A Hitchhiker's Guide to PCA and CCA

Karl Stratos

1 Notation

Vectors and matrices are denoted by boldface letters. The transpose operator is denoted by a superscript $\top$. Vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)} \in \mathbb{R}^n$ are organized into a matrix in two ways:

Matrix $(\mathbf{x}_1 \ldots \mathbf{x}_N) \in \mathbb{R}^{N \times n}$ has $\mathbf{x}_1^\top \ldots \mathbf{x}_N^\top$ as rows.

Matrix $[\mathbf{x}_1 \ldots \mathbf{x}_N] \in \mathbb{R}^{n \times N}$ has $\mathbf{x}_1 \ldots \mathbf{x}_N$ as columns.

2 Principal Component Analysis (PCA)

Consider a random variable $X \in \mathbb{R}^n$ which is an n-dimensional vector. We want to derive a lower dimensional variable $\underline{X} \in \mathbb{R}^m$ $(m \leq n)$ to represent X. Intuitively, $\underline{X}$ should capture as much information about X as possible in these fewer dimensions. PCA finds $\underline{X} = (\underline{X}_1 \ldots \underline{X}_m)$ such that each component $\underline{X}_i \in \mathbb{R}$ is a one-dimensional projection of X with maximum variance that is uncorrelated with the previous components. Specifically, for $i = 1 \ldots m$, it finds

$$\underline{X}_i = \underset{\Psi \in \mathbb{R}}{\arg\max}\, \mathrm{Var}(\Psi)$$

under the constraint that $\Psi = \mathbf{a}^\top X$ for some vector $\mathbf{a} \in \mathbb{R}^n$ with $||\mathbf{a}||^2 = 1$ and that

$$\mathrm{Cor}(\Psi, \underline{X}_j) = 0$$

for $j = 1 \ldots i - 1$. Note that we need to constrain the length of $\mathbf{a}$; otherwise, $\mathrm{Var}(\mathbf{a}^\top X)$ can be arbitrarily large. This new variable $\underline{X} \in \mathbb{R}^m$ found by PCA can be viewed as an optimal m-dimensional representation of $X \in \mathbb{R}^n$.

2.1 A Derivation of the Algorithm

Let $\mathbf{a}_1 \ldots \mathbf{a}_m$ be the projection vectors used for deriving $\underline{X}_1 \ldots \underline{X}_m$. Finding $\mathbf{a}_1$ can be framed as the following optimization problem.

$$\mathbf{a}_1 = \underset{\mathbf{a} \in \mathbb{R}^n:\ ||\mathbf{a}||^2 = 1}{\arg\max}\ \mathrm{Var}(\mathbf{a}^\top X) \tag{1}$$

Without loss of generality, we can assume that each dimension is centered so that $\mathbf{E}[X_1] = \cdots = \mathbf{E}[X_n] = 0$ since $\mathrm{Var}(X - \mathbf{E}[X]) = \mathrm{Var}(X)$ for any random variable X. Using this assumption, we manipulate the expression as follows.

$$\begin{aligned}\mathrm{Var}(\mathbf{a}^\top X) &= \mathbf{E}[(\mathbf{a}^\top X)^2]\\ &= \mathbf{E}[\mathbf{a}^\top XX^\top \mathbf{a}]\\ &= \mathbf{a}^\top \mathbf{E}[XX^\top]\mathbf{a}\\ &= \mathbf{a}^\top \mathbf{C}_{XX}\mathbf{a}\end{aligned}$$

where $\mathbf{C}_{XX} \in \mathbb{R}^{n\times n}$ is the covariance matrix of X with value $[\mathbf{C}_{XX}]_{i,j} = \mathrm{Cov}(X_i, X_j) = \mathbf{E}[X_iX_j]$. Then the optimization problem in Eq. (1) can be reframed as

$$\mathbf{a}_1 = \underset{\mathbf{a}\in\mathbb{R}^n:\ ||\mathbf{a}||^2=1}{\arg\max}\ \ \mathbf{a}^\top \mathbf{C}_{XX}\mathbf{a} \tag{2}$$

Since $\mathbf{C}_{XX}$ is positive semi-definite, it has non-negative eigenvalues and orthonormal eigenvectors. We will now use an eigenvalue decomposition on $\mathbf{C}_{XX}$ to solve Eq. (2) (see section 5.1).

Proposition 1. $\mathbf{a}_1$ is the unit eigenvector of $\mathbf{C}_{XX}$ with the greatest eigenvalue.

Proof. We use the Lagrangian relaxation to maximize the quantity $\mathbf{a}^\top\mathbf{C}_{XX}\mathbf{a}$ under the normalization constraint $||\mathbf{a}||^2 = 1$:

$$L = \mathbf{a}^\top \mathbf{C}_{XX}\mathbf{a} + \frac{1}{2}\lambda(1 - \mathbf{a}^\top\mathbf{a})$$

When we differentiate with respect to $\mathbf{a}$ and set to zero, we arrive in the equation

$$\mathbf{C}_{XX}\mathbf{a} = \lambda\mathbf{a}$$

which tells us that $\mathbf{a}$ is an eigenvector of $\mathbf{C}_{XX}$ and the Lagrangian multiplier λ is its eigenvalue. Since our objective is to maximize $\mathbf{a}^\top\mathbf{C}_{XX}\mathbf{a} = \lambda$, the claim follows. □

More generally, $\mathbf{a}_1 \ldots \mathbf{a}_m$ are the unit eigenvectors of $\mathbf{C}_{XX}$ that correspond to the top m largest eigenvalues.

Theorem 1. $\mathbf{a}_i$ is the unit eigenvector of $\mathbf{C}_{XX}$ with the i^{th} greatest eigenvalue.

Proof. The base case is given by proposition 1. Let $\mathbf{a}_i$ be the i^{th} unit eigenvector of $\mathbf{C}_{XX}$. Then $\mathbf{a}_i$ maximizes over all $\mathbf{a} \in \mathbb{R}^n$

$$\mathbf{a}^\top\mathbf{C}_{XX}\mathbf{a} = \mathrm{Var}(\mathbf{a}^\top X)$$

while satisfying $\mathbf{a}_i^\top\mathbf{a}_j = 0$ for $j \in \{1 \ldots i-1\}$. This means

$$\mathrm{Cor}(\mathbf{a}_iX, \mathbf{a}_jX) \propto \mathbf{a}_i^\top\mathbf{C}_{XX}\mathbf{a}_j = \lambda_j\mathbf{a}_i^\top\mathbf{a}_j = 0$$

where we used $\mathbf{C}_{XX}\mathbf{a}_j = \lambda_j\mathbf{a}_j$. Hence

$$\mathbf{a}_i = \underset{\substack{\mathbf{a}\in\mathbb{R}^n:\ ||\mathbf{a}||^2=1\\ \mathrm{Cor}(\mathbf{a}^\top X, \mathbf{a}_j^\top X)=0\ \ \forall j\in\{1\ldots i-1\}}}{\arg\max}\ \ \mathrm{Var}(\mathbf{a}^\top X)$$

as desired. □

Theorem 1 states that the eigenvalue λ_i of $\mathbf{C}_{XX}$ is the variance of $\underline{X}_i$:

$$\lambda_i = \mathrm{Var}(\underline{X}_i)$$

Thus we can decide dimension m by inspecting the eigenvalue spectrum of $\mathbf{C}_{XX}$.

We have shown that we can obtain vectors $\mathbf{a}_1 \dots \mathbf{a}_m$ in a single shot via an eigenvalue decomposition on $\mathbf{C}_{XX}$. Once we have these vectors, we can form matrix $\mathbf{A}_m = [\mathbf{a}_1 \dots \mathbf{a}_m] \in \mathbb{R}^{n \times m}$ to project X down to $\underline{X} = \mathbf{A}_m^\top X$.

$$\underline{X} = \mathbf{A}_m^\top X = (\mathbf{a}_1^\top X \dots \mathbf{a}_m^\top X) = (\underline{X}_1 \dots \underline{X}_m)$$

The algorithm to compute this projection is given below.

PCA-PROJECTION
Input: covariance matrix $\mathbf{C}_{XX} \in \mathbb{R}^{n \times n}$ for $X \in \mathbb{R}^n$ where $\mathbf{E}[X_i] = 0$, $m \leq d$
Output: PCA projection $\mathbf{A}_m \in \mathbb{R}^{n \times m}$

1. Compute an eigenvalue decomposition

$$\mathbf{C}_{XX} = [\mathbf{a}_1 \dots \mathbf{a}_n] \times \Lambda \times [\mathbf{a}_1 \dots \mathbf{a}_n]^\top$$

where Λ is a diagonal matrix of n eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n \geq 0$ and $\mathbf{a}_i$ is the corresponding (normalized) eigenvector.

2. Return $\mathbf{A}_m = [\mathbf{a}_1 \dots \mathbf{a}_m]$.

Suppose now we wish to recover X from $\underline{X}$. Since $\mathbf{A}_m \in \mathbb{R}^{n \times m}$ is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$, a natural solution is to let the Moore-Penrose pseudoinverse $\mathbf{A}_m^+ \in \mathbb{R}^{m \times n}$ be a reverse mapping from $\mathbb{R}^m$ to $\mathbb{R}^d$. Then we recover

$$\begin{aligned} \tilde{X} &= (\mathbf{A}_m^+)^\top \underline{X} \\ &= (\mathbf{A}_m \mathbf{A}_m^+)^\top X \end{aligned}$$

The linear operator $\mathbf{A}_m \mathbf{A}_m^+ \in \mathbb{R}^{n \times n}$ is an orthogonal projection onto the subspace spanned by $\mathbf{a}_1 \dots \mathbf{a}_m$. This implies that if we use $m = \mathrm{rank}(\mathbf{C}_{XX})$, we will have $\tilde{X} = X$ exactly. If $m < \mathrm{rank}(\mathbf{C}_{XX})$, $\tilde{X} \approx X$ is not exact; however, the impressive quality of the best-fit subspace will be demonstrated in the experiment section.

2.2 Sample-Based PCA

In practice, we have samples $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)} \in \mathbb{R}^n$ of the random variable $X \in \mathbb{R}^n$ and compute an empirical estimate $\hat{\mathbf{C}}_{XX}$ of the covariance matrix $\mathbf{C}_{XX}$. Since we must have $\mathbf{E}[X_i] = 0$, we center the samples to obtain $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(N)} \in \mathbb{R}^n$ where each dimension has zero mean: for all $k = 1 \dots N$,

$$\mathbf{z}_i^{(k)} = \mathbf{x}_i^{(k)} - \frac{1}{N} \sum_{k=1}^{N} \mathbf{x}_i^{(k)} \quad \forall i \in \{1 \dots n\}$$

If we define $\mathbf{Z} = [\mathbf{z}^{(1)} \dots \mathbf{z}^{(N)}] \in \mathbb{R}^{n \times N}$, then the covariance matrix is estimated as $\hat{\mathbf{C}}_{XX} = \frac{1}{N}\mathbf{Z}\mathbf{Z}^\top \in \mathbb{R}^{n \times n}$ where

$$\begin{aligned}[\hat{\mathbf{C}}_{XX}]_{i,j} &= \frac{1}{N}\sum_{k=1}^{N} \mathbf{z}_i^{(k)}\mathbf{z}_j^{(k)} \\ &\approx \mathbf{E}[(X_i - \mathbf{E}[X_i])(X_j - \mathbf{E}[X_j])] = \mathrm{Cov}(X_i, X_j)\end{aligned}$$

The algorithm for deriving the lower dimensional samples is given below.

```
**PCA-SAMPLE**
**Input**: samples $\mathbf{x}^{(1)} \dots \mathbf{x}^{(N)}$ of $X \in \mathbb{R}^n$, an integer $m \leq n$
**Output**: samples $\underline{\mathbf{x}}^{(1)} \dots \underline{\mathbf{x}}^{(N)}$ of $\underline{X} \in \mathbb{R}^m$

    1. Compute $\mu \in \mathbb{R}^n$ where $\mu_i = \frac{1}{N}\sum_{k=1}^{N} \mathbf{x}_i^{(k)}$. Let $\mathbf{Z} = [\mathbf{z}^{(1)} \dots \mathbf{z}^{(N)}] \in \mathbb{R}^{n \times N}$ where
           $\mathbf{z}^{(k)} = \mathbf{x}^{(k)} - \mu$
    2. $\hat{\mathbf{C}}_{XX} \leftarrow \frac{1}{N}\mathbf{Z}\mathbf{Z}^\top \in \mathbb{R}^{n \times n}$
    3. $\hat{\mathbf{A}}_m \leftarrow$ **PCA-PROJECTION**$(\hat{\mathbf{C}}_{XX}, m)$
    4. Return $\underline{\mathbf{x}}^{(1)}, \dots, \underline{\mathbf{x}}^{(N)} \in \mathbb{R}^m$ where
           $\underline{\mathbf{x}}^{(k)} = \hat{\mathbf{A}}_m^\top \mathbf{z}^{(k)}$
```

The new points $\underline{\mathbf{x}}^{(1)} \dots \underline{\mathbf{x}}^{(N)} \in \mathbb{R}^m$ have zero mean, so when we reverse project using $\hat{\mathbf{A}}_m^+$ to approximate the original points, we must add back the subtracted mean to each dimension:

$$\tilde{\mathbf{x}}^{(k)} = \hat{\mathbf{A}}_m^+ \underline{\mathbf{x}}^{(k)} + \mu$$

3 Canonical Correlation Analysis (CCA)

Consider two random variable $X \in \mathbb{R}^{n_1}$ and $Y \in \mathbb{R}^{n_2}$. We believe that they characterize the same object, each a distinct "view" offering different information. So from them, we want to derive new variables $\underline{X}, \underline{Y} \in \mathbb{R}^m$ (where $m \leq \min(n_1, n_2)$) whose correlation is maximized. CCA finds $\underline{X} = (\underline{X}_1 \dots \underline{X}_m)$ and $\underline{Y} = (\underline{Y}_1 \dots \underline{Y}_m)$ such that

$$\underline{X}_i, \underline{Y}_i = \operatorname*{arg\,max}_{\Phi, \Psi \in \mathbb{R}} \mathrm{Cor}(\Phi, \Psi)$$

under the constraint that $\Phi = \mathbf{a}^\top X$ and $\Psi = \mathbf{b}^\top Y$ for some vectors $\mathbf{a} \in \mathbb{R}^{n_1}$ and $\mathbf{b} \in \mathbb{R}^{n_2}$ and that

$$\begin{aligned}\mathrm{Cor}(\Phi, \underline{X}_j) = 0 \\ \mathrm{Cor}(\Psi, \underline{Y}_j) = 0\end{aligned}$$

for $j = 1 \dots i-1$. Note that we no longer constrain the length of projection vectors because scaling does not affect correlation. These new variables $\underline{X}$ and $\underline{Y}$ found by CCA can be viewed as m-dimensional representations of $X \in \mathbb{R}^{n_1}$ and $Y \in \mathbb{R}^{n_2}$ that have incorporated our prior belief that X and Y are referring to the same object.

3.1 A Derivation of the Algorithm

Let $(\mathbf{a}_1, \mathbf{b}_1) \dots (\mathbf{a}_m, \mathbf{b}_m)$ be the projection vectors used for deriving $(\underline{X}_1, \underline{Y}_1) \dots (\underline{X}_m, \underline{Y}_m)$. Finding $(\mathbf{a}_1, \mathbf{b}_1)$ can be framed as the following optimization problem.

$$(\mathbf{a}_1, \mathbf{b}_1) = \operatorname*{arg\,max}_{\mathbf{a} \in \mathbb{R}^{n_1},\, \mathbf{b} \in \mathbb{R}^{n_2}} \mathrm{Cor}(\mathbf{a}^\top X, \mathbf{b}^\top Y) \tag{3}$$

Again, we will assume that each dimension is centered so that

$$\mathbf{E}[X_1] = \cdots = \mathbf{E}[X_{n_1}] = \mathbf{E}[Y_1] = \cdots = \mathbf{E}[Y_{n_2}] = 0$$

without loss of generality because $\mathrm{Cor}(X - \mathbf{E}[X], Y - \mathbf{E}[Y]) = \mathrm{Cor}(X, Y)$ for any random variables X and Y. Using this assumption, we manipulate the expression as follows.

$$\begin{aligned}
\mathrm{Cor}(\mathbf{a}^\top X, \mathbf{b}^\top Y) &= \frac{\mathbf{E}[(\mathbf{a}^\top X)(\mathbf{b}^\top Y)]}{\sqrt{\mathbf{E}[(\mathbf{a}^\top X)^2]\mathbf{E}[(\mathbf{b}^\top Y)^2]}} \\
&= \frac{\mathbf{a}^\top \mathbf{E}[XY^\top]\mathbf{b}}{\sqrt{\mathbf{a}^\top \mathbf{E}[XX^\top]\mathbf{a}}\sqrt{\mathbf{b}^\top \mathbf{E}[YY^\top]\mathbf{b}}} \\
&= \frac{\mathbf{a}^\top \mathbf{C}_{XY}\mathbf{b}}{\sqrt{\mathbf{a}^\top \mathbf{C}_{XX}\mathbf{a}}\sqrt{\mathbf{b}^\top \mathbf{C}_{YY}\mathbf{b}}}
\end{aligned}$$

where $\mathbf{C}_{XY} \in \mathbb{R}^{n_1 \times n_2}$, $\mathbf{C}_{XX} \in \mathbb{R}^{n_1 \times n_1}$, and $\mathbf{C}_{YY} \in \mathbb{R}^{n_2 \times n_2}$ are the covariance matrices. Maximizing it is equivalent to maximizing only the numerator with an additional constraint $\mathbf{a}^\top \mathbf{C}_{XX}\mathbf{a} = \mathbf{b}^\top \mathbf{C}_{YY}\mathbf{b} = 1$. Thus the optimization problem in Eq. (3) can be reframed as

$$(\mathbf{a}_1, \mathbf{b}_1) = \operatorname*{arg\,max}_{\substack{\mathbf{a} \in \mathbb{R}^{n_1},\, \mathbf{b} \in \mathbb{R}^{n_2}: \\ \mathbf{a}^\top \mathbf{C}_{XX}\mathbf{a} = \mathbf{b}^\top \mathbf{C}_{YY}\mathbf{b} = 1}} \mathbf{a}^\top \mathbf{C}_{XY}\mathbf{b}$$

We can simplify the constraint by defining $\Omega \in \mathbb{R}^{n_1 \times n_2}$, $\mathbf{c} \in \mathbb{R}^{n_1}$, and $\mathbf{d} \in \mathbb{R}^{n_2}$ as

$$\begin{aligned}
\Omega &= \mathbf{C}_{XX}^{-1/2}\mathbf{C}_{XY}\mathbf{C}_{YY}^{-1/2} \\
\mathbf{c} &= \mathbf{C}_{XX}^{1/2}\mathbf{a} \\
\mathbf{d} &= \mathbf{C}_{YY}^{1/2}\mathbf{b}
\end{aligned}$$

For this, we need $\mathbf{C}_{XX}$ and $\mathbf{C}_{YY}$ to be invertible. This corresponds to requiring that each X_i behaves differently from X_j for $j \neq i$ (similarly for Y_i), which is benign in that we can always choose to ignore redundant variables. Now we first solve for

$$(\mathbf{c}_1, \mathbf{d}_1) = \operatorname*{arg\,max}_{\mathbf{c} \in \mathbb{R}^d,\, \mathbf{d} \in \mathbb{R}^{d'}:\, ||\mathbf{c}||^2 = ||\mathbf{d}||^2 = 1} \mathbf{c}^\top \Omega \mathbf{d} \tag{4}$$

Then $\mathbf{a}_1 = \mathbf{C}_{XX}^{-1/2}\mathbf{c}_1$ and $\mathbf{b}_1 = \mathbf{C}_{YY}^{-1/2}\mathbf{d}_1$ are the solution to Eq. (3). We will now use an SVD on Ω to solve Eq. (4) (see section 5.2).

Proposition 2. $\mathbf{c}_1$ and $\mathbf{d}_1$ are respectively the left and right unit singular vectors of Ω with the greatest singular value.

Proof. We again use the Lagrangian relaxation to maximize the quantity $\mathbf{c}^\top \Omega \mathbf{d}$ under the normalization constraint $||\mathbf{c}||^2 = ||\mathbf{d}||^2 = 1$:

$$L = \mathbf{c}^\top \Omega \mathbf{d} + \frac{1}{2}\sigma(1 - \mathbf{c}^\top \mathbf{c}) + \frac{1}{2}\sigma'(1 - \mathbf{d}^\top \mathbf{d})$$

When we differentiate with respect to $\mathbf{c}$ and $\mathbf{d}$ and set to zero, we arrive in the equations

$$\Omega \mathbf{d} = \sigma \mathbf{c}$$
$$\Omega^\top \mathbf{c} = \sigma' \mathbf{d}$$

The fact that $\sigma = \sigma'$ can be seen by multiplying the equations by $\mathbf{c}^\top$ and $\mathbf{d}^\top$, and using the normalization constraint,

$$\mathbf{c}^\top \Omega \mathbf{d} = \sigma$$
$$\mathbf{d}^\top \Omega^\top \mathbf{c} = \sigma'$$

where $\mathbf{c}^\top \Omega \mathbf{d} = \mathbf{d}^\top \Omega \mathbf{c}$. This tells us that $\mathbf{c}$ is the left and $\mathbf{d}$ is the right unit singular vector of Ω with singular value σ. Since our objective is to maximize $\mathbf{c}^\top \Omega \mathbf{d} = \sigma$, the claim follows. □

More generally, $(\mathbf{c}_1, \mathbf{d}_1) \ldots (\mathbf{c}_m, \mathbf{d}_m)$ are the unit singular vectors of Ω that correspond to the top m largest singular values.

Theorem 2. $\mathbf{c}_i$ and $\mathbf{d}_i$ are respectively the left and right unit singular vectors of Ω with the i^{th} greatest singular value.

Proof. The base case is given by proposition 2. Let $(\mathbf{c}_i, \mathbf{d}_i)$ be the i^{th} unit singular vectors of Ω. Then $(\mathbf{c}_i, \mathbf{d}_i)$ maximizes over all $\mathbf{c} \in \mathbb{R}^{n_1}$ and $\mathbf{d} \in \mathbb{R}^{n_2}$

$$\mathbf{c}^\top \Omega \mathbf{d} = \mathrm{Cor}((\mathbf{C}_{XX}^{-1/2}\mathbf{c})^\top X, (\mathbf{C}_{YY}^{-1/2}\mathbf{d})^\top Y)$$

while satisfying $\mathbf{c}_i^\top \mathbf{c}_j = \mathbf{d}_i^\top \mathbf{d}_j = 0$ for $j \in \{1 \ldots i-1\}$. This means

$$\mathrm{Cor}((\mathbf{C}_{XX}^{-1/2}\mathbf{c}_i)^\top X, (\mathbf{C}_{XX}^{-1/2}\mathbf{c}_j)^\top X) = \mathbf{c}_i^\top \mathbf{c}_j = 0$$
$$\mathrm{Cor}((\mathbf{C}_{YY}^{-1/2}\mathbf{d}_i)^\top Y, (\mathbf{C}_{YY}^{-1/2}\mathbf{d}_j)^\top Y) = \mathbf{d}_i^\top \mathbf{d}_j = 0$$

Letting $\mathbf{a}_i = \mathbf{C}_{XX}^{-1/2}\mathbf{c}_i$ and $\mathbf{b}_i = \mathbf{C}_{YY}^{-1/2}\mathbf{d}_i$, we have

$$(\mathbf{a}_i, \mathbf{b}_i) = \underset{\substack{\mathbf{a} \in \mathbb{R}^{n_1},\ \mathbf{b} \in \mathbb{R}^{n_2}: \\ \mathrm{Cor}(\mathbf{a}^\top X, \mathbf{a}_j^\top X) = 0\ \ \forall j \in \{1 \ldots i-1\} \\ \mathrm{Cor}(\mathbf{b}^\top Y, \mathbf{b}_j^\top Y) = 0\ \ \forall j \in \{1 \ldots i-1\}}}{\arg\max} \mathrm{Cor}(\mathbf{a}^\top X, \mathbf{b}^\top Y)$$

as desired. □

Theorem 2 states that the singular value σ_i of Ω is the correlation between $\underline{X}_i$ and $\underline{Y}_i$:

$$\sigma_i = \mathrm{Cor}(\underline{X}_i, \underline{Y}_i)$$

Thus we can decide dimension m by inspecting the singular value spectrum of Ω.

We have shown that we can obtain vectors $(\mathbf{c}_1, \mathbf{d}_1) \dots (\mathbf{c}_m, \mathbf{d}_m)$ in a single shot via an SVD on Ω and then set $\mathbf{a}_i = \mathbf{C}_{XX}^{-1/2}\mathbf{c}_i$ and $\mathbf{b}_i = \mathbf{C}_{YY}^{-1/2}\mathbf{d}_i$ for $i = 1 \dots m$. Once we have these vectors, we can form matrix $\mathbf{A}_m = [\mathbf{a}_1 \dots \mathbf{a}_m] \in \mathbb{R}^{n_1 \times m}$ to project X down to $\underline{X} = \mathbf{A}_m^\top X$ and matrix $\mathbf{B}_m = [\mathbf{b}_1 \dots \mathbf{b}_m] \in \mathbb{R}^{n_2 \times m}$ to project Y down to $\underline{Y} = \mathbf{B}_m^\top X$.

$$\underline{X} = \mathbf{A}_m^\top X = (\mathbf{a}_1^\top X \dots \mathbf{a}_m^\top X) = (\underline{X}_1 \dots \underline{X}_m)$$
$$\underline{Y} = \mathbf{B}_m^\top Y = (\mathbf{b}_1^\top Y \dots \mathbf{b}_m^\top Y) = (\underline{Y}_1 \dots \underline{Y}_m)$$

The algorithm for computing these transformations is given below.

```
CCA-PROJECTIONS
Input: covariance matrices for $X \in \mathbb{R}^{n_1}$ and $Y \in \mathbb{R}^{n_2}$ where $\mathbf{E}[X_i] = \mathbf{E}[Y_i] = 0$
    • $\mathbf{C}_{XY} \in \mathbb{R}^{n_1 \times n_2}$
    • invertible $\mathbf{C}_{XX} \in \mathbb{R}^{n_1 \times n_1}$ and invertible $\mathbf{C}_{YY} \in \mathbb{R}^{n_2 \times n_2}$
    • dimension $m \leq \min(n_1, n_2)$
Output: CCA projections $\mathbf{A}_m \in \mathbb{R}^{n_1 \times m}$ and $\mathbf{B}_m \in \mathbb{R}^{n_2 \times m}$
    1. $\Omega \leftarrow \mathbf{C}_{XX}^{-1/2}\mathbf{C}_{XY}\mathbf{C}_{YY}^{-1/2} \in \mathbb{R}^{d \times d'}$.
    2. Compute an SVD
           $\Omega = [\mathbf{c}_1 \dots \mathbf{c}_{n_1}] \times \Sigma \times [\mathbf{d}_1 \dots \mathbf{d}_{n_2}]^\top$
       where $\Sigma$ is a diagonal matrix of singular values $\sigma_1 \geq \cdots \geq \sigma_{\min(n_1,n_2)} \geq 0$ and
       $(\mathbf{c}_i, \mathbf{d}_i)$ are the corresponding (normalized) left and right singular vectors.
    3. Return $\mathbf{A}_m = \mathbf{C}_{XX}^{-1/2}[\mathbf{c}_1 \dots \mathbf{c}_m]$ and $\mathbf{B}_m = \mathbf{C}_{YY}^{-1/2}[\mathbf{d}_1 \dots \mathbf{d}_m]$.
```

3.2 Sample-Based CCA

In practice, we have samples $(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}) \dots (\mathbf{x}^{(N)}, \mathbf{y}^{(N)})$ of the random variables $X \in \mathbb{R}^{n_1}$ and $Y \in \mathbb{R}^{n_2}$ and compute empirical estimates $\hat{\mathbf{C}}_{XY}$, $\hat{\mathbf{C}}_{XX}$, and $\hat{\mathbf{C}}_{YY}$. Since we must have $\mathbf{E}[X_i] = \mathbf{E}[Y_i] = 0$, we center the samples to obtain $\mathbf{s}^{(1)} \dots, \mathbf{s}^{(N)} \in \mathbb{R}^{n_1}$ and $\mathbf{t}^{(1)}, \dots, \mathbf{t}^{(N)} \in \mathbb{R}^{n_2}$: for all $k = 1 \dots N$,

$$\mathbf{s}_i^{(k)} = \mathbf{x}_i^{(k)} - \frac{1}{N}\sum_{k=1}^{N} \mathbf{x}_i^{(k)} \quad \forall i \in \{1 \dots n_1\}$$
$$\mathbf{t}_i^{(k)} = \mathbf{y}_i^{(k)} - \frac{1}{N}\sum_{k=1}^{N} \mathbf{y}_i^{(k)} \quad \forall i \in \{1 \dots n_2\}$$

If we define $\mathbf{S} = [\mathbf{s}^{(1)} \dots \mathbf{s}^{(N)}] \in \mathbb{R}^{n_1 \times N}$ and $\mathbf{T} = [\mathbf{t}^{(1)} \dots \mathbf{t}^{(N)}] \in \mathbb{R}^{n_2 \times N}$, then the coavriance matrices are estimated as

$$\hat{\mathbf{C}}_{XY} = \frac{1}{N}\mathbf{S}\mathbf{T}^\top \qquad \hat{\mathbf{C}}_{XX} = \frac{1}{N}\mathbf{S}\mathbf{S}^\top \qquad \hat{\mathbf{C}}_{YY} = \frac{1}{N}\mathbf{T}\mathbf{T}^\top$$

The algorithm for deriving the CCA samples is given below. For simplicity, we assume $\hat{\mathbf{C}}_{XX}$ and $\hat{\mathbf{C}}_{YY}$ have full rank, but we can ensure this condition by preprocessing. For instance, we can remove redundant dimensions with PCA:

$$\mathbf{x}^{(1)} \dots \mathbf{x}^{(N)} = \textbf{PCA-SAMPLE}(\mathbf{x}_{\text{old}}^{(1)} \dots \mathbf{x}_{\text{old}}^{(N)}, \text{rank}(\hat{\mathbf{C}}_{XX}))$$
$$\mathbf{y}^{(1)} \dots \mathbf{y}^{(N)} = \textbf{PCA-SAMPLE}(\mathbf{y}_{\text{old}}^{(1)} \dots \mathbf{y}_{\text{old}}^{(N)}, \text{rank}(\hat{\mathbf{C}}_{YY}))$$

```
CCA-SAMPLE
Input: samples $(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}) \dots (\mathbf{x}^{(N)}, \mathbf{y}^{(N)})$ of $X \in \mathbb{R}^{n_1}$ and $Y \in \mathbb{R}^{n_2}$, $m \leq \min(n_1, n_2)$
Output: samples $(\underline{\mathbf{x}}^{(1)}, \underline{\mathbf{y}}^{(1)}) \dots (\underline{\mathbf{x}}^{(N)}, \underline{\mathbf{y}}^{(N)})$ of $\underline{X} \in \mathbb{R}^m$ and $\underline{Y} \in \mathbb{R}^m$
    1. Compute $\mu^X \in \mathbb{R}^{n_1}$ and $\mu^Y \in \mathbb{R}^{n_2}$ where
           $\mu_i^X = \frac{1}{N} \sum_{k=1}^{N} \mathbf{x}_i^{(k)}$        $\mu_i^Y = \frac{1}{N} \sum_{k=1}^{N} \mathbf{y}_i^{(k)}$
       Let $\mathbf{S} = [\mathbf{s}^{(1)} \dots \mathbf{s}^{(N)}] \in \mathbb{R}^{n_1 \times N}$ and $\mathbf{T} = [\mathbf{t}^{(1)} \dots \mathbf{t}^{(N)}] \in \mathbb{R}^{n_2 \times N}$ where
           $\mathbf{s}^{(k)} = \mathbf{x}^{(k)} - \mu^X$
           $\mathbf{t}^{(k)} = \mathbf{y}^{(k)} - \mu^Y$
    2. $\hat{\mathbf{C}}_{XY} \leftarrow \frac{1}{N}\mathbf{S}\mathbf{T}^\top, \hat{\mathbf{C}}_{XX} \leftarrow \frac{1}{N}\mathbf{S}\mathbf{S}^\top, \hat{\mathbf{C}}_{YY} \leftarrow \frac{1}{N}\mathbf{T}\mathbf{T}^\top$
    3. $(\hat{\mathbf{A}}_m, \hat{\mathbf{B}}_m) \leftarrow$ CCA-PROJECTION$(\hat{\mathbf{C}}_{XY}, \hat{\mathbf{C}}_{XX}, \hat{\mathbf{C}}_{YY}, m)$
    4. Return $\underline{\mathbf{x}}^{(1)}, \dots, \underline{\mathbf{x}}^{(N)} \in \mathbb{R}^m$ and $\underline{\mathbf{y}}^{(1)}, \dots, \underline{\mathbf{y}}^{(N)} \in \mathbb{R}^m$ where
           $\underline{\mathbf{x}}^{(k)} = \hat{\mathbf{A}}_m^\top \mathbf{s}^{(k)}$
           $\underline{\mathbf{y}}^{(k)} = \hat{\mathbf{B}}_m^\top \mathbf{t}^{(k)}$
```

4 Experiments

4.1 PCA

A primary usage of PCA is to condense information into a smaller space. By doing so, we often gain better understanding of the data and eliminate noise.

4.1.1 Condensed Information

Our data consists of 435 representatives' voting record on 16 bills, where a vote can be either `yes`, `no`, or `?`.[1] Thus we can model a representative as a 16-dimensional random

[1] Congressional Quarterly Almanac, 98th Congress, 2nd session 1984, Volume XL: Congressional Quarterly Inc. Washington, D.C., 1985

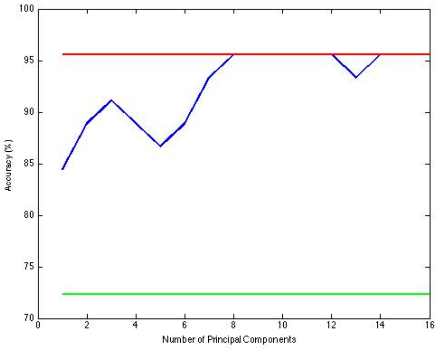


Figure 1: RED: using all of the original 16 features; BLUE: using the top m PCA components; GREEN: using a single original feature, averaged over the 16 features

variable

$$X = (X_1 \dots X_{16}) : \; X_i \in \{\texttt{yes}, \texttt{no}, \texttt{?}\}$$

for which we have 435 samples $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(435)} \in \mathbb{R}^{16}$. The eigenvalue spectrum of the sample covariance matrix $\hat{\mathbf{C}}_{XX} \in \mathbb{R}^{16 \times 16}$ is

1	2	3	4	5	$\cdots$	15	16
2971.3	554.9	441.4	347.2	300.6	$\cdots$	88.5	55.4

Note the eigenvalue mass is heavily concentrated in the first few components. This means given $m \ll 16$, the PCA representation

$$\underline{X} = (\underline{X}_1 \dots \underline{X}_m) : \; X_i \in \mathbb{R}$$

will preserve much of the information in the data.

To demonstrate this, we will predict a representative's party affiliation (either republican or democratic), using the votes as features to an SVM. We train on 390 points and test on 45 points. The classification performance is shown in figure 1. For training and testing,

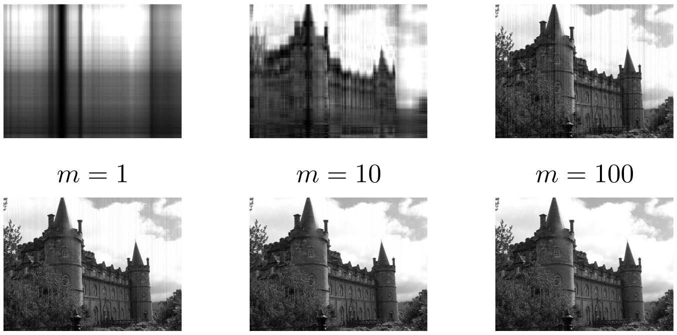
- RED: 16 original bills $X_1, \dots, X_{16} \in \{\texttt{yes}, \texttt{no}, \texttt{?}\}$
- BLUE: m PCA components $X_i, \dots, X_m \in \mathbb{R}$
- GREEN: 1 original bill $X_i \in \{\texttt{yes}, \texttt{no}, \texttt{?}\}$ (averaged)

We see that there is no drastic loss in accuracy with the lower dimensional PCA representation. In particular, note that the top component alone packs an enormous amount of information. In contrast, using a single original feature performs poorly.

4.1.2 Size Reduction

Consider a 768 × 1024 black and white picture of an old castle. We will "squash" each row of the image via PCA. A row is modeled as a 1024-dimensional random variable

$$X = (X_1 \dots X_{1024}) : \quad X_i \in [0, 1]$$



$m = 1$ $m = 10$ $m = 100$

$m = 300$ $m = 767$ original

Table 1: Images restored from varying m values

The eigenvalue spectrum of $\hat{\mathbf{C}}_{XX}$ is

1	2	3	4	5	6	$\cdots$	767	768	$\cdots$	1024
54314	11690	6606	1993	1226	948	$\cdots$	1.2265×10^{-4}	0	$\cdots$	0

The images restored from various m values are shown in table 1. We can see that the approximated recovery with $m = 767$ is almost as good as the original image. This is a 25% reduction in size.

4.2 CCA

CCA is most interesting when the two variables are significantly different. By projecting them down to the most correlated space, CCA derives new variables that have learned from each other.

4.2.1 Multi-view Learning

We again consider the classification task in section 4.1.1 with a twist: we only have the voting record on 7 bills. Thus our view of a person is limited to the following random variable.

$$X = (X_1 \dots X_7) : \; X_i \in \{\texttt{yes}, \texttt{no}, \texttt{?}\}$$

Now, assume we obtain the voting record on the other 9 bills *for training only.* This becomes our second view of a person.

$$Y = (Y_1 \dots Y_9) : \; Y_i \in \{\texttt{yes}, \texttt{no}, \texttt{?}\}$$

Within the training data, CCA learns two projections $\mathbf{A}_m \in \mathbb{R}^{7 \times m}$ and $\mathbf{B}_m \in \mathbb{R}^{9 \times m}$ that map X and Y to an m-dimensional space where their correlation is maximized. For the test data in which the second view is absent, we project each instance (i.e., the votes on 7 bills) with $\mathbf{A}_m$ and then do the classification, essentially amplifying the semantics of the data by leveraging the second view that was available in training. The effect is seen in a boost in performance.

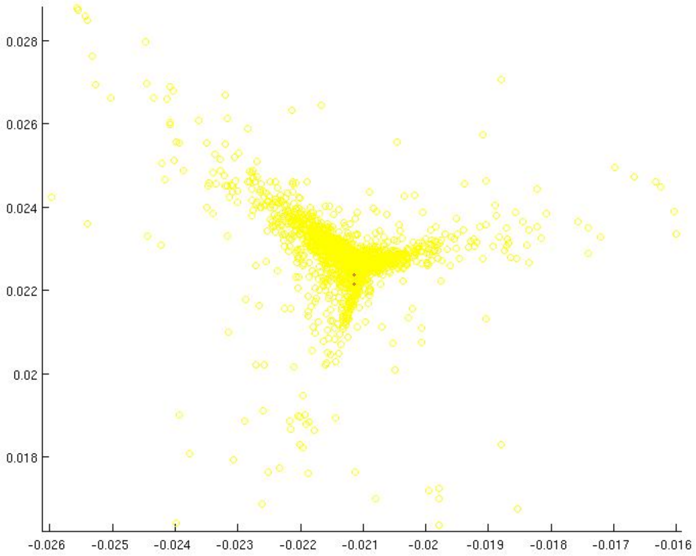


Figure 2: The first two dimensions in a CCA representation of words from Gigaword. The two red dots correspond to "Paul" and "David".

	classification procedure	accuracy
single-view learning	train (X), test (X)	75%
multi-view learning via CCA	train (X, Y), test (X)	78%

As a final example, we derive a CCA represenation of English words. One view of a word is its identity. The other view is its surrounding context. Figure 2 shows the first two CCA components of the words obtained from 4.1 billion tokens in the Gigaword dataset. The horizontal axis corresponds to the first dimension and the vertical axis to the second. The two red dots correspond to words "Paul" and "David". Note their closeness, especially in the first dimension which is far more important than the second. CCA has learned that "Paul" and "David" are similar words from their neighboring context.

Further Reading

Hardoon, D. R., Szedmak, S. R., and Shawe-Taylor, J. R. (2004). Canonical correlation analysis: An overview with application to learning methods.

Kakade, S. and Foster, D. (2007). Multi-view regression via canonical correlation analysis.

D Hsu, S M. Kakade, and Tong Zhang (2009). A spectral algorithm for learning hidden markov models.

Shay Cohen, Karl Stratos, Michael Collins, Dean Foster, and Lyle Ungar (2012). Spectral learning of latent-variable PCFGs.

5 Appendix

5.1 Eigenvalue Decomposition

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a matrix that can be expressed as $\mathbf{A} = \mathbf{B}\mathbf{B}^\top$ for some real matrix $\mathbf{B}$. Then $\mathbf{A}$ is said to be positive semi-definite and has real non-negative eigenvalues. Thus an eigenvalue decomposition of $\mathbf{A}$

$$\underbrace{\mathbf{A}}_{n\times n} = \underbrace{Q}_{n\times n}\underbrace{\Lambda}_{n\times n}\underbrace{Q^\top}_{n\times n} = \begin{bmatrix}\mathbf{q}_1 & \cdots & \mathbf{q}_n\end{bmatrix} \times \begin{bmatrix}\lambda_1 & & \varnothing \\ & \ddots & \\ \varnothing & & \lambda_n\end{bmatrix} \times \begin{bmatrix}\mathbf{q}_1^\top \\ \vdots \\ \mathbf{q}_n^\top\end{bmatrix}$$

with the following properties is always possible.

- We have n eigenvalues $\lambda_1 \geq \ldots \geq \lambda_n \geq 0$.
- We have n eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n \in \mathbb{R}^n$ such that
  - ◇ Each $\mathbf{q}_i$ corresponds to λ_i.
  - ◇ Each $\mathbf{q}_i$ yields $\mathbf{q}_i^\top \mathbf{q}_i = 1$.
  - ◇ Each $\mathbf{q}_i$ yields $\mathbf{q}_i^\top \mathbf{q}_j = 0$ for $j \neq i$.

5.2 Singular Value Decomposition

Let $\mathbf{A} \in \mathbb{R}^{n\times m}$ be any real matrix and $r = \min(n, m)$. A singular value decomposition (SVD) of $\mathbf{A}$ has the form

$$\underbrace{\mathbf{A}}_{n\times m} = \underbrace{U}_{n\times n}\underbrace{\Sigma}_{n\times m}\underbrace{V^\top}_{m\times m} = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_n\end{bmatrix} \times \begin{bmatrix}\sigma_1 & & \varnothing \\ & \ddots & \\ & & \sigma_r \\ \varnothing & & \end{bmatrix} \times \begin{bmatrix}\mathbf{v}_1^\top \\ \vdots \\ \mathbf{v}_m^\top\end{bmatrix}$$

with the following properties

- We have r singular values $\sigma_1 \geq \ldots \geq \sigma_r \geq 0$.
- We have n left singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_n \in \mathbb{R}^n$ and m right singular vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{R}^m$ such that
  - ◇ Each pair $(\mathbf{u}_i, \mathbf{v}_i)$ corresponds to σ_i for $j = 1 \ldots r$.
  - ◇ Each pair $(\mathbf{u}_i, \mathbf{v}_i)$ yields $\mathbf{u}_i^\top \mathbf{u}_i = \mathbf{v}_i^\top \mathbf{v}_i = 1$.
  - ◇ Each pair $(\mathbf{u}_i, \mathbf{v}_i)$ yields $\mathbf{u}_i^\top \mathbf{u}_j = \mathbf{v}_i^\top \mathbf{v}_j = 0$ for $j \neq i$.

The singular values of $\mathbf{A} \in \mathbb{R}^{n\times m}$ are related to the eigenvalues of $\mathbf{A}\mathbf{A}^\top \in \mathbb{R}^{n\times n}$ as

$$\sigma_i = \sqrt{\lambda_i}$$

where $\lambda_1 \ldots \lambda_r$ are the top r eigenvalues of $\mathbf{A}\mathbf{A}^\top$ sorted in descending order.