An overview of the generalized eigenvector formulation of multi-view CCA including regularization and kernels

Iain Carmichael

October 15, 2020

Contents

1 Introduction

A multi-block (or multi-view) data set $X^{(1)}, \ldots, X^{(B)}$, $X^{(b)} \in \mathbb{R}^{n \times d^{(b)}}$, is a set of B data matrices on a fixed set of n observations that each have different sets of variables.

Canonical correlation analysis (CCA) is a classical dimensionality reduction algorithm for $B = 2$ data blocks (Hotelling, 1992). There are many extensions of CCA to more than two views including (Kettenring, 1971; Bach and Jordan, 2002; Tenenhaus and Tenenhaus, 2011; Klami et al., 2013; Zhao et al., 2014; Gaynanova and Li, 2019). We present a simple multi-view CCA framework built on the classical SUMCORR-AVGVAR MCCA (see Chapter 10 of Asendorf (2015)) that can handle regularization and kernels.

Section 2 defines a multi-view CCA problem through a *generalized eigenvector problem*. For the purpose of exposition we refer to this method as MCCA. Notation and naming conventions are not standardized in the literature and Section 2.1 presents our conventions for the various MCCA quantities of interest. Section 3 presents a few special cases of MCCA. Section 4 presents a kernel formulation of MCCA. Finally, section A gives a brief overview of generalized eigenvector problems.

For an overview of related regularized CCA and multi-CCA methods see (Kettenring, 1971; Bach and Jordan, 2002; Nielsen, 2002; Hardoon et al., 2004; De Bie et al., 2005; Bach and Jordan, 2005; Fukumizu et al., 2007; Witten et al., 2009; Asendorf, 2015; Bilenko and Gallant, 2016).

2 MCCA generalized eigenvector formulation

Suppose we have a set of multi-view data matrices, $X^{(1)}, \ldots, X^{(B)}$, as above. We assume the data matrices have been first mean centered. Let

$$\widehat{\Sigma}^{(ab)} = {X^{(a)}}^T X^{(b)} \in \mathbb{R}^{d^{(a)} \times d^{(b)}} \tag{1}$$

be the cross-covariance matrix for data blocks a and b for each $a, b \in [B] := \{1, \ldots, B\}$ (the $\frac{1}{n}$ factor is unimportant) . Also let

$$\widehat{\Sigma}_\gamma^{(bb)} = (1-\gamma){X^{(b)}}^T X^{(b)} + \gamma I_{d^{(b)}} \in \mathbb{R}^{d^{(b)} \times d^{(b)}} \tag{2}$$

be a (possibly regularized) estimate of the covariance matrix for the bth data block where $\gamma \in [0,1]$ is the regularization parameter. Each block can have its own regularization value, but for the sake of exposition we use a single value for each block.

The rank K MCCA problem is the following optimization problem for $1 \leq K \leq D$ where $D := \sum_{b=1}^{D} d^{(b)}$.

$$\begin{aligned} \underset{W \in \mathbb{R}^{D \times K}}{\text{maximize}} \quad & \text{Tr}\left(W^T \begin{bmatrix} \widehat{\Sigma}_\gamma^{(11)} & \widehat{\Sigma}^{(12)} & \ldots & \widehat{\Sigma}^{(1B)} \\ \widehat{\Sigma}^{(21)} & \widehat{\Sigma}_\gamma^{(22)} & \ldots & \widehat{\Sigma}^{(2B)} \\ \vdots & & & \vdots \\ \widehat{\Sigma}^{(B1)} & & \ldots & \widehat{\Sigma}_\gamma^{(BB)} \end{bmatrix} W \right) \\ \text{subject to} \quad & W^T \begin{bmatrix} \widehat{\Sigma}_\gamma^{(bb)} & & \\ & \ddots & \\ & & \widehat{\Sigma}_\gamma^{(BB)} \end{bmatrix} W = I_K \end{aligned} \tag{3}$$

Let $\widehat{\Sigma}_\gamma \in \mathbb{R}^{D \times D}$ be the matrix in the objective function of (3) and let $\widehat{\Sigma}_{\text{block-diag},\gamma} \in \mathbb{R}^{D \times D}$ be the block diagonal matrix in the constraints of (3). Problem (3) is easily computed via a generalized eigenvector problem with these two matrices.

Proposition 2.1. *Assume $\widehat{\Sigma}_{block\text{-}diag,\gamma}$ is full rank. A global solution to* (3) *is given by any matrix $W \in \mathbb{R}^{D \times K}$ whose columns are K leading generalized eigenvectors of the matrix pencil $\left(\widehat{\Sigma}_\gamma, \widehat{\Sigma}_{block\text{-}diag,\gamma}\right)$. The optimal value of this problem is given by the sum of the largest K generalized eigenvalues.*

This is an immediate consequence of Proposition A.1.

With no regularization, $\gamma = 0$, (3) is equivalent to the classical SUMCORR-AVGVAR multi-CCA. Regularizing a covariance matrix by shrinking it towards the identify is well known to improve estimation for high-dimensional data (Ledoit and Wolf, 2004). In the context of CCA with $B = 2$, $\gamma = 0$, $\max(d^{(1)}, d^{(2)}) > n$ the CCA estimates are known to break down. Regularized CCA, however, still works in high-dimensions (Hardoon et al., 2004; Bach and Jordan, 2002; Fukumizu et al., 2007).

2.1 Quantities of interest

Depending on the application, there are a variety of quantities of interest for MCCA.

- The *concatenated loadings* matrix, $W \in \mathbb{R}^{D \times K}$, is a solution to (3). The columns of W are the leading K generalized eigenvectors of $\left(\widehat{\Sigma}_\gamma, \widehat{\Sigma}_{\text{block-diag},\gamma}\right)$.
- The *block loadings* are $W^{(b)} \in \mathbb{R}^{d^{(b)} \times K}$ are then rows of W corresponding to the bth block for each $b \in [B]$ (i.e. W is the vertical concatenation of the block loadings).
- The *block scores* are
$$S^{(b)} := X^{(b)} W^{(b)} \in \mathbb{R}^{n \times K},$$
for each $b \in [B]$.
- The (unnormalized) *common scores* matrix is the sum of the block scores
$$F_{\text{un}} := XW = \sum_{b=1}^{B} S^{(b)} \in \mathbb{R}^{n \times K}.$$
We refer to the *common normalized scores* matrix $F \in \mathbb{R}^{n \times K}$ as the result of normalized the columns of F_{un}.

- The *MCCA eigenvalues* $\lambda_1 \geq \cdots \geq \lambda_D$ are the generalized eigenvalues of $\left(\widehat{\Sigma}_\gamma, \widehat{\Sigma}_{\text{block-diag},\gamma}\right)$.

Remark 2.1. *The concatenated loadings are orthonormal in the inner product induced by* $\widehat{\Sigma}_{block\text{-}diag,\gamma}$ *i.e.*

$$W^T \widehat{\Sigma}_{block\text{-}diag,\gamma} W = I_K.$$

If $\gamma = 0$ *the common normalized scores are orthonormal i.e* $F^T F = I_K$*. If* $\gamma = 1$ *the concatenated loadings are orthonormal i.e.* $W^T W = I_K$*.*

2.2 Informative-MCCA: PCA first, then MCCA

Dimensionality reduction is an alternative option to regularization for using CCA on high-dimensional data (Asendorf and Nadakuditi, 2015; Asendorf, 2015; Song et al., 2016; Asendorf and Nadakuditi, 2017). For example, suppose we first compute a low rank PCA of each data matrix then run MCCA on the reduced data (the PCA scores). Following the convention of (Asendorf and Nadakuditi, 2015; Asendorf, 2015) we call this PCA-MCCA procedure *informative-MCCA* or i-MCCA.

For i-MCCA we need to adjust the definitions of some of the quantities of interest. Suppose we compute a rank $r^{(b)}$ PCA for the bth data matrix. Let $U^{(b)}, \sigma^{(b)}, V^{(b)}$ be the resulting PCA scores, singular values and loadings. Next suppose we compute a rank K i-MCCA on the (unnormalized) PCA scores $\{U^{(b)}\text{diag}(\sigma^{(b)})\}_{b=1}^B$. Let $\widetilde{W}^{(b)} \in \mathbb{R}^{r^{(b)} \times K}$ be the resulting bth block loadings for the PCA reduced data (similarly for the block scores $\widetilde{S}^{(b)}$ and common normalized scores $\widetilde{F}$).

- We refer to the block scores of the reduced data as i-MCCA block scores. Similarly for the common normalized scores and eigenvalues.
- We refer to the matrix $W^{(b)} := V^{(b)}\widetilde{W}^{(b)}$ as the i-MCCA block loadings. This matrix maps the original data to the block scores of the reduced data i.e.

$$X^{(b)} W^{(b)} = \widetilde{S}^{(b)}.$$

Typically i-MCCA focuses on the case of no regularization (Asendorf and Nadakuditi, 2015). However, i-MCCA can still be apply in the regularized case $\gamma > 0$ and with kernels (e.g. apply kernel-PCA first then compute MCCA).

3 Special cases

3.1 SUMCORR-AVGVAR CCA

When $\gamma = 0$, Problem (3) is equivalent to the SUMCORR-AVGVAR multi-CCA extension (e.g. see Chapter 10 of Asendorf (2015)). We can check the block scores and MCCA eigenvalues depend only on the subspaces spanned by the columns of the data blocks.

MCCA has a nice geometric interpretation via the subspace *flag mean*, which is a notion of the "most central" subspace of a collection of subspaces in $\mathbb{R}^n$ (Draper et al., 2014). In the case of SUMCORR-AVGVAR, the common normalized scores, F, are the flag mean of the subspaces spanned by the columns of the data blocks.

The name "SUMCORR" at first appears to be a misnomer since the objective function of (3) depends on pairwise covariances between the blocks. However, we can check the first component of (3) is equivalent to the following problem

$$\begin{aligned} \underset{f \in \mathbb{R}^n, w^{(b)} \in \mathbb{R}^{d^{(b)}}}{\text{maximize}} \quad & \sum_{b=1}^B \text{corr}\left(f, X^{(b)} w^{(b)}\right)^2 \\ \text{subject to} \quad & \sum_{b=1}^B ||X^{(b)} w^{(b)}||_2^2 = 1, \\ & ||f||_2 = 1. \end{aligned} \tag{4}$$

In other words, we are maximizing the correlation between the common scores the "most central" vector among the blocks. It can be checked that the solution f is the first component of the common normalized scores, which is in turn the first flag mean component $\{\text{col-span}(X^{(b)})\}_{b=1}^{B}$ (Draper et al., 2014).

The MCCA eigenvalues have the following geometric interpretation,

$$\lambda_k = \sum_{b=1}^{B} \cos^2\left(\text{angle}(S_k^{(b)}, F_k)\right),$$

where $\text{angle}(S_k^{(b)}, F_k)$ is the angle between the bth block scores and the common normalized scores for the kth component.

Note that unlike $B = 2$ view CCA the data blocks are not necessarily "weighted equally." For example, the block scores do not have the same norms in general i.e. $||S_k^{(a)}|| \neq ||S_k^{(b)}||$ and $\text{angle}(S_k^{(a)}, F_k) \neq \text{angle}(S_k^{(b)}, F_k)$ or $a \neq b$. In other words, when $B \geq 3$, SUMCORR-AVGVAR can down-weight or even ignore some of the data blocks.

3.2 Canonical correlation analysis ($B = 2$ blocks)

When $B = 2$ and $\gamma = 0$ (3) is equivalent to the standard canonical correlation analysis. Our definition of the block scores is equivalent (up to a multiplicative factor) to the CCA scores (similarly for the block loadings). In this CCA case the block scores are orthonormal (after multiplying by a constant) i.e.

$$\sqrt{2}{S^{(1)}}^T S^{(1)} = \sqrt{2}{S^{(2)}}^T S^{(2)} = I_K.$$

This is different from then general SUMCORR-AVGVAR case when only the common normalized scores are orthonormal, not the block scores.

The first $\min(d^{(1)}, d^{(2)})$ MCCA eigenvalues are related to the *canonical correlations* $\rho_1 \geq \rho_2, \ldots$ via

$$\lambda_k = 1 + \rho_k^2, \text{ for } 1 \leq k \leq \min(d^{(1)}, d^{(2)}).$$

Standard CCA only defines $K \leq \min(d^{(1)}, d^{(2)})$ components, but we allow up to $d^{(1)} + d^{(2)}$ components. While we allow additional components in our definition, the components $k \geq \min(d^{(1)}, d^{(2)})$ are not interesting. Note that for $B \geq 2$ the MCCA components may be interesting for larger values of k.

3.3 Multi-view PLS

When $\gamma = 1$ and $B = 2$, Problem (3) is equivalent to the classical *partial least squares*-SVD (PLS-SVD) method (De Bie et al., 2005). Here the objective function aims to maximize the cross-covariances of the block scores

$$\text{cov}(X^{(1)}w^{(1)}, X^{(2)}w^{(2)}).$$

and the block loadings are orthonormal (up to a constant) i.e.

$$\sqrt{2}{W^{(1)}}^T W^{(1)} = \sqrt{2}{W^{(2)}}^T W^{(2)} = I_K$$

The case of $\gamma = 1$ and $B \geq 2$ can be seen as a natural multi-view generalization of PLS-SVD. Again the objective function is a sum of cross-covariances. Note only the concatenated loadings are orthonormal ($W^T W = I_K$), not the block loadings in general.

4 Kernel regularized SUMCORR-AVGVAR

Suppose we have a kernel matrix for each block $K^{(1)}, \ldots, K^{(B)} \in \mathbb{R}^{n \times n}$. For example, with the linear kernel $K^{(b)} = X^{(b)}{X^{(b)}}^T$. We assume the kernel matrices have been mean centered. The following quantities are the kernel analogs of the cross-covariance matrix and regularized covariance matrices respectively,

$$K^{(ab)} = K^{(a)}K^{(b)} \in \mathbb{R}^{n \times n} \tag{5}$$

and

$$K_\gamma^{(bb)} = (1-\gamma){K^{(b)}}^2 + \gamma K^{(b)} \in \mathbb{R}^{n\times n}. \tag{6}$$

We can kernelize (3) by solving the following generalized eigenvector problem for the concatenated *dual variables* $A \in \mathbb{R}^{nB\times K}$.

$$\begin{aligned} \underset{A\in\mathbb{R}^{nB\times K}}{\text{maximize}} \quad & \text{Tr}\left(A^T \begin{bmatrix} K_\gamma^{(11)} & \dots & K^{(1B)} \\ & & \\ K^{(B1)} & \dots & K_\gamma^{(BB)} \end{bmatrix} A\right) \\ \text{subject to} \quad & A^T \begin{bmatrix} K_\gamma^{(11)} & & \\ & \ddots & \\ & & K_\gamma^{(BB)} \end{bmatrix} A = I_K \end{aligned} \tag{7}$$

Splitting up the rows of A into the *block dual variables* $A^{(b)} \in \mathbb{R}^{n\times K}$ we see the block scores are given by $S^{(b)} = K^{(b)}A^{(b)}$.

Regularization is important for kernel methods because the data are often embedded in a high-dimensional space. Unfortunately if $K^{(b)}$ is singular then $K_\gamma^{(bb)}$ will also be singular, even if $\gamma > 0$. We can work around this issue by computing the SVD of each $K^{(b)}$, retaining only the non-zero singular vectors and solving an eigen-problem related to (7). The details of this are beyond the scope of this note.

Some papers work around this singularity issue in a different way by replacing $K_\gamma^{(bb)}$ with a another matrix. For example, (Bach and Jordan, 2002) suggests

$$K_{\gamma,\text{B}}^{(bb)} := (1-\gamma)\left(K^{(b)} + \frac{n}{2}\frac{\gamma}{1-\gamma}I_n\right)^2, \tag{8}$$

for $\gamma \in [0,1)$. For $\gamma = 1$ we set $K_{1,\text{B}}^{(bb)} = I_n$. Another option from (Bilenko and Gallant, 2016) is

$$K_{\gamma,\text{C}}^{(bb)} := (1-\gamma){K^{(b)}}^2 + \gamma I_n \tag{9}$$

When $\gamma = 0$ each of these three options reduce to ${K^{(b)}}^2$. For $\gamma > 0$, $K_{\gamma,\text{B}}^{(bb)}$ and $K_{\gamma,\text{C}}^{(bb)}$ are invertible even if $K^{(b)}$ is not. Additionally, when $\gamma = 1$ both of these matrices reduce to the identity.

A Generalized eigenvectors

Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric and assume B is positive definite. A generalized eigenvalue/vector pair (λ, v) of the *matrix pencil* (A, B) satisfies

$$Av = \lambda Bv.$$

where we assume the normalization $v^TBv = 1$ by convention. Generalized eigenvalue/vectors of a pencil are eigenvalue/vectors of a related matrix,

$$\begin{aligned} &(\lambda, v) \text{ is an generalized eigenvalue/vector of } (A, B) \\ \iff &(\lambda, B^{1/2}v) \text{ is an eigenvalue/vector of } B^{-1/2}AB^{-1/2} \end{aligned} \tag{10}$$

Generalized vectors show up as solutions to optimization problems of the following form.

Proposition A.1. *A global solution to*

$$\begin{aligned} \underset{W\in\mathbb{R}^{n\times K}}{\text{maximize}} \quad & Tr(W^TAW) \\ \textit{subject to} \quad & W^TBW = I_K \end{aligned} \tag{11}$$

is given by a matrix $W \in \mathbb{R}^{n\times K}$ whose columns are the leading K generalized eigenvectors of (A, B). The optimal value is the sum of the largest generalized eigenvalues, $\sum_{k=1}^K \lambda_k$. If maximize if replaced with minimize a similar statement can be make about the smallest generalized eigenvectors/values.

This proposition is an extension of the famous Fan's theorem and its proof can be obtained from Proposition 20.A.2.a of (Marshall et al., 1979).

References

Asendorf, N. and Nadakuditi, R. R. (2015). Improving multiset canonical correlation analysis in high dimensional sample deficient settings. In *2015 49th Asilomar Conference on Signals, Systems and Computers*, pages 112–116. IEEE.

Asendorf, N. and Nadakuditi, R. R. (2017). Improved detection of correlated signals in low-rank-plus-noise type data sets using informative canonical correlation analysis (icca). *IEEE Transactions on Information Theory*, 63(6):3451–3467.

Asendorf, N. A. (2015). *Informative data fusion: Beyond canonical correlation analysis*. PhD thesis.

Bach, F. R. and Jordan, M. I. (2002). Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48.

Bach, F. R. and Jordan, M. I. (2005). A probabilistic interpretation of canonical correlation analysis.

Bilenko, N. Y. and Gallant, J. L. (2016). Pyrcca: regularized kernel canonical correlation analysis in python and its applications to neuroimaging. *Frontiers in neuroinformatics*, 10:49.

De Bie, T., Cristianini, N., and Rosipal, R. (2005). Eigenproblems in pattern recognition. In *Handbook of Geometric Computing*, pages 129–167. Springer.

Draper, B., Kirby, M., Marks, J., Marrinan, T., and Peterson, C. (2014). A flag representation for finite collections of subspaces of mixed dimensions. *Linear Algebra and its Applications*, 451:15–32.

Fukumizu, K., Bach, F. R., and Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(Feb):361–383.

Gaynanova, I. and Li, G. (2019). Structural learning and integrative decomposition of multi-view data. *Biometrics*, 75(4):1121–1132.

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.

Hotelling, H. (1992). Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer.

Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451.

Klami, A., Virtanen, S., and Kaski, S. (2013). Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14(Apr):965–1003.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411.

Marshall, A. W., Olkin, I., and Arnold, B. C. (1979). *Inequalities: theory of majorization and its applications*, volume 143. Springer.

Nielsen, A. A. (2002). Multiset canonical correlations analysis and multispectral, truly multitemporal remote sensing data. *IEEE transactions on image processing*, 11(3):293–305.

Song, Y., Schreier, P. J., Ramírez, D., and Hasija, T. (2016). Canonical correlation analysis of high-dimensional data with very small sample support. *Signal Processing*, 128:449–458.

Tenenhaus, A. and Tenenhaus, M. (2011). Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Zhao, S., Gao, C., Mukherjee, S., and Engelhardt, B. E. (2014). Bayesian group latent factor analysis with structured sparse priors. *arXiv preprint arXiv:1411.2698*, 208.