Chalk-Talk Notes for Demixed PCA

Alex Williams

April 27, 2016

This was presented at the computational neuroscience journal club at Stanford (sponsored by the center for Mind, Brain, and Computation): `https://web.stanford.edu/group/mbc/JournalClub/`. These notes were used as introductory/review material to Kobak et al. (2016) "Demixed principalcomponent analysis of neural population data." *eLife*, `http://dx.doi.org/10.7554/eLife.10989`

1 Basic notation and setup

- We have a $n \times p$ matrix A.
- Let row i of A be denoted $\mathbf{a}_i$. We will use the convention that we care about each row of the matrix as a unit of data. In machine learning parlance the rows are "observations" and the columns are "features" or measurements.
- I may refer to each $\mathbf{a}_i$ as a "datapoint." This is because you can visualize each row as a point in p-dimensional Cartesian coordinate space.
- For the applications we consider today, each $\mathbf{a}_i$ will be a neural firing rate trace.

2 Linear Algebra Review

- Basic idea of linear dimensionality reduction:

  $$A \approx WC^T$$

  where $W \in \mathbf{R}^{n \times r}$ and $C \in \mathbf{R}^{p \times r}$. I may refer to C as the "components," and W as the "weightings" or "loadings."

- Two views of matrix-vector multiplication:
  - Produces a linear combination of the columns (vector on the right) or rows (vector on the left) of the matrix. In the context of this discussion, this is relevant because each datapoint is approximated as a linear combination of the rows of C^T:

    $$\mathbf{a}_i \approx \sum_k W_{ik} \mathbf{c}_k$$

  - Each element in the output vector is a dot product with each row (vector on right) or column (vector on the left) of the matrix. When the vector is of unit length, this is the projection of each row/column onto the vector. In this context, this is relevant because you get dimension reduction by linear projection. If each column of C is unit length, then multiplying A by C gives you a lower-dimensional representation of your dataset L:

    $$AC = L$$

- *Corollary:* one can view linear dimension reduction from an information theoretic perspective. Consider an encoder matrix $E \in \mathbf{R}^{p \times r}$ that reduces the dimension of each datapoint from p to r, as well as a decoder matrix $D \in \mathbf{R}^{p \times r}$ that restores the full dimensionality. Our general goal is to find E and D that minimize the reconstruction error:

$$\underset{E,D}{\text{minimize}} \quad \text{loss}(A, AED^T)$$

- Nice properties of symmetric matrices
  - Consider a real, symmetric matrix $S = S^T$
  - *Theorem:* S can be diagonalized by an orthogonal matrix, $S = U\Lambda U^T$. All eigenvalues, $\boldsymbol{\lambda}$ where $\Lambda = \text{diag}(\boldsymbol{\lambda})$, are real. The eigenvectors are the columns of U, or equivalently, the rows of U^T.
  - *Theorem:* If each column of A is mean-centered, then the covariance matrix $\Sigma = A^T A$ is symmetric and positive semidefinite, meaning that all eigenvalues are nonnegative.
  - *Bonus:* A skew symmetric matrix, defined as a matrix $S^T = -S$, and has purely imaginary eigenvalues. Doing a dimension reduction on a skew symmetric matrix leads to jPCA [1].
- Fitting a multivariate Gaussian to data

$$p(\mathbf{x}) \propto \exp\left(-\frac{1}{2} \underbrace{\mathbf{x}^T \Sigma^{-1} \mathbf{x}}_{\text{Mahalanobis dist.}}\right)$$

  - Another name for the inverse covariance matrix Σ^{-1} is the *precision matrix*. It has the same eigenbasis as the covariance matrix:

$$\Sigma^{-1} = (U \cdot \text{diag}(\boldsymbol{\lambda}) \cdot U^T)^{-1} = U \cdot \text{diag}\left(\frac{1}{\boldsymbol{\lambda}}\right) \cdot U^T$$

    Where we have used the identity $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$.
  - This distribution is centered at the origin (since we mean-centered the data). The eigenvectors (columns of U) form orthogonal axes of a hyperellipsoid. The orthogonal axes of the ellipsoid are given by the eigenvectors (columns of U) and the relative lengths of the ellipsoid axes are given by the eigenvalues (larger axes have larger λ_i)

3 Two views of PCA

- **Classic view:** preserving maximal variance in the projection.

$$\begin{aligned} \underset{\mathbf{C}}{\text{maximize}} \quad & \mathbf{C}^T A^T A \mathbf{C} \\ \text{subject to} \quad & \mathbf{C}^T \mathbf{C} = I \end{aligned}$$

  Seminal work [2] shows that the solution is to set C to be the top r eigenvectors of the covariance matrix (i.e. the columns of U with the largest eigenvalues). Intuitively, this fact arises from the multivariate Gaussian perspective: take the longest axes of the hyperellipsoid and throw away the smaller ones that contain less variance.

  Technical note: We can find the top r eigenvectors by computing the singular value decomposition (SVD) on the data matrix A.

- **Equivalent view:** minimizing projection distance (more generalizable view).

$$\begin{aligned} \underset{\mathbf{C}}{\text{minimize}} \quad & \|A - ACC^T\|_F^2 \\ \text{subject to} \quad & \mathbf{C}^T\mathbf{C} = I \end{aligned}$$

We $\|\cdot\|_F$ denotes the Frobenius norm. Note that $W = AC$, or (from the information theory persepctive) $E = C$, $D = C^T$. The objective function can be re-expressed as minimizing the reconstruction error in the least-squares sense:

$$\underset{W,C}{\text{minimize}} \quad \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{1}{2}\left(A_{ij} - \sum_{k=1}^{r} W_{ik}C_{jk}\right)^2$$

Proving the equivalence of these two views of PCA involves simple geometric principles like the Pythagorean theorem.

4 Equivalent solutions for PCA:

The constraint $C^TC = I$ ensures that solution to PCA is unique up to a permutation of the rows and columns of C. When we solve PCA via an eigendecomposition of the covariance matrix, C is naturally orthogonal since the eigenvectors of a postive semi-definite matrix are orthogonal.

We can write re-write the PCA objective without this constraint. In this case the W and C we find are not unique since we can multiply them by any invertible matrix Q and get the same reconstruction error:

$$\|A - WC^T\|_F^2 = \|A - WQ^{-1}QC^T\|_F^2 = \|A - W'C'^T\|_F^2$$

Here we applied a fairly general linear transformation Q to the principal components and recovered the same reconstruction error by applying the inverse transformation to W (the loadings). By doing this, we came up with a new low-dimensional representation of the data represented by W' and C'.

5 Variants on PCA:

- Generalized Low-Rank Model [3]

$$\underset{W,C}{\text{minimize}} \quad \sum_{i=1}^{n}\sum_{j=1}^{p}\ell_j(A_{ij}, \sum_{k=1}^{r} W_{ik}C_{jk}) \;+\; \gamma\sum_{i=1}^{n} r_w^{(i)}(\mathbf{w}_i) \;+\; \sum_{j=1}^{p} r_c^{(j)}(\mathbf{c}_j)$$

- Example: logistic PCA

$$\underset{W,C}{\text{minimize}} \quad \sum_{i=1}^{n}\sum_{j=1}^{p}\log(1 + \exp(-A_{ij} \cdot \sum_{k=1}^{r} W_{ik}C_{jk}))$$

- Example: sparse PCA

$$\underset{W,C}{\text{minimize}} \quad \|A - WC^T\|_F^2 \;+\; \gamma\sum_{i=1}^{n}\|\mathbf{w}_i\|_1 \;+\; \gamma\sum_{j=1}^{p}\|\mathbf{c}_j\|_1$$

6 Noise models:

- Classic least-squares regression estimates a dependent variable y from an independent variable x (assumed to be measured without noise). The model is $y = mx + b + \eta$, where the noise η is normally distributed with constant variance as a function of x.
- PCA can be understood as a generalization where there is noise in both y and x. Specifically the noise is isotropic. See `https://liorpachter.wordpress.com/2014/05/26/what-is-principal-component-analysis/`

7 Demixed PCA:

7.1 Figure 1:

- Task is to compare stimulus frequencies separated by three second delay.
- Task complexity is governed by difference in frequencies and whether first or second frequency is higher.
- **Panel b** - Four example neurons
- **Panel e** - Goal is to accentuate differences across conditions. Find neurons that significantly correlate with stimulus frequency (left plot, differences across colors) or the decision (right plot, differences between solid vs. dashed lines). Then trial average their PSTHs.
- **Panel h** - Same goal as panel e, but use linear regression coefficients at each time point across N neurons. Neurons that don't correlated with stim freq. or with decision will have regression coefficients near zero, others will contribute to the ultimate PSTH.
- **Panel k** - New goal, capture variance in the data. Use PCA. The authors say "PCA paints a much more complex picture of the population activity, dominated by strong temporal dynamics, with several stimulus- and decision-related components."

7.2 Methods:

- The data is organized into a 5-way array (N neurons $\times$ T times $\times$ S stimuli $\times$ D decisions $\times$ K trials). They collect this data into a big matrix $\mathbf{X}$ which is $N \times KSQT$.
- They then decompose this matrix by averaging over all combinations of time, stimulus, and decision (and their interactions) this follows the decomposition of variance in a factorial ANOVA. Figure 8 shows a nice graphical example. The decomposition has the form:

$$\mathbf{X} = \sum_{\phi} \mathbf{X}_{\phi} + \mathbf{X}_{\text{noise}}$$

- The core idea will be to find separate low-dimensional models that reconstruct each $\mathbf{X}_{\phi}$ individually.
  - By virtue of doing this, you achieve a low-dimensional model that encodes features about each ϕ which is the "demixing" part of this paper.
  - Additionally, since the decomposition of $\mathbf{X}$ preserves variance, if we build low-dimensional models that capture variance in each $\mathbf{X}_{\phi}$ then we also capture the bulk of the variance in the full dataset.
- Concretely, the total loss function is:

$$L = \sum_{\phi} \frac{1}{2} \|\mathbf{X}_{\phi} - \mathbf{F}_{\phi}\mathbf{D}_{\phi}\mathbf{X}\|_F^2$$

  Where we need to optimize over the "encoder" and "decoder" matrices $\mathbf{F}_{\phi}$ and $\mathbf{D}_{\phi}$ for each subproblem. The subproblems are referred to as *reduced-rank regression* since the matrix $\mathbf{F}_{\phi}\mathbf{D}_{\phi}$ is rank q.

- There is a cute way to find the solution to each subproblem by first solving $\mathbf{AX} = \mathbf{X}_\phi$ for $\mathbf{A}$ by least-squares. This produces a full-rank solution for $\mathbf{A}$. Then do PCA on A and keep the top q components. $\mathbf{F}_\phi$ is simply the top q loadings and $\mathbf{D}_\phi$ is simply the top q components. (We respectively called these matrices "W" and "C" in the introduction.)

References

[1] Mark M. Churchland, John P. Cunningham, Matthew T. Kaufman, Justin D. Foster, Paul Nuyujukian, Stephen I. Ryu, and Krishna V. Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 07 2012.

[2] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[3] Madeleine Udell, Corinne Horn, Reza Zadeh, and Stephen Boyd. Generalized low rank models. *arXiv preprint arXiv:1410.0342*, 2014.