Using sparse mCCA to parse multi-omics data: A primer

Will Patterson

July 2020

1 Introduction to Canonical Correlation Analysis

We first introduce canonical correlation analysis (CCA). This technique, attributed to Hotelling [1], is used to determine the relationship between two sets of variables. For instance, if we are given two sets of observations on the same n observations, $\mathbf{X}_1$ and $\mathbf{X}_2$ (of dimensions $n \times p_1$ and $n \times p_2$ respectively), we can use CCA to find weighted linear composite for each variate that will maximize correlation with each other. The weights $\mathbf{w}_1 \in \mathbb{R}^{p_1}$ and $\mathbf{w}_1 \in \mathbb{R}^{p_2}$ are optimized to maximize the objective function or CCA criterion, which is given:

$$\text{maximize}_{w_1,w_2} \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2 \text{ subject to } \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_1 \mathbf{w}_1 = \mathbf{w}_2^T \mathbf{X}_2^T \mathbf{X}_2 \mathbf{w}_2 = 1 \quad (1)$$

In the above equation, we just also make sure that both $\mathbf{X}_1$ and $\mathbf{X}_2$ are standardized to mean zero and standard deviation of one.

However, if we wish to apply CCA to any high dimensional data (such as genomics data), we are stymied – as the number of features of the data greatly outranks the sample size. In order to apply CCA to these data, we need to instead use a penalized CCA, such as was proposed by Written et al. (2009) [2]. The objective function in this penalized version takes the form:

$$\begin{aligned} &\text{maximize}_{w_1,w_2} \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2 \\ &\text{subject to } ||\mathbf{w}_1||^2 \leq 1, ||\mathbf{w}_2||^2 \leq 1, P_1(\mathbf{w}_1) \leq c_1, P_2(\mathbf{w}_2) \leq c_2 \end{aligned} \quad (2)$$

where P_1 and P_2 are penalty functions (usually lasso, L_1). For the case where P_i is an L_1 penalty, Written and Tibshirani (2009) [3] suggest that for $\mathbf{w}_1$ sparse, c_1 is chosen such that $1 \leq c_1 \leq \sqrt{p_1}$.

2 Introduction to sparse multiple CCA

From above, we have determined how to integrate two data sets in an analysis. Sparse multiple CCA (or sparse mCCA) is an extensible form of sparse CCA, generalizable for any K data sets $\mathbf{X}_1, ..., \mathbf{X}_K$, where data set k contains p_k features. The objective function then takes the form:

$$\text{maximize}_{\mathbf{w}_1,...,\mathbf{w}_K} \sum_{i<j} \mathbf{w}_i^T \mathbf{X}_i^T \mathbf{X}_j \mathbf{w}_j \text{ subject to } \mathbf{w}_k^T \mathbf{X}_k^T \mathbf{X}_k \mathbf{w}_k = 1 \forall k \tag{3}$$

where the vector $\mathbf{w}_K \in \mathbb{R}^{p_k}$. This follows, as when we set $K = 2$ for Equation 3 it simplifies to Equation 1.

As we did above, we can overcome issues with high dimensional data by imposing sparsity constraints on the objective function. We must first ensure that each variable has mean zero and standard deviation one. Satisfying these two criteria, we then know that $\mathbf{X}_k^T \mathbf{X}_k = \mathbf{I}$ for each k[1] and (within each data set) features are independent. Therefore, our objective function becomes:

$$\text{maximize}_{\mathbf{w}_1,...,\mathbf{w}_K} \sum_{i<j} \mathbf{w}_i^T \mathbf{X}_i^T \mathbf{X}_j \mathbf{w}_j \text{ subject to } ||\mathbf{w}_i||^2 \leq 1, P_i(\mathbf{w}_i) \leq c_i, \forall i \tag{4}$$

where P_i is the lasso penalty ensuring sparsity of $\mathbf{w}_i$, as described previously.

3 Extension of sparse mCCA with a binary outcome

Witten and Tibshirani (2009) [3] suggest an extension of sparse mCCA that allows for the incorporation of a two-class outcome. Their method simply treats this $\mathbb{R}^{n\times 1}$ matrix as a third data set. Their objective function takes the form:

$$\begin{aligned} &\text{maximize}_{\mathbf{w}_1,\mathbf{w}_2,\mathbf{w}_3} \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2 + \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{y} \mathbf{w}_3 + \mathbf{w}_2^T \mathbf{X}_2^T \mathbf{y} \mathbf{w}_3 \\ &\text{subject to } ||\mathbf{w}_i||^2 \leq 1, P_i(\mathbf{w}_i) \leq c_i, \forall i \end{aligned} \tag{5}$$

but since we know that $p_y = 1$ and $\mathbf{w}_3 \in \mathbb{R}^{p_y}$, therefore $\mathbf{w}_3 \in \mathbb{R}^1$ and $\mathbf{w}_3 = 1$. This simplifies the objective function:

$$\begin{aligned} &\text{maximize}_{\mathbf{w}_1,\mathbf{w}_2} \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2 + \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{y} + \mathbf{w}_2^T \mathbf{X}_2^T \mathbf{y} \\ &\text{subject to } ||\mathbf{w}_1||^2 \leq 1, ||\mathbf{w}_2||^2 \leq 1, P_1(\mathbf{w}_1) \leq c_1, P_2(\mathbf{w}_2) \leq c_2 \end{aligned} \tag{6}$$

4 Applying sparse mCCA to our case

In our proposed project, we will have three data sets (which I believe are DNAseq, RNAseq, and methylation data) as well as binary outcome data. In

[1] For some random variable x, $\text{Var}[x] = E[x^2] - E[x]^2$. This can then be extended into two dimensions with random vector $\mathbf{X}$, which takes the form $\text{Var}(\mathbf{X}) = E(\mathbf{X}^T\mathbf{X}) - E(\mathbf{X}^T)E(\mathbf{X})$. In our special case, since we established each variable has mean (expectation) zero and standard deviation one, we can simplify such that $\mathbf{X}^T\mathbf{X} = \mathbf{I}$.

order to integrate these four data sets, we will simply further extend Equation 7. This should take the form:

$$\begin{aligned} &\text{maximize}_{\mathbf{w}_1,\mathbf{w}_2,\mathbf{w}_3} \mathbf{w}_1^T\mathbf{X}_1^T\mathbf{X}_2\mathbf{w}_2 + \mathbf{w}_1^T\mathbf{X}_1^T\mathbf{X}_3\mathbf{w}_3 + \mathbf{w}_2^T\mathbf{X}_2^T\mathbf{X}_3\mathbf{w}_3 \\ &\qquad + \mathbf{w}_1^T\mathbf{X}_1^T\mathbf{y} + \mathbf{w}_2^T\mathbf{X}_2^T\mathbf{y} + \mathbf{w}_2^T\mathbf{X}_2^T\mathbf{y} \\ &\text{subject to } ||\mathbf{w}_1||^2 \leq 1, ||\mathbf{w}_2||^2 \leq 1, ||\mathbf{w}_3||^2 \leq 1, \\ &\qquad P_1(\mathbf{w}_1) \leq c_1, P_2(\mathbf{w}_2) \leq c_2, P_3(\mathbf{w}_3) \leq c_3 \end{aligned} \tag{7}$$

References

[1] Harold Hotelling. “Relations Between Two Sets of Variates”. en. In: *Biometrika* 28.3-4 (Dec. 1936). Publisher: Oxford Academic, pp. 321–377. ISSN: 0006-3444. DOI: 10.1093/biomet/28.3-4.321. URL: https://academic.oup.com/biomet/article/28/3-4/321/220073 (visited on 07/20/2020).

[2] Daniela M. Witten, Robert Tibshirani, and Trevor Hastie. “A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis”. eng. In: *Biostatistics (Oxford, England)* 10.3 (July 2009), pp. 515–534. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxp008.

[3] Daniela M Witten and Robert J. Tibshirani. “Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data”. In: *Statistical Applications in Genetics and Molecular Biology* 8.1 (Jan. 2009), pp. 1–27. ISSN: 1544-6115. DOI: 10.2202/1544-6115.1470. URL: https://www.degruyter.com/view/j/sagmb.2009.8.1/sagmb.2009.8.1.1470/sagmb.2009.8.1.1470.xml (visited on 07/20/2020).