Unsupervised Kernel Dimension Reduction Supplemental Material

Meihong Wang
Dept. of Computer Science
U. of Southern California
Los Angeles, CA 90089
meihongw@usc.edu

Fei Sha
Dept. of Computer Science
U. of Southern California
Los Angeles, CA 90089
`feisha@usc.edu`

Michael I. Jordan
Dept. of Statistics
U. of California
Berkeley, CA
`jordan@cs.berkeley.edu`

1 Relationship between $\hat{J}_{YY|X}$ and $\hat{J}_{XY}$ and Proof of Proposition 1

We start by noting that conditional independence $X \perp\!\!\!\perp Y \mid \boldsymbol{B}^\top X$ does not necessarily imply the correlation between $\boldsymbol{B}^\top X$ and Y is maximized. To see this, let X be a Gaussian random vairable with zero mean and diagonal covariance matrix. Assume $\boldsymbol{B}$ is an identity matrix and $Y = X^2 = (\boldsymbol{B}^\top X)^2$ (elementwise square for a vectorial X). The conditional independence is obviously satisfied yet the correlation between $\boldsymbol{B}^\top X$ and Y is zero. This observation is yet another example showing the limitation of Spearson's correlation measures, which detect only linear dependence between random variables. In the following, we show that when measured in the RKHS, the two measures $\hat{J}_{YY|X}$ and $\hat{J}_{XY}$ are equivalent.

Assume we use Gaussian RBF kernel for both $\hat{J}_{YY|X}(\boldsymbol{B}^\top X, Y)$ and $\hat{J}_{XY}(\boldsymbol{B}^\top X, Y)$: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/\sigma_N^2\right)$ where σ_N is the bandwidth. It should be chosen to match roughly the scale of the data. While $\hat{J}_{XY}(\boldsymbol{B}^\top X, Y)$ depends only on σ_N, $\hat{J}_{YY|X}(\boldsymbol{B}^\top X, Y)$ depends on σ_N and ϵ_N, which need to be adjusted in empirical studies. Since both parameters control the smoothness of the kernel matrix $\boldsymbol{K}_{\boldsymbol{B}^\top X}$, it is sensible to adjust the relative scale of the two parameters. Particularly, the following statement establishes the link between the two measures:

Proposition 1. *Let $N \to +\infty$ and $\epsilon_N \to 0$. Additionally, assume the samples are distributed uniformly on the unit sphere. If $\sigma_N \ll \epsilon_N^2$, then up to a constant,*

$$\hat{J}_{YY|X}(\boldsymbol{B}^\top X, Y) \approx -c_0 N^2 \epsilon_N^2 \hat{J}_{XY}(\boldsymbol{B}^\top X, Y) \tag{1}$$

Therefore, It is equivalent to minimizing $\hat{J}_{YY|X}(\boldsymbol{B}^\top X, Y)$ and maximizing $\hat{J}_{XY}(\boldsymbol{B}^\top X, Y)$.

We sketch the proof in the following. The proof consists of two main steps. In the first step, we bound the largest eigenvalue λ_0 of $\boldsymbol{K}_{\boldsymbol{B}^\top X}$ from above. In the second step, we show that applying the bound allows us to approximate the inverse $(\boldsymbol{G}_{\boldsymbol{B}^\top X} + N\epsilon_N \boldsymbol{I}_N)^{-1}$ with $\boldsymbol{K}_{\boldsymbol{B}^\top X}$.

Bound on λ_0. With Gaussian RBF kernel, the elements of the kernel matrix $\boldsymbol{K}_{\boldsymbol{B}^\top X}$ is strictly positive. By Perron-Frobenius theorem, the spectral radius is λ_0 and is bound by

$$\lambda_0 \le \max_i \sum_j \boldsymbol{K}_{\boldsymbol{B}^\top X}(i, j), \tag{2}$$

namely, the maximum of the row sums. The (i, j)-th element of the kernel matrix is given by

$$\boldsymbol{K}_{\boldsymbol{B}^\top X}(i, j) = \exp\left(-\sum_m \|\boldsymbol{B}_m{}^\top \boldsymbol{x}_i - \boldsymbol{B}_m{}^\top \boldsymbol{x}_j\|^2/\sigma_N^2\right) \tag{3}$$

where $\boldsymbol{B}_m$ is the m-th column of the matrix $\boldsymbol{B}$. Since each column contributes nonnegatively to the exponent, it is obvious

$$\lambda_0 \leq \max_i \sum_j \exp\left(-\|{\boldsymbol{B}_m}^\top \boldsymbol{x}_i - \boldsymbol{B}_m \boldsymbol{x}_j\|^2/\sigma_N^2\right) \tag{4}$$

for any m. Furthermore, since $\boldsymbol{x}_i$ are assumed to be uniformly distributed on the unit sphere, we choose an arbitrary coordinate system such that $\boldsymbol{B}_m = [1\ 0\ 0 \cdots\ 0]^\top$. This gives rise to

$$\lambda_0 \leq \sum_j \exp\left(-(x_{i1} - x_{j1})^2/\sigma_N^2\right). \tag{5}$$

Note that we have removed the $\max_i$ operation based on the symmetry argument. The first element x_{i1} of $\boldsymbol{x}_i$ can be set as 1 without loss of generality. Moreover, the first elements x_{j1} of $\boldsymbol{x}_j$ can be parameterized as $\cos\theta_j$ where θ_j is uniformly distributed between 0 and 2π. This leads to

$$\lambda_0 \leq \sum_j \exp\left(-(1 - \cos\theta_j)^2/\sigma_N^2\right). \tag{6}$$

When $N \to +\infty$, the right-hand-size tends to an integral

$$\lambda_0 \leq N \int_0^{2\pi} e^{-\frac{(1-\cos\theta)^2}{\sigma_N^2}}\, d\theta\ . \tag{7}$$

The integral does not have a analytic closed-form. Our intention is to approximate it with a function of σ_N. In particular, σ_N needs to tend to zero when $N \to +\infty$. Therefore, our next step is to identify an asymptotic expansion in σ_N.

After a few variable substitutions ($1 - \cos\theta = 2\cos^2\theta/2$, then $\cos\theta/2 = t$, the integral of eq. (7) is transformed (omitting constants) to

$$I(\sigma_N) = \int_0^1 \frac{1}{\sqrt{1-t^2}} e^{-4t^4/\sigma_N^2}\, dt \tag{8}$$

Applying Watson Lemma [1] to the above integral as $4/\sigma_N^2 \to +\infty$, we obtain the asymptotic expansion up to (and including) the first-order of $1/\sqrt{1-t^2}$,

$$I(\sigma_N) \sim 1/4\ \Gamma(1/4)\sqrt{\sigma_N/2} + O(\sigma_N^{3/2}) \quad (\sigma_N \to 0)\ . \tag{9}$$

This gives us a bound on λ_0

$$\lambda_0 \leq c_0 N \sqrt{\sigma_N} \tag{10}$$

where c_0 is a constant.

Approximate $\hat{J}_{YY|X}(\boldsymbol{B}^\top X, Y)$. We apply the Woodbury matrix inversion lemma to the independence measure. Let $\delta_N = N\epsilon_N$, we have

$$\hat{J}_{YY|X}(\boldsymbol{B}^\top X, Y) \tag{11}$$

$$= \mathrm{Trace}\left[\boldsymbol{G}_Y(\boldsymbol{H}\boldsymbol{K}_{\boldsymbol{B}^\top X}\boldsymbol{H} + \delta_N \boldsymbol{I}_N)^{-1}\right] \tag{12}$$

$$= \mathrm{Trace}\left[\boldsymbol{G}_Y\left\{\delta_N^{-1}\boldsymbol{I}_N - \delta_N^{-1}\boldsymbol{H}(\boldsymbol{K}_{\boldsymbol{B}^\top X}^{-1} + \boldsymbol{H}\delta_N^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}\delta_N^{-1}\right\}\right] \tag{13}$$

$$= -\delta_N^{-2}\mathrm{Trace}\left[\boldsymbol{G}_Y(\boldsymbol{K}_{\boldsymbol{B}^\top X}^{-1} + \delta_N^{-1}\boldsymbol{H})^{-1}\right] + \mathsf{const} \tag{14}$$

where we have used the identities $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$ and $\boldsymbol{H}\boldsymbol{G}_Y\boldsymbol{H} = \boldsymbol{G}_Y$. We have also assumed that $\boldsymbol{K}_{\boldsymbol{B}^\top X}$ is invertible. This is reasonable as we are using Gaussian RBF kernels. If we further assume that $\boldsymbol{B}^\top X$ yields different $\boldsymbol{B}^\top\boldsymbol{x}_1$, $\boldsymbol{B}^\top\boldsymbol{x}_2, \ldots$ and $\boldsymbol{B}^\top\boldsymbol{x}_N$, then the kernel matrix is full ranked and invertible.

We consider the limiting case when $N \to +\infty$. We have shown in above that the largest eigenvalue of $\boldsymbol{K}_{\boldsymbol{B}^\top X}$ is bound from above by $N\sqrt{\sigma_N}$. This means that the smallest eigenvalue $1/\lambda_0$ of $\boldsymbol{K}_{\boldsymbol{B}^\top X}^{-1}$ is bound from below by $1/(N\sqrt{\sigma_N})$. Note that the largest eigenvalue of $\delta_N^{-1}\boldsymbol{H}$ is $1/(N\epsilon_N)$. Therefore, if

$$1/(N\sqrt{\sigma_N}) \gg 1/(N\epsilon_N), \tag{15}$$

we can approximate the inversion with

$$(\boldsymbol{K}_{\boldsymbol{B}^\top X}^{-1} + \delta_N^{-1}\boldsymbol{H})^{-1} \approx (\boldsymbol{K}_{\boldsymbol{B}^\top X}^{-1})^{-1} = \boldsymbol{K}_{\boldsymbol{B}^\top X}\ . \tag{16}$$

The condition eq. (15) corresponds to the condition $\sigma_N \ll \epsilon_N^2$ in the Proposition 1, which leads to (after substituting eq. (16) into eq.(14)),

$$\hat{J}_{YY|X}(\boldsymbol{B}^\top X, Y) = -\delta_N^2 \hat{J}_{XY}(\boldsymbol{B}^\top X, Y) + \mathsf{const} \tag{17}$$

2 Relation between t-SNE and UKDR

t-SNE aims to preserve local conditional probability structure computed in the original space of X [2] . The structure is encoded as the random walk probability of

$$p_{ij} = P(\boldsymbol{x}_i \to \boldsymbol{x}_j) = \frac{\exp\{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/\sigma_i^2\}}{\sum_{j\neq i}\exp\{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/\sigma_i^2\}}\ . \tag{18}$$

The low dimensional coordinates Z are estimated such that the structure computed in the low dimensional space

$$q_{ij} \propto C(\boldsymbol{z}_i, \boldsymbol{z}_j) = \frac{1}{1 + \|\boldsymbol{z}_i - \boldsymbol{z}_j\|^2} \tag{19}$$

has the smallest KL divergence: $\boldsymbol{z} = \arg\min KL(\tilde{\boldsymbol{P}}\|\boldsymbol{Q})$ where $\boldsymbol{Q}$'s elements are q_{ij}. $\tilde{\boldsymbol{P}} = (\boldsymbol{P} + \boldsymbol{P}^\top)/2$ is a symmetric version of $\boldsymbol{P}$. Minimizing the KL divergence is equivalent to maximize the conditional entropy

$$\sum_{ij} \tilde{P}_{ij} \log Q_{ij} = \sum_{ij} \tilde{P}_{ij} \log C(\boldsymbol{z}_i, \boldsymbol{z}_j) - \log \sum_{ij} C(\boldsymbol{z}_i, \boldsymbol{z}_j) = \mathrm{Trace}[\boldsymbol{P} \log \boldsymbol{C}] - \log \boldsymbol{1}^\top \boldsymbol{C}\,\boldsymbol{1} \tag{20}$$

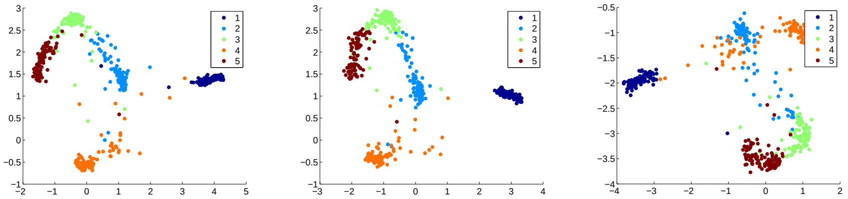
where the logarithm of the matrix $\boldsymbol{C}$ is taken element-wisely. There is a strong analogy of this objective function to $\hat{J}_{XY}(\boldsymbol{B}^\top X, X)$ with random walk kernel over X and Cauchy kernel over $\boldsymbol{B}^\top X$,

$$\hat{J}_{XY}(\boldsymbol{B}^\top X, X) = \mathrm{Trace}[\boldsymbol{P}\boldsymbol{P}^\top \boldsymbol{H}\boldsymbol{C}\boldsymbol{H}] \approx \mathrm{Trace}[\boldsymbol{P}\boldsymbol{P}^\top \boldsymbol{C}] - 1/N\, \boldsymbol{1}^\top \boldsymbol{C}\boldsymbol{1} \tag{21}$$

where N is the number of data points. The approximation is valid if we assume data are well-clustered and the distances between data in the same clusters are roughly the same. We gain further insights about t-SNE by comparing the two objective functions of eq. (20) and eq. (21). Apart from using different yet related "kernels" (tSNE's kernel is not positive semidefinite), both objective functions have two terms which function similarly. The trace terms try to match the similarities between low dimensional coordinates (encoded by $\boldsymbol{C}$ and $\log \boldsymbol{C}$ respectively) with the similarities in the high dimensional coordinates. The normalization terms, to be minimized, try to push data points far away from each other.

3 Effect of sparsity in RSN UKDR

For nonlinear unsupervised kernel dimension reduction, using random and sparse features (RSF) extracted from X has significant computational advantage than using transformed X by radial basis networks. RSF features are sparse and easy to compute. In particular, the dimensionality of RSF features depends on the dimensionality of the data while RBF transformation depends on the number of data points in the training data set. Furthermore, in computing RSF features, the bias constant b can be used to yield sparser feature vectors. In Fig. 1, we investigate the effect of the sparsity level on embeddings. The setup is similar to what is used for generating the embedding in Fig.2(i)(in the main paper) where b is 0 and leads to a feature vector with sparsity level of 50%. We change b to 5, 7 and 10, obtaining sparsity levels of 75%, 82% and 90% in Fig. 1. Through visual inspection, it is clear that a sparsity level of 82% does not bring detrimental effects, while the higher 90% starts to show fragmented clustering of data. Therefore, we deem RSF features as a viable option in handling high dimension data for nonlinear UKDR.



(a) Sparsity: 75% (b) Sparsity: 82% (c) Sparsity: 90%

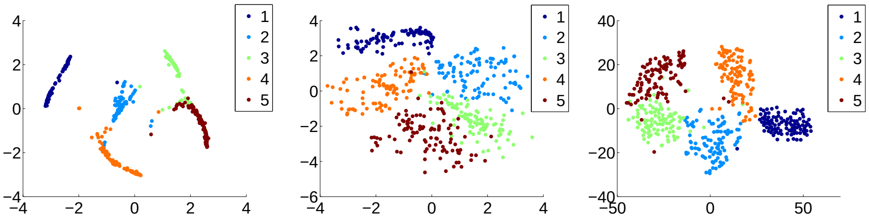
Figure 1: 2D Embedding of USPS-500 dataset, computed using random sparse features with various degrees of sparsity. The higher the sparsity is, the sparser the features are.

4 Comparison to other dimensionality reduction method

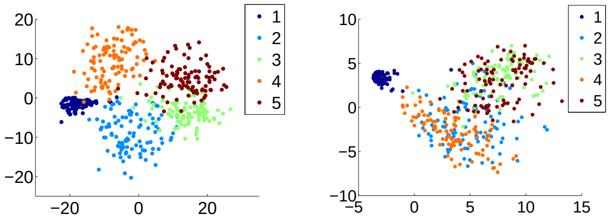
We compare our UKDR method to other dimensionality reduction methods, including SNE [3], t-SNE, MUSHIC (Colored Maximum Variance Unfolding) [4] and PCA. The first dataset is USPS 500, which contains digit 1, 2, 3, 4, and 5. The second dataset is UPSPS 2007, which contains 2007 images of 10 digits. The third dataset is discussion postings from 5 Newsgroups with 100 posting in each group. It is a subset of the 20 Newsgroups used in [4].

4.1 USPS 500

Fig.2 shows the 2D embedding results on dataset USPS 500.



(a) UKDR (Nonparametric embedding) (b) SNE (c) t-SNE



(d) MUHSIC (e) PCA

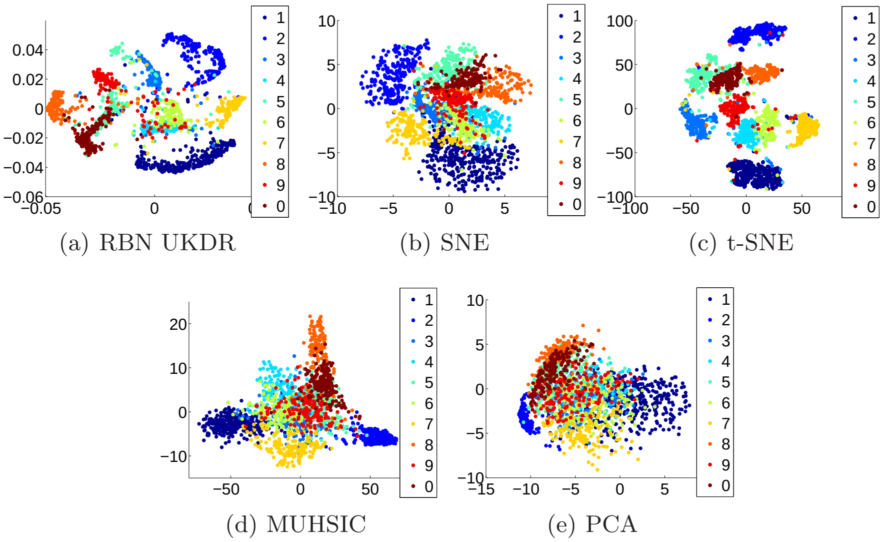
Figure 2: 2D embedding results on USPS 500 by our method UKDR and other methods. UKDR performs much better than other methods

4.2 Results on USPS 2007

Fig.3 shows the 2D embedding results on dataset USPS 2007.

4.3 Results on 5 Newsgroup

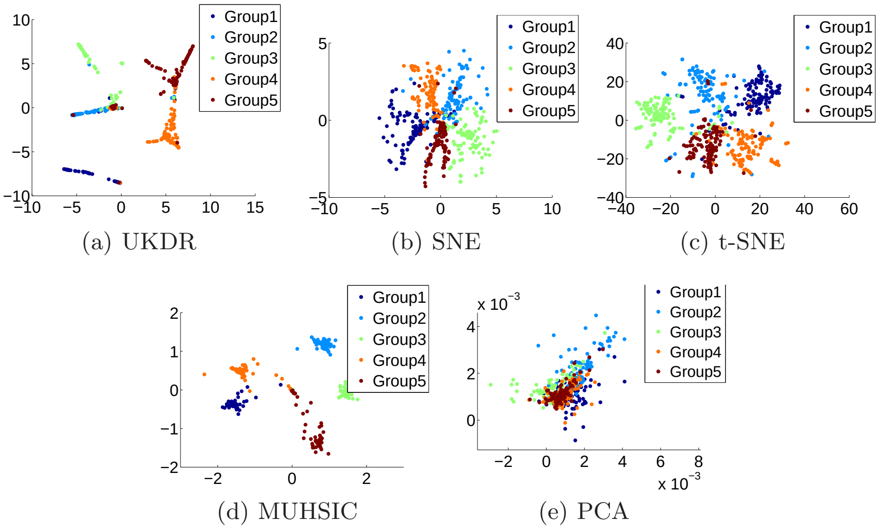
Fig.4 shows the 2D embedding results on dataset 5 Newsgroups



(a) RBN UKDR (b) SNE (c) t-SNE

(d) MUHSIC (e) PCA

Figure 3: 2D embedding results on USPS 2007 by our method UKDR and other methods. Only UKDR and t-SNE separate all classes reasonably well



(a) UKDR (b) SNE (c) t-SNE

(d) MUHSIC (e) PCA

Figure 4: 2D embedding results on 5 Newsgroups by our method UKDR and other methods. UKDR, t-SNE and MUHSIC works well.

References

[1] E. T. Copson. *Asymptotic expansions.* Cambridge University Press, 2004.

[2] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[3] G. Hinton and S. Roweis. Stochastic neighbor embedding. *Advances in Neural Information Processing Systems 15*, pages 857–864, 2003.

[4] L. Song, A. Smola, K. Borgwardt, and A. Gretton. Colored maximum variance unfolding. *Advances in Neural Information Processing Systems 20*, pages 1385–1392, 2008.