Learning Mixtures of Multi-Output Regression Models By Correlation Clustering for Multi-View Data

Eric Lei
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Kyle Miller
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Michael Pinsky
School of Medicine
University of Pittsburgh
Pittsburgh, PA 15213

Artur Dubrawski
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

Multi-view data are an increasingly prevalent type of dataset that allows exploitation of relationships between sets of variables. It is often interesting to analyze the correlation between two views via multi-view component analysis techniques such as Canonical Correlation Analysis (CCA). However, different parts of the data may have their own patterns of correlation, which CCA cannot reveal. To address this challenge, we propose a method called Canonical Least Squares (CLS) clustering. Somewhat like CCA, a single CLS model can be regarded as a multi-output regression model that finds latent variables to connect inputs and outputs. This method, however, also identifies partitions of data that enhance correlations in each partition, which may be useful when different correlation structures appear in different subsets of the data or when nonlinear correlations may be present. Furthermore, we introduce a supervised classification method that relies on CLS clustering. The value of these methods rests in their capability to find interpretable structure in the data to explain their predictions. We demonstrate the potential utility of the proposed approach using an example application in clinical informatics to detect and characterize slow bleeding in patients whose vital signs are monitored at the bedside. We empirically show how the proposed method can help discover and analyze multiple-to-multiple correlations, which could be nonlinear or vary throughout the population, while retaining interpretability of the resulting models.

1 INTRODUCTION

Multi-view data are an increasingly prevalent type of dataset that allows exploitation of relationships between sets of variables. Examples include genes and diseases (Seoane et al., 2014), visuals and text (Rasiwasia et al., 2010), and emotions and personality disorders (Sherry and Henson, 2005). It is often interesting to analyze the correlation between two views. A method of choice is Canonical Correlation Analysis (CCA) from classical statistics (Hotelling, 1936), which analyzes correlations between two views of data. CCA finds linear projections from each data view into a shared latent space such that the projections have maximal correlation. According to Bach and Jordan (2006), the canonical, or latent, variables can be considered the basis of a generative model for the observed views. These canonical variables often have some practical meaning, such as a certain combination of genes that corresponds to a combination of phenotypes (Witten and Tibshirani, 2009). CCA, or more generally component analysis, can therefore be used to analyze complex datasets in an interpretable fashion. In many practical scenarios, however, different parts of the data may have distinct patterns of correlation; namely, important canonical variables might differ between subsets of observations. For instance, certain populations might express a gene combination differently, or distinct subsets of subjects might have a different physiological response to medical trauma. Additionally, there may be global nonlinear correlation structure in the data.

To address these challenges, we propose a method called Canonical Least Squares (CLS) clustering. A single CLS model can be regarded as a multi-output regression model that finds latent variables to connect inputs and outputs, somewhat like CCA. The proposed approach, however, also identifies a clustering of data, which may be useful when different correlation structures appear in different subsets of the data and when nonlinear correlations may be present. Furthermore, we introduce a supervised clas-

sification method that relies on CLS clustering. The value of these methods exists in their capability to find interpretable structure in the data to explain their predictions. The correlation structures found in each cluster are linear, which aids interpretation, and the classification score has a gradient that is straightforward to compute and interpret.

As an example application, we consider an interesting area in medicine concerning the detection of slow bleeding. In variety of clinical settings it is common to encounter datasets of high complexity, reflecting the high variance in physiological responses. Although machine learning models have been applied successfully in a variety of clinical scenarios, medical practitioners are wary of methods that lack explanatory power (Murdoch and Detsky, 2013; Krumholz, 2014; Holzinger and Jurisica, 2014; Obermeyer and Emanuel, 2016; Choi et al., 2016). Recently a growing effort has been made to develop "translucent box" methods that offer some extent of human interpretability of the learned models and predictions they make, improving face validity of data-driven analytics in clinical applications. In an experiment based on bleeding detection, we demonstrate that our approach can help detect and analyze multiple-to-multiple correlations, which could be nonlinear or vary throughout the population, while retaining interpretability.

The main contributions of this paper are

1. A proposed correlation clustering method for multi-view data.
2. A proposed classification method based on the clustering method.
3. A demonstration of how to apply this method in a practical context of clinical importance.

The paper is organized as follows: Section 2 summarizes related work; Section 3 gives necessary background on CCA; Section 4 derives the CLS clustering and classification methods; Section 5 describes experimental results; Section 6 discusses medical uses and properties of the method; and Section 7 concludes the paper.

2 RELATED WORK

Cluster-wise linear regression Späth (1982) introduces a method for clustering the observations in a single-output regression dataset. Like k-means, this method is greedy and iterative and alternates between two steps. Given cluster labels, it fits a linear regression to each cluster. Given regression coefficients, it assigns each observation to the cluster whose regression residual is the smallest for that observation. It is simple to show that this method is a special case of CLS clustering in which the regression inputs are one set of variables and the regression output by itself is the other set—i.e., one view is univariate.

Dependency seeking clustering An interesting approach to correlation clustering is explored by Klami and Kaski (2008) and Rey and Roth (2012). Klami and Kaski (2008) establish a probabilistic generative modeling framework to allow Bayesian inference. They do so by proposing a model of probabilistic families for finding dependency and give a general clustering algorithm for this family. CCA is shown to be a special case. A key assumption is that a linearly transformed Gaussian latent variable produces the variation in the data. However, there may be severe model mismatch when this assumption was violated. To remedy this behavior, Rey and Roth (2012) deploy a copula mixture model to the framework, enabling them to model mixtures of CCA, similar to the clustering setup in this work. A Bayesian clustering algorithm is proposed and shown to perform well on synthetic and real datasets.

Multi-view clustering There has been substantial past work on multi-view clustering. However, many authors, such as Livescu and Stoehr (2009) and Bruno and Marchand-Maillet (2009), work with multi-modal data such as audio-visual data; clinical data is less common. Multi-view versions of k-means and Expectation Maximization were considered by Bickel and Scheffer (2004) and found to outperform the single-view counterparts. A method by Chaudhuri et al. (2009) uses CCA to find the subspace spanned by the means of mixture components. The data are projected down to this subspace and clustered. In Kumar et al. (2011), a multi-view spectral clustering framework is proposed. This framework employs co-regularization to enforce agreement between clusterings in different views. Another line of work by Nie et al. (2011) and Wang et al. (2013) approaches clustering as a regression-like problem of fitting the data to cluster membership probabilities. The work of Wang et al. (2013) applies structured sparsity to weight features in different views by their importance. In addition, a method was introduced by Liu et al. (2013) that uses nonnegative matrix factorization. This method searches for matrix factorizations that give compatible clusters across the views.

Single-view correlation clustering Zimek (2009) considers the problem of clustering data based on patterns of correlation when the variables are not partitioned into two groups. At a high level, the paper's definition of correlation clustering is similar to the definition in the CLS or CCA context: the task of separating observations into clusters that have distinct correlation structure. Unlike

CLS or CCA, however, this work assumes a single view; the correlation refers to correlation between all the variables, not just between two sets. The paper presents a diverse body of algorithms for this task. Related to CLS clustering is a class of methods based on on Principal Components Analysis (PCA) (Zimek, 2009). A key intuition is that the principal components corresponding to lower eigenvalues resemble clusters because they have less variance. They then partition observations into clusters whose lower principal components are close together.

3 CANONICAL CORRELATION ANALYSIS

In this section we summarize CCA, a useful starting point for understanding the proposed methods. CCA analyzes cross-covariance between two sets of variables that have aligned observations. Informally, it finds common signals between the sets. By performing CCA, one can understand how much variance in the sets can be explained by common factors. For example, if one set is genes and the other is diseases, then CCA might connect combinations of genes with certain diseases, potentially corresponding to physiological traits. Formally, it finds maximally correlated linear combinations of each set. Let $X \in \mathbb{R}^{d_X}$ and $Y \in \mathbb{R}^{d_Y}$ be random vectors. Without loss of generality, assume $\mathrm{E}[X] = \mathrm{E}[Y] = 0$. Then CCA solves the problem

$$\max_{u \in \mathbb{R}^{d_X}, v \in \mathbb{R}^{d_Y}} \mathrm{Corr}(X^\mathsf{T} u, Y^\mathsf{T} v). \tag{1}$$

Define $\Sigma_{XY} = \mathrm{Cov}(X, Y)$, $\Sigma_{XX} = \mathrm{Cov}(X)$, and $\Sigma_{YY} = \mathrm{Cov}(Y)$. The solution of (1) is well-understood (Hardoon et al., 2004): u is the leading eigenvector of

$$A = \Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1} \Sigma_{XY}^\mathsf{T}$$

and v the leading eigenvector of

$$B = \Sigma_{YY}^{-1} \Sigma_{XY}^\mathsf{T} \Sigma_{XX}^{-1} \Sigma_{XY}.$$

Subsequent linear combinations can be found under the constraint that the previous components $X^\mathsf{T} u$ and $Y^\mathsf{T} v$, known as canonical variables, are uncorrelated with the new canonical variables. Formally, let u_i, v_i, $i = 1, \ldots, m-1$, be the first $m-1$ solutions. Then u_m and v_m solve

$$\begin{aligned} \max_{u \in \mathbb{R}^{d_X}, v \in \mathbb{R}^{d_Y}} \quad & \mathrm{Corr}(X^\mathsf{T} u, Y^\mathsf{T} v) \\ \text{subject to} \quad & \mathrm{Cov}(Xu, Xu_i) = \mathrm{Cov}(Yv, Yv_i) = 0, \\ & i = 1, \ldots, m-1. \end{aligned} \tag{2}$$

For $m \leq \min(d_X, d_Y, \mathrm{rank}(\Sigma_{XY}))$, the solutions to (2) are known to be the m-th leading eigenvectors of A and B.

A standard reformulation (Hardoon et al., 2004) of (1) is

$$\begin{aligned} \max_{u \in \mathbb{R}^{d_X}, v \in \mathbb{R}^{d_Y}} \quad & u^\mathsf{T} \Sigma_{XY} v \\ \text{subject to} \quad & u^\mathsf{T} \Sigma_{XX} u = v^\mathsf{T} \Sigma_{YY} v = 1. \end{aligned} \tag{3}$$

The objective of (3) can also be expressed with the same constraints as

$$\min_{u \in \mathbb{R}^{d_X}, v \in \mathbb{R}^{d_Y}} \mathrm{E}\left[\|Xu - Yv\|_2^2\right] \tag{4}$$

which resembles a least-squares problem.

4 CANONICAL LEAST SQUARES

In this section we develop a method for correlation clustering called Canonical Least Squares (CLS) clustering. We then describe how the clustering method can serve as the basis of a supervised classification method. Note there already exists a clustering method directly based on CCA introduced by Fern and Friedl (2005). Although our proposal is similar, it has a few theoretical and practical benefits, which are further discussed in Sec. 6.

Now we introduce CLS, an analogue of CCA. Like CCA, CLS takes sets of variables X and Y and produces up to $m \leq \min(d_X, d_Y, \mathrm{rank}(X^\mathsf{T} Y))$ pairs of vectors (u, v) such that the components $X^\mathsf{T} u$ and $Y^\mathsf{T} v$ have some kind of relationship. Unlike CCA, this relationship is not of maximum correlation but of least squared error.

4.1 FIRST COMPONENTS

First consider only the top pair of components ($m = 1$). We redefine $X \in \mathbb{R}^{n \times d_X}$ and $Y \in \mathbb{R}^{n \times d_Y}$ as centered data matrices. Then (4) becomes

$$\begin{aligned} \min_{u \in \mathbb{R}^{d_X}, v \in \mathbb{R}^{d_Y}} \quad & \|Xu - Yv\|_2^2 \\ \text{subject to} \quad & u^\mathsf{T} X^\mathsf{T} X u = v^\mathsf{T} Y^\mathsf{T} Y v = 1. \end{aligned}$$

We propose the following modification, which has the same objective but different constraints:

$$\begin{aligned} \min_{u \in \mathbb{R}^{d_X}, v \in \mathbb{R}^{d_Y}} \quad & \|Xu - Yv\|_2^2 \\ \text{subject to} \quad & v^\mathsf{T} v = 1. \end{aligned} \tag{5}$$

We denote (5) CLS (for the first component). One major difference from CCA is the lack of X or Y in the constraints. This difference enables CLS to form the building block of a clustering method with a well-defined optimization procedure, as will soon be explained. The other difference is the lack of u in the constraints. When only v is constrained, the problem generalizes ordinary least

squares, which does not constrain the coefficients of the independent variables, to multiple outputs.

Next we present the solution to (5). First let v be fixed. The problem becomes ordinary least squares in u, yielding

$$u = (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y v.$$

Let $H = I - X(X^\mathsf{T} X)^{-1} X^\mathsf{T}$, a symmetric idempotent matrix. After substituting for u, the problem in v is given by

$$\min_{v \in \mathbb{R}^{d_Y}} \|HYv\|_2^2 \qquad \text{subject to} \quad v^\mathsf{T} v = 1.$$

This problem resembles PCA except with a minimum instead of maximum. The solution v is the eigenvector with the lowest eigenvalue of $Y^\mathsf{T} H^\mathsf{T} HY = Y^\mathsf{T} HY$.

4.2 MULTIPLE COMPONENTS

In CCA, subsequent canonical variables are uncorrelated with each other. After changing these constraints to be independent of the data, we are left with simple orthogonality constraints between vectors of coefficients. The generalization of (5) to m components is then

$$\begin{aligned} \min_{\substack{U \in \mathbb{R}^{d_X \times m} \\ V \in \mathbb{R}^{d_Y \times m}}} \quad & \|XU - YV\|_\mathcal{F}^2 \\ \text{subject to} \quad & V^\mathsf{T} V = I. \end{aligned} \tag{6}$$

This problem is non-convex in the constraints. It is difficult to solve because all components must be found simultaneously. We instead choose an easier suboptimal solution: let V be the eigenvectors corresponding to the m lowest eigenvalues from the solution to (5), and compute U accordingly. This solution corresponds to greedily solving for each component sequentially under orthogonality. It is an interesting tangent to juxtapose this procedure with Principal Components Analysis (PCA), which solves a similar problem

$$\max_{W \in \mathbb{R}^{d \times d}} \|ZW\|_\mathcal{F}^2 \quad \text{subject to} \quad W^\mathsf{T} W = I$$

where $Z \in \mathbb{R}^{n \times d}$ is a centered data matrix. In PCA, the greedy eigenvector solution is optimal because of the orthogonality constraints between full vectors of coefficients. In CLS, however, only the vectors v_i must be orthogonal, rendering the greedy solution suboptimal.

Separately, in the special case that $m = \min\{d_X, d_Y\}$, then U or V is an orthogonal matrix, so CLS reduces to ordinary least squares on the columns of X or Y respectively.

4.3 CLUSTERING

The CLS clustering algorithm takes matrices X and Y, a number k of clusters, and a number m of components. Let $X^{(i)}$ and $Y^{(i)}$ denote X and Y with rows subsampled to those in cluster i. Let the coefficients corresponding to that cluster be $U^{(i)}$ and $V^{(i)}$. To find cluster labels for each data point, we iterate the following steps until convergence:

- **CLS step** Given cluster labels, for each cluster $i = 1, \ldots, k$: run CLS (6) on $X^{(i)}$ and $Y^{(i)}$ to find $U^{(i)}$ and $V^{(i)}$.
- **Labeling step** Given CLS coefficients $U^{(i)}$ and $V^{i)}$, for each observation (x_ℓ, y_ℓ), $\ell = 1, \ldots, n$: assign it to

$$\text{argmin}_i \, \|y_\ell^\mathsf{T} V^{(i)} - x_\ell^\mathsf{T} U^{(i)}\|_2^2.$$

This procedure is a block coordinate-wise iterative approach, resembling Expectation-Maximization (Dempster et al., 1977), to solving the overall optimization problem

$$\begin{aligned} \sum_i \min_{\substack{U^{(i)} \in \mathbb{R}^{d_1 \times m} \\ V^{(i)} \in \mathbb{R}^{d_2 \times m}}} & \|R^{(i)}(XU^{(i)} - YV^{(i)})\|_\mathcal{F}^2 \\ \text{subject to} \quad & V^{(i)\mathsf{T}} V^{(i)} = I, \; i = 1, \ldots, k. \end{aligned} \tag{7}$$

There is a convergence guarantee when $m = 1$, i.e., when only the first pair of components is used. The CLS step optimizes over the u_i's and v_i's, while the labeling step optimizes over the $R^{(i)}$'s. Thus the objective is non-increasing at every step, so convergence is guaranteed. It is not unreasonable to use $m = 1$ because the first components are often the most meaningful. If $m > 1$, since a greedy approximation of the CLS solution is used, the objective is not guaranteed to be monotonic. Nevertheless in practice it usually is, which aids convergence.

Furthermore, in some applications it is helpful to add must-link constraints that designate sets of points that must appear in the same cluster. These constraints can be encoded by assigning each set of points to the cluster that minimizes the sum of squared errors over the points.

4.4 CLASSIFICATION

It is straightforward to build a supervised classification method on top of CLS clusters. First CLS clusters are learned independently on each class. Then new points are scored for each class according to the best fitting (lowest scoring) cluster in that class's fitted model. More formally, the score, or loss, of point (x, y) in cluster i is given by

$$\frac{1}{2} \|x^\top U^{(i)} - y^\top V^{(i)}\|_\mathcal{F}^2.$$

The minimum of these scores is taken over the clusters in a given class to produce the score for that class. The final classification is the class that has the best fitting cluster overall. In addition, the procedure can be run many times with different random initializations and the scores averaged, which would make this classifier an ensemble method.

In many applications, it is interesting to examine only two of the learned clusters and ask how to decide which of them a new observation should belong to. It is possible to derive a locally linear model of the relevant factors, which should be readily interpretable. Of course, we can only interpret a single weak learner from the ensemble, not the entire ensemble at once, but this difficulty is shared by all ensemble methods. Now, let the observation be given by (x, y) and let z be the vertical concatenation of x and y. Let the two clusters of interest have coefficients $U^{(i)}$ and $V^{(i)}$, where $i \in \{0, 1\}$, and let $W^{(i)}$ be the vertical concatenation of $U^{(i)}$ and $-V^{(i)}$. The loss in cluster i is then $z^\mathsf{T} W^{(i)} W^{(i)\mathsf{T}} z/2$. The classification score between the two clusters is

$$\frac{1}{2} z^\mathsf{T} (W^{(0)} W^{(0)^\mathsf{T}} - W^{(1)} W^{(1)^\mathsf{T}}) z$$

where a higher score indicates membership in cluster 1. We determine the effect of a small change in any individual feature by computing the gradient,

$$(W^{(0)} W^{(0)^\mathsf{T}} - W^{(1)} W^{(1)^\mathsf{T}}) z.$$

Thus we can hypothesize how a pair of waveforms might need to change to alter its classification. Since the effects are of second order, however, and depend on the observation z, a practitioner must be careful not to assume they generalize over examples. This difficulty is similar to the challenge of interpreting the coefficients of a logistic regression, where a change in one feature could have a different magnitude of effect depending on the observation.

4.5 PRACTICALITIES

Intercept An intercept should be incorporated in CLS clustering by augmenting X with a column of 1's.

Data scale CCA is affine-invariant with respect to X and Y. However, CLS is sensitive to scaling because it uses Euclidean distance, similar to k-means. Therefore, we recommend normalizing the column variance in preprocessing.

Initialization Like in all greedy iterative algorithms similar to k-means, random initialization over many runs improves the chance of CLS clustering to achieve a robust solution.

5 RESULTS

5.1 SYNTHETIC DATASET

The method was briefly tested on synthetic data to explore its performance under simple conditions.

5.1.1 Description

We deployed CLS clustering on a medium-sized synthetic dataset produced by a generative model. The dataset consisted of $2k$ equally sized clusters, where $k = 40$, of $n = 25,000$ points each. The clusters heavily overlapped. Each point was in $\mathbb{R}^{2d}$ and was formed by concatenating two points $x, y \in \mathbb{R}^d$, where $d = 20$. We sampled x from a multivariate normal distribution centered at the origin with covariance from a Wishart distribution. We projected x into a noisy m-dimensional subspace, where $m = 5$, and then projected it back to make y. The projections were sampled from spherical Gaussians. The data distribution was held fixed for a given cluster, but its parameters were resampled in each new cluster. Furthermore, the first k clusters were assigned to the "0" class and the second k to the "1" class. After this dataset was created, a second with the exact same composition and distribution was created as test data.

5.1.2 Results

Table 1: Convergence of Loss Function

Iteration	4	8	12	16	20
Loss	.4540	.0212	.0199	.0196	.0196

CLS clusters were found on each class separately with correct choices of k and m. On the test data, the classification accuracy was 99.92%, which is nearly flawless. Further, Tbl. 1 illustrates the smooth, rapid convergence of the loss function on this dataset.

5.2 MEDICAL DATASET

5.2.1 Background

An open question in medicine is whether the presence of bleeding and other conditions can be captured by monitoring central venous pressure (CVP), the blood pressure in a region near the right atrium of the heart. The debate over CVP has produced a large body of literature on the issue of its clinical utility. Most studies appear to post negative results, such as Michard and Teboul (2000), Pinsky and Payen (2005), and Kumar et al. (2004), even though CVP

is often used in practice (Marik and Cavallazzi, 2013); some studies claim positive results, such as Damman et al. (2009) and Boyd et al. (2011). However, little work has been published on investigating CVP in a controlled, low noise setting.

Here we investigate potential utility of CVP signal within a controlled setting by attempting to classify CVP waveforms as indicative of an active bleeding episode vs. periods of no-bleeding. The two views in the data correspond to waveforms during inspiration and expiration phases of the respiratory cycle. We show how CLS can make predictions as well as automate the discovery of insights of potential medical interest. CLS clusters can be interpreted as clinical phenotypes characterizing patients' pre-bleeding or post-bleeding responses. Also, the relationship of bleeding with inspiration and expiration can be interpreted in terms of the original CVP waveforms.

5.2.2 Description

The data were collected from an experiment in which healthy anesthetized pigs were subjected to controlled bleeding. The experimental procedure was similar to that in Pinsky (1984). Thirty-eight Yorkshire pigs were sedated and bled at a constant rate of 20 mL/min. Their central venous pressure (CVP), the blood pressure in a major vein to the heart, was monitored for 20 minutes before bleeding and 30 minutes after its onset. Two CVP waveforms (Fig. 1) were extracted from each respiration cycle, one from the inspiration phase of breathing and the other from the expiration phase. Inspiration and expiration represent the two views of the data in our experiment. The respiration cycles lasted 5.2 seconds each on average, resulting in an average of 556 cycles per pig over the 50 minutes of observation. Thirteen features were extracted from each waveform as averages and ratios between different characteristic points of the CVP waveform. These features included differences between peaks and troughs such as the height SA between points S and A as well as ratios such as VR over CQ (Fig. 1).

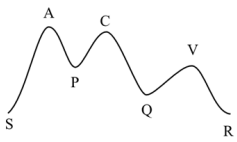

Figure 1: A Typical Central Venous Pressure Waveform

5.2.3 Procedure

The classification task was to decide whether a pair of waveforms from a given time step came from before or after the onset of bleeding. The pigs were partitioned into training and test sets of 25 and 13 pigs respectively. Leave-one-subject-out cross-validation was employed to select the hyperparameters k and m for the two classes. Following the classification algorithm from Sec. 4.4, CLS clusters were learned separately on non-bleeding and bleeding data. Observations from the same pig were constrained to belong to the same cluster. The classification scores on a left-out pig were used to compute the area under the receiver operating characteristic curve (AUC), true positive rate (TPR) at a false positive rate (FPR) of 10% and 1%, and FPR at a TPR of 50%. Hyperparameters were selected by optimizing the AUC. We chose 3 clusters with 6 components for pre-bleeding and 5 clusters with 7 components for post-bleeding.

5.2.4 Results

Table 2 shows performance metrics of the final model on test data. It also gives the performance of a model that learns only one cluster on each class. The sizeable gap in performance demonstrates the benefit of searching for correlations that exist in subsets of the data, as opposite to a global correlation model identifiable in the whole set. The table also compares the performance to a random forest classifier (Breiman, 2001) with 100 trees trained on the combined views. The random forest performs best in most metrics, but its advantage vs. CLS is not statistically significant. This result is acceptable since CLS enables detailed yet interpretable view of discovered structures in data while its performance metrics remain within the confidence interval of random forest. The interpretability of CLS results will be discussed later in this section and in Sec. 6.

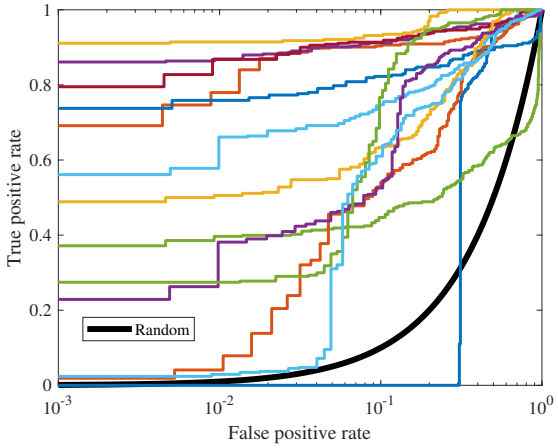

Figure 2: ROC of Individual Pigs in Test Set

Figure 2 illustrates the performance of the classifier on individual pigs. Although there is high variance between the subjects, all but one has significantly better performance

Table 2: Bleeding Classification Performance

	Single cluster CLS	Final CLS	Random forest
AUC	.701 ± .128	.862 ± .064	.891 ± .075
TPR @ .10 FPR	.468 ± .185	.674 ± .145	.762 ± .167
TPR @ .01 FPR	.222 ± .134	.501 ± .185	.610 ± .210
FPR @ .50 TPR	.239 ± .152	.064 ± .055	.073 ± .075

than random guessing.

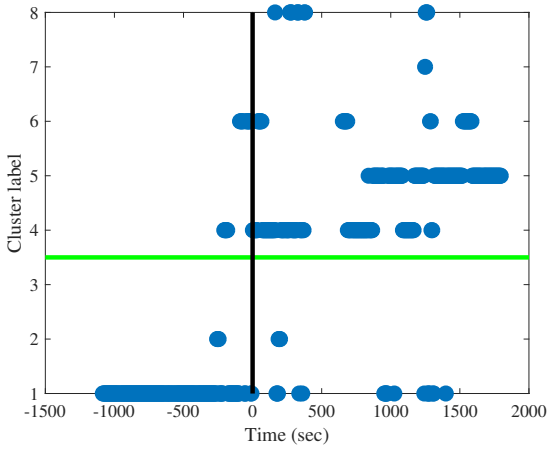

Figure 3: Cluster assignments for one pig. Clusters below the green line correspond to non-bleeding and the rest to bleeding.

Figure 3 shows the cluster assignments for one pig. The assignments appear to be noisy, but there are three dominant clusters: cluster 1 for before bleeding, cluster 4 for the early phase of bleeding after its onset, and cluster 5 for further into bleeding. Although the shown result is only from one weak learner, the structure is shared by much of the ensemble.

We examined latent variables that determine classification in Fig. 4. The latent variables, constructed as linear combinations of each view, are expected to align and have zero residual. This pattern roughly held before bleeding but was violated after its onset, which indicates that this cluster fit only before bleeding, as expected. The heightened residuals are highlighted in red. The timeline of the highlighted residuals is given in Fig. 5.

To interpret the model's decisions, we used the method involving the gradient derived in Sec. 4.4 on weak learners from the pre- and post-bleeding ensembles. We checked the score that determined whether the pig from Fig. 3 belonged to cluster 1 or 4, where cluster 1 was pre-bleeding and cluster 4 was post-bleeding. We computed the gradient of the score on a pre-bleeding observation. The results are displayed in Fig. 6. The original waveform of the observation is plotted on the left. According to the gradient, the most major changes that would make the observation closer to a bleeding waveform were shortening the lengths SA and AP during expiration. Correspondingly, the figure shows on the right an expiration waveform from soon after the onset of bleeding. The two characteristic waveform parameters have shrunk dramatically, and SA has even completely disappeared.

6 DISCUSSION

Bleeding clusters Figure 3 highlights an interesting pattern. For many pigs, there is a dominant cluster before bleeding, but when bleeding starts, a different cluster takes over. This new cluster typically only holds observations from the first ten or fewer minutes after bleeding. Afterward, other clusters become dominant. One interpretation is that the physiological response to bleeding changes as the induced stress escalates. There may be an initial compensation jolt, followed by a more systemic response which can also change in terms of its modality as a function of escalating stress. This hypothesis may be supported by Fig. 4, which shows that the immediate onset of bleeding corresponds to an unusual spike in residuals in the latent variable spaces. This pattern is an example of how the interpretable structure of CLS clustering can lend itself to finding practical insights.

Utility of CVP Although this study showed a link between CVP and bleeding, there still remains considerable work to be done before it can be utilized in practice. In this study, we examined the utility of CVP waveform signal when it was relatively noiseless. However, CVP is known to be highly sensitive to shifts in body position, rendering it quite noisy outside of laboratory conditions and cases when a patient is securely immobilized. It follows that one line of future work would be to identify and filter out this noise in order to leverage the predictive power displayed here.

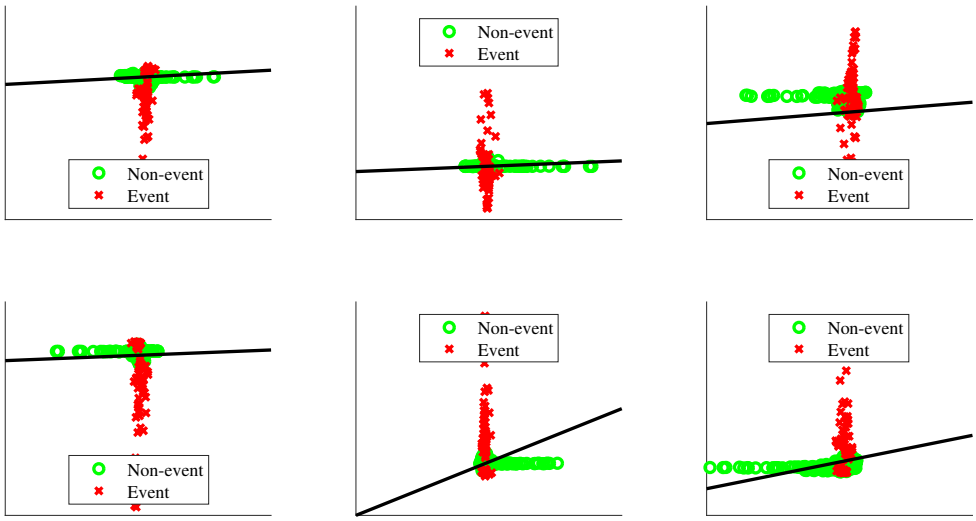

Figure 4: Latent variable residuals in a non-bleeding cluster from the entire timeline for one pig. Residuals diverge from 0 when bleeding starts, indicated by red.

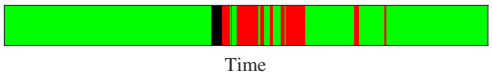

Figure 5: Timeline of highlighted events in Fig. 4. Black is the onset of bleeding.

Comparison to random forest In the above experiments we compared the classification performance of the proposed method to a random forest (Breiman, 2001). Our goal was to illustrate that the proposed method was reasonably close to state-of-the-art. The advantage of the proposed method is not that its classification performance is top-notch; instead, it has more interpretable structure. Although both methods employ ensembles and the individual decision trees in a random forest are somewhat interpretable, a major difference are the types of problems for which they are suited. The proposed method is more suited for multi-view data that are hypothesized to have interesting correlations between the views, especially when those correlations differ between subsets of observations, as random forests do not incorporate any clustering mechanism.

Comparison to CCA clustering The similarity of CLS to CCA was noted in Sec. 4. Indeed, an analogous clustering algorithm was proposed by Fern and Friedl (2005) called CCA clustering. While CCA maximizes correlation between variables in the latent space, CLS minimizes the squared error. These objectives are similar, but CLS can find components with weaker correlation and smaller residuals, which is not necessarily an advantage or disadvantage. However, CLS clustering does have some concrete advantages: its objective function is well-defined, and it is guaranteed to converge when $m = 1$. Recall that the CCA optimization had constraints dependent on data,

$$u^\top X^\top X u = v^\top Y^\top Y v = 1.$$

As a result, when cluster assignments change, the constraints for each cluster's CCA problem change as well. To be consistent, the search space for cluster assignments would also have to satisfy those constraints, but this requirement is infeasible. By removing the dependence on data in constraints, CLS clustering avoids this problem and therefore permits a well-defined optimization.

Link to reduced-rank regression The CLS optimization problem has deep similarities to reduced-rank regression (RRR) (Izenman, 1975). Both consider multivariate latent variable models of the form $x \to z \to y$, where x is an input, y an output, and z a latent variable. While RRR minimizes error in the output space, CLS minimizes it in the latent space. It can be shown that under certain assumptions RRR reduces to CLS. A promising area for future work is to insert CLS into the flexible component analysis framework proposed by De La Torre (2012),

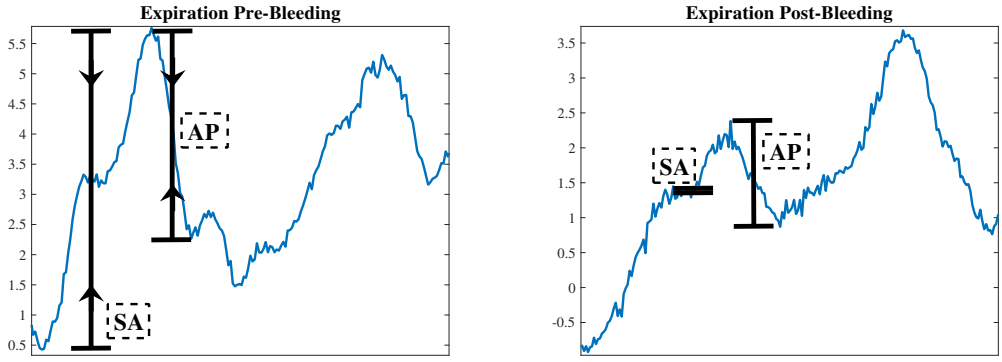

Figure 6: A pair of CVP waveforms from expiration before and after bleeding. The impact of certain features has been labeled on the pre-bleeding side. Arrows indicate lengths that must decrease to appear more like a waveform from after bleeding.

which uses a special type of RRR to unify PCA, CCA, LDA, and others.

Soft clustering We developed a soft clustering extension of this method based on ideas in Hathaway and Bezdek (1993). One way to view this extension is that cluster probabilities of an observation are regularized toward a uniform distribution over clusters. In the soft version, the optimization is much smoother, resulting in more consistent solutions over different runs. In this application, however, it was outperformed by the hard version, even though the assignment step in the hard version is highly non-smooth. A potential avenue for future work would be to analyze this and other theoretical optimization properties of the method.

Spectral interpretation Recall that the solution v for the first component of CLS was given by the last eigenvector of $Z \equiv Y^\mathsf{T}(I - X(X^\mathsf{T}X)^{-1}X^\mathsf{T})Y$. Let $\Sigma_{xx} = X^\mathsf{T}X$, $\Sigma xy = X^\mathsf{T}Y$, and $\Sigma_{yy} = Y^\mathsf{T}Y$. Assuming the data are centered, these variables are covariance and cross-covariance matrices of X and Y. Then $Z = \Sigma_{yy} - \Sigma_{xy}^\mathsf{T}\Sigma_{xx}^{-1}\Sigma_{xy}$ is the Schur complement of the covariance matrix of the joint distribution of X and Y. If this joint distribution is multivariate normal, then Z is the conditional covariance of Y given X. Hence CLS can be interpreted as finding the direction of minimum variance in Y given X. When Y has less variance after controlling for its relationship with X, it is easier to find a better linear fit with X. CLS clustering is similar in this regard to correlation clustering methods by Zimek (2009), which also leverage eigenvectors of lower variance.

7 CONCLUSION

This work considered the problem of discovering interpretable structures in complex datasets. In particular, we proposed a method to learn correlation clusters for multi-view data, where important relationships between the views are discovered. We also proposed a routine to perform supervised classification using the discovered correlation clusters as a basis. The method was tested on a dataset of induced bleeding and was demonstrated to perform well. Furthermore, the experiment showcased the method's capability to produce discover interesting patterns and produce explanations of its predictions.

References

Bach, F. R. and Jordan, M. I. (2006). A probabilistic interpretation of canonical correlation analysis. *Dept Statist Univ California Berkeley CA Tech Rep*, 688:1–11.

Bickel, S. and Scheffer, T. (2004). Multi-view clustering. In *IEEE International Conference on Data Mining*, number December 2004, pages 19–26.

Boyd, J. H., Forbes, J., Nakada, T.-a., Walley, K. R., and Russell, J. A. (2011). Fluid resuscitation in septic shock: a positive fluid balance and elevated central venous pressure are associated with increased mortality. *Critical care medicine*, 39(2):259–265.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Bruno, E. and Marchand-Maillet, S. (2009). Multiview clustering: A late fusion approach using latent models. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, number January, page 736.

Chaudhuri, K., Kakade, S., Livescu, K., and Sridharan, K. (2009). Multi-view clustering via canonical correlation analysis. In *International Conference on Machine Learning*, pages 1–8.

Choi, E., Bahadori, M. T., Kulas, J. A., Schuetz, A., Stewart, W. F., and Sun, J. (2016). RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Neural Information Processing Systems*.

Damman, K., van Deursen, V. M., Navis, G., Voors, A. A., van Veldhuisen, D. J., and Hillege, H. L. (2009). Increased central venous pressure is associated with impaired renal function and mortality in a broad spectrum of patients with cardiovascular disease. *Journal of the American College of Cardiology*, 53(7):582–588.

De La Torre, F. (2012). A least-squares framework for component analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1041–1055.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 39(1):1–38.

Fern, X. and Friedl, M. (2005). Correlation clustering for learning mixtures of canonical correlation models. In *SIAM International Conference on Data Mining*, pages 439–446.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Hathaway, R. J. and Bezdek, J. C. (1993). Switching regression models and fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, 1(3):195–204.

Holzinger, A. and Jurisica, I. (2014). Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, (March):1–18.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4):321–377.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264.

Klami, A. and Kaski, S. (2008). Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72(1):39–46.

Krumholz, H. M. (2014). Big data and new knowledge in medicine: The thinking, training, and tools needed for a learning health system. *Health Affairs*, 33(7):1163–1170.

Kumar, A., Anel, R., Bunnell, E., Habet, K., Zanotti, S., Marshall, S., Neumann, A., Ali, A., Cheang, M., Kavinsky, C., et al. (2004). Pulmonary artery occlusion pressure and central venous pressure fail to predict ventricular filling volume, cardiac performance, or the response to volume infusion in normal subjects. *Critical care medicine*, 32(3):691–699.

Kumar, A., Rai, P., and Daumé, H. (2011). Co-regularized multi-view spectral clustering. *Neural Information Processing Systems*, pages 1413–1421.

Liu, J., Wang, C., Gao, J., and Han, J. (2013). Multi-view clustering via joint nonnegative matrix factorization. In *SIAM International Conference on Data Mining*, pages 252–260.

Livescu, K. and Stoehr, M. (2009). Multi-view learning of acoustic features for speaker recognition. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 82–86.

Marik, P. E. and Cavallazzi, R. (2013). Does the central venous pressure predict fluid responsiveness? an updated meta-analysis and a plea for some common sense. *Critical care medicine*, 41(7):1774–1781.

Michard, F. and Teboul, J.-L. (2000). Using heart-lung interactions to assess fluid responsiveness during mechanical ventilation. *Critical Care*, 4(5):282.

Murdoch, T. and Detsky, A. (2013). The inevitable application of big data to health care. *JAMA*, 309(13):1351–1352.

Nie, F., Zeng, Z., Tsang, I., Xu, D., and Zhang, C. (2011). Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering. *IEEE Transactions on Neural Networks*, 22(11):1796–1808.

Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the future big data, machine learning, and clinical medicine. *N Engl J Med*, 375(13):1216–1219.

Pinsky, M. R. (1984). Instantaneous venous return curves in an intact canine preparation. *Journal of Applied Physiology*, 56(3):765–771.

Pinsky, M. R. and Payen, D. (2005). Functional hemodynamic monitoring. *Critical Care*, 9(6):566.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., and Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In *ACM International Conference on Multimedia*, pages 251–260. ACM.

Rey, M. and Roth, V. (2012). Copula mixture model for dependency-seeking clustering. *International Conference on Machine Learning*.

Seoane, J. A., Campbell, C., Day, I. N., Casas, J. P., and Gaunt, T. R. (2014). Canonical correlation analysis for gene-based pleiotropy discovery. *PLoS Compututational Biology*, 10(10):e1003876.

Sherry, A. and Henson, R. K. (2005). Conducting and interpreting canonical correlation analysis in personality research: A user-friendly primer. *Journal of Personality Assessment*, 84(1):37–48.

Späth, H. (1982). A fast algorithm for clusterwise linear regression. *Computing*, 29(2):175–181.

Wang, H., Nie, F., and Huang, H. (2013). Multi-view clustering and feature learning via structured sparsity. In *International Conference on Machine Learning*, volume 28, pages 352–360.

Witten, D. and Tibshirani, R. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–27.

Zimek, A. (2009). Correlation clustering. *ACM SIGKDD Explorations*, 11(1):53–54.