AC-PCA adjusts for confounding variation in transcriptome data and recovers the anatomical structure of neocortex

Zhixiang Lin, Can Yang, Ying Zhu, John C. Duchi, Yao Fu, Yong Wang, Bai Jiang, Mahdi Zamanighomi, Xuming Xu, Mingfeng Li, Nenad Sestan, Hongyu Zhao* & Wing Hung Wong*

February 20, 2016

Abstract: Microarray and RNA-sequencing technologies have enabled rapid quantification of the transcriptomes in a large number of samples. Although dimension reduction methods are commonly applied to transcriptome datasets for visualization and interpretation of the sample variations, the results can be hindered by confounding factors, either biological or technical. In this study, we propose a Principal Component Analysis-based approach to Adjust for Confounding variation (AC-PCA). We show that AC-PCA can adjust for variations across individual donors present in a human brain exon array dataset. Our approach is able to recover the anatomical structure of neocortex regions, including the frontal-temporal and dorsal-ventral axes, and reveal temporal dynamics of the interregional variation, mimicking the "hourglass" pattern of spatiotemporal dynamics. For gene selection purposes, we extend AC-PCA with sparsity constraints, and propose and implement an efficient algorithm. The top selected genes from this algorithm demonstrate frontal/temporal and dorsal/ventral expression gradients and strong functional conservation.

1 Introduction

The development of microarray and next-generation sequencing technologies has enabled rapid quantification of the mammalian transcriptomes in a large number of samples[9, 45]. Dimension reduction methods, such as Multidimensional Scaling (MDS) and Principal Component Analysis (PCA) are commonly applied to visualize data in a low dimensional space, or/and identify dominant patterns of gene expression (feature extraction) [57, 51, 22, 33, 44, 15]. MDS aims to place each sample in a lower-dimensional space such that the between-sample distances are preserved as much as possible [35]. PCA seeks the linear combinations of the original variables such that the derived variables capture maximal variance [30]. One advantage of PCA is that the principal components

(PCs) are more interpretable by checking the loadings of the variables.

Human neurodevelopment is a dynamic and highly regulated biological process [59]. Abnormalities in neurodevelopment in humans may lead to psychiatric and neurological disorders, such as autism spectrum disorders [21, 64, 56]. Recent transcriptome studies of developing human brain and neocortex provide insights on the spatial or/and temporal dynamics of neurodevelopment [33, 44]. [33] collapsed the neocortex regions and demonstrated the spatial and temporal variation through MDS and PCA on all the samples. However, the spatial variation across neocortex regions was not explored in [33]. In human, the neocortex is involved in higher functions such as sensory perception, generation of motor commands, spatial reasoning, conscious thought and language [40]. Through analyzing the exon array dataset reported in [33], we found that visualization of the neocortex regions is affected by confounding factors, likely originated from the variations across individual donors (Figure 1). Without the adjustment, a) there is no clear pattern among the neocortex regions or/and samples from the same individual donors tend to form clusters, and b) it is challenging to identify neurodevelopmental genes with interregional variation.

Confounding factors, usually originated from experimental artifacts and frequently referred to as "batch effects", are commonly observed in high throughput transcriptome experiments. Various methods have been proposed to remove the unwanted variation through regression models on known confounding factors [28], factor models and surrogate vector analysis for unobserved confounding factors [38, 20, 37, 52, 68]. However, directly removing the confounding variation using these methods may introduce bias, as a result of incorrect model assumption of the confounding variation, and it can also remove the desired biological variation. Moreover, limited work has been done in the context of dimension reduction.

To address the limitations of existing methods, we have developed AC-PCA for simultaneous dimension reduction and adjustment for confounding variation, such as variations across individual donors. Applying AC-PCA to the human brain exon array dataset [33], we are able to recover the anatomical structure of neocortex regions, including the frontal to temporal axis and the dorsal to ventral axis. Our results are able to capture the interregional variation in neocortex and reveal the temporal dynamics of the spatial variation. In PCA, the loadings for the variables are typically nonzero. In high dimensional settings, sparsity constraints have been proposed in PCA for better interpretation of the PCs [31, 70, 66, 49] and better statistical properties, such as asymptotic consistency [29, 32, 41]. We have also developed an efficient and fast algorithm to find sparse solutions for AC-PCA. The genes identified by AC-PCA demonstrate smooth frontal to temporal, dorsal to ventral gradient and strong functional conservation.

2 Methods

2.1 AC-PCA adjusting for variations of individual donors

Let X_i represent the $b \times p$ matrix for the gene expression levels of individual i, where b is the number of brain regions and p is the number of genes. By stacking the rows of $X_1, \cdots, X_n$, X represents the $(n \times b) \times p$ matrix for the gene expression levels of n individuals. We propose the following objective function to adjust for individual variation:

$$\begin{aligned} \underset{v \in \mathbb{R}^p}{\text{maximize}} \quad & v^T X^T X v - \frac{2}{n-1} \lambda \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} v^T (X_j - X_i)^T (X_j - X_i) v \\ \text{subject to} \quad & ||v||_2^2 \leq 1. \end{aligned} \tag{1}$$

In (1), the term $v^T X^T X v$ is the objective function for standard PCA, and the regularization term $-\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} v^T (X_j - X_i)^T (X_j - X_i) v$ encourages the coordinates of the brain regions across individuals to be similar. The factor $\frac{2}{n-1}$ makes the regularization term in formulation (1) scale linearly with the number of individuals. The tuning parameter $\lambda > 0$ controls the strength of regularization. When $\lambda = +\infty$, we are forcing the coordinates of the same brain region across individuals to be the same after projection. Only the labels for brain regions (i.e. labels for the primary variables) are required when implementing formulation (1). We can apply it even if the individual labels of donors (i.e. the confounding variables) are unknown. The connection of formulation (1) with Canonical Correlation Analysis (CCA) is shown in the appendix.

2.2 AC-PCA in a general form

Let X be the $N \times p$ data matrix and Y be the $N \times l$ matrix for l confounding factors. Denote y_i the ith row in Y. We propose the following objective function to adjust for more general confounding variation:

$$\begin{aligned} \underset{v \in \mathbb{R}^p}{\text{maximize}} \quad & v^T X^T X v - \lambda v^T X^T K X v \\ \text{subject to} \quad & ||v||_2^2 \leq 1, \end{aligned} \tag{2}$$

where K is the $N \times N$ kernel matrix, and $K_{ij} = k(y_i, y_j)$. It can be shown that $v^T X^T K X v$ is the same as the empirical Hilbert-Schmidt independence criterion[23, 7] for Xv and Y, when linear kernel is applied on Xv (Appendix). In the objective function, we are penalizing the dependence between Xv (i.e. extracted feature) and the confounding factors. Formulation (1) is a special case of (2), where linear kernel (i.e. YY^T) is applied on Y, and Y has the following structure: in each column of Y, there are only two non-zero entries, $\sqrt{2/(n-1)}$ and $-\sqrt{2/(n-1)}$), corresponding to a pair of samples from the same brain region but different individuals. Implementation examples for formula (2) are provided in the simulation section (settings 3 and 4).

Denote $Z = X^TX - \lambda X^TKX$. Problem (2) can be rewritten as:

$$\underset{v\in\mathbb{R}^p}{\text{maximize}}\ v^TZv \quad \text{subject to } ||v||_2^2 \leq 1. \tag{3}$$

Therefore it can be solved directly by implementing eigendecomposition on Z.

2.3 AC-PCA with sparse loading

Denote $H = X^TKX$. It can be shown that solving (2) is equivalent to solving:

$$\underset{v\in\mathbb{R}^p}{\text{maximize}}\ v^TX^TXv \quad \text{subject to } v^THv \leq c_1,\ ||v||_2^2 \leq 1, \tag{4}$$

where c_1 is a constant depending on λ. A sparse solution for v can be achieved by adding ℓ_1 constraint:

$$\underset{v\in\mathbb{R}^p}{\text{maximize}}\ v^TX^TXv \quad \text{subject to } v^THv \leq c_1,\ ||v||_1 \leq c_2,\ ||v||_2^2 \leq 1. \tag{5}$$

Following [66], this is equivalent to:

$$\underset{u,v\in\mathbb{R}^p}{\text{minimize}}\ -u^TXv \quad \text{subject to } v^THv \leq c_1,\ ||v||_1 \leq c_2,\ ||v||_2^2 \leq 1,\ ||u||_2^2 \leq 1. \tag{6}$$

Problem (6) is biconvex in u and v, and it can be solved by iteratively updating u and v. At the kth iteration, the update for u is simply $\frac{Xv_{(k-1)}}{||Xv_{(k-1)})||_2^2}$. To update v, we need to solve:

$$\underset{v\in\mathbb{R}^p}{\text{minimize}}\ -u_{(k)}^TXv \quad \text{subject to } v^THv \leq c_1,\ ||v||_1 \leq c_2,\ ||v||_2^2 \leq 1. \tag{7}$$

Because of the quadratic constraint on v, it is hard to solve (7) directly. We propose to use the bisection method, solving the following feasibility problem iteratively:

$$\text{find } v \quad \text{subject to } -u_k^TXv \leq t,\ v^THv \leq c_1,\ ||v||_1 \leq c_2,\ ||v||_2^2 \leq 1, \tag{8}$$

where t is an upper-bound for $-u_k^TXv$ and is updated in each iteration. Note that H is positive semidefinite and (8) can be solved by alternating projection on the convex sets. Details for the projection are included in the Appendix. In summary, the algorithm to update v is as follows:

Algorithm 1: Bisection method for solving problem (7) and updating v

1. Initialize $t_{up} = 0$ and $t_{low} = -||X^Tu_k||_2$
2. Iterate until convergence:
   (a) $t^* = (t_{up} + t_{low})/2$
   (b) $t_{up} \leftarrow t^*$ if (8) is feasible for t^*
   (c) $t_{low} \leftarrow t^*$ if (8) is not feasible for t^*

3. Let $t = t_{up}$ and find v by solving (8)

The algorithm to solve (6) is as follows:
Algorithm 2: Finding the sparse principal component

1. Initialize v to be the solution of (2)
2. Iterate until convergence:
   (a) $u \leftarrow \frac{Xv_{(k-1)}}{||Xv_{(k-1)}||_2^2}$
   (b) Update v by implementing algorithm 1

2.4 Multiple principal components

In (2), obtaining multiple principal components is straightforward, as they are just the eigenvectors of Z; for the sparse solution, (6) can only obtain the first sparse principal component. To obtain the other sparse principal components, we can update X sequentially with $X_{(i+1)} = X_{(i)}(I - \hat{v}_i\hat{v}_i^T)$ for $i = 1, \cdots$, and implement (6) on $X_{(i+1)}$, where $X_{(1)} = X$ and $\hat{v}_i$ is the ith principal component.

2.5 Data preprocessing

The human brain exon array dataset was downloaded from the Gene Expression Omnibus (GEO) database under the accession number GSE25219. The dataset was generated from 1,340 tissue samples collected from 57 developing and adult post-mortem brain[33]. Details for the quality control (QC) of the dataset are described as in [33]. After the QC procedures, noise reduction was accomplished by removing genes that were not expressed [39], leaving 13,720 genes in the dataset. We next selected the top 5,000 genes sorted by coefficient of variation and centered the expression levels of this consort.

2.6 Conservation and heterozygosity scores

The dN/dS score for cross-species conservation was calculated using Ensembl BioMart [14]. The heterozygosity score was calculated using 1,000 Genomes phase 1 version 3 [12]. Let $f_1, \cdots f_p$ denote the allele frequencies for the p non-synonymous coding SNPs in a gene, and let l denote the maximum transcript length over all isoforms of that gene. The heterozygosity score was calculated as: $2\sum_{i=1}^{p} f_i(1 - f_i)/l$. For a gene with low heterozygosity score, the non-synonymous variants in that gene tend to be rare, which indicates the functional importance of that gene.

3 Tuning parameters

3.1 Tuning λ

Let X denote the $N \times p$ data matrix. $l = v^T X^T K X v$ can be treated as a loss function to be minimized. We do 5-fold cross-validation to tune λ:

(a) From X, we construct 5 data matrices $X_1, \cdots, X_5$, each of which is missing a non-overlapping one-fifth of the rows of X

(b) For each X_i, $i = 1, \cdots, 5$, implement (2) and obtain v_i

(c) Calculate l_{cv}. In Xv, we use v_i and the missing rows that are left out in X_i, for $i = 1, \cdots, 5$

(d) When λ increases from 0, l_{cv} usually decreases sharply and then either increases or becomes flat. In practice, we choose λ to be the "elbow" point of l_{cv}. When there is no or little confounding variation, l_{cv} tends to be flat when λ changes, and we can use regular PCA.

3.2 Tuning c_1 and c_2

Because c_1 and c_2 capture different aspects of the data, we propose a two-step approach: first tune c_1, and then tune c_2 with c_1 fixed. This also greatly reduces the computational cost since tuning c_2 can be slow. To tune c_1:

(a) We follow the previous procedure to tune λ

(b) The best λ is used to calculate v, as in (2)

(c) Let $c_1 = v^T Y v$

To tune c_2, we follow Algorithm 5 in [66], which is based on matrix completion:

(a) From X, we construct 5 data matrices $X_1, \cdots, X_5$, each of which is missing a non-overlapping one-fifth of the elements of X

(b) For $X_1, \cdots, X_5$, fit (6) and obtain $\hat{X}_i = duv^T$, the resulting estimate of X_i and $d = u^T X_i v$

(c) Calculate the mean squared errors of $\hat{X}_i$, for $i = 1, \cdots, 5$, using only the missing entries

(d) Choose c_2 that minimizes the sum of mean squared errors

4 Results

The human brain exon array dataset reported in [33] includes the transcriptomes of 16 brain regions comprising 11 areas of the neocortex, the cerebellar cortex, mediodorsal nucleus of the thalamus, striatum, amygdala and hippocampus. Because the time period system defined in [33] had varying numbers of donors across developmental epochs, we grouped samples from every 6 donors, sorted by age and beginning from period 3. While the last time window had only 5 donors, this reorganization more evenly distributed sample sizes and allowed improved comparisons across time (Table 1).

Table 1: Span of the windows in this study.

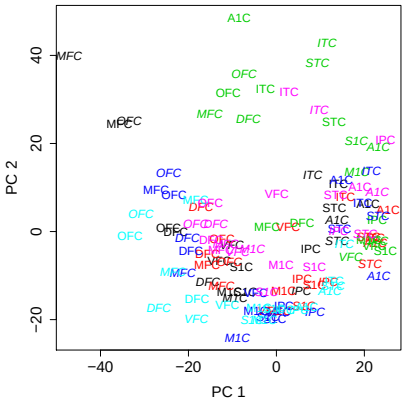
Window	Age
1	12PCW≤Age<16PCW
2	16PCW≤Age<21PCW
3	21PCW≤Age<35PCW
4	35PCW≤Age<10M
5	10M≤Age<8Y
6	8Y≤Age<19Y
7	19Y≤Age<30Y
8	30Y ≤Age<42Y
9	42Y≤Age

M: postnatal months
PCW: post-conception weeks
Y: postnatal years.

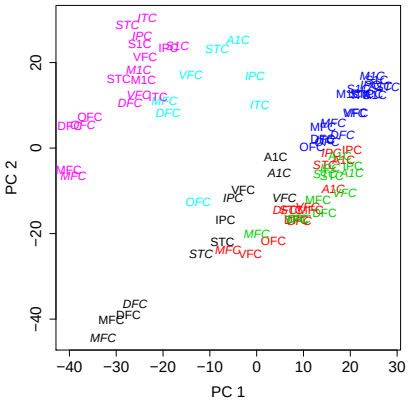
In the analysis, we used samples from 10 regions in the neocortex. V1C was excluded from the analysis as the distinct nature of this area relative to other neocortical regions tended to compress the other 10 regions into a single cluster. We conducted traditional Principal Component Analysis (PCA) for windows 1 and 2, as shown in Figures 1a and 1b. At first glance, neither analysis produced any clear patterns among neocortical regions. However, closer observation of these plots suggested some underlying structure: we performed PCA on just the right hemisphere of donor HSB113 and found that the gross morphological structure of the hemisphere was largely recapitulated (Figure 1c). The gross structure tends to be consistent between hemispheres and across donors within time windows when PCA is performed simultaneously, but the pattern is largely distorted, likely due to the small sample size and noisy background (Supplementary Materials). If we first regress the gene expression levels on the individual labels and then perform PCA on the residuals, the result is better but still not satisfactory (Supplementary Materials).

In contrast, when we applied AC-PCA to see the effectiveness of our approach in adjusting confounding effects from individual donors, we were able to recover the anatomical structure of neocortex (Figure 1f; Supplementary Figures 1a, 1b, 2a and 2b in [33]). We treated the left and right hemispheres from the same donor as different individuals when implementing (1). Every region tended to form a smaller cluster, when λ is larger than the optimal tuning value. Variation of the principal components shrinks and the overall pattern remains consistent (Supplementary Materials).

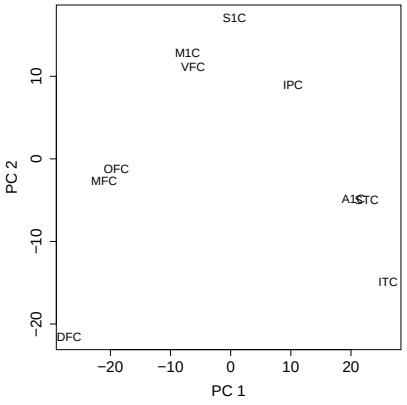
Next, we explored the temporal dynamics of the principal components (PC) (Figure 2). The pattern is similar from windows 1 to 5, with PC1 representing the frontal to temporal gradient, which follows the contour of developing cortex [44], and PC2 representing the dorsal to ventral gradient. Starting from window 6, these two components reversed order. In windows 6 to 9, MFC shows increasing distance from the other regions, and primary areas (M1C, S1C, A1C) are



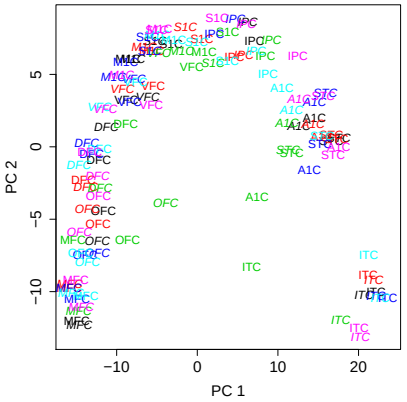
(a) Window 1, PCA



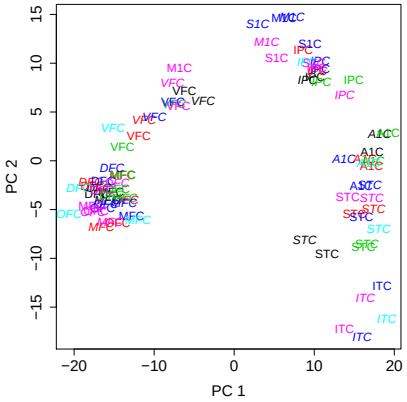
(b) Window 2, PCA



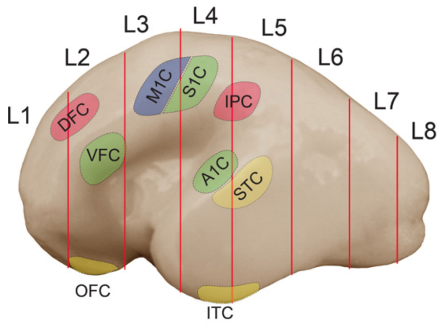
(c) Window 1, PCA for HSB113(R)



(d) Window 1, AC-PCA, $\lambda^* = 3$

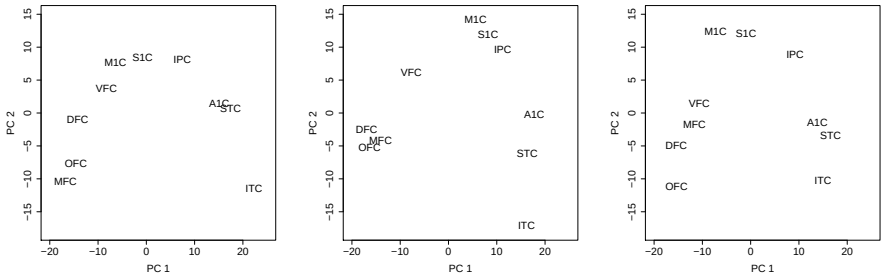


(e) Window 2, AC-PCA, $\lambda^* = 3$



(f) Representative fetal human brain, lateral surface of hemisphere [33]

Figure 1: Visualization of the brain exon-array data [33], windows 1 and 2. Each color represents a donor. Samples from the right hemisphere are labeled as italic. For both windows, the best tuning parameter λ^* was 3. In (f), MFC is not visible on the lateral surface, and it belongs to slice L2.



(a) Window 1 (b) Window 2 (c) Window 3

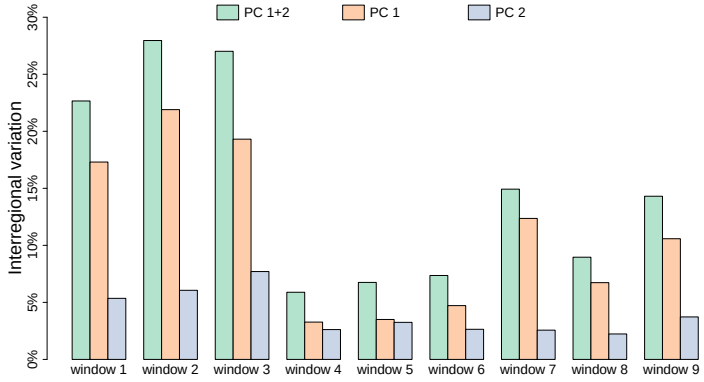
(d) Window 4 (e) Window 5 (f) Window 6

(g) Window 7 (h) Window 8 (i) Window 9

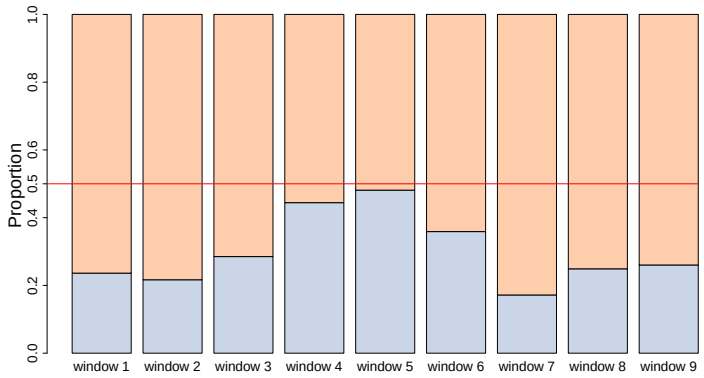
Figure 2: Temporal dynamics of the principal components. Median was taken across individuals. λ was fixed to 2 for all windows when implementing Formulation (1).

separated from other regions in PC1, implying the prominence of transcriptional changes during specification of regional functions.

We also calculated the interregional variation explained by the PCs (Figure 3). In the first three windows, PC1 explains about 20% of the interregional variation. The variation explained by the first two PCs decreases close to birth (window 4) and then increases in later time windows, similar to the "hourglass" pattern previously reported based on cross-species comparison and differential expression [47, 39]. Interestingly, if we compare the proportion of variation



(a) Interregional variation explained by the first two PCs



(b) Proportion of variation explained by the first two PCs

Figure 3: Interregional variation explained by the principal components. Interregional variation is calculated to be the sum over the variance across regions among the individuals.

explained by the first two PCs, a smooth pattern is revealed: compared with PC1, the variation explained by PC2 first increases, in window 5, it nearly equals to that of PC1, then it decreases and finally slightly increases.

We then implemented Formulation (6) to select genes associated with the PCs. The number of genes with non-zero loadings are shown in Figure 4, along with the interregional variation explained in the regular PCs. Interestingly, the trends tend to be consistent: when the regular PC explains more variation, more genes are selected in the corresponding sparse PC; compared with PC1, fewer genes tend to be selected in PC2. The numbers of selected genes are quite dynamic. At both early developmental periods prior to birth and later postnatal periods, more genes tend to be selected in both PCs. For PC1 in windows 1, 2, 3 and 9, more than $4,000$ genes are selected, which indicates global trends in

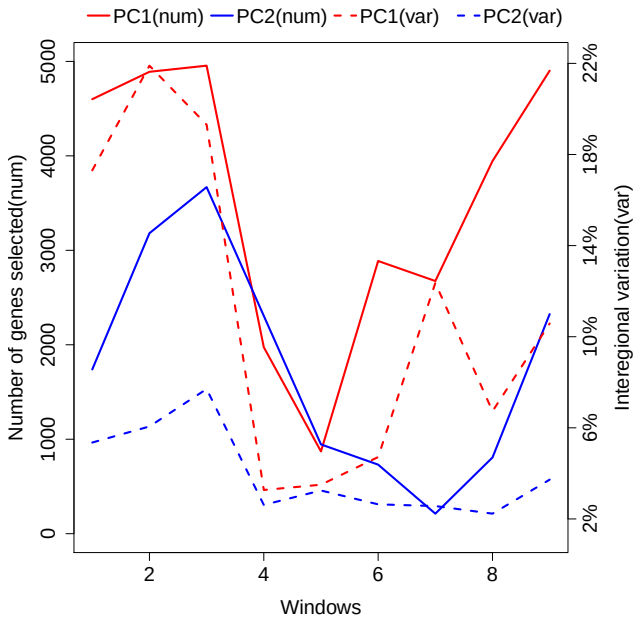


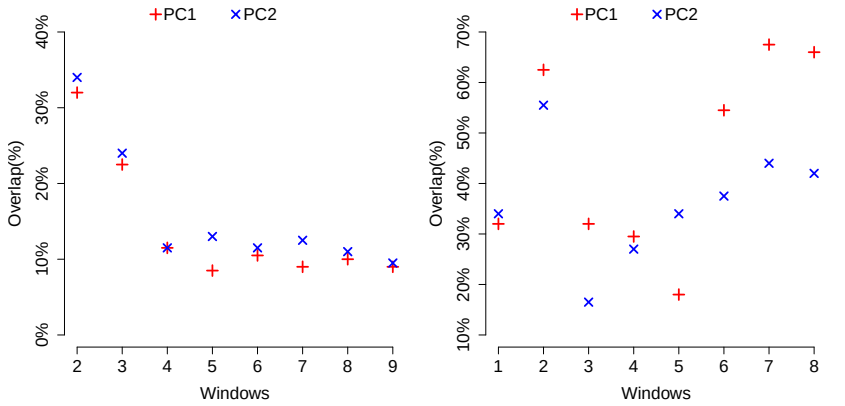
Figure 4: Number of genes selected in the sparse PCs and the interregional variation explained by the regular PCs.

neocortex that affect large numbers of genes.

To produce more stringent and comparable gene lists, we chose c_2 such that 200 genes are selected in each window. The overlap of gene lists across windows is moderate (Figure 5) and, as expected, the overlap with the first window decreases over time. The overlap between adjacent windows tend to be larger in later time windows, indicating that interregional differences become stable. Interestingly, the overlap between windows 2 and 3 is also large. Results of pathway enrichment analysis using DAVID bioinformatics resources [27, 58] are available in the Supplementary Tables.

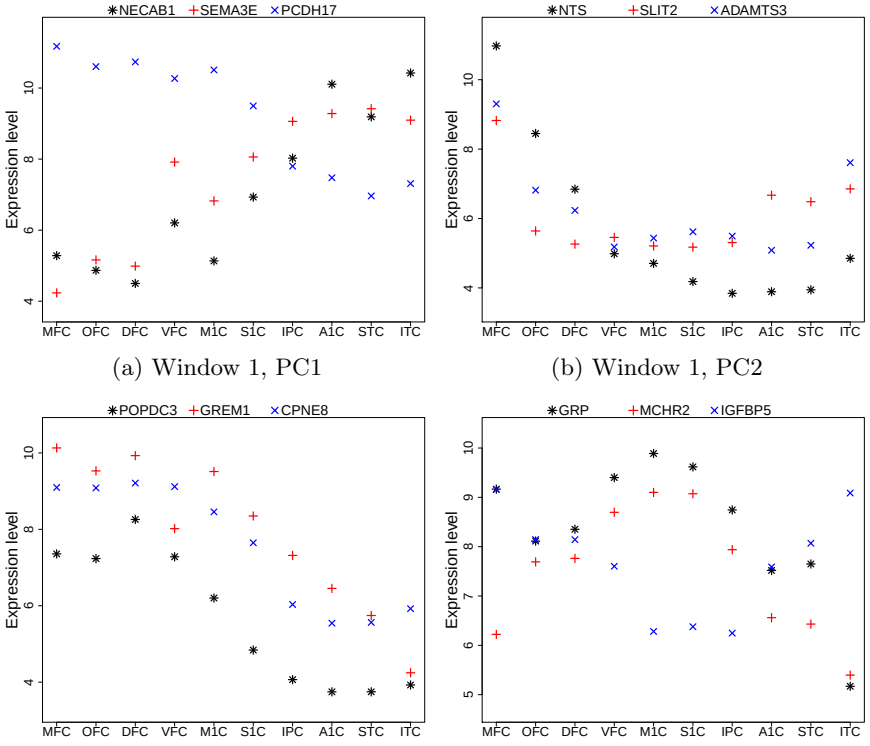
In windows 1 and 3, genes with the largest loadings demonstrate interesting spatial patterns (Figure 6). For PC1, the top genes follow the frontal to temporal gradient; while for PC2, they tend to follow the dorsal to ventral gradient. A brief overview of the functions of these genes are listed in Table 2.

Finally, we demonstrate the functional conservation of the 200 genes selected in PC1 and PC2 (Figure 7). These genes tend to have low dN/dS scores for human vs. macaque comparison, even lower than the complete list of all essential genes. In the human vs. mouse comparison, we observed a similar trend (Supplementary Materials). Parallel to the cross-species conservation, we also observed that these genes tend to have low heterozygosity scores, a measure of functional conservation in human (Figure 7).



(a) Overlap with window 1 (b) Overlap between adjacent windows

Figure 5: Overlap of the top 200 genes in PC1 and PC2.



(a) Window 1, PC1

(b) Window 1, PC2

(c) Window 3, PC1

(d) Window 3, PC2

Figure 6: Expression levels for the genes with top loadings in windows 1 and 3. The top 3 genes are shown and each point represents the median over the individuals.

Table 2: Top genes in the PCs and their functions.

Window 1, PC1	
NECAB1	calcium ion binding [60, 19]
SEMA3E	semaphorin, control vascular morphogenesis, serve as axon guidance ligands [24, 55, 11]
PCDH17	calcium ion binding, establishment and function of specific cell-cell connections in the brain [67, 25]
Window 1, PC2	
NTS	dopamine signaling [34]
SLIT2	axonal guidance, midline guidance in the forebrain [6, 5, 48]
ADAMTS3	extracellular matrix proteases, cleaves the propeptides of type II collagen [62, 18]
Window 3, PC1	
POPDC3	important in heart development [8, 4]
GREM1	BMP Antagonist, may play a role in regulating organogenesis, body patterning, and tissue differentiation [26]
CPNE8	calcium-dependent phospholipid binding [63, 13]
Window 3, PC2	
GRP	gastrin-releasing peptide, regulates the gastrointestinal and central nervous systems [43, 10, 61]
MCHR2	receptor for Melanin-concentrating hormone (MCH), important in feeding behaviors and energy metabolism [3, 65]
IGFBP5	insulin-like growth factor(IGF) binding protein, essential for growth and development [1, 53]

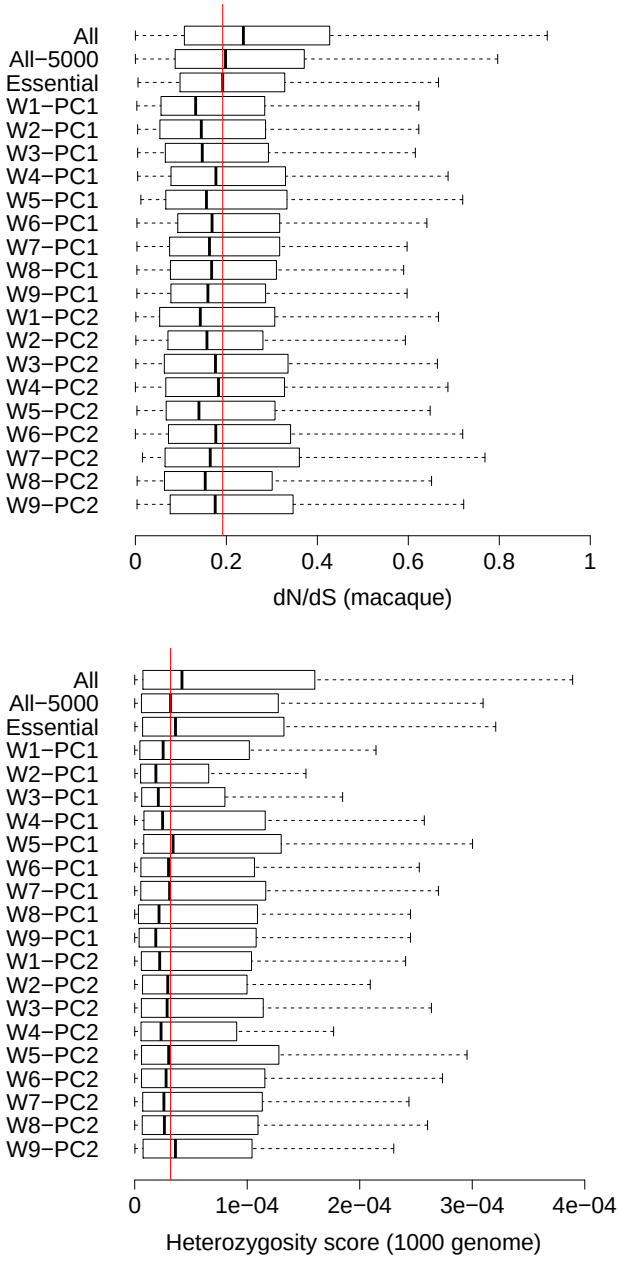


Figure 7: Conservation and heterozygosity scores for the genes in PC1 and PC2. "All" represents the total 17,568 genes in the exon array experiment. "All-5000" represents the 5,000 genes used in the analysis. "Essential" represents a list of 337 genes that are functionally conserved and essential, obtained from the Database of Essential Genes (DEG) version 5.0 [69]. "Window" is abbreviated as "W".

5 Simulation

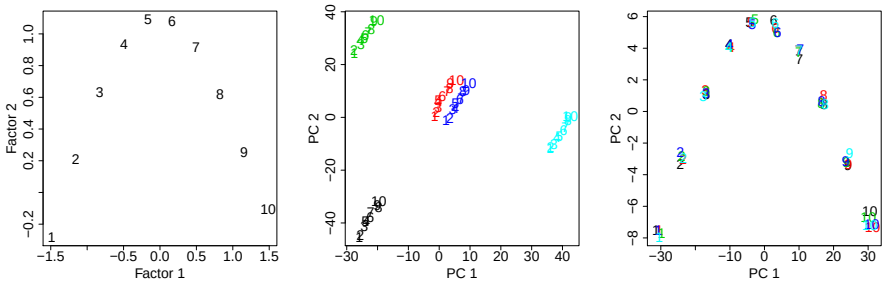
We first tested the performance of formulation (1) on simulated datasets with individual variation. We considered $n = 5$, $b = 10$ and $p = 400$.
Simulation setting 1: for the ith individual, the $b \times p$ matrix $X_i = Wh + \alpha B s_i + \epsilon_i$. Wh represents the shared variation across individuals. $\alpha B s_i$ corresponds to individual variation and ϵ_i is noise. $W = (w_1\ w_2)$ is a $b \times 2$ matrix, representing the latent structure of the shared variation. For visualization purpose, we assumed that it is smooth and has rank 2. Let $\mu = (1, \cdots, b)'$ and w_1 is the normalized μ, with mean 0 and variance 1. $w_2 \sim \mathcal{N}(0, 0.25 \cdot \Sigma)$, where $\Sigma_{ij} = \exp(-\frac{(w_{i1} - w_{j1})^2}{4})$. h is a $2 \times p$ matrix and the rows in h are generated from $\mathcal{N}(0, I_p)$, where I_p is the $p \times p$ identity matrix. B is a $b \times 1$ matrix with all 1s. s_i is generated from $\mathcal{N}(0, I_p)$. α is a scalar indicating the strength of confounding variation, we set $\alpha = 2.5$. The rows in ϵ_i are generated from $\mathcal{N}(0, 0.25 \cdot I_p)$.
Simulation setting 2: for the ith individual, let $X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix}$, where X_{1i} represents the data matrix for the first 200 genes and X_{2i} represents that of the other 200 genes. $X_{1i} = Wh + \epsilon_{1i}$ and $X_{2i} = \alpha W_i h_i + \epsilon_{2i}$. W, h, h_i, ϵ_{1i} and ϵ_{2i} are generated similarly as that in setting 1. $\alpha = 2.5$. The first column in W_i is generated from $\mathcal{N}(0, I_b)$, and the second column is generated from $\mathcal{N}(0, 0.25 \cdot I_b)$, where I_b is the $b \times b$ identity matrix.

Settings 1 and 2 represent two different scenarios: in setting 1, the individual variation is represented by a global trend for all the genes; while in setting 2, we assumed that for some genes, the variation is shared among individuals, and it is not shared for the other genes. The results of data visualization are presented in Figure **??**.

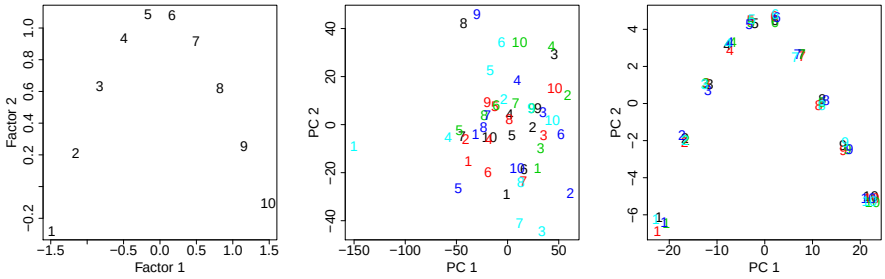
Then we tested the performance of formulation (2) on simulated dataset with other confounding structure:
Simulation setting 3: Let $N = 10$ and $p = 400$. The $N \times p$ matrix $X = Wh + \alpha B s + \epsilon$. W, h are the same as that in setting 1. We set $\alpha = 2.5$. $B = (b_1\ b_2)$ is a $N \times 2$ matrix: the entries in b_1 have 0.3 probability of being 0, otherwise the entries are set to 1; b_2 equals $1 - b_1$. s is a $2 \times p$ matrix, where the rows are generated from $\mathcal{N}(0, I_p)$. ϵ is an $N \times p$ matrix, where the rows are generated from $\mathcal{N}(0, 0.25 \cdot I_p)$.
Simulation setting 4: Let $N = 10$ and $p = 400$. The $N \times p$ matrix $X = Wh + \alpha \tilde{w}_1 s + \epsilon$. W, h are the same as that in setting 1. We set $\alpha = 2.5$. $\tilde{w}_1$ is a permutation of w_1. s is generated from $\mathcal{N}(0, I_p)$. ϵ is an $N \times p$ matrix, where the rows are generated from $\mathcal{N}(0, 0.25 \cdot I_p)$.

Setting 3 represents an experiment with two "batches", contributing globally to the gene expression levels. Setting 4 represents an experiment with a continuous confounding factor (e.g. age). We implemented formulation (2) with linear kernels, and set $Y = B$ and $\tilde{w}_1$ for settings 3 and 4, correspondingly. The visualization results are shown in Figure 8.

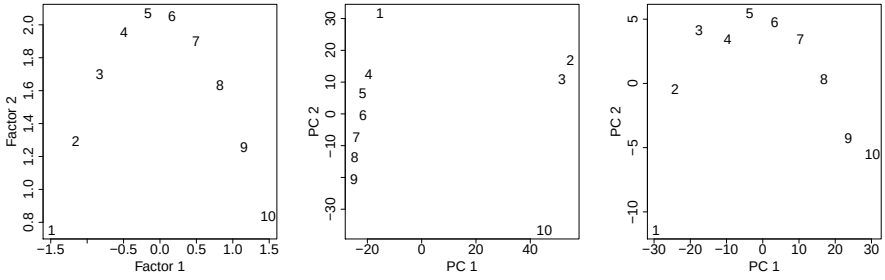
Finally, we tested the performance of formulation (6) for variable selection when the true loading is sparse. For simplicity, we assumed that the latent



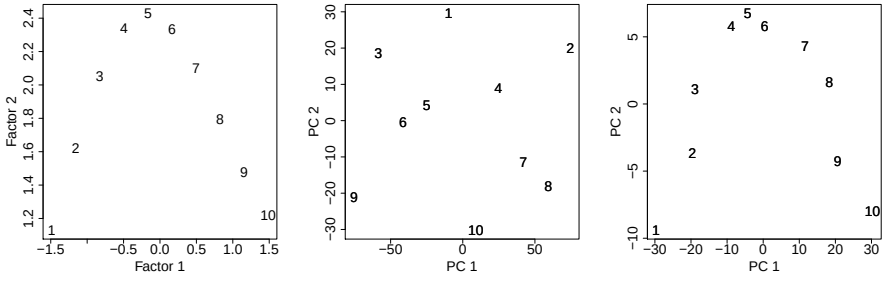
(a) Setting 1, true pattern (b) Setting 1, PCA (c) Setting 1, AC-PCA



(d) Setting 2, true pattern (e) Setting 2, PCA (f) Setting 2, AC-PCA



(g) Setting 3, true pattern (h) Setting 3, PCA (i) Setting 3, AC-PCA



(j) Setting 4, true pattern (k) Setting 4, PCA (l) Setting 4, AC-PCA

Figure 8: Visualization of simulated data with confounding variation. The parameter λ was tuned as described in the previous section.

Table 3: Estimation of sparsity and sensitivity

		True non-zeros=100			True non-zeros=40	
		$\alpha = 1.5$	$\alpha = 2$	$\alpha = 2.5$	$\alpha = 2$	$\alpha = 2^*$
Estimated non-zeros	$\sigma = 0.2$	113.7(34.2)	88.7(20.9)	81.4(20.5)	37.4(14.7)	46.7(16.5)
	$\sigma = 0.5$	120.6(42.5)	91.6(31.1)	79.4(24.4)	38.9(15.6)	54.6(19.0)
Sensitivity	$\sigma = 0.2$	0.85(0.06)	0.83(0.08)	0.83(0.08)	0.74(0.11)	0.77(0.10)
	$\sigma = 0.5$	0.80(0.07)	0.79(0.08)	0.80(0.05)	0.62(0.10)	0.69(0.11)

To calculate sensitivity, c_2 was chosen such that the estimated non-zero entries equals the true number. In $\alpha = 2^*$, wh was scaled with 2.5 to let the variation match with the less sparse setting.

factor is of rank 1.

Simulation setting 5: for the ith individual, let $X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix}$, where X_{1i} represents the expression levels of the first 200 genes and X_{2i} represents that of the other 200 genes. $X_{1i} = wh + \epsilon_{1i}$ and $X_{2i} = \alpha W_i h_i + \epsilon_{2i}$. w is the same as w_1 in setting 1. Some entries in h are set to 0 to reflect sparsity, the other entries are generated from $\mathcal{N}(0, I)$. W_i and h_i are the same as that in setting 2. The rows in ϵ_{1i} and ϵ_{2i} are generated from $\mathcal{N}(0, \sigma^2 \cdot I)$, where σ is a scalar indicating the noise level.

Results of simulation setting 5 are shown in Table 3. Larger noise (σ) leads to lower sensitivity, larger standard error for the estimated non-zeroes, but does not affect the mean much. Smaller confounding variation (α) leads to overestimate of the non-zeroes, but does not affect the sensitivity much.

6 Discussion

Dimension reduction methods are commonly applied to visualize datapoints in a lower dimensional space and identify dominant patterns in the data. Confounding variation, technically and biologically originated, may affect the performance of these methods, and hence the visualization and interpretation of the results (Figure 1).

In this study, we have proposed AC-PCA for simultaneous dimension reduction and adjustment for confounding variation, such as variations of the individual donors, and demonstrated its good performance through the analysis of human brain developmental exon array dataset [33] and simulated data as well. We showed that AC-PCA is able to recover the anatomical structure of the neocortex regions. In the first five time windows, PC1 captures the frontal to temporal variation and PC2 captures the dorsal to ventral variation. Because of the structural complexity of primate neocortex, physical distance may not be a good measure for the true similarity between regions. Our results show that AC-PCA is able to reconstruct the regional map in neocortex based on transcriptome data alone. The developmental gradients in neocortex are likely a result of intrinsic signaling, controlled in part by graded expression of transcrip-

tion factors during early cortical development, followed by extrinsic signaling from thalamic afferents after the start of corticogenesis [44, 42, 46, 50, 54, 59]. For better interpretation of the PCs and gene selection purpose, we proposed to incorporate sparsity constraints in AC-PCA. The selected genes demonstrate frontal to temporal gradient and dorsal to ventral gradient. These genes are of potential interest for studying cortical patterning. They also tend to be functionally important, as indicated by the cross-species conservation and the heterozygosity scores calculated from the 1000 Genomes data.

One feature of AC-PCA is its simplicity. There is no need to specify any analytical forms for the confounding variation. Instead, we extended the objective function of regular PCA with penalty on the dependence between the extracted features (i.e. Xv) and the confounding factors. AC-PCA is designed to capture the desired biological variation, even when the confounding factors are unobserved, as long as the labels for the primary variable of interest are known.

The application of AC-PCA is not limited to transcriptome datasets. Dimension reduction methods have been applied to other types of genomics data for various purposes, such as feature extraction for methylation prediction [16], classifying yeast mutants using metabolic footprinting [2], and classifying immune cells using DNA methylome [36], etc. AC-PCA is applicable to these datasets to capture the desired variation, adjust for potential confounders, and select the relevant features. AC-PCA can serve as an exploratory tool and be combined with other methods. For example, the extracted features can be implemented in regression models for more rigorous statistical inference. The R package, Matlab source code, and application examples will be available on Bioconductor and Gtihub.

7 Acknowledgement

Zhixiang Lin and Hongyu Zhao were supported in part by the National Science Foundation grant DMS-1106738 and the National Institutes of Health grants R01 GM59507 and P01 CA154295. Ying Zhu, Xuming Xu, Mingfeng Li and Nenad Sestan were supported by the National Institutes of Health grants P50 MH106934 and U01 MH103339. We also thank Matthew W. State for the partial financial support of Zhixiang Lin. Can Yang was supported by grant No.61501389 from National Science Funding of China, grant No.22302815 from the Hong Kong Research Grant Council, and grant FRG2/14-15/069 from Hong Kong Baptist University the Hong Kong RGC grant HKBU. All computations were performed on the Yale University Biomedical High Performance Computing Center.

References

[1] Susanne V Allander, Catharina Larsson, Ewa Ehrenborg, Adisak Suwanichkul, Gunther Weber, Sheila L Morris, Svetlana Bajalica, Michael C Kiefer, Holger Luthman, and David R Powell. Characterization of the chromosomal gene and promoter for human insulin-like growth factor binding protein-5. *Journal of Biological Chemistry*, 269(14):10891–10898, 1994.

[2] Jess Allen, Hazel M Davey, David Broadhurst, Jim K Heald, Jem J Rowland, Stephen G Oliver, and Douglas B Kell. High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nature biotechnology*, 21(6):692–696, 2003.

[3] Songzhu An, Gene Cutler, Jack Jiagang Zhao, Shu-Gui Huang, Hui Tian, Wanbo Li, Lingming Liang, Miki Rich, Amy Bakleh, Juan Du, et al. Identification and characterization of a melanin-concentrating hormone receptor. *Proceedings of the National Academy of Sciences*, 98(13):7576–7581, 2001.

[4] Birgit Andrée, Tina Hillemann, Gania Kessler-Icekson, Thomas Schmitt-John, Harald Jockusch, Hans-Henning Arnold, and Thomas Brand. Isolation and characterization of the novel popeye gene family expressed in skeletal muscle and heart. *Developmental biology*, 223(2):371–382, 2000.

[5] Kim T Nguyen Ba-Charvet, Katja Brose, Le Ma, Kuan H Wang, Valérie Marillat, Constantino Sotelo, Marc Tessier-Lavigne, and Alain Chédotal. Diversity and specificity of actions of slit2 proteolytic fragments in axon guidance. *The Journal of Neuroscience*, 21(12):4281–4289, 2001.

[6] Kim Tuyen Nguyen Ba-Charvet, Katja Brose, Valérie Marillat, Tom Kidd, Corey S Goodman, Marc Tessier-Lavigne, Constantino Sotelo, and Alain Chédotal. Slit2-mediated chemorepulsion and collapse of developing forebrain axons. *Neuron*, 22(3):463–473, 1999.

[7] Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7):1357–1371, 2011.

[8] Thomas Brand. The popeye domain-containing gene family. *Cell biochemistry and biophysics*, 43(1):95–103, 2005.

[9] Patrick O Brown and David Botstein. Exploring the new world of the genome with dna microarrays. *Nature genetics*, 21:33–37, 1999.

[10] PL Brubaker and DJ Drucker. Minireview: glucagon-like peptides regulate cell proliferation and apoptosis in the pancreas, gut, and central nervous system. *Endocrinology*, 145(6):2653–2659, 2004.

[11] Sophie Chauvet, Samia Cohen, Yutaka Yoshida, Lylia Fekrane, Jean Livet, Odile Gayet, Louis Segu, Marie-Christine Buhot, Thomas M Jessell, Christopher E Henderson, et al. Gating of sema3e/plexind1 signaling by neuropilin-1 switches axonal repulsion to attraction during brain development. *Neuron*, 56(5):807–822, 2007.

[12] 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[13] Carl E Creutz, Jose L Tomsig, Sandra L Snyder, Marie-Christine Gautier, Feriel Skouri, Janine Beisson, and Jean Cohen. The copines, a novel class of c2 domain-containing, calciumdependent, phospholipid-binding proteins conserved from paramecium to humans. *Journal of Biological Chemistry*, 273(3):1393–1402, 1998.

[14] Fiona Cunningham, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, et al. Ensembl 2015. *Nucleic acids research*, 43(D1):D662–D669, 2015.

[15] Spyros Darmanis, Steven A Sloan, Ye Zhang, Martin Enge, Christine Caneda, Lawrence M Shuer, Melanie G Hayden Gephart, Ben A Barres, and Stephen R Quake. A survey of human brain transcriptome diversity at the single cell level. *Proceedings of the National Academy of Sciences*, page 201507125, 2015.

[16] Rajdeep Das, Nevenka Dimitrova, Zhenyu Xuan, Robert A Rollins, Fatemah Haghighi, John R Edwards, Jingyue Ju, Timothy H Bestor, and Michael Q Zhang. Computational prediction of methylation status in human genomic sequences. *Proceedings of the National Academy of Sciences*, 103(28):10713–10716, 2006.

[17] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM, 2008.

[18] Russell J Fernandes, Satoshi Hirohata, J Michael Engle, Alain Colige, Daniel H Cohn, David R Eyre, and Suneel S Apte. Procollagen ii amino propeptide processing by adamts-3 insights on dermatosparaxis. *Journal of Biological Chemistry*, 276(34):31502–31509, 2001.

[19] Alexandra P Few, Nathan J Lautermilch, Ruth E Westenbroek, Todd Scheuer, and William A Catterall. Differential regulation of cav2. 1 channels by calcium-binding protein 1 and visinin-like protein-2 requires n-terminal myristoylation. *The Journal of neuroscience*, 25(30):7071–7080, 2005.

[20] Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

[21] Daniel H Geschwind and Pat Levitt. Autism spectrum disorders: developmental disconnection syndromes. *Current Opinion in Neurobiology*, 17(1):103–111, 2007.

[22] Thomas J Giordano, Rork Kuick, Tobias Else, Paul G Gauger, Michelle Vinco, Juliane Bauersfeld, Donita Sanders, Dafydd G Thomas, Gerard Doherty, and Gary Hammer. Molecular classification and prognostication of adrenocortical tumors by transcriptome profiling. *Clinical Cancer Research*, 15(2):668–676, 2009.

[23] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005.

[24] Chenghua Gu, Yutaka Yoshida, Jean Livet, Dorothy V Reimert, Fanny Mann, Janna Merte, Christopher E Henderson, Thomas M Jessell, Alex L Kolodkin, and David D Ginty. Semaphorin 3e and plexin-d1 control vascular pattern independently of neuropilins. *Science*, 307(5707):265–268, 2005.

[25] Naosuke Hoshina, Asami Tanimura, Miwako Yamasaki, Takeshi Inoue, Ryoji Fukabori, Teiko Kuroda, Kazumasa Yokoyama, Tohru Tezuka, Hiroshi Sagara, Shinji Hirano, et al. Protocadherin 17 regulates presynaptic assembly in topographic corticobasal ganglia circuits. *Neuron*, 78(5):839–854, 2013.

[26] David R Hsu, Aris N Economides, Xiaorong Wang, Peter M Eimon, and Richard M Harland. The xenopus dorsalizing factor gremlin identifies a novel family of secreted proteins that antagonize bmp activities. *Molecular cell*, 1(5):673–683, 1998.

[27] Dawei Huang, Brad T Sherman, Richard A Lempicki, et al. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44–57, 2008.

[28] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[29] Iain M Johnstone and Arthur Yu Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486), 2009.

[30] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2002.

[31] Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547, 2003.

[32] Sungkyu Jung, JS Marron, et al. Pca consistency in high dimension, low sample size context. *The Annals of Statistics*, 37(6B):4104–4130, 2009.

[33] Hyo Jung Kang, Yuka Imamura Kawasawa, Feng Cheng, Ying Zhu, Xuming Xu, Mingfeng Li, André MM Sousa, Mihovil Pletikos, Kyle A Meyer, Goran Sedmak, et al. Spatio-temporal transcriptome of the human brain. *Nature*, 478(7370):483–489, 2011.

[34] Patrick Kitabgi. Neurotensin modulates dopamine neurotransmission at several levels along brain dopaminergic pathways. *Neurochemistry international*, 14(2):111–119, 1989.

[35] Joseph B Kruskal and Myron Wish. *Multidimensional scaling*, volume 11. Sage, 1978.

[36] Marta Kulis, Simon Heath, Marina Bibikova, Ana C Queirós, Alba Navarro, Guillem Clot, Alejandra Martínez-Trillos, Giancarlo Castellano, Isabelle Brun-Heath, Magda Pinyol, et al. Epigenomic analysis detects widespread gene-body dna hypomethylation in chronic lymphocytic leukemia. *Nature genetics*, 44(11):1236–1242, 2012.

[37] Jeffrey T Leek, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6):882–883, 2012.

[38] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):1724–1735, 2007.

[39] Zhixiang Lin, Stephan J Sanders, Mingfeng Li, Nenad Sestan, Hongyu Zhao, et al. A markov random field-based approach to characterizing human brain development using spatial–temporal transcriptome data. *The Annals of Applied Statistics*, 9(1):429–451, 2015.

[40] Jan H Lui, David V Hansen, and Arnold R Kriegstein. Development and evolution of the human neocortex. *Cell*, 146(1):18–36, 2011.

[41] Zongming Ma et al. Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, 41(2):772–801, 2013.

[42] Antonello Mallamaci and Anastassia Stoykova. Gene networks controlling early cerebral cortex arealization. *European Journal of Neuroscience*, 23(4):847–856, 2006.

[43] TJ McDonald, H Jörnvall, G Nilsson, M Vagne, M Ghatei, SR Bloom, and V Mutt. Characterization of a gastrin releasing peptide from porcine non-antral gastric tissue. *Biochemical and biophysical research communications*, 90(1):227–233, 1979.

[44] Jeremy A Miller, Song-Lin Ding, Susan M Sunkin, Kimberly A Smith, Lydia Ng, Aaron Szafer, Amanda Ebbert, Zackery L Riley, Joshua J Royall, Kaylynn Aiona, et al. Transcriptional landscape of the prenatal human brain. *Nature*, 508(7495):199–206, 2014.

[45] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

[46] Dennis DM O'Leary. Do cortical areas emerge from a protocortex? *Trends in neurosciences*, 12(10):400–406, 1989.

[47] Mihovil Pletikos, Andre MM Sousa, Goran Sedmak, Kyle A Meyer, Ying Zhu, Feng Cheng, Mingfeng Li, Yuka Imamura Kawasawa, and Nenad Sestan. Temporal specification and bilaterality of human neocortical topographic gene expression. *Neuron*, 81(2):321–332, 2014.

[48] Andrew S Plump, Lynda Erskine, Christelle Sabatier, Katja Brose, Charles J Epstein, Corey S Goodman, Carol A Mason, and Marc Tessier-Lavigne. Slit1 and slit2 cooperate to prevent premature midline crossing of retinal axons in the mouse visual system. *Neuron*, 33(2):219–232, 2002.

[49] Xin Qi, Ruiyan Luo, and Hongyu Zhao. Sparse principal component analysis by choice of norm. *Journal of multivariate analysis*, 114:127–160, 2013.

[50] Pasko Rakic. Specification of cerebral cortical areas. *Science*, 241(4862):170–176, 1988.

[51] Markus Ringnér. What is principal component analysis? *Nature biotechnology*, 26(3):303–304, 2008.

[52] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896–902, 2014.

[53] Dervis AM Salih, Gyanendra Tripathi, Cathy Holding, Tadge AM Szestak, M Ivelisse Gonzalez, Emma J Carter, Laura J Cobb, Joan E Eisemann, and Jennifer M Pell. Insulin-like growth factor-binding protein 5 (igfbp5) compromises survival, growth, muscle development, and fertility in mice. *Proceedings of the National Academy of Sciences*, 101(12):4314–4319, 2004.

[54] Stephen N Sansom and Frederick J Livesey. Gradients in the brain: the control of the development of form and function in the cerebral cortex. *Cold Spring Harbor Perspectives in Biology*, 1(2):a002519, 2009.

[55] Guido Serini, Donatella Valdembri, Sara Zanivan, Giulia Morterra, Constanze Burkhardt, Francesca Caccavari, Luca Zammataro, Luca Primo, Luca Tamagnone, Malcolm Logan, et al. Class 3 semaphorins control vascular morphogenesis by inhibiting integrin function. *Nature*, 424(6947):391–397, 2003.

[56] Nenad Sestan and Matthew W State. The emerging biology of autism spectrum disorders. *Science (New York, NY)*, 337(6100):1301, 2012.

[57] Alexei A Sharov, Yulan Piao, Ryo Matoba, Dawood B Dudekula, Yong Qian, Vincent VanBuren, Geppino Falco, Patrick R Martin, Carole A Stagg, Uwem C Bassey, et al. Transcriptome analysis of mouse stem cells and early embryos. *PLoS Biol*, 1(3):E74, 2003.

[58] Brad T Sherman, Richard A Lempicki, et al. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, 37(1):1–13, 2009.

[59] John C Silbereis, Sirisha Pochareddy, Ying Zhu, Mingfeng Li, and Nenad Sestan. The cellular and molecular landscapes of the developing human central nervous system. *Neuron*, 89(2):248–268, 2016.

[60] S Sugita, A Ho, and TC Südhof. Necabs: A family of neuronal ca 2+-binding proteins with an unusual domain structure and a restricted expression pattern. *Neuroscience*, 112(1):51–63, 2002.

[61] Yan-Gang Sun and Zhou-Feng Chen. A gastrin-releasing peptide receptor mediates the itch sensation in the spinal cord. *Nature*, 448(7154):700–703, 2007.

[62] Bor Luen Tang. Adamts: a novel family of extracellular matrix proteases. *The international journal of biochemistry & cell biology*, 33(1):33–44, 2001.

[63] Jose Luis Tomsig and Carl E Creutz. Biochemical characterization of copine: a ubiquitous ca2+-dependent, phospholipid-binding protein. *Biochemistry*, 39(51):16163–16175, 2000.

[64] Christopher A Walsh, Eric M Morrow, and John LR Rubenstein. Autism and brain development. *Cell*, 135(3):396–400, 2008.

[65] Suke Wang, Jiang Behan, Kim O'Neill, Blair Weig, Steven Fried, Thomas Laz, Marvin Bayne, Eric Gustafson, and Brian E Hawes. Identification and pharmacological characterization of a novel human melanin-concentrating hormone receptor, mch-r2. *Journal of Biological Chemistry*, 276(37):34664–34670, 2001.

[66] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, page kxp008, 2009.

[67] Takeshi Yagi and Masatoshi Takeichi. Cadherin superfamily genes: functions, genomic organization, and neurologic diversity. *Genes & development*, 14(10):1169–1180, 2000.

[68] Can Yang, Lin Wang, Shuqin Zhang, and Hongyu Zhao. Accounting for non-genetic factors by low-rank representation and sparse regression for eqtl mapping. *Bioinformatics*, 29(8):1026–1034, 2013.

[69] Ren Zhang and Yan Lin. Deg 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic acids research*, 37(suppl 1):D455–D458, 2009.

[70] Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Appendix A Connection with Canonical Correlation Analysis (CCA)

Without loss of generality, let $n = 2$, then the objective function in formulation (1) becomes:

$$v^T X^T X v - 2\lambda v^T (X_2 - X_1)^T (X_2 - X_1) v = (1 - 2\lambda) v^T X^T X v + 4\lambda v^T X_2^T X_1 v$$

In Canonical Correlation Analysis (CCA), there are two datasets X and Y, and the goal is to maximize the correlation between X and Y after projections. The objective function in CCA is $a^T X^T Y b$, where a and b are two column vectors. Note that $a^T X^T Y b$ and $v^T X_2^T X_1 v$ have similar forms: in $v^T X_2^T X_1 v$, the projection vectors a and b are the same as v. Therefore the objective function in (1) represents a balance between regular PCA and CCA. When $\lambda > 0.5$, the weight on PCA is negative.

Appendix B Hilbert-Schmidt independence criterion

The linear kernel of Xv is $L = Xvv^T X^T$. Let K be the kernel of Y. Let $H = I - N^{-1} ee^T$, where e is a column vector with all 1s. Then $H^T X = HX = X$, as X is centered. The empirical Hilbert-Schmidt independence criterion for Xv and Y is:

$$\begin{aligned} \mathrm{Tr}(HLHK) &= \mathrm{Tr}(HXvv^T X^T HK) \\ &= \mathrm{Tr}(v^T X^T HKHXv) \\ &= \mathrm{Tr}(v^T X^T KXv) \\ &= v^T X^T KXv \end{aligned}$$

Appendix C Details for the projections

Projection onto ℓ_2 ball

First do decomposition $K = \Delta\Delta^T$, and let $M = \Delta^T X$. We have $M^T M = X^T \Delta\Delta^T X = X^T KX$.

We need to solve the following optimization problem:

$$\underset{v}{\text{minimize}} \quad \|v - v_0\|_2^2 \quad \text{subject to} \quad \|Mv\|_2^2 \leq c_1$$

It is equivalent to the following Lagrangian problem:

$$L(\lambda, v) = \|v - v_0\|_2^2 + \lambda(\|Mv\|_2^2 - c_1), \text{ where } \lambda \geq 0.$$

For any fixed λ, $L(\lambda, v)$ is a convex differentiable function of v. By taking derivative, it can be shown that $v^* \equiv v^*(\lambda) = (I_p + \lambda M^T M)^{-1} v_0$ minimizes the Lagrangian. Then we have

$$g(\lambda) \equiv \inf_v L(\lambda, v) = L(\lambda, v^*)$$

By singular value decomposition, $M = U\Sigma V^T$, where $U^T U = UU^T = I_N$ and $V^T V = VV^T = I_p$. Then we have

$$v^* = (I_p + \lambda V \Sigma^T \Sigma V^T)^{-1} v_0 = V(I_p + \lambda \Sigma^T \Sigma)^{-1} V^T v_0 \tag{9}$$

and

$$\begin{aligned} g(\lambda) &= \|V(I_p - (I_p + \lambda\Sigma^T\Sigma)^{-1})V^T v_0\|_2^2 + \lambda\|U\Sigma(I_p + \lambda\Sigma^T\Sigma)^{-1}V^T v_0\|_2^2 - \lambda c_1 \\ &= \|(I_p - (I_p + \lambda\Sigma^T\Sigma)^{-1})V^T v_0\|_2^2 + \lambda\|\Sigma(I_p + \lambda\Sigma^T\Sigma)^{-1}V^T v_0\|_2^2 - \lambda c_1 \\ &= \sum_{i=1}^{d} (\frac{\lambda\sigma_i^2}{1 + \lambda\sigma_i^2})^2 (V^T v_0)_i^2 + \lambda \sum_{i=1}^{d} (\frac{\sigma_i}{1 + \lambda\sigma_i^2})^2 (V^T v_0)_i^2 - \lambda c_1 \\ &= \sum_{i=1}^{d} \frac{\lambda\sigma_i^2}{1 + \lambda\sigma_i^2} (V^T v_0)_i^2 - \lambda c_1, \end{aligned}$$

where $d(<= N)$ is the number of non-zero singular values of M, σ_i is the ith non-zero singular value, and $(\cdot)_i$ represents the ith element of a vector. When $\lambda \geq 0$, $g(\lambda)$ is concave and the optimal value λ^* can be found by Newton-Raphson's Method. With $\lambda = \lambda^*$, the projection v can be calculated with (9), where the inversion part is a diagonal matrix. It can be shown that only the first d columns of V affect the projection and typically we have $d \ll p$.

Projection onto ℓ_1 ball

$$\underset{v}{\text{minimize}} \quad \|v - v_0\|_2^2 \quad \text{subject to} \quad \|v\|_1 \leq c_2$$

This can be solved efficiently by the algorithm presented in [17].

Projection onto hyperplane

$$\underset{v}{\text{minimize}} \quad \|v - v_0\|_2^2 \quad \text{subject to} \quad -u_k^T X v \leq t$$

The solution is:

$$v = v_0 - \frac{u_k^T X v_0 + t}{\|X^T u_k\|_2^2} X^T u_k$$