Technical Report
TTI-TR-2008-4

Multi-View Dimensionality Reduction via Canonical Correlation Analysis

Dean P. Foster
University of Pennsylvania
Sham M. Kakade
Toyota Technological Institute at Chicago
Tong Zhang
Rutgers University

ABSTRACT

We analyze the multi-view regression problem where we have two views $X = (X^{(1)}, X^{(2)})$ of the input data and a target variable Y of interest. We provide sufficient conditions under which we can reduce the dimensionality of X (via a projection) without loosing predictive power of Y. Crucially, this projection can be computed via a Canonical Correlation Analysis only on the unlabeled data. The algorithmic template is as follows: with unlabeled data, perform CCA and construct a certain projection; with the labeled data, do least squares regression in this lower dimensional space. We show how, under certain natural assumptions, the number of labeled samples could be significantly reduced (in comparison to the single view setting) — in particular, we show how this dimensionality reduction does not loose predictive power of Y (thus it only introduces little bias but could drastically reduce the variance).

We explore two separate assumptions under which this is possible and show how, under either assumption alone, dimensionality reduction could reduce the labeled sample complexity. The two assumptions we consider are a *conditional independence* assumption and a *redundancy* assumption. The typical conditional independence assumption is that conditioned on Y the views $X^{(1)}$ and $X^{(2)}$ are independent — we relax this assumption to be conditioned on some hidden state H the views $X^{(1)}$ and $X^{(2)}$ are independent. Under the redundancy assumption, we have that the best predictor from each view is roughly as good as the best predictor using both views.

1 Introduction

In recent years, the “multi-view” approach has been receiving increasing attention as a paradigm for semi-supervised learning. In the two view setting, there are views (sometimes in a rather abstract sense) $X^{(1)}$ and $X^{(2)}$ of the data, which co-occur, and there is a target variable Y of interest. The setting is one where it is easy to obtain unlabeled samples $(X^{(1)}, X^{(2)})$ but the labeled samples $(X^{(1)}, X^{(2)}, Y)$ are more scarce. The goal is to implicitly learn about the target Y via the relationship between $X^{(1)}$ and $X^{(2)}$.

We work in a setting where we have a joint distribution over $(X^{(1)}, X^{(2)}, Y)$, where all $X^{(1)}$ and $X^{(2)}$ are vectors (of arbitrarily large dimension) and $Y \in \mathbb{R}$.

This work focuses on the underlying assumptions in the multi-view setting and provides an algorithm which exploits these assumptions. We *separately* consider two natural assumptions, a conditional independence assumption and a redundancy assumption. Our work here builds upon the work in Ando and Zhang [2007] and Kakade and Foster [2007], summarizing the close connections between the two.

Ando and Zhang [2007] provide an analysis under only a conditional independence assumption — where $X^{(1)}$ and $X^{(2)}$ are conditionally independent of Y (in a multi-class setting, where Y is one of k outcomes). The common criticism of these conditional independence assumption is that it is far too stringent to assume that $X^{(1)}$ and $X^{(2)}$ are independent just conditioned on a the rather low dimensional target variable Y. We relax this assumption by only requiring that $X^{(1)}$, $X^{(2)}$, and Y all be independent conditioned on some hidden state H. Roughly speaking, we think of H as being the augmented information required to make $X^{(1)}$, $X^{(2)}$, and Y conditionally independent.

The other assumption we consider (and we consider it separately from the previous one) is based on redundancy, as in Kakade and Foster [2007]. Here, we assume that the best linear predictor from each view is roughly as good as the best linear predictor based on both views. This assumption is weak in the sense that it only requires, on average, for the optimal linear predictors from each view to agree.

There are many natural applications for which either of these underlying assumptions are applicable. For example, consider a setting where it is easy to obtain pictures of objects from different camera angles and say our supervised task is one of object recognition. Here, the first assumption holds, and, intuitively, we can think of unlabeled data as providing examples of viewpoint invariance. If there is no occlusion, then we expect our second assumption to hold as well. One can even consider multi-modal views, with one view being a video stream and the other an audio stream, and the task might be to identify properties of the speaker (e.g. recognition) — here conditioned on the speaker identity, the views may be uncorrelated. In NLP, an example would be a paired document corpus, consisting of a document and its translation into another language, and the supervised task could be understanding some high level property of the document (here both assumptions may hold). The motivating example in Blum and Mitchell [1998] is a webpage classification task, where one view was the text in the page and the other was the hyperlink structure.

It turns out that under *either* assumption, Canonical Correlation Analysis (CCA) provides a dimensionality reduction method, appropriate for use in a regression algorithm (see Hardoon et al. [2004] for a review of CCA with applications to machine learning). In particular, the semi-supervised algorithm is:

1. Using unlabeled data $\{(X^{(1)}, X^{(2)})\}$, perform a CCA.
2. Construct a projection Π that projects $(X^{(1)}, X^{(2)})$ to the most correlated lower dimensional subspace (as specified in Theorems 3 and 5).
3. With a labeled dataset $\{(X^{(1)}, X^{(2)}, Y)\}$, do a least squares regression (with MLE estimates) in this lower dimensional subspace, i.e. regress Y with $(\Pi X^{(1)}, \Pi X^{(2)})$.

Algorithm 1: Regression in a CCA Subspace

Our main results show that (under either assumption) we lose little predictive information by using this lower dimensional CCA subspace – the gain is that our regression problem has a lower sample complexity due to the lower dimension.

1.1 Related Work

Both of these assumptions have been considered together in the co-training framework of Blum and Mitchell [1998] (in a rather strong sense).

In Ando and Zhang [2007], the conditional independence assumption was with respect to a multi-class setting, where Y is discrete, i.e. $Y \in [k]$. In our generalization, we let Y be real valued and we relax the assumption in that we need only independence with respect to some hidden state. Our proof is similar in spirit to that in Ando and Zhang [2007].

The redundancy assumption we consider here is from Kakade and Foster [2007], where two algorithms were proposed: one based on "shrinkage" (a form of regularization) and one based on dimensionality reduction. The results were stronger for the shrinkage based algorithm. Here, we show that the dimensionality reduction based algorithm works just as well as the proposed "shrinkage" algorithm.

2 Multi-View Assumptions

All vectors in our setting are column vectors. We slightly abuse notation and write $(X^{(1)}, X^{(2)})$, which really denotes the column vector of $X^{(1)}$ concatenated with $X^{(2)}$.

Recall the definition of the R^2, the coefficient of determination, between Y and X. Let $\beta \cdot X$ be the best linear prediction of Y with X. Recall that:

$$R^2_{X,Y} := \text{correlation}(\beta \cdot X, Y)^2$$

In the words, $R^2_{Y,X}$ is the proportion of variability in Y that is accounted for by the best linear prediction with X, i.e.

$$R^2_{X,Y} = 1 - \frac{\text{loss}(\beta)}{\text{var}(Y)}$$

where loss is the square loss.

2.1 Independence and Predictability of Hidden States

First, let us present the definition of a hidden state H. Intuitively, we think of hidden states as those which imply certain independence properties with respect to our observed random variables.

Definition We say that a random vector H is a *hidden state* for $X^{(1)}$, $X^{(2)}$ and Y if, conditioned on H, we have that $X^{(1)}$, $X^{(2)}$, and Y are all uncorrelated.

Note there always exits an H which satisfies this uncorrelated property. We say H is a *linear hidden state* if we also have that $\mathbb{E}[X^{(1)}|H]$, $\mathbb{E}[X^{(2)}|H]$, and $\mathbb{E}[Y|H]$ are linear in H.

Instead of dealing with independence with respect to Y (which is typically far too stringent), our assumption will be with respect to H, which always exists. Also note that the above definition only requires uncorrelatedness rather than independence.

Assumption 1. (Hidden State Predictability) *Let H be a linear hidden state such that both $X^{(1)}$ and $X^{(2)}$ are non-trivially predictive of H. More precisely, assume that for all directions $w \in \mathbb{R}^{dim(H)}$:*

$$R^2_{X^{(1)},w\cdot H} > 0,\ R^2_{X^{(2)},w\cdot H} > 0$$

Intuitively, this multi-view assumption is that both $X^{(1)}$ and $X^{(2)}$ are informative of the hidden state. However, they need not be good predictors.

2.2 Redundancy

The other assumption we consider is one based on redundancy.

Assumption 2. **(ϵ-Redundancy)** *Assume that the best linear predictor from each view is roughly as good as the best linear predictor based on both views. More precisely, we have:*

$$R^2_{X^{(1)},Y} \geq R^2_{X,Y} - \epsilon$$
$$R^2_{X^{(2)},Y} \geq R^2_{X,Y} - \epsilon$$

Note this equivalent to:

$$\text{loss}(\beta_1) - \text{loss}(\beta) \leq \epsilon \ \text{var}(Y)$$
$$\text{loss}(\beta_2) - \text{loss}(\beta) \leq \epsilon \ \text{var}(Y)$$

where β_1, β_2 and β are the best linear predictors with $X^{(1)}$, $X^{(2)}$, and X, respectively. This is the form of the assumption stated in Kakade and Foster [2007].

3 CCA and Projections

We say that $\{U_i\}_i$ and $\{V_i\}_i$ are canonical coordinate systems for $X^{(1)}$ and $X^{(2)}$ if they are an orthonormal basis for each view and they satisfy

$$\text{correlation}(U_i \cdot X^{(1)}, V_j \cdot X^{(2)}) = \begin{cases} \lambda_i & \text{if } i = j \\ 0 & \text{else} \end{cases}$$

CCA finds such a basis (which always exists). Without loss of generality, assume that:

$$1 \geq \lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots . \geq 0$$

We refer to U_i and V_i as the i-th canonical directions and λ_i as the i-th canonical value.

Let Π_{CCA} be the projection operator which projects into the CCA subspaces, which are *strictly* correlated. More precisely, define the strictly correlated subspaces as

$$\mathcal{U} = \text{span}(\{U_i : \lambda_i > 0\}_i), \ \mathcal{V} = \text{span}(\{V_i : \lambda_i > 0\}_i) \,. \tag{1}$$

Define $\Pi_{\text{CCA}} X^{(1)}$ and $\Pi_{\text{CCA}} X^{(2)}$ to be the projection (using Euclidean distance) of $X^{(1)}$ and $X^{(2)}$ into $\mathcal{U}$ and $\mathcal{V}$, respectively. In particular, $\Pi_{\text{CCA}} X^{(1)} = \sum_{i:\lambda_i > 0} (X \cdot U_i) U_i$. These projection operators take each view into the subspace that is strictly correlated with the other view.

Now let us define Π_λ as the projection which takes vectors into the subspace which has a correlation (to the other view) no less than λ. More precisely, define:

$$\mathcal{U}_\lambda = \text{span}(\{U_i : \lambda_i \geq \lambda\}_i), \ \mathcal{V}_\lambda = \text{span}(\{V_i : \lambda_i \geq \lambda\}_i) \,.$$

Similarly, let $\Pi_\lambda X^{(1)}$ and $\Pi_\lambda X^{(2)}$ be the projection (using Euclidean distance) of $X^{(1)}$ and $X^{(2)}$ into $\mathcal{U}_\lambda$ and $\mathcal{V}_\lambda$, respectively. This projection Π_λ is useful since sometimes we deal with subspaces that are sufficiently correlated (to the tune of λ).

4 Dimensionality Reduction Under Conditional Independence

We now present our main theorems, under our Hidden State Predictability Assumption. These theorems shows that after dimensionality reduction (via CCA), we have not lost any predictive power of our target variable.

Theorem 3. *Suppose that Assumption 1 holds and that the dimension of H is k. Then Π_{CCA} is projection into a subspace of dimension precisely k and the following three statements hold:*

1. *This best linear predictor of Y with $X^{(1)}$ is equal to the best linear predictor of Y with $\Pi_{CCA}X^{(1)}$.*
2. *This best linear predictor of Y with $X^{(2)}$ is equal to the best linear predictor of Y with $\Pi_{CCA}X^{(2)}$.*
3. *This best linear predictor of Y with $(X^{(1)}, X^{(2)})$ is equal to the best linear predictor of Y with $\Pi_{CCA}X = (\Pi_{CCA}X^{(1)}, \Pi_{CCA}X^{(2)})$.*

where the best linear predictor is measured with respect to the square loss.

This lemma shows that we need only concern ourselves with a k dimensional regression problem, after the CCA projection, and we have not lost any predictive power. Note that the prediction error in each of these cases need *not* be the same. In particular, with both views, one could potentially obtain significantly lower error.

In addition to direct CCA reduction, one may derive a similar result using bilinear functions of $X^{(1)}$ and $X^{(2)}$. Let d_1 be the dimension of $X^{(1)}$ and d_2 be the dimension of $X^{(2)}$. We define the tensor product $X^{(1)} \circ X^{(2)}$ as the vector $[X_i^{(1)} X_j^{(2)}]_{i,j} \in R^{d_1 d_2}$.

Theorem 4. *Suppose that Assumption 1 holds and that the dimension of H is k. Let $Z = X^{(1)} \circ X^{(2)}$. Then the best linear predictor of Y with Z is equal to the best linear predictor of Y with the following k^2 projected variables*

$$Z^\top (\mathbb{E} Z Z^\top)^{-1} ((\mathbb{E} X^{(1)} {X^{(1)}}^\top U_i) \circ (\mathbb{E} X^{(2)} {X^{(2)}}^\top V_j)),$$

where $U_i \in \mathcal{U}$ and $V_j \in \mathcal{V}$ are CCA basis vectors for the two views respectively, as defined in Equation 1 (so $i, j = 1, \ldots, k$).

Note that $\mathbb{E} Z Z^\top$, $\mathbb{E} X^{(1)} {X^{(1)}}^\top$, and $\mathbb{E} X^{(2)} {X^{(2)}}^\top$ can be computed from unlabeled data. Therefore Theorem 4 says that we can compute a k^2 dimensional subspace that contains the best linear predictor using the tensor product of the two views. If the representation for each view contains a complete basis for functions defined on the corresponding view, then the tensor product gives a complete basis for functions that depend on both views. In such a case, Theorem 4 implies consistency. That is, the optimal predictor using $[X^{(1)}, X^{(2)}]$ is equal to the best linear predictor with the k^2 projections given by the theorem. Section 4.1 contains two examples, showing that for some models, Theorem 3 is sufficient, while for other models, it is necessary to apply Theorem 4.

4.1 Examples: Hidden State Prediction Models

We consider two concrete conditional independence probability models where CCA can be applied. The general graphical model representation is given in Figure 1, which implies that $P(X^{(1)}, X^{(2)}|H) = P(X^{(1)}|H)P(X^{(2)}|H)$. Let us also assume that Y is a contained in H (e.g. say Y is the first coordinate H_1). The first model is a two view Gaussian model, similar to that of Bach and Jordan [2005], and the second is a discrete model, similar to Ando and Zhang [2007]. The optimal predictor of Y (using both views) is linear in the first model, and thus the CCA reduction in Theorem 3 is sufficient. In the second model, the optimal predictor of Y is linear in the tensor product of $X^{(1)}$ and $X^{(2)}$, and thus Theorem 4 is needed.

4.1.1 Two view Gaussian model

We consider the following model, similar to Bach and Jordan [2005], but with a more general Gaussian prior on H:

$$\begin{aligned} P(X^{(\ell)}|H) =& N(W_\ell^\top H, \Sigma_\ell) \qquad (\ell \in \{1, 2\}), \\ P(H) =& N(\mu_0, \Sigma_0), \end{aligned}$$

where μ_0, W_ℓ, and Σ_ℓ are unknowns.

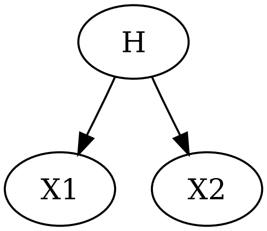

Figure 1: Two View Conditional Independence Model

Since $\mathbb{E}[X^{(\ell)}|H] = W_\ell^\top H$ ($\ell \in \{1,2\}$), the result of Section 4 can be applied. In this model, we have

$$P(H|X^{(1)}, X^{(2)}) \propto \exp\left[-\frac{1}{2}\sum_{\ell=1}^{2}(X^{(\ell)} - W_\ell^\top H)^\top \Sigma_\ell^{-1}(X^{(\ell)} - W_\ell^\top H) - \frac{1}{2}(H-\mu_0)^\top \Sigma_0^{-1}(H-\mu_0)\right],$$

which is Gaussian in H. Moreover, the optimal prediction of H based on $(X^{(1)}, X^{(2)})$ is the conditional posterior mean $\mathbb{E}[H|X^{(1)}, X^{(2)}]$, which is clearly linear in $(X^{(1)}, X^{(2)})$. Therefore Theorem 3 implies that the Bayes optimal prediction rule is a linear predictor with $2k$ dimensional CCA projections of $X^{(1)}$ and $X^{(2)}$. Importantly, note that we do not have to estimate the model parameters W_ℓ, Σ_ℓ, and μ_0, which could rather high dimensional quantities.

4.1.2 Two view discrete probability model

We consider the following model, which is a simplified case of Ando and Zhang [2007]. Each view $X^{(\ell)}$ represents a discrete observation in a finite set Ω_ℓ, where $|\Omega_\ell| = d_\ell$ (where $\ell \in \{1,2\}$). We may encode each view as a d_ℓ dimensional 0-1 valued indicator vector: $X^{(\ell)} \in \{0,1\}^{d_\ell}$, where only one component has value 1 (which indicates the index of the value in Ω_ℓ being observed), and the others have values 0. Similarly, assume the hidden state variable H is discrete and takes on one of k values, and we represent this by a length k binary vector (with the a-th entry being 1 iff $H = a$). Each hidden state induces a probability distribution over Ω_ℓ for view ℓ. That is, the conditional probability model is given by

$$P([X^{(\ell)}]_i = 1|H) = [W_\ell^\top H]_i \qquad (\ell \in \{1,2\}).$$

i.e. each row a of W_ℓ is the probability vector for $X^{(\ell)}$ conditioned on the underlying discrete hidden state being a. Hence, $\mathbb{E}[X^{(\ell)}|H] = W_\ell^\top H$, so H is a linear hidden state. Moreover, since the two views are discrete, the vector $(X^{(1)}, X^{(2)})$ is uniquely identified with $X^{(1)} \circ X^{(2)} \in R^{d_1 \times d_2}$ that contains only one nonzero component, and any arbitrary function of $(X^{(1)}, X^{(2)})$ is trivially a linear function of $X^{(1)} \circ X^{(2)}$. This means that Theorem 4 can be applied to reduce the overall dimension to k^2 and that the Bayes optimal predictor is linear in this reduced k^2 dimensional space. Moreover, in this case, it can be shown that the reduced dimensions are given by the tensor products of the CCA basis for the two views.

5 Dimensionality Reduction Under Redundancy

In this Section, we assume that Y is a scalar. We also use the projection Π_λ, which projects to the subspace which has correlation at least λ (recall the definition of Π_λ from Section 3). The follow theorem shows that using $\Pi_\lambda X^{(1)}$ instead of $X^{(1)}$ for linear prediction does *not* significantly degrade performance.

Theorem 5. *Suppose that Assumption 2 holds (recall, ϵ is defined in this Assumption) and that $Y \in \mathbb{R}$. For all $0 \leq \lambda \leq 1$, we have that:*

$$R^2_{\Pi_\lambda X^{(1)},Y} \geq R^2_{X^{(1)},Y} - \frac{4\epsilon}{1-\lambda}$$
$$R^2_{\Pi_\lambda X^{(2)},Y} \geq R^2_{X^{(2)},Y} - \frac{4\epsilon}{1-\lambda}$$

Note that this implies that these R^2's are also close to $R^2_{X,Y}$.

Clearly, if we chose $\lambda = \frac{1}{2}$, then our loss in error (compared to the best linear prediction) is at most 8ϵ. However, we now only have to deal with estimation in the subspace spanned by $\{U_i : \lambda_i \geq \frac{1}{2}\}_i$, a potentially much lower dimensional space. In fact, the following corollary bounds the dimension of this space in terms of the spectrum.

Corollary 6. *Assume that we choose $\lambda = \frac{1}{2}$. Let d be the dimension of the space that Π_λ projects to, i.e. d is the number of i such that $\lambda_i \geq \frac{1}{2}$. For all $\alpha > 0$, we have:*

$$d \leq 2^\alpha \sum_i \lambda_i^\alpha$$

In particular, this implies that:

$$d \leq 2 \sum_i \lambda_i \text{ and } d \leq 4 \sum_i \lambda_i^2$$

Note that unlike the previous setting, this spectrum need not ever have a finite number of nonzero entries, so we may require a larger power of α to make the sum finite.

Proof. Using that $\lambda_i \geq \frac{1}{2}$, we have:

$$d = \sum_{i=1}^{d} 1 = \sum_{i=1}^{d} \frac{\lambda_i^\alpha}{\lambda_i^\alpha} \leq 2^\alpha \sum_{i=1}^{d} \lambda_i^\alpha \leq 2^\alpha \sum_{i=1}^{\infty} \lambda_i^\alpha$$

where the second to last step follows from the fact that $\lambda_i \leq \lambda$ by definition of d. □

6 Proofs

Let us denote:

$$\begin{aligned}
\Sigma_{11} &= \mathbb{E}[X^{(1)}(X^{(1)})^\top] \\
\Sigma_{22} &= \mathbb{E}[X^{(2)}(X^{(2)})^\top] \\
\Sigma_{12} &= \mathbb{E}[X^{(1)}(X^{(2)})^\top] \\
\Sigma_{HH} &= \mathbb{E}[HH^\top] \\
\Sigma_{1H} &= \mathbb{E}[X^{(1)}H^\top] \\
\Sigma_{2H} &= \mathbb{E}[X^{(2)}H^\top] \\
\Sigma_{1Y} &= \mathbb{E}[X^{(1)}Y^\top] \\
\Sigma_{2Y} &= \mathbb{E}[X^{(2)}Y^\top]
\end{aligned}$$

and $\Sigma_{12}^\top = \Sigma_{21}$, $\Sigma_{H1} = \Sigma_{1H}^\top$, etc.

Without loss of generality, we assume that we have the following isotropic conditions:

$$\Sigma_{11} = \text{Identity}, \ \Sigma_{22} = \text{Identity}, \ \Sigma_{HH} = \text{Identity}, \ \text{var}(Y) = 1$$

This is without loss of generality as our algorithm does not make use of any particular coordinate system (the algorithm is only concerned with the subspaces themselves). This choice of coordinate system eases the notational burden in our proofs. Furthermore, note that we can still have $H_1 = Y$ in this coordinate system.

Under these conditions, CCA corresponds to an SVD of Σ_{12}. Let the SVD decomposition of Σ_{12} be:

$$\Sigma_{12} = UDV^\top$$

where U and V are orthogonal and D is diagonal. Let

$$D = \text{diag}(\lambda_1, \lambda_2, \ldots)$$

Without loss of generality, assume that the SVD is ordered such that:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots .$$

Here, the column vectors of U and V form the CCA basis, and note that for a column U_i of U and V_j of V

$$\mathbb{E}[(U_i \cdot X^{(1)})(V_j \cdot X^{(2)})] = \begin{cases} \lambda_i & \text{if } i = j \\ 0 & \text{else} \end{cases} \tag{2}$$

which implies that

$$0 \leq \lambda_i \leq 1$$

since we are working in the coordinate system where $X^{(1)}$ and $X^{(2)}$ are isotropic.

6.1 Proof of Theorem 3

Throughout this subsection, we let $\beta_1 \cdot X^{(1)}$ be the best linear prediction of H with $X^{(1)}$ (that which minimizes the square loss), let $\beta_2 \cdot X^{(2)}$ be the best linear prediction of H with $X^{(2)}$, and let $\beta \cdot (X^{(1)}, X^{(2)})$ be the best linear prediction of H with both $(X^{(1)}, X^{(2)})$.

With the aforementioned isotropic conditions:

$$\beta_1 = \Sigma_{1H},\ \beta_2 = \Sigma_{1H},\ \beta = (\mathbb{E}[XX^\top])^{-1}\mathbb{E}[XH^\top] \tag{3}$$

which follows directly from the least squares solution. Note that $\mathbb{E}[XX^\top]$ is not diagonal.

Our proof consists of showing that the best linear prediction of H with X is equal to the best linear prediction of H with $\Pi_{\text{CCA}}X$. This implies that the best linear prediction of Y with X is equal to the best linear prediction of Y with $\Pi_{\text{CCA}}X$, by the following argument. Since $\mathbb{E}[Y|H]$ is linear in H (by assumption), we can do a linear transformation of H such that $\mathbb{E}[Y|H] = H_1$ (where H_1 is the first coordinate of H). By Assumption 1, it is follows that for all $\beta \in \mathbb{R}^{\dim(X)}$,

$$\mathbb{E}(Y - \beta \cdot X)^2 = \mathbb{E}(Y - H_1)^2 + \mathbb{E}(H_1 - \beta \cdot X)^2 .$$

Hence, our proof need only be concerned with the linear prediction of H.

The following lemma shows the imposed structure on the covariance matrix Σ_{12} for any linear hidden state.

Lemma 7. *If H is a linear hidden state, then we have that:*

$$\Sigma_{12} = \Sigma_{1H}\Sigma_{H2}$$

which implies that the rank of Σ_{12} is at most k.

Proof. By the linear mean assumption we have that:

$$\mathbb{E}[X^{(1)}|H] = \Sigma_{1H}H$$
$$\mathbb{E}[X^{(2)}|H] = \Sigma_{2H}H$$

which follows from the fact that $\Sigma_{1H}H$ is the least squares prediction of $X^{(1)}$ with H (and this least squares prediction is the expectation, by assumption).

Recall, we are working with H in an isotropic coordinate system. Hence,

$$\begin{aligned}
\Sigma_{12} &= \mathbb{E}[X^{(1)}(X^{(2)})^\top] \\
&= \mathbb{E}_H[\mathbb{E}[X^{(1)}(X^{(2)})^\top|H]] \\
&= \mathbb{E}_H[\mathbb{E}[X^{(1)}|H]\mathbb{E}[(X^{(2)})^\top|H]] \\
&= \mathbb{E}_H[\Sigma_{1H}HH^\top\Sigma_{H2}] \\
&= \Sigma_{1H}\Sigma_{H2}
\end{aligned}$$

which completes the proof. □

Now we are ready to complete the proof of Theorem 3.

Proof. Now Assumption 1 implies that both Σ_{1H} and Σ_{2H} are rank k. This implies that Σ_{12} is also rank k by the previous lemma. Hence, we have the equality

$$\Sigma_{1H} = \Sigma_{12}(\Sigma_{H2})^{-1}$$

(where the inverse is the pseudo-inverse). Now, the optimal linear predictor $\beta_1 = \Sigma_{1H}$, so we have

$$\beta_1 = \Sigma_{12}(\Sigma_{H2})^{-1} .$$

Hence,

$$\begin{aligned}
\beta_1 \cdot X^{(1)} &= (\Sigma_{2H})^{-1}\Sigma_{21}X^{(1)} \\
&= (\Sigma_{2H})^{-1}VDU^\top X^{(1)} \\
&= (\Sigma_{2H})^{-1}VDU^\top \Pi_{\text{CCA}}X^{(1)} \\
&= \beta_1\Pi_{\text{CCA}}X^{(1)} .
\end{aligned}$$

The second to last step follows due to that $DU^\top X^{(1)} = DU^\top\Pi_{\text{CCA}}X^{(1)}$ (since $\lambda_i = 0$ for all directions in which Π_{CCA} does not project to). This completes the proof of the first claim, and the proof of the second claim is analogous.

Now we prove the third claim. Let $\tilde{\beta}$ be the weights for the best linear prediction of H with $\Pi_{\text{CCA}}X := (\Pi_{\text{CCA}}X^{(1)}, \Pi_{\text{CCA}}X^{(2)})$. The optimality (derivative) conditions on $\tilde{\beta}$ imply that:

$$\mathbb{E}[(H - \tilde{\beta}^\top\Pi_{\text{CCA}} \cdot X)(\Pi_{\text{CCA}}X)^\top] = 0 \tag{4}$$

If we show that:

$$\mathbb{E}[(H - \tilde{\beta}^\top\Pi_{\text{CCA}} \cdot X)X^\top] = 0$$

then this proves the result (as the derivative conditions for $\tilde{\beta}^\top$ being optimal are satisfied). To prove the above, it is sufficient to show that, for all vectors α

$$\mathbb{E}[(H - \tilde{\beta}^\top\Pi_{\text{CCA}} \cdot X)(\alpha \cdot X)] = 0 \tag{5}$$

Let us decompose α as $\alpha = (u + u_\perp, v + v_\perp)$, where u is in $\mathcal{U} = \text{span}(\{U_i : \lambda_i > 0\}_i)$ and $u_\perp$ is in $\mathcal{U}_\perp = \text{span}(\{U_i : \lambda_i = 0\}_i)$. Clearly u and $u_\perp$ are orthogonal. Similarly, define v and $v_\perp$. Since

$$\alpha \cdot X = u \cdot X^{(1)} + u_\perp \cdot X^{(1)} + v \cdot X^{(2)} + v_\perp \cdot X^{(2)}$$

To prove Equation 5, it is sufficient to show that:

$$\mathbb{E}[(H - \tilde{\beta}^\top\Pi_{\text{CCA}} \cdot X)(u \cdot X^{(1)})] = 0 \tag{6}$$

$$\mathbb{E}[(H - \tilde{\beta}^\top\Pi_{\text{CCA}} \cdot X)(v \cdot X^{(2)})] = 0 \tag{7}$$

and

$$\mathbb{E}[(H - \tilde{\beta}^\top \Pi_{\text{CCA}} \cdot X)(u_\perp \cdot X^{(1)})] = 0 \tag{8}$$
$$\mathbb{E}[(H - \tilde{\beta}^\top \Pi_{\text{CCA}} \cdot X)(v \perp \cdot X^{(2)})] = 0 \tag{9}$$

To prove Equation 6, simply note that $u \cdot X^{(1)} = u \cdot \Pi_{\text{CCA}} X^{(1)}$ by construction of u, so the result follows from Equation 4. Equation 7 is proven identically.

Now we prove Equation 8. First note that:

$$\mathbb{E}[\Pi_{\text{CCA}} X^{(1)}(u_\perp \cdot X^{(1)})] = 0$$

by our isotropic assumption and since $u_\perp$ is orthogonal to $\mathcal{U}$. Also,

$$\mathbb{E}[\Pi_{\text{CCA}} X^{(2)}(u_\perp \cdot X^{(1)})] = 0$$

from Equation 2 and by construction of $u_\perp$. These two imply that:

$$\mathbb{E}[\Pi_{\text{CCA}} X(u_\perp \cdot X^{(1)})] = 0$$

We also have that:

$$\begin{aligned} \mathbb{E}[H(u_\perp \cdot X^{(1)})] &= \mathbb{E}[H(X^{(1)})^\top] u_\perp \\ &= \Sigma_{H1} u_\perp \\ &= (\Sigma_{2H})^{-1} \Sigma_{21} u_\perp \\ &= 0 \end{aligned}$$

where we have used the full rank condition on Σ_{12} in the second to last step. An identical arguments proves Equation 9. This completes the proof. □

6.2 Proof of Theorem 4

The best linear predictor of Y with $Z = X^{(1)} \circ X^{(2)}$ is given by $\beta_*^T Z$, where

$$\beta_* = \arg\min_\beta \mathbb{E}_{Z,Y}(\beta_*^\top Z - Y)^2.$$

That is,

$$\beta_*^T Z = Z^T(\mathbb{E}_Z ZZ^T)^{-1} \mathbb{E}_{Z,Y}(ZY). \tag{10}$$

Now, for each index i, j and $Z_{i,j} = X_i^{(1)} X_j^{(2)}$, there exists $\alpha_i^{(1)} = [\alpha_{i,1}^{(1)}, \ldots, \alpha_{i,k}^{(1)}]$ and $\alpha_j^{(2)} = [\alpha_{j,1}^{(2)}, \ldots, \alpha_{j,k}^{(2)}]$ such that

$$\mathbb{E}[X_i^{(1)} | H] = H^\top \alpha_i^{(1)}, \quad \mathbb{E}[X_j^{(2)} | H] = H^\top \alpha_j^{(2)}$$

by assumption. Therefore, taking expectations over Z and Y,

$$\begin{aligned} \mathbb{E}[Y Z_{i,j}] =& \mathbb{E}_H \; \mathbb{E}[Y X_i^{(1)} X_j^{(2)} | H] \\ =& \mathbb{E}_H \; [\mathbb{E}[Y|H] \; \mathbb{E}[X_i^{(1)}|H] \; \mathbb{E}[X_j^{(2)}|H]] \\ =& \mathbb{E}_H \; [\mathbb{E}[Y|H] \; (H^\top \alpha_i^{(1)})(H^\top \alpha_j^{(2)})] \\ =& {\alpha_i^{(1)}}^\top Q \alpha_j^{(2)} = \sum_{a=1}^k \sum_{b=1}^k Q_{a,b} \alpha_{i,a}^{(1)} \alpha_{j,b}^{(2)}, \end{aligned}$$

where $Q = \mathbb{E}_H[YHH^\top]$. Let $\alpha^{(1)}_{\cdot,a} = [\alpha^{(1)}_{i,a}]_i$ and $\alpha^{(2)}_{\cdot,b} = [\alpha^{(2)}_{j,b}]_j$, then

$$\mathbb{E}_{Z,Y}[ZY] = \sum_{a=1}^{k} \sum_{b=1}^{k} Q_{a,b} \alpha^{(1)}_{\cdot,a} \circ \alpha^{(2)}_{\cdot,b}.$$

Since we are working with certain isotropic coordinates (see the beginning of Section 6), each $\alpha^{(1)}_{\cdot,a}$ is a linear combination of the CCA basis U_i ($i = 1, \ldots, k$) and each $\alpha^{(2)}_{\cdot,b}$ is a linear combination of the CCA basis V_j ($j = 1, \ldots, k$). Therefore we can find $Q'_{i,j}$ such that

$$\mathbb{E}_{Z,Y}[ZY] = \sum_{i=1}^{k} \sum_{j=1}^{k} Q'_{i,j} U_i \circ V_j.$$

From (10), we obtain that

$$\beta_*^T Z = \sum_{i=1}^{k} \sum_{j=1}^{k} Q'_{i,j} Z^T (\mathbb{E}_Z Z Z^T)^{-1} U_i \circ V_j.$$

By changing to an arbitrary basis in $X^{(1)}$ and in $X^{(2)}$, we obtain the desired formula.

6.3 Proof of Theorem 5

Here, β_1, β_2 and β are the best linear predictors of Y with $X^{(1)}$, $X^{(2)}$, and X, respectively. Also, in our isotropic coordinates, var$(Y) = 1$, so we will prove the claim in terms of the loss, which implies that statements about R^2.

The following lemma is useful to prove Theorem 5.

Lemma 8. *Assumption 2 implies that*

$$\sum_i (1 - \lambda_i)(\beta_\nu \cdot U_i)^2 \le 4\epsilon$$

for $\nu \in \{1, 2\}$.

With this lemma, the proof of our Theorem follows.

Proof of Theorem 5. Let β_{CCA} be the weights of the best linear predictor using only $\Pi_\lambda X^{(1)}$. Since $X^{(1)}$ is isotropic, it follows that $\beta_{\text{CCA}} \cdot U_i = \beta_1 \cdot U_i$ for $\lambda_i \ge \lambda$, as these directions are included in Π_λ. First, note that since the norm of a vector is unaltered by a rotation, we have:

$$\text{loss}(\beta_{\text{CCA}}) - \text{loss}(\beta_1) = ||\beta_{\text{CCA}} - \beta_1||_2^2 = \sum_i ((\beta_{\text{CCA}} - \beta_1) \cdot U_i)^2$$

since U is rotation matrix. Hence, we have that:

$$\begin{aligned}
\text{loss}(\beta_{\text{CCA}}) - \text{loss}(\beta_1) &= \sum_i ((\beta_{\text{CCA}} - \beta_1) \cdot U_i)^2 \\
&= \sum_{i:\lambda_i < \lambda} ((\beta_{\text{CCA}} - \beta_1) \cdot U_i)^2 \\
&= \sum_{i:\lambda_i < \lambda} \frac{1 - \lambda_i}{1 - \lambda_i} ((\beta_{\text{CCA}} - \beta_1) \cdot U_i)^2 \\
&\le \frac{1}{1 - \lambda} \sum_{i:\lambda_i < \lambda} (1 - \lambda_i)((\beta_{\text{CCA}} - \beta_1) \cdot U_i)^2 \\
&\le \frac{4\epsilon}{1 - \lambda}
\end{aligned}$$

where the first line follows from algebraic manipulations for the square loss. □

The following lemma is useful for proving Lemma 8:

Lemma 9. *Assumption 2 implies that*

$$\mathbb{E}[(\beta_1 \cdot X^{(1)} - \beta_2 \cdot X^{(2)})^2] \leq 4\epsilon\,.$$

Proof. Let β be the best linear weights using $X = (X^{(1)}, X^{(2)})$. By Assumption 2

$$\begin{aligned}
\epsilon &\geq \mathbb{E}(\beta_1 \cdot X^{(1)} - Y)^2 - \mathbb{E}(\beta \cdot X - Y)^2 \\
&= \mathbb{E}(\beta_1 \cdot X^{(1)} - \beta \cdot X + \beta \cdot X - Y)^2 - \mathbb{E}(\beta \cdot X - Y)^2 \\
&= \mathbb{E}(\beta_1 \cdot X^{(1)} - \beta \cdot X)^2 - 2\mathbb{E}[(\beta_1 \cdot X^{(1)} - \beta \cdot X)(\beta \cdot X - Y)]
\end{aligned}$$

Now the first derivative conditions for the optimal linear predictor β implies that:

$$\mathbb{E}[X(\beta \cdot X - Y)] = 0$$

which implies that:

$$\mathbb{E}[\beta \cdot X(\beta \cdot X - Y)] = 0$$
$$\mathbb{E}[\beta_1 \cdot X^{(1)}(\beta \cdot X - Y)] = 0$$

Hence,

$$\mathbb{E}[(\beta_1 \cdot X^{(1)} - \beta \cdot X)(\beta \cdot X - Y)] = 0$$

A similar argument proves the identical statement for β_2.

We have shown that:

$$\mathbb{E}(\beta_1 \cdot X^{(1)} - \beta \cdot X)^2 \leq \epsilon$$
$$\mathbb{E}(\beta_2 \cdot X^{(2)} - \beta \cdot X)^2 \leq \epsilon$$

The triangle inequality states that:

$$\begin{aligned}
&\mathbb{E}(\beta_1 \cdot X^{(1)} - \beta_2 \cdot X^{(2)})^2 \\
\leq\ &\left(\sqrt{\mathbb{E}(\beta_1 \cdot X^{(1)} - \beta \cdot X)^2} + \sqrt{\mathbb{E}(\beta_2 \cdot X^{(2)} - \beta \cdot X)^2}\right)^2 \\
\leq\ &(2\sqrt{\epsilon})^2
\end{aligned}$$

which completes the proof. □

Now we prove Lemma 8.

Proof of Lemma 8. Let us write $[\beta_1]_i = \beta_1 \cdot U_i$ and $[\beta_2]_i = \beta_2 \cdot V_i$. From Lemma 9, we have:

$$\begin{aligned}
4\epsilon &\geq \mathbb{E}\left[(\beta_1 \cdot X^{(1)} - \beta_2 \cdot X^{(2)})^2\right] \\
&= \sum_i \big(([\beta_1]_i)^2 + ([\beta_2]_i)^2 - 2\lambda_i[\beta_1]_i[\beta_2]_i\big) \\
&= \sum_i \big((1-\lambda_i)([\beta_1]_i)^2 + (1-\lambda_i)([\beta_2]_i)^2 + \lambda_i(([\beta_1]_i)^2 + ([\beta_2]_i)^2 - 2[\beta_1]_i[\beta_2]_i)\big) \\
&= \sum_i \big((1-\lambda_i)([\beta_1]_i)^2 + (1-\lambda_i)([\beta_2]_i)^2 + \lambda_i([\beta_1]_i - [\beta_2]_i)^2\big) \\
&\geq \sum_i \big((1-\lambda_i)([\beta_1]_i)^2 + (1-\lambda_i)([\beta_2]_i)^2\big) \\
&\geq \sum_i (1-\lambda_i)([\beta_\nu]_i)^2
\end{aligned}$$

where the last step holds for either $\nu = 1$ or $\nu = 2$.

□

References

Rie Kubota Ando and Tong Zhang. Two-view feature generation model for semi-supervised learning. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 25–32, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: http://doi.acm.org/10.1145/1273496.1273500.

Francis R. Bach and Michael I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, U.C. Berkeley, 2005.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, pages 92–100, 1998.

David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664, 2004. ISSN 0899-7667.

Sham M. Kakade and Dean P. Foster. Multi-view regression via canonical correlation analysis. In Nader H. Bshouty and Claudio Gentile, editors, *COLT*, volume 4539 of *Lecture Notes in Computer Science*, pages 82–96. Springer, 2007.