Consistent and Efficient Nonparametric Different-Feature Selection – Supplementary Material –

Satoshi Hara
National Institute of Informatics, Japan
JST, ERATO, Kawarabayashi Large Graph Project
satohara@nii.ac.jp

Takayuki Katsuki **Hiroki Yanagisawa**
IBM Research – Tokyo, Japan
kats@jp.ibm.com, yanagis@jp.ibm.com

Takafumi Ono
University of Bristol, UK
takafumi.ono@bristol.ac.uk

Ryo Okamoto **Shigeki Takeuchi**
Kyoto University, Japan
okamoto.ryo.4w@kyoto-u.ac.jp, takeuchi@kuee.kyoto-u.ac.jp

7 Relation to Current Methods

The proposed method can be interpreted as a generalization of our previous method [1], which is the first algorithm that uses the sparsest k-subgraph problem for different-feature selection. Unlike the proposed method, the previous method has limited applicability due to the Gaussian assumption. In the previous method, we assumed Gaussian distributions on p and q, and defined the matrix $\hat{L}$ by $\hat{L}_{dd'} := |C^{\mathcal{P}}_{dd'} - C^{\mathcal{Q}}_{dd'}|$, where the matrices $C^{\mathcal{P}}$ and $C^{\mathcal{Q}}$ are the covariance or precision matrices of the datasets $\mathcal{P}$ and $\mathcal{Q}$, respectively. The method is particularly relevant to the proposed method when covariances are used as matrices $C^{\mathcal{P}}$ and $C^{\mathcal{Q}}$. Indeed, the previous method corresponds to minimizing the lower bound of (3) in a specific case described in the next proposition.

Proposition 1 *Suppose p and q are Gaussian distributions with the same mean $\boldsymbol{\mu} \in \mathbb{R}^D$: $p(\boldsymbol{x}) := \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and $q(\boldsymbol{x}) := \mathcal{N}(\boldsymbol{\mu}, \Gamma)$. When both Σ and Γ are invertible and have diagonal components equal to one, $|\Sigma_{dd'} - \Gamma_{dd'}|$ is a lower bound of the KL-divergence $\mathrm{KL}[p(x_d, x_{d'})||q(x_d, x_{d'})]$ up to a constant term.*

8 Baseline Methods

We present the detail of the baseline methods in Section 5.

Notation: $\boldsymbol{\mu}_{\mathcal{P}}, \boldsymbol{\mu}_{\mathcal{Q}} \in \mathbb{R}^d$ and $\Sigma_{\mathcal{P}}, \Sigma_{\mathcal{Q}} \in \mathbb{R}^{D\times D}$ denote empirical averages and covariances of datasets $\mathcal{P}$ and $\mathcal{Q}$, respectively. $\Lambda_{\mathcal{P}}, \Lambda_{\mathcal{Q}} \in \mathbb{R}^{D\times D}$ denote estimated precision matrices of datasets $\mathcal{P}$ and $\mathcal{Q}$ using the Tikhonov-regularized method, respectively. That is, we define $\Lambda_{\mathcal{P}} := (\Sigma_{\mathcal{P}} + \kappa I_D)^{-1}$ and $\Lambda_{\mathcal{Q}} := (\Sigma_{\mathcal{Q}} + \kappa I_D)^{-1}$ with a regularization parameter κ. The value of κ is chosen from 11 different parameter candidates between 10^{-4} and 10^{1} using 3-fold cross validation. For a square matrix $U \in \mathbb{R}^{D\times D}$ and a set $S \subseteq [D]$, we denote the submatrix by $U_S := \{U_{dd'} \mid d, d' \in S\}$.

[MT] [2] We adopted the simplified version of MT for ease of computation. We used combinatorial optimization instead of the F-test in the original MT. The estimated feature set $\hat{S}_{\mathrm{MT}}$ is given by solving the next problem:

$$\hat{S}_{\mathrm{MT}} = \mathrm{argmin}_{S\subseteq[D]} \left| D - \alpha - \mathrm{tr}\left[\Gamma_{S^{\mathrm{c}}} C^{-1}_{S^{\mathrm{c}}}\right] \right|, \quad \text{subject to } |S| = \alpha, \tag{4}$$

where $\Gamma := \frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{y}^{(m)} - \boldsymbol{\mu}_{\mathcal{P}})(\boldsymbol{y}^{(m)} - \boldsymbol{\mu}_{\mathcal{P}})^{\top}$ and $C := \Lambda^{-1}_{\mathcal{P}}$. Because the number α is unknown, we used the greedy scoring method (Algorithm 1) to solve the problem (4), where we defined $f(S) := \left| |S^{\mathrm{c}}| - \mathrm{tr}\left[\Gamma_{S^{\mathrm{c}}} C^{-1}_{S^{\mathrm{c}}}\right] \right|$.

[Idé'09] [3] In Idé'09, the score of the d-th feature $\hat{s}_d$ is given by

$$\hat{s}_d := \max\{\hat{s}^{\mathcal{P}\mathcal{Q}}_d, \hat{s}^{\mathcal{Q}\mathcal{P}}_d\},$$

$$\hat{s}^{\mathcal{P}\mathcal{Q}}_d := \boldsymbol{w}^{\top}_{\mathcal{P}}(\boldsymbol{\ell}_{\mathcal{Q}} - \boldsymbol{\ell}_{\mathcal{P}}) + \frac{1}{2}\left\{\frac{\boldsymbol{\ell}^{\top}_{\mathcal{Q}} W_{\mathcal{P}} \boldsymbol{\ell}_{\mathcal{Q}}}{\lambda_{\mathcal{Q}}} - \frac{\boldsymbol{\ell}^{\top}_{\mathcal{P}} W_{\mathcal{P}} \boldsymbol{\ell}_{\mathcal{P}}}{\lambda_{\mathcal{P}}}\right\} + \frac{1}{2}\left\{\log\frac{\lambda_{\mathcal{P}}}{\lambda_{\mathcal{Q}}} + \sigma_{\mathcal{P}}(\lambda_{\mathcal{P}} - \lambda_{\mathcal{Q}})\right\},$$

where the matrices are partitioned as

$$\Lambda_{\mathcal{P}} = \begin{bmatrix} L_{\mathcal{P}} & \boldsymbol{\ell}_{\mathcal{P}} \\ \boldsymbol{\ell}^{\top}_{\mathcal{P}} & \lambda_{\mathcal{P}} \end{bmatrix}, \ \Lambda^{-1}_{\mathcal{P}} = \begin{bmatrix} W_{\mathcal{P}} & \boldsymbol{w}_{\mathcal{P}} \\ \boldsymbol{w}^{\top}_{\mathcal{P}} & \sigma_{\mathcal{P}} \end{bmatrix}.$$

Here, we assume that the rows and columns of $\Lambda_{\mathcal{P}}$ and $\Lambda^{-1}_{\mathcal{P}}$ are permuted so that their original d-th rows and

columns are located at the last rows and columns of the matrix. The matrices $\Lambda_{\mathcal{Q}}$ and $\Lambda_{\mathcal{Q}}^{-1}$ are partitioned in the same manner.

[Hara'15] [1] Hara'15 uses the sparsest k-subgraph problem (3) similar to the proposed method. The matrix $\hat{L}$ is given by $\hat{L}_{dd'} := |\Sigma_{\mathcal{P},dd'} - \Sigma_{\mathcal{Q},dd'}|$. We used the greedy scoring method (Algorithm 1) to solve the problem.

[SPARDA] [4] The solution of SPARDA $\hat{\boldsymbol{\beta}}$ can be derived by solving the next max-min problem:

$$\max_{\boldsymbol{\beta}\in\mathcal{B}} \min_{U\in\mathcal{M}} \sum_{n=1}^{N}\sum_{m=1}^{M} (\boldsymbol{\beta}^\top \boldsymbol{x}^{(n)} - \boldsymbol{\beta}^\top \boldsymbol{y}^{(m)})^2 U_{nm} - \lambda\|\boldsymbol{\beta}\|_1,$$

where $\mathcal{B} := \{\boldsymbol{\beta} \in \mathbb{R}^D \mid \|\boldsymbol{\beta}\| \leq 1, \beta_1 \geq 0\}$ and $\mathcal{M} = \{U \in \mathbb{R}_+^{N\times M} \mid \forall m, \sum_{n=1}^{N} U_{nm} = 1/M$ and $\forall n, \sum_{m=1}^{M} U_{nm} = 1/N\}$. The minimization term corresponds to computing the Wasserstein distance between the distributions. We implemented SPARDA using C++ based on the MATLAB code `fastSPARDA.m` available on the author's Website [1]. Because the relax and tighten procedure proposed by [4] was too slow, we used the projected gradient ascent which runs in $O(D(N+M) + \log N + \log M)$ per iteration. In our preliminary experiment, we observed that the projected gradient ascent ran more than ten times faster than the relax and tighten procedure. Because the projected gradient ascent tends to be trapped by local optima, we used five random restarts. We set the parameter candidate for λ as $\{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and selected the optimal one using 5-fold cross validation. After we derived the solution $\hat{\boldsymbol{\beta}}$, we set the score of each feature as $\hat{s}_d = |\hat{\beta}_d|$.

9 Extra Results on UCI Datasets

Table 5, 6, and 7 show the results on the UCI datasets for all five changes. These results are similar to Covariance Change (Table 3), and we can observe three important results also, as discussed in Section 5.2.

1. The AUROC of the proposed method attained the best average score among the five methods or comparable scores with the best result for many cases. Even in the complex data (Covariance Change (Conditional)), the proposed method marked the top or top-tier AUROCs for 13 cases out of 15 cases.

2. The proposed method was faster than SPARDA. In particular, the proposed method was from 3 to more than 70 times faster than the entire runtime of SPARDA.

3. The proposed method with the greedy scoring method attained comparable results with the exact solution on the sparsest k-subgraph problem (3). That is, the greedy scoring method could find a nearly global optima.

We note that the proposed method with the greedy scoring method sometimes scored better AUROC than the exact method. This is because the exact method score each feature with 0 or 1. In the exact method, if one feature is misspecified (i.e., scored as 0 instead of 1), that feature is ranked equally with the other features with no distribution differences. This induces a substantial decrease of AUROC because only the order of the scores is important when computing AUROC. By contrast, the greedy scoring method is less sensitive to such a misspecification. Some features may be scored lower than the ideal because of misspecification, but the score on such a feature can still remain a bit high and thus tends to remain in a higher order than the other features with no distribution differences. Hence, the decrease of AUROC will be limited.

10 Quantum Data: Experimental Setup

We experimentally obtained density matrices, for which we use a two photon polarization entangled state, $|\psi(\phi)\rangle = (|H;H\rangle_{a,b} + |V;V\rangle_{a,b})/\sqrt{2}$, where H and V represent horizontally and vertically polarized photon respectively, and a and b denote spatial modes. In this physical experiment, 300 healthy density matrices were derived each of which are 4×4 Hermitian matrices. 50 anomalous matrices were also derived where the effect of decoherence can be found in their $(1,4)$-th entry which is off diagonal term between $|H;H\rangle$ and $|V;V\rangle$ of the density matrices. We note that the detection of the change in $(1,4)$-th entry is equivalent to the detection of the change in the quantity of entanglement with the assumption that local polarization does not flip between $|H\rangle$ and $|V\rangle$.

11 Proof of Theorems

11.1 Preliminary

We first give two lemmas we use in the proof of Theorem 2. Here, we define the solution to the problem (3) under the true KL-divergence matrix as

$$S_0 := \operatorname*{argmin}_{S\subseteq[D]} \sum_{d,d'\in S^{\mathrm{c}}} L_{dd'},\ \text{s.t. } |S| = \alpha.$$

[1] `http://www.mit.edu/~jonasm/`

Table 5: Results with UCI Datasets: Left – average AUROC ± standard deviation on 20 random data realizations. Proposed (exact) is a referential result with the exact solution of (3) derived using IBM ILOG CPLEX. The highest AUROC among five methods is shown in bold letters. The best results and other results were compared using a t-test (5%), and results that were not rejected are also highlighted; Right – average runtime ± standard deviation of the proposed method and SPARDA with ten-thread parallelization. The smaller runtime is highlighted.

[Mean Shift]

		AUROC						Runtime (sec)	
	c	Proposed (exact)	Proposed	MT [2]	Idé'09 [3]	Hara'15 [1]	SPARDA [4]	Proposed	SPARDA
CASP	.1	.89 ± .16	**.89 ± .16**	.50 ± .21	.46 ± .22	.48 ± .24	.63 ± .19	**0.37 ± 0.05**	13.4 ± 7.32
	.3	.99 ± .05	**.93 ± .09**	.56 ± .23	.46 ± .22	.48 ± .24	**.86 ± .20**	**0.38 ± 0.05**	5.13 ± 2.39
	.5	1.0 ± .00	**.97 ± .07**	.60 ± .21	.46 ± .22	.48 ± .24	**.96 ± .08**	**0.34 ± 0.07**	2.31 ± 0.78
CBM	.1	.97 ± .08	**.93 ± .13**	.42 ± .12	.51 ± .19	.50 ± .21	.66 ± .33	**0.40 ± 0.06**	25.8 ± 11.7
	.3	1.0 ± .00	**.93 ± .12**	.52 ± .06	.51 ± .19	.50 ± .21	**.93 ± .19**	**0.41 ± 0.06**	5.24 ± 1.87
	.5	1.0 ± .00	.95 ± .10	.49 ± .07	.51 ± .19	.50 ± .21	**1.0 ± .00**	**0.42 ± 0.07**	2.03 ± 0.51
Diag nosis	.1	.87 ± .12	**.87 ± .13**	.41 ± .14	.47 ± .17	.50 ± .18	.49 ± .18	**1.34 ± 0.08**	29.5 ± 48.2
	.3	.96 ± .07	**.97 ± .06**	.43 ± .14	.47 ± .17	.50 ± .18	.60 ± .26	**1.35 ± 0.08**	28.3 ± 49.6
	.5	.96 ± .07	**.99 ± .03**	.47 ± .12	.47 ± .17	.50 ± .18	.68 ± .29	**1.37 ± 0.11**	26.8 ± 49.5
Mini BooNE	.1	.96 ± .09	**.97 ± .06**	.17 ± .03	.47 ± .15	.43 ± .16	.60 ± .32	**1.61 ± 0.07**	127 ± 57.1
	.3	.98 ± .05	**1.0 ± .00**	.17 ± .08	.47 ± .15	.43 ± .16	.73 ± .32	**1.62 ± 0.10**	126 ± 55.7
	.5	1.0 ± .00	**1.0 ± .00**	.25 ± .13	.47 ± .15	.43 ± .16	.74 ± .36	**1.63 ± 0.11**	123 ± 55.6
Stat log	.1	1.0 ± .00	**1.0 ± .00**	.28 ± .09	.51 ± .15	.52 ± .17	.66 ± .27	**0.68 ± 0.07**	12.9 ± 6.03
	.3	1.0 ± .00	**1.0 ± .00**	.43 ± .10	.51 ± .15	.52 ± .17	**.91 ± .24**	**0.67 ± 0.06**	4.62 ± 1.34
	.5	1.0 ± .00	**1.0 ± .00**	.49 ± .04	.51 ± .15	.52 ± .17	**1.0 ± .00**	**0.67 ± 0.08**	2.02 ± 0.45

[Variance Change]

		AUROC						Runtime (sec)	
	c	Proposed (exact)	Proposed	MT [2]	Idé'09 [3]	Hara'15 [1]	SPARDA [4]	Proposed	SPARDA
CASP	.1	.76 ± .15	**.83 ± .19**	.48 ± .18	.50 ± .20	.56 ± .24	.62 ± .17	**0.38 ± 0.06**	13.9 ± 7.08
	.3	.96 ± .09	**.93 ± .09**	.58 ± .25	.75 ± .17	.82 ± .13	.67 ± .21	**0.35 ± 0.08**	14.4 ± 8.23
	.5	.98 ± .07	**.95 ± .08**	.67 ± .21	.77 ± .17	**.93 ± .07**	.70 ± .22	**0.37 ± 0.06**	13.5 ± 6.21
CBM	.1	1.0 ± .00	**.97 ± .06**	.37 ± .11	.88 ± .07	.85 ± .07	.30 ± .15	**0.43 ± 0.06**	43.1 ± 12.9
	.3	1.0 ± .00	**.97 ± .07**	.51 ± .09	**.92 ± .12**	**.94 ± .07**	.26 ± .18	**0.40 ± 0.06**	36.1 ± 11.0
	.5	1.0 ± .00	**.97 ± .08**	.61 ± .14	.86 ± .18	**.99 ± .03**	.62 ± .33	**0.41 ± 0.06**	21.4 ± 7.19
Diag nosis	.1	.73 ± .17	**.79 ± .16**	.40 ± .13	.56 ± .15	.52 ± .19	.41 ± .11	**1.37 ± 0.08**	27.8 ± 45.8
	.3	.87 ± .12	**.90 ± .11**	.41 ± .17	.77 ± .22	.61 ± .16	.47 ± .12	**1.35 ± 0.08**	29.6 ± 46.5
	.5	.90 ± .12	**.93 ± .10**	.49 ± .18	.79 ± .23	.70 ± .14	.48 ± .14	**1.36 ± 0.08**	31.1 ± 50.2
Mini BooNE	.1	.94 ± .10	**.98 ± .03**	.14 ± .04	.73 ± .26	.80 ± .13	.42 ± .23	**1.62 ± 0.12**	128 ± 70.6
	.3	.94 ± .10	**1.0 ± .00**	.17 ± .20	.84 ± .32	.86 ± .16	.34 ± .23	**1.63 ± 0.11**	131 ± 79.0
	.5	.96 ± .08	**1.0 ± .00**	.41 ± .17	**.84 ± .34**	**.92 ± .17**	.35 ± .27	**1.61 ± 0.12**	129 ± 78.5
Stat log	.1	1.0 ± .00	**1.0 ± .00**	.26 ± .12	.66 ± .12	.55 ± .18	.58 ± .20	**0.71 ± 0.05**	14.9 ± 6.39
	.3	1.0 ± .00	**1.0 ± .00**	.56 ± .08	**1.0 ± .00**	.89 ± .12	.56 ± .21	**0.70 ± 0.08**	14.0 ± 5.77
	.5	1.0 ± .00	**1.0 ± .00**	.46 ± .04	**1.0 ± .00**	**1.0 ± .00**	.63 ± .24	**0.68 ± 0.07**	16.2 ± 8.04

Lemma 1 *$S = S_0$ holds if one of (S1) and (S2) holds for $d \in [D]$.*

(Proof of Lemma 1) Let $A := S^{\mathrm{c}} \setminus S_0^{\mathrm{c}}$, $B := S_0^{\mathrm{c}} \setminus S^{\mathrm{c}}$, and $C := S^{\mathrm{c}} \cap S_0^{\mathrm{c}}$. When $|A| = |B| > 0$, we have

$$\begin{aligned}\sum_{d,d' \in S_0^{\mathrm{c}}} L_{dd'} &= \sum_{d,d' \in S^{\mathrm{c}}} L_{dd'} + 2 \sum_{d \in C, d' \in B} L_{dd'} + \sum_{d,d' \in B} L_{dd'} \\ &\quad \underbrace{- 2 \sum_{d \in C, d' \in A} L_{dd'} - \sum_{d,d' \in A} L_{dd'}}_{=0\,(\because \forall d,d' \in S^{\mathrm{c}},\, L_{dd'}=0)} \\ &> \sum_{d,d' \in S^{\mathrm{c}}} L_{dd'}.\end{aligned}$$

In the inequality, we used (S1) and (S2) that $L_{dd'} > 0$ holds when at least one of d and d' is involved in S. This result contradicts the definition of S_0, and we thus have $S = S_0$. □

Lemma 2 ([1], Theorem 1) *Let $\eta := \min_{S' \subseteq [D], S' \neq S} \sum_{d,d' \in S'^{\mathrm{c}}} L_{dd'}$, s.t. $|S'| = \alpha$. We have $\hat{S} = S_0$ when $\|\!|\hat{L} - L|\!\|_\infty < \eta/((D-\alpha)^2 + D^2)$, where $\|\!|\cdot|\!\|_\infty$ denotes an element-wise infinity norm $\|\!|U|\!\|_\infty := \max_{i,j} |U_{ij}|$.*

Table 6: Results with UCI Datasets: Left – average AUROC ± standard deviation on 20 random data realizations. Proposed (exact) is a referential result with the exact solution of (3) derived using IBM ILOG CPLEX. The highest AUROC among five methods is shown in bold letters. The best results and other results were compared using a t-test (5%), and results that were not rejected are also highlighted; Right – average runtime ± standard deviation of the proposed method and SPARDA with ten-thread parallelization. The smaller runtime is highlighted.

[Covariance Change]

		AUROC						Runtime (sec)	
	c	Proposed (exact)	Proposed	MT [2]	Idé'09 [3]	Hara'15 [1]	SPARDA [4]	Proposed	SPARDA
CASP	.1	.80 ± .19	**.79 ± .15**	.49 ± .22	.68 ± .12	.62 ± .23	.63 ± .18	**0.35 ± 0.07**	13.6 ± 7.73
	.3	.92 ± .11	**.93 ± .07**	.61 ± .20	.86 ± .10	.84 ± .14	.75 ± .18	**0.37 ± 0.07**	8.51 ± 4.93
	.5	.98 ± .07	**.98 ± .03**	.66 ± .18	.90 ± .06	.92 ± .11	.77 ± .19	**0.36 ± 0.08**	4.98 ± 3.77
CBM	.1	.84 ± .13	**.87 ± .12**	.43 ± .18	.70 ± .12	.69 ± .14	.48 ± .11	**0.38 ± 0.07**	36.2 ± 11.2
	.3	.95 ± .09	**.96 ± .07**	.41 ± .15	.81 ± .15	.82 ± .11	.62 ± .15	**0.41 ± 0.06**	19.3 ± 9.74
	.5	.96 ± .09	**.98 ± .04**	.45 ± .18	.82 ± .14	.84 ± .11	.70 ± .13	**0.41 ± 0.04**	12.7 ± 10.5
Diag nosis	.1	.80 ± .14	**.82 ± .15**	.55 ± .15	.71 ± .16	.58 ± .19	.41 ± .14	**1.32 ± 0.06**	36.7 ± 68.8
	.3	.90 ± .09	**.94 ± .06**	.45 ± .16	.82 ± .13	.79 ± .13	.47 ± .14	**1.35 ± 0.08**	31.1 ± 54.5
	.5	.95 ± .08	**.97 ± .04**	.50 ± .11	.87 ± .11	.87 ± .12	.62 ± .24	**1.33 ± 0.07**	24.3 ± 40.2
Mini BooNE	.1	.64 ± .13	**.73 ± .16**	.48 ± .15	.53 ± .13	.49 ± .13	.51 ± .18	**1.61 ± 0.10**	133 ± 58.6
	.3	.74 ± .14	**.94 ± .05**	.45 ± .13	.60 ± .16	.54 ± .13	.55 ± .19	**1.64 ± 0.12**	153 ± 58.8
	.5	.88 ± .12	**.98 ± .02**	.44 ± .13	.65 ± .19	.58 ± .15	.56 ± .20	**1.68 ± 0.10**	138 ± 57.0
Stat log	.1	.95 ± .08	**.94 ± .12**	.66 ± .17	**.96 ± .06**	.70 ± .19	.56 ± .27	**0.71 ± 0.08**	13.4 ± 5.39
	.3	1.0 ± .00	**1.0 ± .00**	.42 ± .18	**1.0 ± .00**	.95 ± .07	.67 ± .26	**0.67 ± 0.06**	10.4 ± 4.32
	.5	.98 ± .05	**.98 ± .07**	.27 ± .09	**1.0 ± .00**	**.99 ± .03**	.82 ± .21	**0.68 ± 0.07**	6.60 ± 5.07

[Covariance Change (Conditional)]

		AUROC						Runtime (sec)	
	c	Proposed (exact)	Proposed	MT [2]	Idé'09 [3]	Hara'15 [1]	SPARDA [4]	Proposed	SPARDA
CASP	.1	.61 ± .17	**.66 ± .22**	.50 ± .25	**.53 ± .20**	**.53 ± .24**	**.63 ± .15**	**0.34 ± 0.08**	14.3 ± 8.52
	.3	.71 ± .14	**.79 ± .18**	.50 ± .24	.65 ± .16	.63 ± .22	.67 ± .19	**0.36 ± 0.06**	10.5 ± 5.25
	.5	.86 ± .14	**.89 ± .10**	.50 ± .20	.69 ± .13	.69 ± .22	.63 ± .16	**0.34 ± 0.05**	7.72 ± 5.35
CBM	.1	.63 ± .17	**.68 ± .18**	.46 ± .17	**.59 ± .13**	**.57 ± .19**	.51 ± .09	**0.39 ± 0.07**	40.5 ± 13.2
	.3	.78 ± .15	**.86 ± .14**	.40 ± .17	.69 ± .14	.66 ± .16	.55 ± .14	**0.41 ± 0.06**	32.3 ± 9.06
	.5	.86 ± .12	**.94 ± .06**	.39 ± .16	.74 ± .13	.73 ± .15	.57 ± .17	**0.41 ± 0.06**	23.9 ± 10.5
Diag nosis	.1	.58 ± .12	**.66 ± .19**	.54 ± .13	**.58 ± .13**	**.55 ± .19**	.42 ± .14	**1.32 ± 0.06**	30.9 ± 48.1
	.3	.67 ± .15	**.78 ± .15**	.46 ± .16	**.75 ± .20**	.68 ± .17	.47 ± .15	**1.35 ± 0.07**	26.6 ± 40.4
	.5	.74 ± .15	**.86 ± .10**	.42 ± .15	**.79 ± .19**	.77 ± .14	.50 ± .17	**1.39 ± 0.08**	25.2 ± 42.3
Mini BooNE	.1	.54 ± .09	**.65 ± .17**	.49 ± .14	.51 ± .13	.48 ± .13	.53 ± .17	**1.64 ± 0.09**	128 ± 60.3
	.3	.61 ± .11	**.72 ± .16**	.46 ± .13	.53 ± .15	.53 ± .13	.51 ± .15	**1.62 ± 0.08**	150 ± 60.4
	.5	.63 ± .10	**.79 ± .13**	.43 ± .12	.57 ± .18	.55 ± .13	.48 ± .16	**1.62 ± 0.10**	148 ± 52.1
Stat log	.1	.55 ± .12	.31 ± .27	.51 ± .17	**.67 ± .19**	.52 ± .21	**.58 ± .23**	**0.68 ± 0.06**	14.3 ± 6.02
	.3	.67 ± .19	.66 ± .28	.43 ± .17	**.87 ± .12**	.68 ± .17	.62 ± .22	**0.66 ± 0.08**	12.8 ± 5.80
	.5	.87 ± .14	**.89 ± .17**	.40 ± .15	**.93 ± .10**	.76 ± .16	.64 ± .23	**0.66 ± 0.09**	9.56 ± 6.03

11.2 Proofs

Proof of Theorem 1: Suppose there exists $\delta \in S$ such that (N1′) and (N2′) holds:

$$\text{(N1}') \; L_{\delta\delta} = 0,$$
$$\text{(N2}') \; \forall d \in [D] \setminus \{\delta\}, \; L_{\delta d} = 0.$$

Then, for any $\delta' \in S^{\mathrm{c}}$,

$$\begin{aligned}
0 &= \sum_{d,d' \in S^{\mathrm{c}}} L_{dd'} \\
&= \sum_{d,d' \in (S^{\mathrm{c}} \cup \{\delta\}) \setminus \{\delta'\}} L_{dd'} + 2 \underbrace{\sum_{d \in S^{\mathrm{c}} \setminus \{\delta'\}} L_{\delta' d} + L_{\delta'\delta'}}_{=0 \; (\because \forall d,d' \in S^{\mathrm{c}},\, L_{dd'}=0)} \\
&\quad \underbrace{- 2 \sum_{d \in S^{\mathrm{c}} \setminus \{\delta'\}} L_{\delta d} - L_{\delta\delta}}_{=0 \; (\because \text{N1}', \text{N2}')} \\
&= \sum_{d,d' \in (S^{\mathrm{c}} \cup \{\delta\}) \setminus \{\delta'\}} L_{dd'}
\end{aligned}$$

Table 7: Results with UCI Datasets: Left – average AUROC ± standard deviation on 20 random data realizations. Proposed (exact) is a referential result with the exact solution of (3) derived using IBM ILOG CPLEX. The highest AUROC among five methods is shown in bold letters. The best results and other results were compared using a t-test (5%), and results that were not rejected are also highlighted; Right – average runtime ± standard deviation of the proposed method and SPARDA with ten-thread parallelization. The smaller runtime is highlighted.

[Covariance Change (No Variance Change)]

		AUROC						Runtime (sec)	
	c	Proposed (exact)	Proposed	MT [2]	Idé'09 [3]	Hara'15 [1]	SPARDA [4]	Proposed	SPARDA
CASP	.1	$.71 \pm .19$	$\mathbf{.76 \pm .22}$	$.51 \pm .21$	$\mathbf{.68 \pm .12}$	$\mathbf{.62 \pm .23}$	$.59 \pm .17$	$\mathbf{0.35 \pm 0.08}$	13.4 ± 7.79
	.3	$.90 \pm .12$	$\mathbf{.92 \pm .09}$	$.62 \pm .19$	$\mathbf{.86 \pm .10}$	$\mathbf{.84 \pm .14}$	$.69 \pm .17$	$\mathbf{0.36 \pm 0.06}$	11.3 ± 6.10
	.5	$.98 \pm .07$	$\mathbf{.96 \pm .05}$	$.68 \pm .20$	$.90 \pm .06$	$\mathbf{.92 \pm .11}$	$.72 \pm .14$	$\mathbf{0.36 \pm 0.06}$	4.71 ± 3.45
CBM	.1	$.83 \pm .13$	$\mathbf{.87 \pm .13}$	$.54 \pm .14$	$.70 \pm .12$	$.69 \pm .14$	$.51 \pm .15$	$\mathbf{0.41 \pm 0.07}$	41.3 ± 13.7
	.3	$.96 \pm .09$	$\mathbf{.97 \pm .06}$	$.57 \pm .16$	$.81 \pm .15$	$.82 \pm .11$	$.58 \pm .14$	$\mathbf{0.40 \pm 0.05}$	20.3 ± 9.74
	.5	$.96 \pm .09$	$\mathbf{.98 \pm .05}$	$.54 \pm .17$	$.82 \pm .14$	$.84 \pm .11$	$.73 \pm .12$	$\mathbf{0.41 \pm 0.07}$	12.0 ± 10.0
Diag nosis	.1	$.72 \pm .16$	$\mathbf{.75 \pm .19}$	$.43 \pm .11$	$\mathbf{.71 \pm .16}$	$.58 \pm .19$	$.43 \pm .13$	$\mathbf{1.33 \pm 0.07}$	32.7 ± 60.6
	.3	$.82 \pm .13$	$\mathbf{.85 \pm .14}$	$.39 \pm .14$	$\mathbf{.82 \pm .13}$	$\mathbf{.79 \pm .13}$	$.47 \pm .16$	$\mathbf{1.33 \pm 0.07}$	29.0 ± 42.7
	.5	$.88 \pm .13$	$\mathbf{.91 \pm .12}$	$.49 \pm .18$	$\mathbf{.87 \pm .11}$	$\mathbf{.87 \pm .12}$	$.61 \pm .18$	$\mathbf{1.37 \pm 0.08}$	29.1 ± 49.8
Mini BooNE	.1	$.63 \pm .14$	$\mathbf{.72 \pm .17}$	$.48 \pm .15$	$.53 \pm .13$	$.49 \pm .13$	$.52 \pm .17$	$\mathbf{1.59 \pm 0.11}$	134 ± 60.2
	.3	$.73 \pm .14$	$\mathbf{.93 \pm .06}$	$.45 \pm .14$	$.60 \pm .16$	$.54 \pm .13$	$.53 \pm .20$	$\mathbf{1.62 \pm 0.09}$	147 ± 56.6
	.5	$.88 \pm .13$	$\mathbf{.98 \pm .03}$	$.44 \pm .14$	$.65 \pm .19$	$.58 \pm .15$	$.56 \pm .19$	$\mathbf{1.60 \pm 0.10}$	137 ± 50.8
Stat log	.1	$.92 \pm .11$	$\mathbf{.92 \pm .16}$	$.52 \pm .22$	$\mathbf{.96 \pm .06}$	$.70 \pm .19$	$.58 \pm .22$	$\mathbf{0.69 \pm 0.06}$	13.8 ± 5.90
	.3	$1.0 \pm .00$	$\mathbf{1.0 \pm .00}$	$.48 \pm .18$	$\mathbf{1.0 \pm .00}$	$.95 \pm .07$	$.66 \pm .26$	$\mathbf{0.69 \pm 0.07}$	11.4 ± 4.81
	.5	$1.0 \pm .00$	$\mathbf{1.0 \pm .00}$	$.41 \pm .17$	$\mathbf{1.0 \pm .00}$	$\mathbf{.99 \pm .03}$	$.82 \pm .19$	$\mathbf{0.66 \pm 0.07}$	5.98 ± 5.58

holds. This shows that both S and $(S \cup \{\delta\}) \setminus \{\delta'\}$ can be the optimal solutions, which contradicts the consistency of $\hat{S}$. □

Proof of Theorem 2: Let η be a parameter defined in Lemma 2. From Theorem 2 of *Wang et al.* [5], for any $\zeta > 0$ there exists $Q^{\zeta}_{dd'} > 0$ such that for all $N, M > Q^{\zeta}_{dd'}$,

$$P\left(|\hat{L}_{dd'} - L_{dd'}| \geq \frac{\eta}{(D-\alpha)^2 + D^2}\right) < \frac{\zeta}{D^2},$$

holds. Hence, we have

$$\begin{aligned} P(S = \hat{S}) &= P(S_0 = \hat{S}) \\ &\geq P\left(\|\hat{L} - L\|_{\infty} < \frac{\eta}{(D-\alpha)^2 + D^2}\right) \\ &\geq 1 - \sum_{d,d'} P\left(|\hat{L}_{dd'} - L_{dd'}| \geq \frac{\eta}{(D-\alpha)^2 + D^2}\right) \\ &\geq 1 - \zeta, \end{aligned}$$

for all $N, M > Q^{\zeta} := \max_{d,d'} Q^{\zeta}_{dd'}$, where we used Lemma 1 for the equality in the first line and Lemma 2 for the inequality in the second line. □

Proof of Theorem 3: Let $p(\boldsymbol{x}) := \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and $q(\boldsymbol{x}) := \mathcal{N}(\boldsymbol{\nu}, \Gamma)$. Suppose there exists $\delta \in S$ such that both (N1') and (N2') hold. Condition (N1′) is equivalent to $p(x_\delta) = q(x_\delta)$ which implies

$$\mu_\delta = \nu_\delta,\ \Sigma_{\delta\delta} = \Gamma_{\delta\delta}.$$

Similarly, Condition (N2′) is equivalent to $p(x_\delta, x_d) = q(x_\delta, x_d)$ for any $d \in [D] \setminus \{\delta\}$ which implies

$$\Sigma_{\delta d} = \Gamma_{\delta d}.$$

From these results, we have

$$p(\boldsymbol{x}_{(S\setminus\{\delta\})^c}) = q(\boldsymbol{x}_{(S\setminus\{\delta\})^c}),$$

which contradicts with Condition (2). Hence, there exists no $\delta \in S$ that satisfies Conditions (N1') and (N2'). □

Proof of Proposition 1: For the bivariate KL-divergence, when the specified condition

$$\begin{aligned} &\mathrm{KL}[p(x_d, x_{d'})||q(x_d, x_{d'})] \\ &= \frac{1}{2}\left\{\frac{2 - 2\Sigma_{dd'}\Gamma_{dd'}}{1 - \Gamma^2_{dd'}} - \log\frac{1 - \Sigma^2_{dd'}}{1 - \Gamma^2_{dd'}} - 2\right\} \\ &= \frac{1}{2}\left\{\frac{(\Sigma_{dd'} - \Gamma_{dd'})^2}{1 - \Gamma^2_{dd'}} + \frac{1 - \Sigma^2_{dd'}}{1 - \Gamma^2_{dd'}} - \log\frac{1 - \Sigma^2_{dd'}}{1 - \Gamma^2_{dd'}} - 1\right\} \\ &\geq \frac{1}{2}\frac{(\Sigma_{dd'} - \Gamma_{dd'})^2}{1 - \Gamma^2_{dd'}} \geq \frac{1}{2}|\Sigma_{dd'} - \Gamma_{dd'}| - \frac{1}{8}, \end{aligned}$$

holds, where we used the assumption that Σ and Γ are invertible which implies $\Sigma^2_{dd'}, \Gamma^2_{dd'} < 1$, and $t - \log t \geq 1$ for $t > 0$. □

References

[1] S. Hara, T. Morimura, T. Takahashi, H. Yanagisawa, and T. Suzuki. A consistent method for

graph based anomaly localization. *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 333–341, 2015.

[2] G. Taguchi and J. Rajesh. New trends in multivariate diagnosis. *The Indian Journal of Statistics, Series B*, pages 233–248, 2000.

[3] T. Idé, A. C. Lozano, N. Abe, and Y. Liu. Proximity-based anomaly detection using sparse structure learning. *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 97–108, 2009.

[4] J. W. Mueller and T. Jaakkola. Principal differences analysis: Interpretable characterization of differences between distributions. *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

[5] Q. Wang, S. R. Kulkarni, and S. Verdú. Divergence estimation for multidimensional densities via-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.