Lecture 13 — May 12

Lecturer: Lester Mackey *Scribe: Jessy Hwang, Minzhe Wang*

13.1 Canonical correlation analysis

13.1.1 Recap

CCA is a linear dimensionality reduction procedure for paired or two-viewed data. Given two mean-centered datasets X and Y, each with n rows, the CCA objective is

$$\max_{u,v} \frac{u^T X^T Y v}{\sqrt{u^T X^T X u v^T Y^T Y v}}.$$

This is a generalized eigenvalue problem, which is solved by a generalized eigenvector $(u^*, v^*)^T$ satisfying

$$\begin{pmatrix} 0 & X^T Y \\ Y^T X & 0 \end{pmatrix} \begin{pmatrix} u^* \\ v^* \end{pmatrix} = \lambda^* \begin{pmatrix} X^T X & 0 \\ 0 & Y^T Y \end{pmatrix} \begin{pmatrix} u^* \\ v^* \end{pmatrix}.$$

As in PCA, after obtaining the first pair of canonical directions in this way, subsequent pairs $(u_2, v_2), \ldots, (u_k, v_k)$ may be extracted by solving the CCA optimization problem subject to the constraint that the new canonical variables be uncorrelated with prior ones, i.e.,

$$\text{Corr}(u_j^T x, u_l^T x) = \text{Corr}(v_j^T y, v_l^T y) = 0 \quad \forall\ l < j,$$

where x is a random vector taking on values $x_1, \ldots, x_n$ with probability $1/n$ each, and y plays an equivalent role.

13.1.2 Degeneracy in CCA

There are several situations in which the CCA solution exhibits degeneracy:

- If $x_i = Ay_i$ for all i, then any u is an optimal CCA direction (with correlation = 1). This can be seen by choosing $v = A^\top u$. So the CCA direction is meaningless in this case.
- If the coordinates of x and y are uncorrelated, i.e., $\frac{1}{n} X^T Y$ is identically 0, then any (u, v) is optimal with correlation = 0.
- CCA is sometimes applied without centering X and Y; this corresponds to using a cosine similarity objective instead of correlation. Then if $\text{rank}(X) = n$, any v is optimal, with correlation = 1, via $u = X^T(XX^T)^{-1}Yv$. We have the same problem if $\text{rank}(Y) = n$.

Regularization is often introduced into CCA to break such degeneracy ties in favor of higher-variance directions and to control overfitting. The standard regularized CCA objective is given by

$$\max_{u,v} \frac{u^T X^T Y v}{\sqrt{u^T (X^T X + \lambda_1 I) u v^T (Y^T Y + \lambda_2 I) v}}$$

for some $\lambda_1, \lambda_2 > 0$. This is akin to penalizing or constraining the ℓ_2 norms of u and v as in ridge regression.

13.1.3 Kernel CCA

As a final note, one may derive a kernelized version of CCA in much the same way that we derived kernel PCA. In this case, we work with two kernel functions k_X and k_Y.

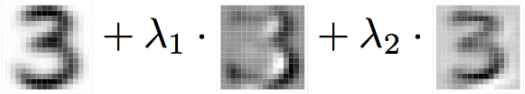
13.2 Sparse unsupervised learning

13.2.1 Motivation

When performing dimensionality reduction or latent feature modeling, we often want to interpret the recovered component loadings $u_j \in \mathbb{R}^p$ in terms of the input coordinates. Here are two examples.

- Recall the PCA decomposition of handwritten 3s discussed in ESL. We were able to visualize our decomposition in terms of the extracted component loadings, the "eigen-3s":

$$\begin{aligned} \hat{f}(\lambda) &= \bar{x} + \lambda_1 v_1 + \lambda_2 v_2 \\ &= \end{aligned}$$



  These intuitive visualizations of the loadings allowed us to interpret the data features being captured by the component directions (in this case, long-tailedness and thickness of the 3, respectively).

- When studying gene expression in cancer patients (so that each coordinate of x_i is a gene), we would like to declare that a few genes are responsible for most of the variation observed in the data. We would like the extracted loadings $u_1, \ldots, u_k \in \mathbb{R}^p$ to be sparse so that all variance is attributed to those few non-zero coordinates.

Unfortunately, the loadings obtained from most of the latent feature modeling methods discussed so far are typically dense: most if not all coordinates are nonzero. This is similar to the dense regression coefficients that result from least squares linear regression.

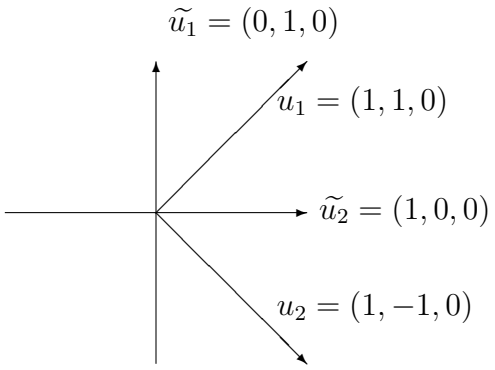
This leads us to the question, "How can we modify a latent feature modeling procedure like PCA to obtain sparse loadings?" In the next subsection, we describe some early approaches to answering this question in the context of PCA.

13.2.2 Earliest attempts

Rotating the loadings matrix

One way to improve the sparsity of a collection of loadings without sacrificing any of the variance explained is to rotate the k loadings to be sparser and more interpretable, i.e., we transform U to $U_{\text{rot}} = UR$ where U_{rot} exhibits sparsity (see, e.g., Richman, 1986). Essentially we are finding a more interpretable basis for the subspace spanned by the original PC loadings. The resulting U_{rot} captures the same data variance overall as the original U. Here is a simple example.

Example 1. Take $p = 3$. In the following picture, (u_1, u_2) represents our original pair of principal component loadings. Each vector is only 2-sparse. Meanwhile, $(\widetilde{u_1}, \widetilde{u_2})$, a rotation of (u_1, u_2), is sparser than (u_1, u_2) but also spans exactly the same subspace of $\mathbb{R}^3$ (namely, the plane in which this sheet of paper lies).



This method has a couple of drawbacks. First, there is no guarantee that a sparse rotation exists. Second, the rotated loadings lose the successive variance maximization property; that is, $\tilde{u}_1$ may not be the direction of maximum variance if $k > 1$.

Thresholding

This approach involves thresholding the small-magnitude entries in the loading vectors (see, e.g., Cadima and Jolliffe, 1995), i.e., setting those small magnitude entries to 0. Unfortunately the results lose orthonormality, and they may not be optimal for a target sparsity level.

13.2.3 An optimization approach

Continuing to focus on the PCA setting, we will next examine the possibility of directly constraining or penalizing the cardinality of the loadings in the context of a PCA optimization problem. Define

$$\text{card}(u) = \sum_{i=1}^{p} \mathbb{I}\left(u_i \neq 0\right) = ||u||_0.$$

Note that $||u||_0$ is often called the l_0 "norm" of u, although this is actually not a norm.

Adding a cardinality constraint to the PCA objective leads to the so-called sparse PCA optimization problem, which is the subject of the next section.

13.3 Sparse PCA optimization problem

13.3.1 Equivalent (single component) objectives

When we first introduced PCA, we presented two different motivations, namely variance maximization and reconstruction error minimization, which led to two equivalent optimization problems. We will now add cardinality constraints to both of these optimization problems.

- **Variance maximization**

$$\max_{u_1} u_1^T \frac{X^T X}{n} u_1 \quad \text{s.t.} \quad ||u_1||_2 = 1, \quad ||u_1||_0 \leq c_1 \tag{V}$$

  where c_1 is the cardinality bound. Note that we could equivalently penalize $||u_1||_0$ in the objective instead of constraining.

- **Reconstruction error minimization**

  In the original PCA setting, we had the objectives

$$\min_{u_1 \text{ s.t. } ||u_1||_2=1} \sum_{i=1}^n ||x_i - u_1 u_1^T x_i||_2^2 \tag{R1}$$

$$\iff \min_{u_1, v_1 \text{ s.t. } ||v_1||_2=1} \sum_{i=1}^n ||x_i - v_1 u_1^T x_i||_2^2 + \lambda ||u_1||_2^2 \tag{R2}$$

  For $\lambda > 0$, (R1) and (R2) are equivalent (up to the scale of u_1) since $v_1^* = \frac{u_1^*}{\|u_1^*\|_2}$ at the solution.

  We showed earlier that, in the absence of a cardinality constraint, variance maximization and reconstruction error minimization are equivalent optimization problems. Perhaps surprisingly, they are still equivalent here: adding the cardinality constraint $||u_1||_0 \leq c_1$ to (R1) or (R2) is equivalent to (V) (up to the scale of u_1)!

  Exercise. Prove the above statement. (Hint: $v_1^* \propto X^T X u_1^*$ for (R2).)

13.3.2 Solutions for sparse PCA

A potential difficulty in the above optimization problem is selecting the cardinality c_1. A more daunting obstacle is that any of these cardinality-constrained optimization problems is NP-hard. We have several options for dealing with this obstacle.

- **Option 1.** We could use standard combinatorial optimization techniques like branch-and-bound to solve the problem exactly (Moghaddam et al., 2006). This is optimal but in practice may take a very, very long time.

- **Option 2.** We could take a greedy approach to variable selection like the GSPCA (greedy sparse PCA) approach of Moghaddam et al. (2006). They propose a **greedy bidirectional search** consisting of both a forward pass and a backward pass:
  - In the **forward pass**, we start with no selected variables and successively add in the variable that improves the objective the most.
  - In the **backward pass**, we start with all variables and successively remove the one that reduces the objective the least.

  Then, for a given target cardinality c, we choose the better solution of the forward and backward passes. This is not guaranteed to yield an optimal solution in general, but it works well in practice.

- **Option 3.** Our third option is the most commonly explored. We will attempt to solve a (hopefully more tractable) surrogate optimization problem. The idea is to replace the hard combinatorial problem with a simpler non-combinatorial problem. Many, many authors have followed this path; we will explore some of the more popular solutions like SCoTLASS of Jolliffe et al. (2003), SPCA of Zou et al. (2006), and Direct SPCA (DSPCA) of d'Aspremont et al. (2004).

13.3.3 SCoTLASS

The SCoTLASS procedure replaces $||u_1||_0$ with the convex surrogate $||u_1||_1 = \sum_{j=1}^{p} |u_{1j}|$ in the variance maximization problem (V). The ℓ_1 norm is selected, because it is known to induce sparse solutions (e.g., as it does in Lasso regression). The resulting optimization problem is

$$\max_{u_1} \frac{u_1 X^T X u_1}{n} \quad \text{s.t.} \quad ||u_1||_2 = 1, \quad ||u_1|| \leq t_1$$

where t_1 is a tuning parameter that controls the resulting sparsity level. The good news is that we now have a continuous optimization problem. The bad is that this problem is still nonconvex, and the proposed algorithm (projected gradient) has been shown to be slow, computationally expensive, and prone to local maxima. We will continue our discussion of surrogate optimization problems in the next lecture.

Bibliography

Cadima, J. and I. T. Jolliffe (1995). Loading and correlations in the interpretation of principal components. *Journal of Applied Statistics 22*(2), 203–214.

d'Aspremont, A., L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet (2004). A direct formulation for sparse PCA using semidefinite programming. In *NIPS*, Volume 16, pp. 41–48.

Jolliffe, I. T., N. T. Trendafilov, and M. Uddin (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics 12*(3), 531–547.

Moghaddam, B., Y. Weiss, and S. Avidan (2006). Spectral bounds for sparse PCA: Exact and greedy algorithms. *Advances in Neural Information Processing Systems 18*, 915.

Richman, M. B. (1986). Rotation of principal components. *Journal of Climatology 6*(3), 293–335.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics 15*(2), 265–286.