Appendix to Supervised Feature Selection via Dependence Estimation

Le Song, Alex Smola, Arthur Gretton and Karsten Borgwardt

1 Proof of Theorem 1

Proof. Define the Pochammer symbol as $(m)_n = \frac{m!}{(m-n)!}$. Also recall that $\mathbf{K}_{ii} = \mathbf{L}_{ii} = 0$. We prove Theorem 1 by constructing unbiased estimator for each term in equation (3) of the main text:

$$\begin{aligned}
\mathsf{E}_{xx'yy'}[k(x,x')l(y,y')] &= \mathsf{E}_Z\left[\frac{1}{(m)_2}\sum_{i\neq j}\mathbf{K}_{ij}\mathbf{L}_{ij}\right] = \mathsf{E}_Z\left[\frac{1}{(m)_2}\mathsf{Tr}(\mathbf{KL})\right] \\
\mathsf{E}_{xx'}[k(x,x')]\mathsf{E}_{yy'}[l(y,y')] &= \mathsf{E}_Z\left[\frac{1}{(m)_4}\sum_{i}\sum_{j\neq i}\sum_{q\neq i,j}\sum_{r\neq i,j,q}\mathbf{K}_{ij}\mathbf{L}_{qr}\right] \\
&= \mathsf{E}_Z\left[\frac{1}{(m)_4}\left(\mathbf{1}^\mathsf{T}\mathbf{K}\mathbf{1}\mathbf{1}^\mathsf{T}\mathbf{L}\mathbf{1} - 4\mathbf{1}^\mathsf{T}\mathbf{KL}\mathbf{1} + 2\mathsf{Tr}(\mathbf{KL})\right)\right] \qquad (1) \\
\mathsf{E}_{xy}[\mathsf{E}_{x'}[k(x,x')]\mathsf{E}_{y'}[l(y,y')]] &= \mathsf{E}_Z\left[\frac{1}{(m)_3}\sum_{i}\sum_{j\neq i}\sum_{q\neq i,j}\mathbf{K}_{ij}\mathbf{L}_{iq}\right] \\
&= \mathsf{E}_Z\left[\frac{1}{(m)_3}\left(\mathbf{1}^\mathsf{T}\mathbf{KL}\mathbf{1} - \mathsf{Tr}(\mathbf{KL})\right)\right]
\end{aligned}$$

Then the relation between $\mathrm{HSIC}(\mathcal{F},\mathcal{G},\mathrm{Pr}_{xy})$ and its unbiased estimator $\mathrm{HSIC}(\mathcal{F},\mathcal{G},Z)$ is

$$\begin{aligned}
\mathrm{HSIC}(\mathcal{F},\mathcal{G},\underset{xy}{\mathrm{Pr}}) =\ & \mathsf{E}_{xx'yy'}[k(x,x')l(y,y')] + \mathsf{E}_{xx'}[k(x,x')]\mathsf{E}_{yy'}[l(y,y')] \\
& - 2\mathsf{E}_{xy}[\mathsf{E}_{x'}[k(x,x')]\mathsf{E}_{y'}[l(y,y')]] \\
& = \mathsf{E}_Z\left[\frac{1}{(m)_2}\mathsf{Tr}(\mathbf{KL})\right] + \mathsf{E}_Z\left[\frac{1}{(m)_4}\left(\mathbf{1}^\mathsf{T}\mathbf{K}\mathbf{1}\mathbf{1}^\mathsf{T}\mathbf{L}\mathbf{1} - 4\mathbf{1}^\mathsf{T}\mathbf{KL}\mathbf{1} + 2\mathsf{Tr}(\mathbf{KL})\right)\right] \\
& - \mathsf{E}_Z\left[\frac{2}{(m)_3}\left(\mathbf{1}^\mathsf{T}\mathbf{KL}\mathbf{1} - \mathsf{Tr}(\mathbf{KL})\right)\right] \\
& = \mathsf{E}_Z\left[\frac{(m-1)(m-2)}{(m)_4}\mathsf{Tr}(\mathbf{KL})\right] + \mathsf{E}_Z\left[\frac{1}{(m)_4}\mathbf{1}^\mathsf{T}\mathbf{K}\mathbf{1}\mathbf{1}^\mathsf{T}\mathbf{L}\mathbf{1}\right] \\
& - \mathsf{E}_Z\left[\frac{2(m-1)}{(m)_4}\mathbf{1}^\mathsf{T}\mathbf{KL}\mathbf{1}\right] \\
& = \mathsf{E}_Z\left[\frac{1}{m(m-3)}\left(\mathsf{Tr}(\mathbf{KL}) + \frac{1}{(m-1)(m-2)}\mathbf{1}^\mathsf{T}\mathbf{K}\mathbf{1}\mathbf{1}^\mathsf{T}\mathbf{L}\mathbf{1} - \frac{2}{m-2}\mathbf{1}^\mathsf{T}\mathbf{KL}\mathbf{1}\right)\right] \\
& = \mathsf{E}_Z\left[\mathrm{HSIC}(\mathcal{F},\mathcal{G},Z)\right] \qquad (2)
\end{aligned}$$

□

2 Proof of Theorem 2

Proof. Denote by $\mathbf{P}_Z$ the probability with respect to m independent observations (x_i, y_i) drawn from Pr_{xy}. Moreover, we split t into $\alpha t + \beta t + (1-\alpha-\beta)t$ where $\alpha, \beta > 0$ and $\alpha + \beta < 1$. The probability of a positive deviation t has bound

$$\begin{aligned}
&\mathbf{P}_Z\{\mathrm{HSIC}(\mathcal{F},\mathcal{G},Z)-\mathrm{HSIC}(\mathcal{F},\mathcal{G},\Pr_{xy})\geq t\}\\
&\leq \mathbf{P}_Z\{\frac{1}{(m)_2}\sum_{i\neq j}\mathbf{K}_{ij}\mathbf{L}_{ij}-\mathsf{E}_{xx'yy'}[k(x,x')l(y,y')]\geq \alpha t\}\\
&+\mathbf{P}_Z\{\frac{1}{(m)_4}\sum_{i}\sum_{j\neq i}\sum_{q\neq i,j}\sum_{r\neq i,j,q}\mathbf{K}_{ij}\mathbf{L}_{qr}-\mathsf{E}_{xx'}[k(x,x')]\mathsf{E}_{yy'}[l(y,y')]\geq \beta t\}\\
&+\mathbf{P}_Z\{-\frac{2}{(m)_3}\sum_{i}\sum_{j\neq i}\sum_{q\neq i,j}\mathbf{K}_{ij}\mathbf{L}_{iq}+2\mathsf{E}_{xy}[\mathsf{E}_{x'}[k(x,x')]\mathsf{E}_{y'}[l(y,y')]]\geq (1-\alpha-\beta)t\}
\end{aligned} \tag{3}$$

Using the shorthand $\mathbf{z}=(x,y)$, we define the kernels of the U-statistics in the three expressions above as $g(\mathbf{z}_i,\mathbf{z}_j)=\mathbf{K}_{ij}\mathbf{L}_{ij}, g(\mathbf{z}_i,\mathbf{z}_j,\mathbf{z}_r)=\mathbf{K}_{ij}\mathbf{L}_{jr}$ and $g(\mathbf{z}_i,\mathbf{z}_j,\mathbf{z}_q,\mathbf{z}_r)=\mathbf{K}_{ij}\mathbf{L}_{qr}$. Frinally employing Hoeffding's Theorem allows us to bound the three probabilities as

$$e^{-2mt^2\frac{\alpha^2}{2}}, e^{-2mt^2\frac{\beta^2}{4}}, \text{ and } e^{-2mt^2\frac{(1-\alpha-\beta)^2}{3}}, \tag{4}$$

Setting the argument of all three exponentials equal yields $\alpha>0.28$: consequently, the positive deviation probability is bounded from above by $3e^{-mt^2\alpha^2}$. Similarly, the negative deviation probability is also bounded by $3e^{-mt^2\alpha^2}$. Thus the overall probability is bounded by doubling this quantity. Solving for t yields the desired result. □

3 Proof of Theorem 3

Proof. We prove this theorem by first relating the biased estimator of HSIC with the biased estimator of MMD. The biased estimator of HSIC is $\frac{1}{(m-1)^2}\mathsf{Tr}(\mathbf{KHLH})$, where $\mathbf{H}=\mathbf{I}-m^{-1}\mathbf{1}\mathbf{1}^{\mathsf{T}}$, and the bias is bounded by $\mathcal{O}(m^{-1})$ [1]; The biased estimator of MMD is

$$\mathrm{MMD}(\mathcal{F},Z)=\frac{1}{m_+^2}\sum_{i,j}^{m_+}k(\mathbf{x}_i,\mathbf{x}_j)+\frac{1}{m_-^2}\sum_{i,j}^{m_-}k(\mathbf{x}_i,\mathbf{x}_j)-\frac{2}{m_+m_-}\sum_{i}^{m_+}\sum_{j}^{m_-}k(\mathbf{x}_i,\mathbf{x}_j)=\mathsf{Tr}(\mathbf{KL}) \tag{5}$$

and the bias is also bounded by $\mathcal{O}(m^{-1})$[2]. Expanding $\mathsf{Tr}(\mathbf{KHLH})$ using $\mathbf{H}$, we obtain $\mathsf{Tr}(\mathbf{KL})+m^{-2}\mathbf{1}^{\mathsf{T}}\mathbf{K}\mathbf{1}\mathbf{1}^{\mathsf{T}}\mathbf{L}\mathbf{1}-2m^{-1}\mathbf{1}^{\mathsf{T}}\mathbf{KL}\mathbf{1}$. Due to our choice of $l(y,y')=\rho(y)\rho(y')$, both $\mathbf{1}^{\mathsf{T}}\mathbf{L}$ and $\mathbf{L}\mathbf{1}$ are zero matrices. Then the last two terms in the expansion vanish, and we obtain $\mathsf{Tr}(\mathbf{KHLH})=\mathsf{Tr}(\mathbf{KL})$ in this case. This means that the biased estimator of HSIC is exactly $(m-1)^{-2}$ times of the biased estimator of MMD. Since the biased estimators of HSIC and MMD both deviates from their unbiased version by only $\mathcal{O}(m^{-1})$, the difference between the unbiased estimators of HSIC and MMD will also be bounded by $\mathcal{O}(m^{-1})$.

The second part is quite similar. Since the empirical KTA is computed as $\langle\mathbf{K},\mathbf{L}\rangle_F$, it is equivalent to $\mathsf{Tr}(\mathbf{KL})$. This means that the empirical KTA is the same as the biased estimator of MMD in this case. □

4 Description of Microarray Datasets

A summary of the microarray datasets and their sources is as follows:

- The six datasets studied in [3]. Three of them deal with breast cancer [4, 5, 6] (numbered 1, 2 and 3), two with lung cancer [7, 8] (4, 5), and one with hepatocellular carcinoma [9] (6). The B cell lymphoma dataset [10] is not used because none of the tested methods produce classification errors lower than 40%.
- The six datasets studied in [11]. Two prostate cancer [12, 13] (7, 8), two breast cancer [14, 15] (9, 10), and two leukaemia [16, 17] (16, 17).
- Five commonly used bioinformatics benchmark datasets on colon cancer [18] (11), ovarian cancer [19] (12), leukaemia [20](13), lymphoma [21](18), and one yeast dataset [22](19).

- Nine datasets from the NCBI GEO database. The GDS IDs and reference numbers for this paper are GDS1962 (20), GDS330 (21), GDS531 (14), GDS589 (22), GDS968 (23), GDS1021 (24), GDS1027 (25), GDS1244 (26), GDS1319 (27), GDS1454 (28), and GDS1490 (15), respectively.

5 Feature Selection via Continuous Relaxation and Zero Norm

Besides the backward elimination algorithm, feature selection using HSIC can also proceed by converting problem (1) in the main text into a continuous optimization problem. This second approach is to jointly optimse a relaxed zero norm of a weight vector over the features, which corresponds to jointly selecting over all features according to a sparsity constraint. This second approach, however, does not perform as good as the the backward elimination procedure proposed in the main text. Hence we postpone its description to this appendix and its experimental result is not reported.

We introduce a weighting $\mathbf{w} \in \mathbb{R}^n$ on the dimensions of the data: $x \longmapsto \mathbf{w} \circ x$, where $\circ$ denotes element-wise product. Thus feature selection using HSIC becomes an optimization problem with respective to $\mathbf{w}$ (for convenience we denote HSIC as function of $\mathbf{w}$, $\mathrm{HSIC}(\mathbf{w})$). To obtain a sparse solution of the selected features, the zero "norm" $\|\mathbf{w}\|_0$ is also incorporated into our objective function. $\|\mathbf{w}\|_0$ computes the number of non-zero entries in $\mathbf{w}$ and the sparsity is achieved by imposing heavier penalty on solutions with large number of non-zero entries. Mathematically, feature selection using HSIC is formulated as:

$$\mathbf{w} = \arg\max_{\mathbf{w}} \ \mathrm{HSIC}(\mathbf{w}) + \lambda||\mathbf{w}||_0, \quad \mathbf{w} \in [0, \infty)^n \tag{6}$$

The zero norm, however, is not a continuous function. It is a sum of a set of step function, and can be approximated with a concave function ($\alpha = 5$ works well in practice):

$$||\mathbf{w}||_0 \approx \mathbf{1}^\mathsf{T}(\mathbf{1} - e^{-\alpha\mathbf{w}}) \tag{7}$$

The optimization problem in (6) is non-convex in general. If we choose a Gaussian kernel for the data and use the biased estimator of HSIC, relatively more efficient optimization can be carried out using the convex-concave procedure (CCCP) [23]. With the added weighting, the Gaussian kernel becomes a convex function $k(x, x') = \exp(-\sigma\|\mathbf{w} \circ x - \mathbf{w} \circ x'\|^2)$ of $\mathbf{w} \in [0, \infty)^n$. Then the objective function in (6) can be decomposed into the difference of two convex functions $g(\mathbf{w})$ and $h(\mathbf{w})$:

$$\begin{aligned}\mathrm{HSIC}(\mathbf{w}) + \lambda\|\mathbf{w}\|_0 &\approx \mathsf{Tr}(\mathbf{K}(\mathbf{I} - m^{-1}\mathbf{1}\mathbf{1}^\mathsf{T})\mathbf{L}(\mathbf{I} - m^{-1}\mathbf{1}\mathbf{1}^\mathsf{T})) + \lambda\mathbf{1}^\mathsf{T}(\mathbf{1} - e^{-\alpha\mathbf{w}}) \\ &= \underbrace{\mathsf{Tr}(\mathbf{K}\mathbf{L} + m^{-2}\mathbf{K}\mathbf{1}\mathbf{1}^\mathsf{T}\mathbf{L}\mathbf{1}\mathbf{1}^\mathsf{T})}_{g(\mathbf{w})} - \underbrace{(\mathsf{Tr}(2m^{-1}\mathbf{K}\mathbf{1}\mathbf{1}^\mathsf{T}\mathbf{L}) - \lambda\mathbf{1}^\mathsf{T}(\mathbf{1} - e^{-\alpha\mathbf{w}}))}_{h(\mathbf{w})}\end{aligned} \tag{8}$$

$g(\mathbf{w})$ consists of a positive combination of entries in the kernel matrix $\mathbf{K}$ and $\mathbf{L}$. If we restrict the kernels to be nonnegative and bounded by 1, we can guarantee the convexity of $g(\mathbf{w})$. $h(\mathbf{w})$ is also convex since it is the difference of a convex part and a concave part. The overall algorithm is presented in Algorithm 2:

Algorithm 1 Feature Selection via CCCP Algorithm

```
Input: training data $Z = \{(x_1, y_1), \ldots, (x_\ell, y_\ell)\}$
Output: the set $\mathcal{T}$ of selected features
initialize $\mathbf{w}$ randomly
repeat
    $\mathbf{w}_0 \leftarrow \mathbf{w}$
    $\mathbf{w} \leftarrow \arg\max_{\mathbf{w}} \ g(\mathbf{w}) - \mathbf{w}^\mathsf{T}\nabla h(\mathbf{w}_0), \quad \mathbf{w} \in [0, \infty)^n$
until $\min(\|\mathbf{w} - \mathbf{w}_0\|, \|\mathbf{1} - \mathbf{w} \circ \mathbf{w}_0^{-1}\|) \leq \varepsilon$
```

References

[1] A. Gretton, O. Bousquet, A.J. Smola, and B. Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In S. Jain and W.-S. Lee, editors, *Proceedings Algorithmic Learning Theory*, 2005.

[2] K. M. Borgwardt, A. Gretton, M.J. Rasch, H.P. Kriegel, B. Schölkopf, and A.J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. In *Proc. of Intelligent Systems in Molecular Biology (ISMB)*, Fortaleza, Brazil, 2006.

[3] L. Ein-Dor, O. Zuk, and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A*, 103(15):5923–5928, Apr 2006.

[4] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[5] M. J. van de Vijver, Y. D. He, L. J. van 't Veer, et al. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, 247:1999–2009, 2002.

[6] Y. Wang, J. G. Klijn, Y. Zhang, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365:671–679, 2005.

[7] A. Bhattacharjee, W. G. Richards, W. G. Staunton, et al. Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci.*, 98:13790–13795, 2001.

[8] D. G. Beer, S. L. Kardia, S. L. Huang, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat. Med.*, 8:816–824, 2002.

[9] N. Iizuka, M. Oka, H. Yamada-Okabe, et al. Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *Lancet*, 361:923–929, 2003.

[10] A. Rosenwald, G. Wright, G. Chan, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *N. Engl. J. Med.*, 346:1937–1947, 2002.

[11] P. Warnat, R. Eils, and B. Brors. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6:265, Nov 2005.

[12] S. M. Dhanasekaran, T. R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan. Delineation of prognostic biomarkers in prostate cancer. *Nature*, 412(6849):822–826, Aug 2001.

[13] J. B. Welsh, L. M. Sapinoso, A. I. Su, S. G. Kern, J. Wang-Rodriguez, C. A. Moskaluk, J. r. Frierson HF, and G. M. Hampton. Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res*, 61(16):5974–5978, Aug 2001.

[14] S. Gruvberger, M. Ringner, Y. Chen, S. Panavally, L. H. Saal, A. Borg, M. Ferno, C. Peterson, and P. S. Meltzer. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res*, 61(16):5979–5984, Aug 2001.

[15] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H Zuzan, J.A. Olson Jr, J.R.Marks, and J.R.Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS*, 98(20), 2001.

[16] L. Bullinger, K. Dohner, E. Bair, S. Frohling, R. F. Schlenk, R. Tibshirani, H. Dohner, and J. R. Pollack. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *N Engl J Med*, 350(16):1605–1616, Apr 2004.

[17] P. J. Valk, R. G. Verhaak, M. A. Beijen, C. A. Erpelinck, S. Barjesteh van Waalwijk van Doorn-Khosrovani, J. M. Boer, H. B. Beverloo, M. J. Moorhouse, P. J. van der Spek, B. Lowenberg, and R. Delwel. Prognostically useful gene-expression profiles in acute myeloid leukemia. *N Engl J Med*, 350(16):1617–1628, Apr 2004.

[18] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci.*, 96:6745–6750, 1999.

[19] A. Berchuck, E. Iversen, and J. Lancaster et al. Patterns of gene expression that characterize long-term survival in advanced stage serous ovarian cancers. *Clin. Cancer Res.*, 11:3686–3696, 2005.

[20] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct 1999.

[21] A. Alizadeh, M. Eisen, R. Davis, et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[22] M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci.*, 97:262–267, 2000.

[23] A.L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15:915–936, 2003.