Robust Multiple Kernel k-means Clustering using Min-Max Optimization

Written by AAAI Press Staff[1] *
AAAI Style Contributions by Pater Patel Schneider,
Sunil Issar, J. Scott Penberthy, George Ferguson, Hans Guesgen
[1] Association for the Advancement of Artificial Intelligence
2275 East Bayshore Road, Suite 160
Palo Alto, California 94303
publications20@aaai.org

Abstract

Multiple kernel learning is a type of multiview learning that combines different data modalities by capturing view-specific patterns using kernels. Although supervised multiple kernel learning has been extensively studied, until recently, only a few unsupervised approaches have been proposed. In the meanwhile, adversarial learning has recently received much attention. Many works have been proposed to defend against adversarial examples. However, little is known about the effect of adversarial perturbation in the context of multiview learning, and even less in the unsupervised case. In this study, we show that adversarial features added to a view can make the existing approaches with the $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ formulation in multiple kernel clustering yield unfavorable clusters. To address this problem and inspired by recent works in adversarial learning, we propose a multiple kernel clustering method with the $\min_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ framework that aims to be robust to such adversarial perturbation. We evaluate the robustness of our method on simulation data under different types of adversarial perturbations and show that it outperforms several compared existing methods. In the real data analysis, We demonstrate the utility of our method on two real-world problems.

1 Introduction

In recent years, multiview (or multimodal) learning approaches have been developed to integrate abundant yet diverse data modality. Integrating diverse modalities is challenging because data from different sources (called *views*) have different statistical properties. To address this problem, multiple kernel learning uses view-specific kernels to capture diverse patterns of multiple views (Lanckriet et al. 2004a). Then, it integrates views as a linear sum of multiple kernels weighted by kernel coefficients $\boldsymbol{\theta}$, and applies a standard classification or clustering algorithm to the combined kernel. Driven by advantages of using kernels, it has witnessed successes in various domains such as computer vision (Gehler and Nowozin 2009) and document classification (Lanckriet et al. 2004a).

While supervised multiple kernel learning has been extensively studied, only a few unsupervised approaches have been proposed until recently, among which, multiple kernel k-means clustering is one of the commonly used approaches. For simplicity, we limit our discussion here to the case of multiple kernel k-means clustering. Although details vary, they find clusters by alternately optimizing the kernel coefficients $\boldsymbol{\theta}$ and clustering assignment $\mathbf{H}$ as shown in Figure S1. Existing works employ a $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ (or $\max_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$) framework. In particular, they first find a combination of views that reveals low within-cluster variance, and then find clusters minimizing such variance (Gönen and Margolin 2014; Liu et al. 2017; 2016; Yu et al. 2012; Yao and Chen 2018).

Meanwhile, adversarial learning has received much attention in recent years. Plenty of studies have demonstrated that very small changes to input can make a model, in particular a deep learning model, to produce incorrect predictions (Biggio et al. 2013; Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015). This phenomenon is so-called *adversarial example phenomenon*. Many studies have proposed defence mechanisms resistant to adversarial example (Madry et al. 2018; Sinha, Namkoong, and Duchi 2018; Zhang et al. 2019), in which they aim to minimize a loss under the maximum adversary. In particular, they use min-max framework that first finds adversarial examples that maximize a loss and then finds the model parameters that minimize the adversarial loss. In the context of deep learning, this min-max framework has become an effective approach to learn a robust model against adversarial attacks.

Despite all these works in adversarial learning, little is known about adversary and robustness in the context of multiview learning, and even less in the unsupervised case. Inspired by recent works in adversarial learning, in this study, we show that adversarial features, e.g., a number of random noise or redundant variables, added to a certain view can deceive the existing $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ methods. In particular, they make $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ methods to ignore the view and find clusters that are largely determined by other views. For simplicity, we denote such features as *adversarial perturbation*.

To address this problem, we propose a multiple kernel clustering method, *multiple kernel k-means clustering with*

*Primarily Mike Hamilton of the Live Oak Press, LLC, with help from the AAAI Publications Committee

$\min_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ *and l_2 regularization* (MML-MKKC). It aims to be robust to adversarial perturbation by using the $\min_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ formulation. In particular, the inner maximization finds a combination of multiple views in favor of a view (or views) that reveals high within-cluster variance, whereas the outer minimization finds clusters that minimize such variance. By capturing such variance while adversary is present, our method can mitigate the effect of adversarial perturbation (see details in Section 2.3).

We evaluate our method on the simulated multiview data with adversarial perturbations that allow us to assess robustness of our method. The result shows that our method outperforms the compared existing multiple kernel clustering methods and yields clusters by making good use of all views, including the view with the added perturbation. We also demonstrate the utility of our method on real-world problems–one is to identify cancer subtypes, and the other is to identify response patterns of asthma patients to medical treatment.

Our main contributions are as follow.

- To the best of our knowledge, this is the first work that studies adversarial perturbation in a unsupervised multiview setting. In particular, we examine the effect of potential adversaries on existing multiview clustering models.
- We found out that adversarial perturbation can make existing multiview clustering methods with the $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ framework yield unfavorable results. They tend to ignore the view with adversarial perturbation and find clusters by relying largely on other views.
- We propose a multiple kernel k-means clustering method MML-MKKC using a $\min_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ framework that aims to be robust to adversarial perturbations. This is achieved by minimizing the within-cluster variance in a combination of the views that reveals high within-cluster variance.

2 Method

In this section, we begin with introducing prior works. We then propose a multiple kernel clustering method that aims to be robust against adversarial perturbation.

2.1 Kernel k-means clustering

Let $\mathbf{x} \in \mathcal{R}^p$ be a sample instance and $\boldsymbol{\phi} : \mathcal{R}^p \to \mathcal{F}$ is a nonlinear mapping of $\mathbf{x}$ onto a reproducing kernel Hilbert space $\mathcal{F}$. By mapping to a higher dimensional feature space using $\boldsymbol{\phi}$, kernel k-means clustering linearly separates samples that were only non-linearly separable in the input space (Girolami 2002). The optimization problem of kernel k-means clustering is the same as k-means clustering but replacing $\mathbf{x}$ with a nonlinear mapping $\boldsymbol{\phi}(\mathbf{x}) \in \mathcal{F}$, which is:

$$\underset{\mathbf{Z}\in\{0,1\}^{n\times k}}{\text{minimize}} \sum_{c=1}^{k}\sum_{i=1}^{n} z_{ic}\|\boldsymbol{\phi}(\mathbf{x}_i) - \boldsymbol{\mu}_c\|_2^2 \quad \text{s.t.} \sum_{c=1}^{k} z_{ic} = 1$$

where $\mathbf{x}_i$ is i-th sample instance, z_{ic} is a binary cluster assignment for i-th sample and cluster c; $\boldsymbol{\mu}_c = \sum_{i=1}^n z_{ic}\boldsymbol{\phi}(\mathbf{x}_i)/n_c$ is cluster center; $n_c = \sum_{i=1}^n z_{ic}$ is the size of cluster c; and n is the number of samples. This is viewed as to minimize within-cluster variance in the feature space. This problem can be reformulated as a trace minimization (Zha et al. 2002):

$$\underset{\mathbf{Z}\in\{0,1\}^{n\times k}}{\text{minimize}} \ \mathbf{tr}\left(\mathbf{K} - \mathbf{L}^{1/2}\mathbf{Z}^\top\mathbf{K}\mathbf{Z}\mathbf{L}^{1/2}\right) \quad \text{s.t. } \mathbf{Z}\mathbf{1}_k = \mathbf{1}_n$$

where $\mathbf{Z} = [z_{ic}]_{n\times k}$, $\mathbf{L} = \text{diag}\,[1/n_1, \cdots, 1/n_k]$, and $\mathbf{K} = [\boldsymbol{\phi}(\mathbf{x}_i)\cdot\boldsymbol{\phi}(\mathbf{x}_j)]_{n\times n}$. Unfortunately, this problem is NP-hard (Michael and David 1979). Note that $\mathbf{H} = \mathbf{Z}\mathbf{L}^{1/2}$ represents normalized clustering assignment. Hence, we solve it by eliminating the discrete constraint on $\mathbf{H}$ while keeping the orthogonal constraint on $\mathbf{H}$:

$$\underset{\mathbf{H}\in\mathcal{R}^{n\times k}}{\text{minimize}}\ \mathbf{tr}\left(\mathbf{K} - \mathbf{H}^\top\mathbf{K}\mathbf{H}\right) \quad \text{s.t. } \mathbf{H}^\top\mathbf{H} = \mathbf{I}_k \qquad (1)$$

This is solved by a well-known result from Fan (1949) (see Theorem S1). The optimal solution is given by $\mathbf{H} = \mathbf{U}_k\mathbf{Q}$ where each column of $\mathbf{U}_k = [\mathbf{u}_1, \cdots, \mathbf{u}_k]$ is eigenvectors of $\mathbf{K}$ involved with k largest eigenvalues $\lambda_1 \geq \cdots \geq \lambda_k$ and Q is an arbitrary orthogonal matrix. That is, the k eigenvalues are one of the continuous solutions to the discrete cluster assignment (Ding and He 2004). After obtaining the continuous solution, the hard clustering assignment $\mathbf{Z}$ is recovered by QR decomposition on $\mathbf{H}$ (Zha et al. 2002) or by k-means clustering on normalized $\mathbf{H}$ (Ng, Jordan, and Weiss 2002).

2.2 Existing multiple kernel k-means clustering

Multiple kernel k-means clustering extends kernel k-means clustering, which has an additional procedure to combine multiple views. It captures view-specific similarity with different kernels and combines multiple kernels weighted by kernel coefficient $\boldsymbol{\theta}$. For example, it uses $\mathbf{K}_{\boldsymbol{\theta}} = \sum_{v=1}^m \theta^{(v)}\mathbf{K}^{(v)}$ or $\mathbf{K}_{\boldsymbol{\theta}} = \sum_{v=1}^m {\theta^{(v)}}^2\mathbf{K}^{(v)}$ where $\boldsymbol{\theta}^{(v)}$ is a (non-negative) kernel coefficient for view v. For a given $\mathbf{K}_{\boldsymbol{\theta}}$, it finds clusters that minimize within-cluster variance in the combined space. The problem is defined as follow:

$$\underset{\mathbf{H}\in\mathcal{R}^{n\times k}}{\text{minimize}}\ \underset{\boldsymbol{\theta}}{\text{minimize}}\ \mathbf{tr}\left(\mathbf{K}_{\boldsymbol{\theta}} - \mathbf{H}^\top\mathbf{K}_{\boldsymbol{\theta}}\mathbf{H}\right) \qquad (2)$$

$$\text{s.t. } \mathbf{H}^\top\mathbf{H} = \mathbf{I}_k,\ \boldsymbol{\theta} \geq \mathbf{0}, f(\boldsymbol{\theta}) \leq \mathbf{0}$$

where $\boldsymbol{\theta} = \left[\theta^{(1)}, \cdots, \theta^{(m)}\right]^\top \in \mathcal{R}_+^m$, and $f(\boldsymbol{\theta}) \leq \mathbf{0}$ is an *appropriate* constraint on $\boldsymbol{\theta}$; without such constraint the inner minimization will have a trivial solution $\boldsymbol{\theta} = \mathbf{0}$. This problem is solved by alternately optimizing kernel coefficients $\boldsymbol{\theta}$ and clustering assignment matrix $\mathbf{H}$ given each other.

Existing methods are similar in that they all use the $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ (or $\max_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$) framework. Gönen and Margolin (2014) captured the sample-specific characteristics by using sample-specific kernel coefficients. Liu et al. (2017) extended Gonen's approach to perform clustering under incomplete kernel matrices. Liu et al. (2016) used a matrix-induced l_2 regularization on $\boldsymbol{\theta}$ to avoid redundancy and improve the diversity of multiple kernels. Yao and Chen (2018) incorporated a representative kernel selection process into multiple kernel k-means clustering to reduce redundancy and enhance the diversity of kernels. Yu et al. (2012) aimed to maximize between-cluster variance, hence, they used $\max_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$, instead of $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$.

2.3 Robust multiple kernel k-means clustering

We propose a multiple kernel k-means clustering method, MML-MKKC, that aims to be robust against adversarial perturbation. In order to achieve this, we use a $\min_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ formulation that combines views in a way to reveal high within-cluster variance in the combined space $\mathbf{K}_{\boldsymbol{\theta}}$ and then updates clusters by minimizing such variance. The optimization problem of our method is:

$$\underset{\mathbf{H}\in\mathcal{R}^{n\times k}}{\text{minimize}}\ \underset{\boldsymbol{\theta}}{\text{maximize}}\ \mathbf{tr}\left(\mathbf{K}_{\boldsymbol{\theta}} - \mathbf{H}^\top\mathbf{K}_{\boldsymbol{\theta}}\mathbf{H}\right) \qquad (3)$$
$$\text{s.t. } \mathbf{H}^\top\mathbf{H} = \mathbf{I}_k,\ \boldsymbol{\theta}^\top\boldsymbol{\theta} \leq 1,\ \boldsymbol{\theta} \geq \mathbf{0}$$

where $\mathbf{K}_{\boldsymbol{\theta}} = \sum_{v=1}^{m} \theta^{(v)}\mathbf{K}^{(v)}$. This problem can also be solved by alternately optimizing $\boldsymbol{\theta}$ and $\mathbf{H}$ given each other. Note that we employ l_2 regularization on $\boldsymbol{\theta}$ to avoid sparse solutions. The advantages of using an l_2 constraint were described previously in situations when the sources of data were carefully selected and carried complementary information (Yu et al. 2010; Kloft et al. 2009; 2011).

The $\min_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ framework is more favorable than $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ in the context of multiview clustering for the following reason. At every iteration, the inner maximization finds a combination of the views that maximizes within-cluster variance of the combined view given previous clusters, while the outer minimization updates clusters that minimizes such variance. We argue that by revealing high within-cluster variance in the combined space, our method can capture more comprehensive patterns of multiple views and thus has a better opportunity to find 'true' clusters. In the presence of adversarial perturbation, $\max_{\boldsymbol{\theta}}$ is particularly important because the effect of such perturbation can be mitigated when the method can tolerate a high within-cluster variance.

In contrast, the $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ framework aims to find a combination of the views that minimize within-cluster variance of the combined view given previous clusters, and then updates clusters that minimize such variance. This can be problematic when adversarial perturbation is present (which can commonly cause higher within-cluster variance of the view) because this $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ approach is not designed to tolerate the view with high within-cluster variance.

In Section 3.3, we illustrate with an example how adversarial perturbation affects multiview clustering and how the $\min_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ and $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ frameworks behave under adversarial perturbations.

3 Algorithm

We alternately optimize the kernel coefficients $\boldsymbol{\theta}$ and the continuous cluster assignment matrix $\mathbf{H}$ given each other: (i) given $\mathbf{H}$, $\boldsymbol{\theta}$ is optimized by solving a quadratically constrained linear programming (QCLP) problem, and (ii) given $\boldsymbol{\theta}$, $\mathbf{H}$ is optimized by solving the problem (1). R package implemented our method is freely available at https://github.com/BLINDED.

Before the iteration, we center the combined mapping function $\boldsymbol{\phi}_{\boldsymbol{\theta}}(\mathbf{x}_i)$ by using a kernel trick $\mathbf{K}_{\boldsymbol{\theta}} \leftarrow \mathbf{K}_{\boldsymbol{\theta}} - \mathbf{J}_n\mathbf{K}_{\boldsymbol{\theta}} - \mathbf{K}_{\boldsymbol{\theta}}\mathbf{J}_n + \mathbf{J}_n\mathbf{K}_{\boldsymbol{\theta}}\mathbf{J}_n$ where $\mathbf{J}_n = \mathbf{1}_n\mathbf{1}_n^\top/n$ (Schölkopf, Smola, and Müller 1998). We scale each kernel matrix before combining them by $\mathbf{K}^{(v)} \leftarrow \mathbf{K}^{(v)}/\mathbf{tr}\left(\mathbf{K}^{(v)}\right)$ to make multiple views comparable to each other (Ong and Zien 2008; Kloft et al. 2011). We refer to Text S1 for a detailed discussion about centering and scaling.

3.1 Estimation of $\boldsymbol{\theta}$

Given $\mathbf{H}$, the optimization problem (3) is reformulated as:

$$\underset{\boldsymbol{\theta}}{\text{maximize}} \sum_{v=1}^{m} \theta^{(v)}\mathbf{tr}\left(\mathbf{K}^{(v)} - \mathbf{H}^\top\mathbf{K}^{(v)}\mathbf{H}\right) \qquad (4)$$
$$\text{s.t. } \frac{1}{2}\boldsymbol{\theta}^\top\mathbf{Q}_m\boldsymbol{\theta} \leq 1,\ \boldsymbol{\theta} \geq \mathbf{0}$$

where $\mathbf{Q}_m = \text{diag}\,[2, \cdots, 2]$. Since $\mathbf{Q}_m$ is a diagonal matrix, this problem is separable. Hence, the entire problem is solved as a conic quadratic program (i.e. second order cone program). It usually performs better than QCLP and is based on more solid duality theory (Andersen 2016). Therefore, we translate QCLP to the conic formulation as follows:

$$\underset{\boldsymbol{\theta}}{\text{maximize}}\ \mathbf{c}^\top\boldsymbol{\theta} \quad \text{s.t. } [p, \boldsymbol{\theta}]^\top \in \mathcal{K}^q,\ p = 1,\ \mathbf{0} \leq \mathbf{I}_m\boldsymbol{\theta}$$

where $\mathbf{c}^\top{=}[\mathbf{tr}\big(\mathbf{K}^{(v)}{-}\,\mathbf{H}^\top\mathbf{K}^{(v)}\mathbf{H}\big)\ \cdots,\ \mathbf{tr}\big(\mathbf{K}^{(m)}{-}\,\mathbf{H}^\top\mathbf{K}^{(m)}\mathbf{H}\big)]$ and $\mathcal{K}^q{=}\Big\{p{\geq}\sqrt{\sum_{v=1}^{m}\theta^{(v)2}}\Big\}$. This problem is analytically solved by existing software such as `mosek` (MOSEK-ApS 2017). In fact, the optimization problem has a closed form solution (See Proposition S3 and S4 for proof):

$$\boldsymbol{\theta}{=}\left(\frac{g^{(1)}}{\sqrt{\left(g^{(1)}\right)^2{+}\cdots{+}\left(g^{(m)}\right)^2}},\ldots,\frac{g^{(m)}}{\sqrt{\left(g^{(1)}\right)^2{+}\cdots{+}\left(g^{(m)}\right)^2}}\right)$$

where $g^{(v)}(\mathbf{H}) = \mathbf{tr}\left(\mathbf{K}^{(v)} - \mathbf{H}^\top\mathbf{K}^{(v)}\mathbf{H}\right)$ is the within-cluster variance in view v and $g^{(v)}(\mathbf{H}) \geq 0$. More precisely, $g^{(v)}(\mathbf{H})$ is a sum of variance and covariance of view v that are not explained by the previous clusters $\mathbf{H}$. Therefore, a view will have larger $\theta^{(v)}$ if its variability is not well explained by previous clsuters; and a combined view weighted by such $\boldsymbol{\theta}$ will have higher within-cluster variance by doing so, it updates $\boldsymbol{\theta}$ to find a combination of views with higher within-cluster variance.

We mathematically prove it by showing

$$g^{(v)}(\mathbf{H}){=}\mathbf{tr}\Big(\mathbf{X}^{(v)}\mathbf{X}^{(v)\top}\Big){-}\Big(\mathbf{tr}\Big(\mathbf{V}_{1:k}^{(v)\top}\mathbf{X}^{(v)\top}\mathbf{X}^{(v)}\mathbf{V}_{1:k}^{(v)}\Big)$$
$$+\textstyle\sum_{w\neq v}\mathbf{tr}\Big(\mathbf{V}_{1:k}^{(v)\top}\mathbf{X}^{(v)\top}\mathbf{X}^{(w)}\mathbf{V}_{1:k}^{(w)}\Big)\Big)$$

where $\mathbf{X}^{(v)}$ is a $n \times p_v$ centered data matrix for view v, $\mathbf{V}_{1:k}^{(1)}$ is a matrix including the first p_1 rows of an orthogonal matrix $\mathbf{V}_{1:k}$ whose columns are the first k right-singular vectors of $\mathbf{X}$, $\mathbf{V}_{1:k}^{(2)}$ is a matrix including the next p_2 rows of $\mathbf{V}_{1:k}$, and so on. Without loss of generality, we assume $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$. See Proposition S2 for proof.

This equation provides a more precise description about how $\boldsymbol{\theta}$ is estimated. Note that $\mathbf{tr}\Big(\mathbf{X}^{(v)}\mathbf{X}^{(v)T}\Big)$ is viewed as total variance of view v; $\mathbf{tr}\Big(\mathbf{V}_{1:k}^{(v)T}\mathbf{X}^{(v)T}\mathbf{X}^{(v)}\mathbf{V}_{1:k}^{(v)}\Big)$ is viewed as variance of view v explained by $\mathbf{H}$; $\mathbf{tr}\Big(\mathbf{V}_{1:k}^{(v)T}\mathbf{X}^{(v)T}\mathbf{X}^{(w)}\mathbf{V}_{1:k}^{(w)}\Big)$ is viewed as covariance of view v and view w explained by $\mathbf{H}$. Consequently, above equation tells that $g^{(v)}(\mathbf{H})$ is a sum of unexplained variance and covariance of view v given previous clusters $\mathbf{H}$. Considering $\theta^{(v)}$ is proportional to $g^{(v)}(\mathbf{H})$, we conclude that a view has a greater $\theta^{(v)}$ when its variability is not well explained by previous clusters.

3.2 Estimation of H

Given $\boldsymbol{\theta}$, the optimization problem (3) is reduced to a simple kernel k-means clustering problem. This is the same with the problem (1) and the optimal solution is $\mathbf{H} = \mathbf{U}_k\mathbf{Q}$. Columns of $\mathbf{U}_k$ are eigenvectors of $\mathbf{K}$ corresponding to the k largest eigenvalues, and $\mathbf{Q}$ is an arbitrary orthogonal matrix. Hence, any spectral clustering methods can be used to restore the binary clustering assignment matrix $\mathbf{Z}$ from the continuous clustering assignment matrix $\mathbf{H}$. Here, we use a spectral clustering method proposed by Ng, Jordan, and Weiss (2002).

3.3 Illustration with an example

We illustrate with an example how an adversarial feature affects multiview clustering and how the $\min_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ and $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ frameworks behave under adversarial perturbation. Consider a two-view data $\{(\mathbf{x}_i^{(A)}, \mathbf{x}_i^{(B)})\}_{i=1}^N$ and unobserved cluster labels $\{c_i\}_{i=1}^N$ $(c = 1, 2, 3)$ for samples in the data where the two views, view A and view B, have complementary patterns from each other. More precisely, we consider:

$$\mathbf{x}^{(A)} \mid c \sim \mathcal{N}(\boldsymbol{\mu}_1 \cdot 1_{\{c=1\}} + \boldsymbol{\mu}_2 \cdot 1_{\{c\neq1\}},\ \Sigma^2)$$
$$\mathbf{x}^{(B)} \mid c \sim \mathcal{N}(\boldsymbol{\mu}_1 \cdot 1_{\{c=3\}} + \boldsymbol{\mu}_2 \cdot 1_{\{c\neq3\}},\ \Sigma^2)$$

where $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$; $\mathbf{x}_i^{(A)} \in \mathcal{R}^{p_A}$ and $\mathbf{x}_i^{(B)} \in \mathcal{R}^{p_B}$ where $p_A = p_B = p$ for simplicity. Note that $\mathbf{x}_i^{(A)}$ can separate cluster 1 from others and $\mathbf{x}_i^{(B)}$ can separate cluster 3 from others, hence views A and B together can separate all three clusters. We then add an adversarial feature,

$$\epsilon \sim \mathcal{N}(0, 1)$$

to view A, hence view A has $p + 1$ features, $\mathbf{x}_i^{(A)}$ and ϵ. Without loss of generality, we assume that $\phi(\mathbf{x}) = \mathbf{x}$. We also center and scale features so that each feature has zero sample mean and unit variance.

Following Ong and Zien; Kloft et al. (2008; 2011), we scale the kernels (discussed in Text S1), so that the two views have the same total variance by doing $\mathbf{K}^{(v)} \leftarrow \mathbf{K}^{(v)}/\mathbf{tr}\left(\mathbf{K}^{(v)}\right)$. This is important because it allows the two views to be comparable to each other. Thus, we get $\mathbf{K}^{(A)} = \left[\frac{1}{p+1}\left(\mathbf{x}_i^{(A)} \cdot \mathbf{x}_j^{(A)} + \epsilon_i\epsilon_j\right)\right]_{ij}$ and $\mathbf{K}^{(B)} = \left[\frac{1}{p}\mathbf{x}_i^{(B)} \cdot \mathbf{x}_j^{(B)}\right]_{ij}$. At the initial step, with uniform kernel coefficients $\theta^{(A)} = \theta^{(B)} = 1/2$, we have:

$$\mathbf{K}_{\boldsymbol{\theta}} = \left[\frac{1}{p+1}\left(\mathbf{x}_i^{(A)} \cdot \mathbf{x}_j^{(A)} + \epsilon_i\epsilon_j\right) + \frac{1}{p}\mathbf{x}_i^{(B)} \cdot \mathbf{x}_j^{(B)}\right]_{ij} \times \frac{1}{2}. \qquad (5)$$

Note that each element of $\mathbf{K}_{\boldsymbol{\theta}}$ represents dissimilarity between a pair of samples in the combined space. Eq. (5) shows that dissimilarity measured by $\mathbf{x}^{(B)}$ contributes more than that measured by $\mathbf{x}^{(A)}$, which can make the initial clusters that are largely based on $\mathbf{x}^{(B)}$. This leads to the between-cluster variance of $\mathbf{x}^{(B)}$ greater than that of $\mathbf{x}^{(A)}$, i.e., $\sum_{c=1}^3 N_c\overline{\mathbf{x}_c}^{(B)} \cdot \overline{\mathbf{x}_c}^{(B)} \geq \sum_{c=1}^3 N_c\overline{\mathbf{x}_c}^{(A)} \cdot \overline{\mathbf{x}_c}^{(A)}$ (in probability) where $\overline{\mathbf{x}_c} = \sum_{c_i=c} \mathbf{x}_i/N_c$ is the cluster center and N_c is the size of cluster c.

Note that within-cluster variance in each view is given as:

$$g^{(A)}(\mathbf{H}) = 1 - \frac{1}{p+1}\sum_{c=1}^3 N_c\left(\overline{\mathbf{x}}_c^{(A)} \cdot \overline{\mathbf{x}}_c^{(A)} + \overline{\epsilon_c}^2\right)$$
$$g^{(B)}(\mathbf{H}) = 1 - \frac{1}{p}\sum_{c=1}^3 N_c\overline{\mathbf{x}_c}^{(B)} \cdot \overline{\mathbf{x}}_c^{(B)}.$$

From above, we can infer that $g^{(A)}(\mathbf{H}) \geq g^{(B)}(\mathbf{H})$ (in probability) from $\sum_{c=1}^3 N_c\overline{\mathbf{x}_c}^{(B)} \cdot \overline{\mathbf{x}_c}^{(B)} \geq \sum_{c=1}^3 N_c\overline{\mathbf{x}_c}^{(A)} \cdot \overline{\mathbf{x}_c}^{(A)}$ and $\overline{\epsilon_c} = 0$ where $\overline{\epsilon_c} = \sum_{c_i=c} \epsilon_i/N_c$ is the cluster mean of ϵ. On the other hand, if there is no adversarial perturbation, the within-cluster variance of view A tends to be equal to that of view B, i.e., $g^{(A)}(\mathbf{H}) = g^{(B)}(\mathbf{H})$ in probability. Thus, the adversarial feature added to view A causes the disparity of the within-cluster variance between views A and B (i.e. $g^{(A)}(\mathbf{H})$ and $g^{(B)}(\mathbf{H})$).

Under the disparity originated from the adversarial feature added to view A, the $\min_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ first updates $\theta^{(A)}$ to have a larger value than $\theta^{(B)}$. This makes variance of view A is magnified (relative to view B) in the combined space so that $\mathbf{H}$ can explain more variability of view A than B. At every later iteration, $\min_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ alternately magnifies variance of each view while alleviating the disparity. As a result, it yields clusters at a saddle point where both views are almost equally favored, thus finding all three clusters by using complementary patterns from both views.

On the other hand, the $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ first updates $\theta^{(B)}$ to have a larger value than $\theta^{(A)}$. This makes variance of view B is magnified (relative to view A) in the combined space so that $\mathbf{H}$ can explain more variability of view B than A. At every later iteration, $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ keep magnifying variance of view B while aggravating the disparity. As a result, it yields clusters that are largely determined by view B.

Furthermore, from this example, we can see that the $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ framework is not preferable when: i) a view is adversarially perturbed; and/or ii) true clusters are determined by complement views that together provide comprehensive patterns about the clusters.

This above-described difference between the $\min_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ and $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ is also observed in simulation experiments (Figure 1).

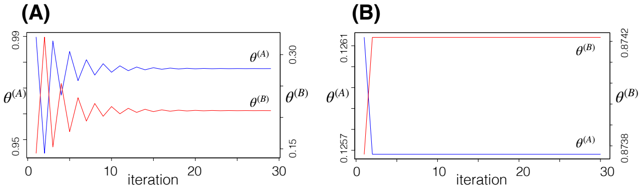

Figure 1: Kernel coefficients $\boldsymbol{\theta}$ are updated by (A) the $\min_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ (our method) and (B) the $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ (Gonen's MKK is shown here, but other compared methods show similar behaviors). One of our simulation data (B-Noise with five noise variables) is used.

3.4 Proof of convergence

We prove that if our alternating strategy converges, it will converge to the global optimal solution, which motivates our alternating strategy. First, recall the $\min_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ problem (3) and the optimization function $f(\mathbf{H}, \boldsymbol{\theta}) = \mathbf{tr}\left(\mathbf{K}_{\boldsymbol{\theta}} - \mathbf{H}^\top\mathbf{K}_{\boldsymbol{\theta}}\mathbf{H}\right)$ where $\mathbf{H}^\top\mathbf{H} = \mathbf{I}_k$, $\boldsymbol{\theta}^\top\boldsymbol{\theta} \leq 1$, and

$\boldsymbol{\theta} \geq \mathbf{0}$. A saddle point $(\mathbf{H}^*, \boldsymbol{\theta}^*)$ of this problem is defined as follows:

$$f(\mathbf{H}^*, \boldsymbol{\theta}^*) \geq f(\mathbf{H}^*, \boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta}$$
$$f(\mathbf{H}^*, \boldsymbol{\theta}^*) \leq f(\mathbf{H}, \boldsymbol{\theta}^*) \text{ for all } \mathbf{H}.$$

That is, given $\mathbf{H}^*$, $\boldsymbol{\theta}^*$ is the maximum among all $\boldsymbol{\theta}$ and given $\boldsymbol{\theta}^*$, $\mathbf{H}^*$ is the minimum among all $\mathbf{H}$. From this definition, it is clear that if the alternating strategy converges, then it converges to a saddle point. Moreover, it is known that if $(\mathbf{H}^*, \boldsymbol{\theta}^*)$ is a saddle point, then: (1) $\mathbf{H}^*$ is a globally optimal solution for $\min_{\mathbf{H}} f_1(\mathbf{H})$ where $f_1(\mathbf{H}) = \max_{\boldsymbol{\theta}} f(\mathbf{H}, \boldsymbol{\theta})$ and (2) $\boldsymbol{\theta}^*$ is a globally optimal solution for $\max_{\boldsymbol{\theta}} f_2(\boldsymbol{\theta})$ where $f_2(\boldsymbol{\theta}) = \min_{\mathbf{H}} f(\mathbf{H}, \boldsymbol{\theta})$. Therefore, if we want to find a global minimizer $\mathbf{H}$, we can try to find a saddle point $(\mathbf{H}^*, \boldsymbol{\theta}^*)$ whose definition motivates our alternating strategy. This alternating strategy is practically efficient and easy to implement because each of the two steps ($\min_{\mathbf{H}}$, $\max_{\boldsymbol{\theta}}$) has a closed-form solution which requires fewer iterations than a gradient approach to converge. Such alternating approach has also been used in other works such as Generative Adversarial Net (Goodfellow et al. 2014), and robust models against adversarial examples (Madry et al. 2018; Sinha, Namkoong, and Duchi 2018).

4 Simulation experiments

4.1 Adversarial perturbations

We evaluate robustness of our method against two types of adversarial features added to a view:

- Noise variables that are independently sampled from Gaussian distribution with zero-mean and unit-variance. We add different numbers ($N_{noise} = 0, 1, \cdots$) of noise variables.
- Redundant variables that are correlated with original variables. We add different numbers ($N_{redun} = 1, 2, \cdots$) of variables having different correlations ($cor = 1, 0.97, 0.90, 0.72, 0.45$) with the original variables.

Under these perturbations, we examine how our method makes use of complementary patterns of multiple views. For this purpose, we first generated multiview data in three scenarios A–C having two or three views that have complementary patterns necessary for identifying true clusters. Scenario A is composed of a complete view that has complete information to detect the three clusters and a partial view that only conveys partial information. Scenario B is composed of two different partial views so that each view alone cannot completely detect the three clusters. Both scenarios A & B aim to test how the compared methods use the complementary information in two views. Scenario C is composed of two different partial views and a noise view. It aims to test further whether the methods robustly use complementary information from views even when one of the views contains only noise variables.

Then, as illustrated in Figure 2, we added different types and levels of adversarial features to one of the views. We denote the simulation data with the noise variables by A-Noise, B-Noise, and C-Noise, and the data with the redundant variables by A-Redun, B-Redun, and C-Redun. For detailed explanation about data and preprocessing, see Text S2.

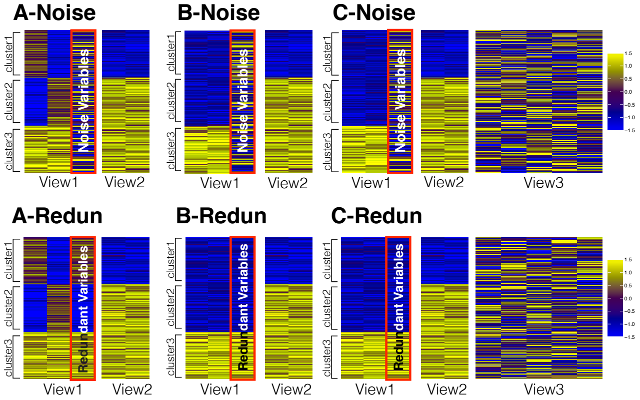

Figure 2: Simulation scenarios with two types of adversarial perturbations, noise and redundancy. The samples in the heatmaps are ordered by their true cluster index. Each simulation data contains 300 samples with 100 samples in each cluster, i.e. three true clusters for each data.

4.2 Compared methods

We compare our method with seven other methods: two baseline methods, **Single Best** and **Uniform Weight**; four multiple kernel k-means clustering methods, **Gonen's MKK** and **LMKK** (Gönen and Margolin 2014), **Liu's MKK-MIR** (Liu et al. 2016), and **Yu's OKKC** (Yu et al. 2012)); and one variant of our method, **MinMax-MinC**. Single Best uses the best view that has the smallest within-cluster variance. Uniform Weight gives the same weights to all views. Gonen's MKK, Gonen's LMKK, and Liu's MKK-MIR combine multiple kernels by $\mathbf{K}_{\boldsymbol{\theta}} = \sum_{v=1}^{m} {\theta^{(v)}}^2 \mathbf{K}^{(v)}$, with l_1 constraint on $\boldsymbol{\theta}$, and use the $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ framework. Yu's OKKC combines multiple kernels by $\mathbf{K}_{\boldsymbol{\theta}} = \sum_{v=1}^{m} \theta^{(v)} \mathbf{K}^{(v)}$ and uses l_p constraint on $\boldsymbol{\theta}$ where $p \geq 1$, and uses the $\max_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ framework. MinMax-MinC is the l_1-regularization version of our method, which is included to examine the effect of l_2-regularization in our method. For a detailed description, see Text S2.

We evaluate how robustly the methods recover the true cluster against such perturbations. For evaluation, we use three metrics: Adjusted Rand Index (AdjRI, (Hubert and Arabie 1985)), Normalized Mutual Information (NormMI, (Strehl and Ghosh 2002)), and Purity (Manning, Raghavan, and Schütze 2008).

4.3 Simulation results

We illustrate the results in Figures 3 & 4. As shown in Figure 3, when there is no adversarial perturbation, all methods accurately identifies the true clusters; however, when the adversarial perturbations are present, our method identifies the true clusters more robustly than others. Note that existing $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ methods perform even worse than Uniform Weight. Figure 4 illustrates that our result agrees well with the ground truth, which indicates that it better uses the complementary information in both views. However, the other $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ methods fail to distinguish the first two clusters in Figure 4A and the last two clusters in Figure 4B, which

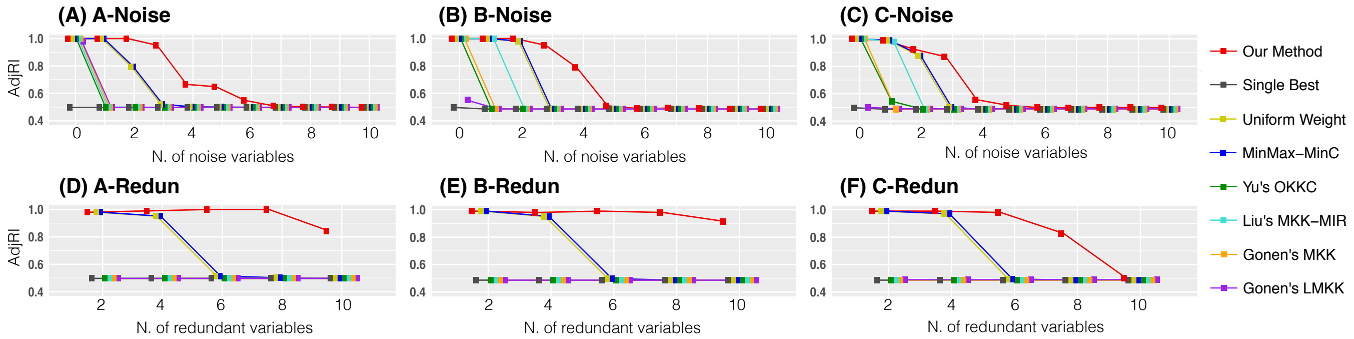

Figure 3: Clustering performance. AdjRI versus the number of the noise (A–C) or redundant variables (D–F, $cor = 0.90$) added to view 1. The identified clusters are compared to the true clusters.

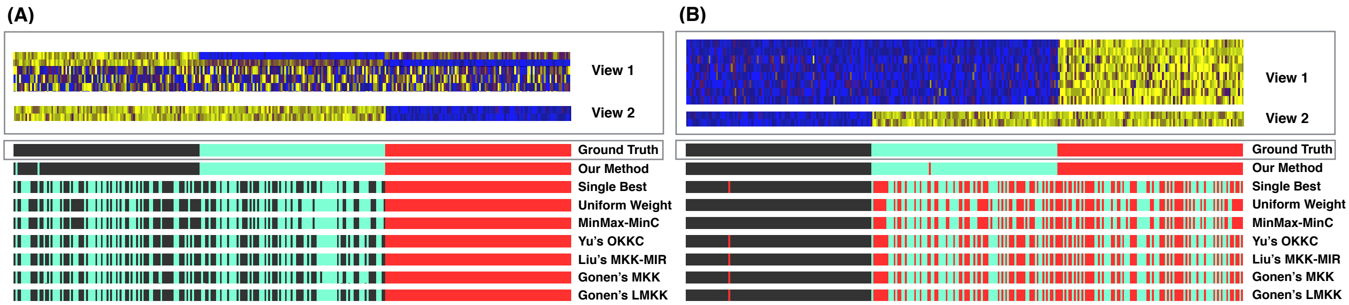

Figure 4: Clustering results of (A) A-Noise with three noise variables and (B) B-Redun with three redundant variables and $cor = 0.7$. The heatmaps at the top panel illustrate the two-view data where the rows and columns represent variables and samples, respectively. The tile plots at the bottom panel illustrate the clustering results. The first tile plot shows the ground truth; the second to the last plots show results of the methods. Identified clusters are labeled in different colors.

suggests that they identify clusters mainly based on View 2 that is not perturbed. For further results, see Table S1– S6.

5 Real data analysis

With the advent of various genome-wide technologies, a wide array of biomedical data has been available, which includes clinical characteristics, DNA copy number, and gene expression profiling. They contain complementary information that can together provide a comprehensive understanding and novel insight into biomedical problems. In this section, we present application of our method to two biomedical problems–one is to identify cancer subtypes, and the other is to identify patients' response patterns to asthma treatment, which demonstrates the utility of our method on real-world problems.

5.1 Identification of cancer subtype

We compared our method with other methods using two TCGA multi-omics cancer datasets. Each dataset includes 468 patients with human breast invasive carcinoma (BRCA) and 251 patients with glioblastoma multiforme (GBM), respectively (Weinstein et al. 2013). BRCA has three views: mRNA sequencings, miRNA sequencings, and copy number variations. GBM also has three views: gene expression microarray profiling, copy number variation, and methylation data. A radial basis function kernel is used for all views as suggested by Lanckriet et al. (2004b). See Text S3 for details about data preprocessing.

Since there is no ground-truth subtype, we compared clinical properties of identified clusters that are observed independently from data and examined how distinct they are. In BRCA, we compared the *AJCC neoplasm disease stage* (which describes the extent of both malignant and benign growths). In GBM, we compared the survival time (days to death) and the *Karnofsky performance score* (a patient's prognosis by measuring a patient's ability to function).

Further, to help understand biological mechanisms underlying the clusters, we identified differentially expressed genes (DEGs) for each cluster. We used RNA sequencing data for BRCA and gene expression microarray data for GBM and performed the two-sample t-test. The p-values are adjusted using a Benjamini-Hochberg procedure to address multiple hypothesis testing problems (Benjamini and Hochberg 1995). We performed gene set enrichment analysis (Subramanian et al. 2005) to find out the KEGG pathways enriched among the DEGs in each cluster. Then, we compared these enriched pathways with the BRCA- or GBM-related biological pathways provided by the KEGG Pathway Database (https://www.kegg.jp) that are defined independently from data (See Table S7).

Results: For ***BRCA***, we identified five clusters (92, 86, 83, 137, and 70 subjects for each cluster) using each methods. Table 1 shows that our method identified clusters that

Table 1: **Evaluation of clustering results.**

Cancer Type	**BRCA**	**GBM**	
Method	AJCC disease stage (p-value)	Survival time (p-value)	Karnofsky score (p-value)
Single Best	0.54	1.13×10^{-1}	0.89
Uni. Weight	0.22	$\mathbf{1.37 \times 10^{-5}}$	0.13
MinMax-MinC	0.22	$\mathbf{1.37 \times 10^{-5}}$	0.13
Yu's OKKC	0.53	2.51×10^{-5}	0.26
Liu's MKK-MIR	0.42	2.21×10^{-5}	0.16
Gonen's MKK	0.48	1.46×10^{-3}	0.88
Gonen's LMKK	0.56	6.68×10^{-2}	0.74
MML-MKKC	**0.09**	$\mathbf{1.37 \times 10^{-5}}$	**0.02**

Note: for BRCA, differences in the AJCC neoplasm disease stages among the clusters are compared using the chi-square test. For GBM, differences in the survival curves and the Karnofsky performance scores among the clusters are compared using the log-rank test (Mantel 1966) and the chi-square test, respectively.

have the most distinct disease stages. Further, our method identified clusters that have many enriched pathways such as *cell cycle* and *alanine* that are consistent with the BRCA-related pathways in the KEGG pathway database (See Table S8). For ***GBM***, we identified five clusters (58, 38, 39, 46, and 70 subjects for each cluster). Table 1 shows that our method identified clusters that have the most distinct survival time and the patients' ability to carry daily activities (as measured by the Karnofsky score). Further, our method identified clusters that have many enriched pathways such as *adherens junction* and *calcium signaling pathway* that are consistent with the GBM-related pathways in the KEGG pathway database (See Table S9). Together, these results show that, compared to the other methods, our method better identifies distinct clusters that are relevant to the pathobiological mechanisms underlying the diseases.

5.2 An application of MML-MKKC to identifying response patterns of asthma patients to corticosteroids

Corticosteroids (CSs) are the most effective treatment for asthma, however they cause many harmful side effects. To avoid unnecessary harmful effects, it is greatly beneficial to asthma patients if clinicians know in advance which patients respond to CSs or not before the medications are given to them. For this purpose, we recently applied our method to cluster 346 asthma patients in the Severe Asthma Research Program (SARP). The data has clinical, physiologic, and inflammatory variables that are complex and heterogeneous in nature, with 70 baseline variables measured only before the CS treatment, and 15 dynamic variables measured both before and after CSs.

To cluster the patients, we first assigned the variables to three different views, based on i) the clinical importance of the variables according to the opinion of domain experts, and ii) whether a variable is static or dynamic. Then using our method MML-MKKC, we identified four clusters of asthma patients with distinct CS response patterns, which were validated using an independent SARP cohort. Our clinical findings are recently published in the American Journal of Respiratory and Critical Care Medicine (Anonymous-Author 2019)[1] This application demonstrates that our method can help clinicians make a better decision and provide novel insights to real-world problems.

6 Conclusion

In this paper, we investigate the effects of adversarial perturbation on multiple kernel k-means clustering. We show that such perturbation can make the existing methods with the $\min_{\mathbf{H}}$-$\min_{\boldsymbol{\theta}}$ formulation ignore the perturbed view and find clusters largely depend on other view(s). To address this problem, we propose a multiple kernel k-means clustering method, MML-MKKC, which aims to be robust to adversarial perturbation by using the $\min_{\mathbf{H}}$-$\max_{\boldsymbol{\theta}}$ formulation.

Our algorithm is practically efficient and easy to implement because it alternately optimizes $\boldsymbol{\theta}$ and $\mathbf{H}$ where each of the two steps ($\min_{\mathbf{H}}$ and $\max_{\boldsymbol{\theta}}$) has a closed-form solution that requires fewer iterations than a gradient approach to converge. In simulation experiments, we showed that our method is more robust to adversarial perturbation than other methods. In real data analysis, our method identified the most distinct clusters of cancer patients, as well as uncovered clusters of asthma patient showing differential response patterns to corticosteroid. Together, these findings demonstrate the utility of our method on real-world problems.

References

Andersen, E. D. 2016. *On formulating quadratic functions in optimization models.*

Anonymous-Author, A. 2019. Blinded due to the double-blinding policy. *Am J Respir Crit Care Med.*

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSS B* 289–300.

Biggio, B.; Corona, I.; Maiorca, D.; Nelson, B.; Šrndić, N.; Laskov, P.; Giacinto, G.; and Roli, F. 2013. Evasion attacks against machine learning at test time. In *ECML PKDD*, 387–402. Springer.

Ding, C., and He, X. 2004. K-means clustering via principal component analysis. In *Proceedings of the 21st International Conference on Machine learning*, 29. ACM.

Fan, K. 1949. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proc Natl Acad Sci* 35(11).

Gehler, P., and Nowozin, S. 2009. On feature combination for multiclass object classification. In *ICCV*, 221–228. IEEE.

Girolami, M. 2002. Mercer kernel-based clustering in feature space. *IEEE Trans Neural Netw* 13(3):780–784.

Gönen, M., and Margolin, A. A. 2014. Localized data fusion for kernel k-means clustering with application to cancer biology. *NeurIPS.*

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *NeurIPS.*

[1] Blinded due to the double-blinding policy.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. *ICLR*.

Hubert, L., and Arabie, P. 1985. Comparing partitions. *Journal of Classification* 2(1):193–218.

Kloft, M.; Brefeld, U.; Laskov, P.; Müller, K.-R.; Zien, A.; and Sonnenburg, S. 2009. Efficient and accurate l_p-norm multiple kernel learning. *NeurIPS*.

Kloft, M.; Brefeld, U.; Sonnenburg, S.; and Zien, A. 2011. l_p-norm multiple kernel learning. *J Mach Learn Res* 12(Mar):953–997.

Lanckriet, G. R.; Cristianini, N.; Bartlett, P.; Ghaoui, L. E.; and Jordan, M. I. 2004a. Learning the kernel matrix with semidefinite programming. *J Mach Learn Res* 5(Jan):27–72.

Lanckriet, G. R.; De Bie, T.; Cristianini, N.; Jordan, M. I.; and Noble, W. S. 2004b. A statistical framework for genomic data fusion. *Bioinformatics* 20(16):2626–2635.

Liu, X.; Dou, Y.; Yin, J.; Wang, L.; and Zhu, E. 2016. Multiple kernel k-means clustering with matrix-induced regularization. *AAAI* 1888–1894.

Liu, X.; Li, M.; Wang, L.; Dou, Y.; Yin, J.; and Zhu, E. 2017. Multiple kernel k-means with incomplete kernels. *AAAI*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In *ICLR*.

Manning, C. D.; Raghavan, P.; and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Mantel, N. 1966. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Reports* 50:163–170.

Michael, R. G., and David, S. J. 1979. Computers and intractability: a guide to the theory of np-completeness. *WH Free. Co., San Fr* 90–91.

MOSEK-ApS. 2017. *MOSEK Rmosek Package Release 8.0.0.81*.

Ng, A. Y.; Jordan, M. I.; and Weiss, Y. 2002. On spectral clustering: Analysis and an algorithm. In *NeurIPS*.

Ong, C., and Zien, A. 2008. An automated combination of kernels for predicting protein subcellular localization. *Algorithms in Bioinformatics* 186–197.

Schölkopf, B.; Smola, A.; and Müller, K.-R. 1998. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* 10(5):1299–1319.

Sinha, A.; Namkoong, H.; and Duchi, J. 2018. Certifying some distributional robustness with principled adversarial training. *ICLR*.

Strehl, A., and Ghosh, J. 2002. Cluster ensembles—a knowledge reuse rramework for combining multiple partitions. *J Mach Learn Res* 3(Dec):583–617.

Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci* 102(43):15545–15550.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. *ICLR*.

Weinstein, J. N.; Collisson, E. A.; Mills, G. B.; Shaw, K. R. M.; Ozenberger, B. A.; Ellrott, K.; Shmulevich, I.; Sander, C.; Stuart, J. M.; Network, C. G. A. R.; et al. 2013. The cancer genome atlas pan-cancer analysis project. *Nature Genet* 45(10):1113.

Yao, Y., and Chen, H. 2018. Multiple kernel k-means clustering by selecting representative kernels. *arXiv:1811.00264*.

Yu, S.; Falck, T.; Daemen, A.; Tranchevent, L.-C.; Suykens, J. A.; De Moor, B.; and Moreau, Y. 2010. L_2-norm multiple kernel learning and its application to biomedical data fusion. *BMC Bioinformatics* 11(1):309.

Yu, S.; Tranchevent, L.; Liu, X.; Glanzel, W.; Suykens, J. A.; De Moor, B.; and Moreau, Y. 2012. Optimized data fusion for kernel k-means clustering. *IEEE Trans Pattern Anal Mach Intell* 34(5):1031–1039.

Zha, H.; He, X.; Ding, C.; Gu, M.; and Simon, H. D. 2002. Spectral relaxation for k-means clustering. *NeurIPS*.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically principled trade-off between robustness and accuracy. *arXiv:1901.08573*.