, M_n) \to (H, M)$ in distribution, where H has the law $\mathcal{N}(\mu, \Sigma)$. Furthermore, due to the independence of $\widehat{\Sigma}_n$ from (H_n, M_n) we can easily extend the convergence to $(H_n, M_n, \widehat{\Sigma}_n) \to (H, M, \Sigma)$. Ultimately, since A and b are almost surely continuous, the CMT yields $(H_n, M_n, \widehat{\Sigma}_n, A_nH_n - b_n) \to (H, M, \Sigma, AH - b)$ in distribution and we define $\Gamma_n := (H_n, M_n, \widehat{\Sigma}_n)$.
We arrange the components of Γ_n in a vector, fix an arbitrary $x \in \mathbb{R}^{p+2p^2}$ and analyse the asymptotics of the conditional distribution $(\Gamma_n | A_nH_n \le b_n)$

$$\begin{aligned}\mathbb{P}\left(\Gamma_n \le x \,|\, A_nH_n - b_n \le 0\right) &= \frac{\mathbb{P}\left(\Gamma_n \le x, A_nH_n - b_n \le 0\right)}{\mathbb{P}\left(A_nH_n - b_n \le 0\right)} \\ &\to \frac{\mathbb{P}\left(\Gamma \le x, AH - b \le 0\right)}{\mathbb{P}\left(AH - b \le 0\right)} = \mathbb{P}\left(\Gamma \le x \,|\, AH - b \le 0\right), \quad \text{as } n \to \infty.\end{aligned}$$

This is true because both the numerator and denominator converge to the respective probabilities and the denominator is bounded away from zero as the interior of $\{AH \le b\}$ is not empty. Thus, we have shown that

$$(H_n, M_n, \widehat{\Sigma}_n)|\{A_nH_n \le b_n\} \xrightarrow{\mathrm{D}} (H, M, \Sigma)|\{AH \le b\}, \quad \text{as } n \to \infty. \tag{5}$$

In order to use this convergence result for

$$F^{[\mathcal{V}^-(Z_n), \mathcal{V}^+(Z_n)]}_{\eta_n^T\mu, \eta_n^T\widehat{\Sigma}_n\eta_n}(\eta_n^T H_n)|\{A_nH_n \le b_n\}, \tag{6}$$

where Z_n is defined according to Lemma 1, we have to verify that this expression is an a.s. continuous function of $(H_n, M_n, \widehat{\Sigma}_n)$. The cdf of a truncated Gaussian random variable $F^{[x_4,x_5]}_{x_1,x_2}(x_3)$ has five arguments and is continuous, if $x_4 < x_5$ holds.
In (6), we plug in $\eta_n^T\mu$, $\eta_n^T\widehat{\Sigma}_n\eta_n$ and $\eta_n^T H_n$ for x_1, x_2 and x_3, respectively, which are, by assumption, a.s. continuous with respect to Γ_n. The truncation points $\mathcal{V}^-(Z_n)$ and $\mathcal{V}^+(Z_n)$ are the maximum and minimum of finite sets of continuous functions. We denote these two sets G^- and G^+. Hence, all discontinuity points of $\mathcal{V}^-(Z_n)$ and $\mathcal{V}^+(Z_n)$ are contained in $E = \bigcup_{j=1}^{p}\{\mathrm{e}_j^T A_nC_n = 0\}$, i.e. the set of points where the functions contained in G^- and/or G^+ change. As a finite union of lower-dimensional subspaces, E is a null set. Therefore, $\mathcal{V}^-(Z_n)$ and $\mathcal{V}^+(Z_n)$ are almost surely continuous functions of Γ_n. Moreover, we deduce from the definition of the truncation points that

$$\mathcal{V}^-(Z) = \mathcal{V}^+(Z) \quad \Leftrightarrow \quad \eta^T H = \frac{b_j - (AZ)_j}{(AC)_j} \quad \forall j \in J,$$

where J is defined as $\{j\colon (AC)_j \neq 0\}$. Rearranging these equations, we arrive at $\{\mathcal{V}^-(Z) = \mathcal{V}^+(Z)\} = \{(AH)_j = b_j\,\forall j \in J\}$. As a lower-dimensional subspace this set has measure zero and, consequently, $\mathcal{V}^-(Z) < \mathcal{V}^+(Z)$ holds almost surely. In summary, (6) depends on $(H_n, M_n, \widehat{\Sigma}_n)$ in an a.s. continuous fashion. Using (5), the CMT and Theorem 3, we obtain

$$F^{[\mathcal{V}^-(Z_n), \mathcal{V}^+(Z_n)]}_{\eta_n^T\mu, \eta_n^T\widehat{\Sigma}_n\eta_n}(\eta_n^T H_n)|\{A_nH_n \le b_n\} \xrightarrow{\mathrm{D}} F^{[\mathcal{V}^-(Z), \mathcal{V}^+(Z)]}_{\eta^T\mu, \eta^T\Sigma\eta}(\eta^T H)|\{AH \le b\} \sim \mathrm{Unif}\,(0,1).$$

□

C.4. Proof of Theorem 10

In order to characterise the selection event $\{\hat{S} = S\}$, we assume w.l.o.g. that the first $|S|$ covariates of $\{X_1, \ldots, X_p\}$ were included into the model. We rely on the Karush-Kuhn-Tucker (KKT) conditions, cf. Section 5.5.3 of (Boyd & Vandenberghe, 2004), that identify the solution of an optimisation problem by a set of equations and inequalities. Since the function to be minimised is convex due to the positive definiteness of M and Slater's condition, cf. Section 5.2.3 of (Boyd & Vandenberghe, 2004), the KKT conditions provide an equivalent characterisation of the solution of the HSIC-Lasso problem. We obtain

$$\begin{aligned} 0 &= -H + M\hat{\beta} + \lambda w - u, \\ \beta_j \geq 0, \qquad u_j \geq 0, \qquad \beta_j u_j &= 0, \qquad \forall j \in \{1, \ldots, p\}. \end{aligned} \tag{7}$$

We partition the upper inequalities along S and S^c and get

$$\hat{\beta}_S = M_{SS}^{-1}(H_S - \lambda\, w_S), \tag{8}$$

$$0 \leq H_{S^c} + (M\hat{\beta})_{S^c} - \lambda\, w_{S^c}. \tag{9}$$

These results translate into two set of inequalities. First, all entries of $\hat{\beta}$ must be non-negative which implies

$$0 \leq M_{SS}^{-1}(H_S - \lambda\, w_S) \quad \Leftrightarrow \quad -\lambda^{-1}\left(M_{SS}^{-1} \,|\, 0\right) H \leq -M_{SS}^{-1}\, w_S.$$

Second, $M\hat{\beta} = M_S\hat{\beta}_S$ holds by definition of $\hat{S}$. Hence, we can plug (8) into (9) and obtain

$$\begin{aligned} & 0 \leq H_{S^c} + M_{SS^c}\big(M_{SS}^{-1}(H_S - \lambda\, w_S)\big) - \lambda\, w_{S^c} \\ \Leftrightarrow \quad & -\lambda^{-1}\left(M_{SS^c}M_{SS}^{-1} \,|\, \mathrm{Id}\right) H \leq w_{S^c} - M_{SS^c}M_{SS}^{-1}\, w_S. \end{aligned}$$

Both these set of inequalities describe the selection in an affine linear fashion. In this setting, we can use the polyhedral lemma to compute the truncation points $\mathcal{V}^-$ and $\mathcal{V}^+$.

For the selection event $\{j \in \hat{S}\}$, we again use the KKT conditions (7). For any $j \in S$, u_j equals zero and we can express H_j as follows

$$H_j = \mathrm{e}_j^T\left(M\hat{\beta} + \lambda w\right) = \mathrm{e}_j^T\left(M\hat{\beta}_{-j} + \lambda w\right) + M_{jj}\hat{\beta}_j > \mathrm{e}_j^T\left(M\hat{\beta}_{-j} + \lambda w\right),$$

where $\hat{\beta}_{-j}$ denotes $\hat{\beta}$ with the j-th entry set to zero. This estimation holds true as $\hat{\beta}_j$ is positive by definition of $\hat{S}$ and $M_{jj} = \mathrm{e}_j^T M \mathrm{e}_j > 0$ because M is positive definite. Rearranging the inequality, we obtain

$$-\mathrm{e}_j^T H < -\mathrm{e}_j^T M\hat{\beta}_{-j} - \lambda w_j.$$

□

Remark 3. Since the selection event $\{j \in \hat{S}\}$ is less complex than $\{\hat{S} = S\}$, it is possible to directly derive the truncation points without the need for the polyhedral lemma.
To this end, we decompose H into a component in direction of η and one perpendicular to η

$$H = (\eta^T H) \cdot C + Z.$$

Again, we apply the KKT conditions (7) and obtain

$$0 = (\eta^T H) \cdot C - Z + M\hat{\beta} + \lambda w - u,$$

with $u \in \mathbb{R}_+^p$. Since $j \notin \hat{S} \Leftrightarrow \hat{\beta}_j = 0$ holds by definition of $\hat{S}$, the inequality

$$0 \leq \mathrm{e}_j^T\left[(\eta^T H) \cdot C - Z + M\hat{\beta}_{-j} + \lambda w\right], \tag{10}$$

ensues for this case. Rearranging (10), we find

$$\eta^T H \leq \frac{1}{\mathrm{e}_j^T C}\left[\mathrm{e}_j^T M\hat{\beta}_{-j} - \mathrm{e}_j^T Z + \lambda w_j\right]. \tag{11}$$

Consequently, for the event $\{j \in \hat{S}\}$ the lower truncation point $\mathcal{V}^-(Z)$ is the RHS of (11) and $\mathcal{V}^+(Z) = \infty$.

D. Pseudocode of the Algorithm

Along with the description of the algorithm in Section 3.3.4, we give a more detailed account on the different steps of our PSI-procedure for HSIC-Lasso in the following.

Algorithm 1 Post-selection inference for HSIC-Lasso selection with HSIC- or partial target

Input: data $(\mathbf{X}^n, \mathbf{Y}^n)$; level α; inference target t; split ratio s; number of screened variables P; screen-, M- and H-estimators $\mathrm{e}_s, \mathrm{e}_M, \mathrm{e}_H$
Output: significant variables I_{sig}
$(\mathbf{X}^{n,1}, \mathbf{Y}^{n,1}), (\mathbf{X}^{n,2}, \mathbf{Y}^{n,2}) \leftarrow \text{split}((\mathbf{X}^n, \mathbf{Y}^n), s)$
{1st fold}
$H^{(1)} \leftarrow \text{estimate}_H(\mathbf{X}^{n,1}, \mathbf{Y}^{n,1}, \mathrm{e}_s)$
$I_{sc} \leftarrow \text{screening}(H^{(1)}, P)$
$M^{(1)} \leftarrow \text{estimate}_M(\mathbf{X}^{n,1}_{I_{sc}}, \mathrm{e}_s)$
$\tilde{M}^{(1)} \leftarrow \text{positive-definite-approximation}(M^{(1)})$
$U_1 \leftarrow \text{cholesky}(\tilde{M}^{(1)});\ Y_1 \leftarrow U_1^{-T} H^{(1)}_{I_{sc}}$
$\lambda \leftarrow \text{cross-validation}(U_1, Y_1)$ {or AIC}
$w \leftarrow \text{weights}(U_1, Y_1)$
{2nd fold}
$H^{(2)} \leftarrow \text{estimate}_H(\mathbf{X}^{n,2}_{I_{sc}}, \mathbf{Y}^{n,2}_{I_{sc}}, \mathrm{e}_H)$
$M^{(2)} \leftarrow \text{estimate}_M(\mathbf{X}^{n,2}_{I_{sc}}, \mathrm{e}_M)$
$\tilde{M}^{(2)} \leftarrow \text{positive-definite-approximation}(M^{(2)})$
$\hat{\Sigma} \leftarrow \text{estimate}_\Sigma(H^{(2)})$
$U_2 \leftarrow \text{cholesky}(\tilde{M}^{(2)});\ Y_2 \leftarrow U_2^{-T} H^{(2)}$
$\hat{\beta} \leftarrow \text{lasso-opimisation}(Y_2, U_2, \lambda, w)$
$S \leftarrow \text{non-zero-indices}(\hat{\beta})$
$I_{sig} \leftarrow \emptyset$
if t is partial target **then**
$$A \leftarrow -\lambda^{-1} \left(\begin{array}{c|c} (\tilde{M}^{(2)}_{SS})^{-1} & 0 \\ \tilde{M}^{(2)}_{S^cS}(\tilde{M}^{(2)}_{SS})^{-1} & \mathrm{Id} \end{array}\right); \quad b \leftarrow \begin{pmatrix} -(\tilde{M}^{(2)}_{SS})^{-1} w_S \\ w_{S^c} - \tilde{M}^{(2)}_{S^cS}(\tilde{M}^{(2)}_{SS})^{-1} w_S \end{pmatrix}$$
end if
for all $j \in S$ **do**
  if t is HSIC-target **then**
    $\eta \leftarrow \mathrm{e}_j$
    $C \leftarrow (\eta^T \hat{\Sigma} \eta)^{-1} \hat{\Sigma}\, \eta; \quad Z \leftarrow (\mathrm{Id} - C\eta^T) H^{(2)}$
    $\mathcal{V}_j^- \leftarrow (\mathrm{e}_j^T C)^{-1} \left[\mathrm{e}_j^T \tilde{M}^{(2)} \hat{\beta}_{-j} - \mathrm{e}_j^T Z + \lambda w_j\right]; \quad \mathcal{V}_j^+ \leftarrow \infty$
  else if t is partial target **then**
    $\eta \leftarrow \mathrm{e}_j^T ((\tilde{M}^{(2)}_{SS})^{-1} \,|\, 0)$
    $\mathcal{V}^-, \mathcal{V}^+ \leftarrow \text{truncation-points}(A, b, \eta, \hat{\Sigma})$
  end if
  $p \leftarrow 1 - F^{[\mathcal{V}^-, \mathcal{V}^+]}_{0, \eta^T \hat{\Sigma} \eta}(\eta^T H)$
  if $p \leq \alpha$ **then**
    $I_{sig} \leftarrow I_{sig} \cup \{j\}$
  end if
end for
return I_{sig}

E. Additional Experiment

We analyse the *Communities and Crime* datset from the UCI Repository, cf. (Dua & Graff, 2017), which was created by (Redmond & Baveja, 2002) and combines socio-economic (1990), law enforcement (1990) and crime data (1995) from 1 994 different US communities.[1] The authors provide 122 (numerical) features to predict the total number of violent crimes per 100 000 inhabitants. Since the law enforcement data is complete for only 319 communities, we carry out two different analyses: First, we examine the subset of 319 datapoints with complete features; then, we consider the whole dataset but only use covariates without missing data. (Moreover, we delete one data point with a missing value in a feature that is not part of the law enforcement data.)

We highlight that our approach does not target causal relationships; instead, it merely concerns the associations between a feature and the absence or presence of violent crimes. Moreover, as we rely on the Hilbert-Schmidt independence criterion, the existence of a relationship only but not its strength or direction is captured. Lastly, we emphasise that our analysis of the dataset focuses on the characteristics of the proposed methods, rather than on drawing sociological conclusions. To the latter end, a more in-depth analysis is necessary.

Since the number of covariates is manageable, we do not screen features but use 25 % and 20 % of the data, respectively, to determine the regularisation parameter of the non-adaptive HSIC-Lasso via 10-fold cross-validation. We employ the Gaussian kernel throughout as the data is continuous and use the unbiased HSIC-estimator for the matrix M. Both the block and the incomplete estimators are examined, each with different sizes, the number of selected variables for Multi is set to $k = 10$ and we use the confidence level $\alpha = 0.05$.

We depict the findings of the two different analyses in Table 1 and 2, respectively. First, we notice that selection via HSIC-Lasso effectively reduces the number of highly correlated features among the selected covariates compared to HSIC-ordering. For instance, the latter chooses the features *racePctBlack* and *racePctWhite*, which describe the percentage of the population which is African American and Caucasian, respectively, and are presumably highly correlated, for every applied HSIC-estimator in both analyses. On the contrary, HSIC-Lasso almost always only selects *racePctWhite*. This behaviour becomes even more evident when we consider the features *pctFam2Par*, *pctKids2Par*, *pctYoungKids2Par* and *pctTeen2Par* which denote the percentage of two parent families with children and children, young children and teenagers in family housing with two parents, respectively. Due to the strong association of these covariates, HSIC-ordering regularly selects all four of them whereas HSIC-Lasso mostly chooses only one of them as they are highly dependent.

In both analyses, we notice that there is only moderate disagreement among the different HSIC-estimators in terms of the selected features; though, it is more pronounced in the first analysis as the sample size is lower and the number of covariates higher. Moreover, we observe that the incomplete HSIC-estimator, in particular with larger sizes, yields a higher number of accepted covariates than the block estimator. We attribute this behaviour to the fact that the accuracy of the estimates grows with the sizes l and B, respectively; however, we cannot choose the block size too large in order to preserve the normality assumption whereas there exists no such restriction for the incomplete HSCI-estimator. For this reason, we argue to particularly focus on the results in the I20 columns.

Furthermore, the *Communities and Crime* dataset highlights the utility of the partial target and thus the benefits of using post-selection inference with HSIC-Lasso instead of HSIC-ordering. Contrary to the HSIC-target which only considers the association between a feature and the outcome, the partial target adjusts for the dependence structure among all selected covariates. In both analyses, this leads to a small set of accepted features that presumably are not dependent on each other to a large degree. Hence, researchers obtain a clearer and more interpretable view on the data compared to using the HSIC-target.

Lastly, we compare the findings of the two different analyses mostly focusing on the results of the partial target for the incomplete HSIC-estimator with size $l = 20$. Some features like *racePctWhite*, *pctKids2Par* and *pctIlleg*, the percent of children born to never married, are agreed upon by both studies. On the other hand, *numStreet*, the number of homeless people counted in the street, *PctVacantBoarded*, the percentage of vacant housing that is boarded up, and *PolicCars*, the number of police cars, are only found significant in the first analysis whereas *pctWInvInc*, the percentage of households with investment/rent income in 1989, *FemalePctDiv*, the percentage of divorced females, *PctHousLess3BR*, the percentage of housing units with less than 3 bedrooms, and *LemasPctOfficDrugUn*, the percentage of officers assigned to drug units, are only signficant in the second analysis. This hints that there might be correlation between features and the presence/absence of missing data stemming from the data collection and/or assimilation giving rise to the discrepancies between the two analyses. Yet, a more in-depth analysis is necessary to dissolve this uncertainty.

[1] The sources of the dataset are (U. S. Department of Commerce, Bureau of the Census), (U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC & Inter-university Consortium for Political and Social Research Ann Arbor, Michigan, 1992), (U.S. Department of Justice, Bureau of Justice Statistics, 1992) and (U.S. Department of Justice, Federal Bureau of Investigation, 1995).

Table 1. Acceptance of the HSIC- and partial target for the subset of data points without missing values ($n = 319, p = 122$). We denote the feature names according to their abbreviation in the UCI Repository and the column names describe the different HSIC estimators and their sizes, e.g. I10 is the shorthand notation for the incomplete HSIC-estimator of size 10. Grey rectangles denote that a feature was selected but rejected to be significant at level $\alpha = 0.05$, whereas a black rectangle means acceptance and significance.

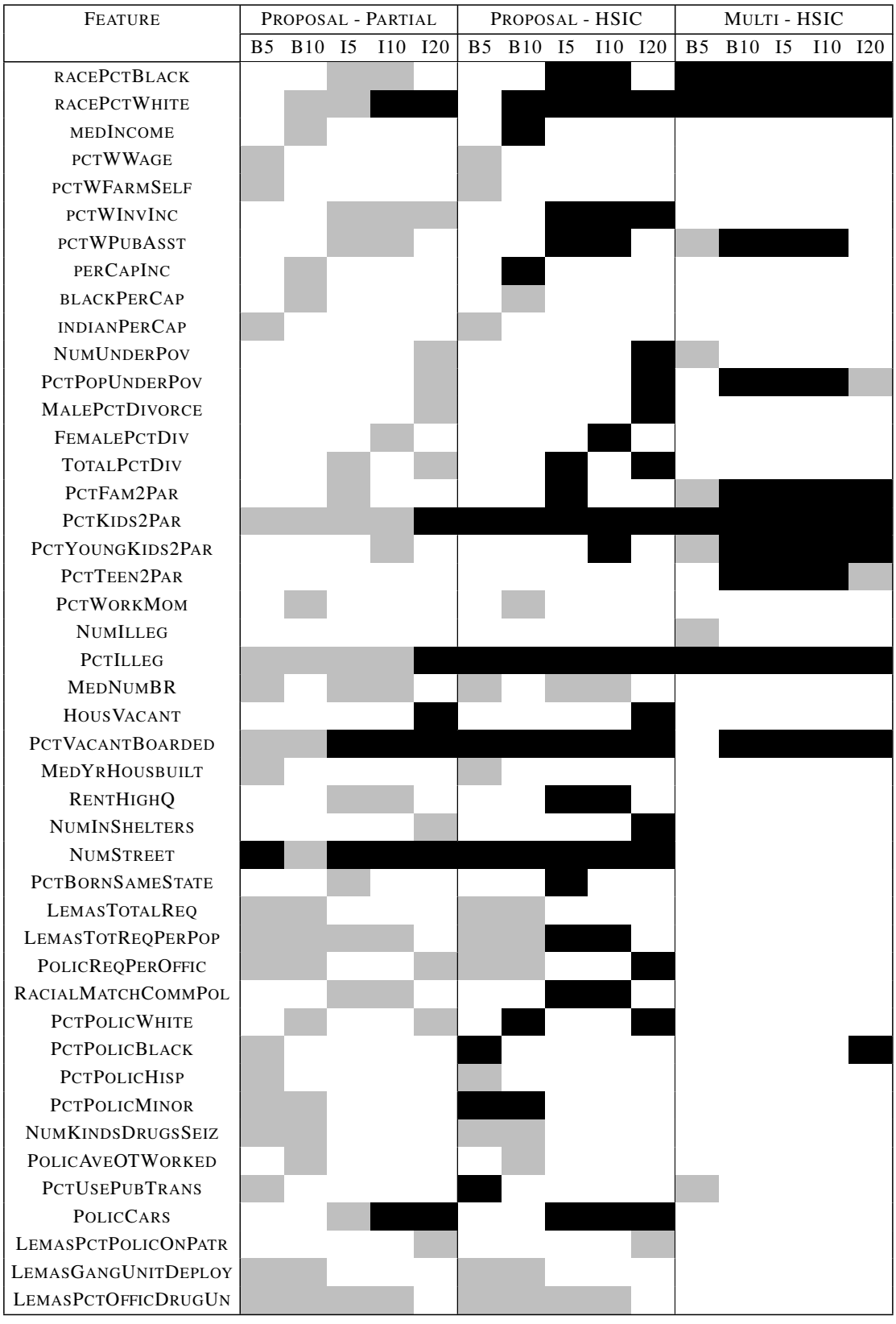
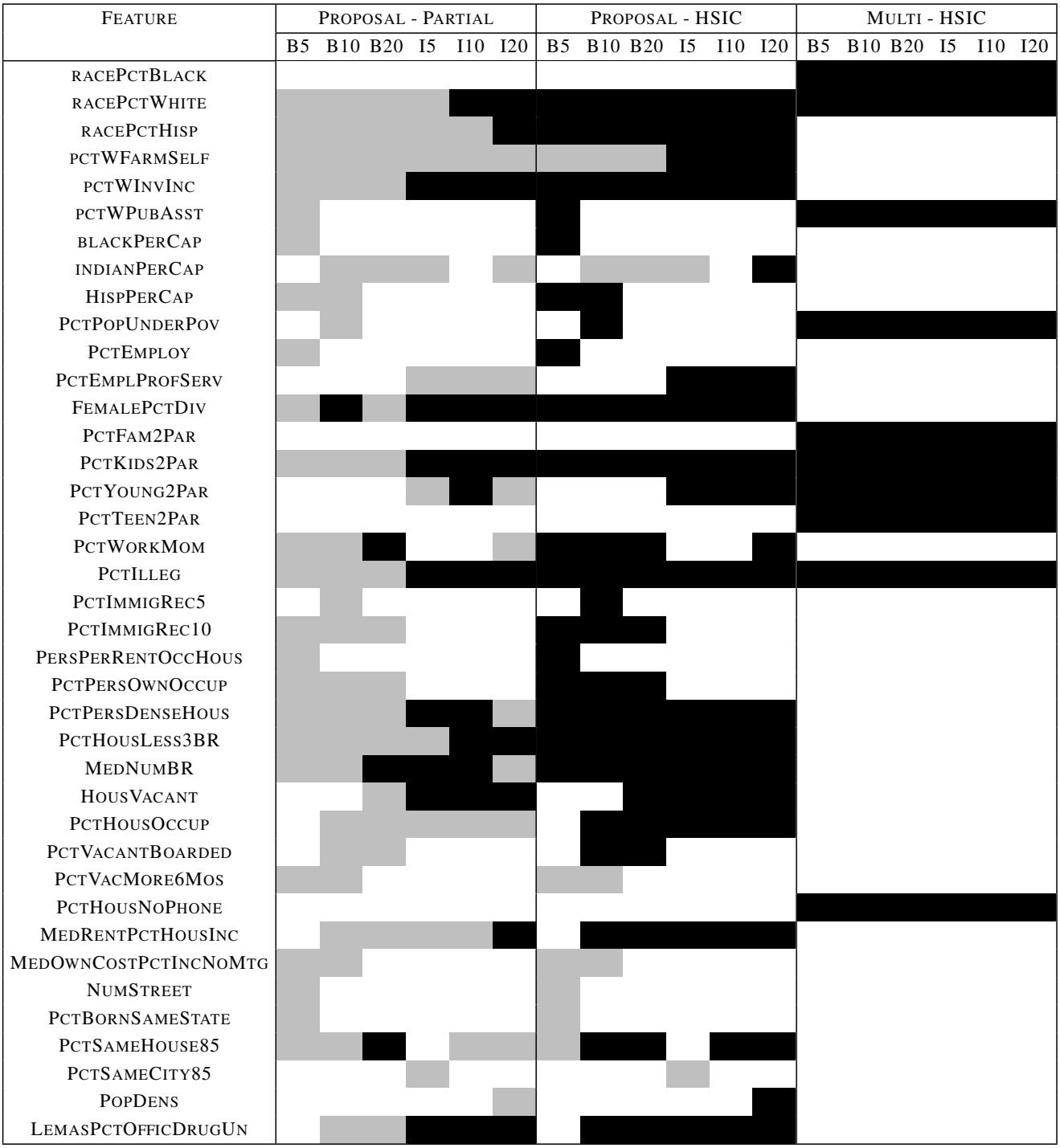


Table 2. Acceptance of the HSIC- and partial target for the subset of data that only contains features without missing data ($n = 1993, p = 100$). We denote the feature names according to their abbreviation in the UCI Repository and the column names describe the different HSIC estimators and their sizes, e.g. I10 is the shorthand notation for the incomplete HSIC-estimator of size 10. Grey rectangles denote that a feature was selected but rejected to be significant at level $\alpha = 0.05$, whereas a black rectangle means acceptance and significance.



References

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

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics*, 22(14):49–57, 2006.

Boyd, S. and Vandenberghe, L. *Convex optimization.* Cambridge University Press, 2004.

Dua, D. and Graff, C. UCI machine learning repository, 2017.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory. 16th international conference, ALT 2005, Singapore, October 8–11, 2005. Proceedings.*, pp. 63–77. Berlin: Springer, 2005a.

Gretton, A., Smola, A., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. Kernel Constrained Covariance for Dependence Measurement. In *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, pp. 1–8, 2005b.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research (JMLR)*, 13:723–773, 2012.

Hoeffding, W. A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, 19:293–325, 1948.

Korolyuk, V. S. and Borovskikh, Y. V. *Theory of U-statistics. Updated and transl. from the Russian by P. V. Malyshev and D. V. Malyshev.* Dordrecht: Kluwer Academic Publishers, 1994.

Lee, A. J. *U-statistics. Theory and practice.*, volume 110. New York etc.: Marcel Dekker, Inc., 1990.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses. 3rd ed.* New York, NY: Springer, 3rd ed. edition, 2005.

Lim, J. N., Yamada, M., Jitkrittum, W., Terada, Y., Matsui, S., and Shimodaira, H. More Powerful Selective Kernel Tests for Feature Selection. volume 108 of *Proceedings of Machine Learning Research*, pp. 820–830. PMLR, 2020.

Redmond, M. and Baveja, A. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *Algorithmic learning theory. 18th international conference, ALT 2007, Sendai, Japan, October 1–4, 2007. Proceedings*, pp. 13–31. Berlin: Springer, 2007.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. Feature selection via dependence maximization. *Journal of Machine Learning Research (JMLR)*, 13:1393–1434, 2012.

Steinwart, I. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research (JMLR)*, 2(1):67–93, 2002.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. Exact Post-Selection Inference for Sequential Regression Procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

Tibshirani, R. J., Rinaldo, A., Tibshirani, R., and Wasserman, L. Uniform asymptotic inference and the bootstrap after model selection. *The Annals of Statistics*, 46(3):1255–1287, 2018.

U. S. Department of Commerce, Bureau of the Census. Census of population and housing 1990 united states: Summary tape file 1a & 3a (computer files).

U.S. Department Of Commerce, Bureau Of The Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan, 1992.

U.S. Department of Justice, Bureau of Justice Statistics. Law enforcement management and administrative statistics (computer file), 1992.

U.S. Department of Justice, Federal Bureau of Investigation. Crime in the united states (computer file), 1995.

van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge: Cambridge Univ. Press, 1998.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. Large-scale kernel methods for independence testing. *Statistics and Computing*, 28(1):113–130, 2018.