STAT 154 Notes (2019) – math behind PCA

Raaz Dwivedi and Bin Yu

UC Berkeley

February 5, 2019

Math behind PCA: Eigendecomposition

- For our positive semidefinite sample covariance matrix $\mathbf{G} = \mathbf{X}^T\mathbf{X}$, we have the eigendecomposition

$$\mathbf{G} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$$

  where $\mathbf{U}$ is an orthonormal matrix $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ and $\mathbf{D}$ is a diagonal matrix with non-negative entries
- Columns of $\mathbf{U}$ are the eigenvectors of the matrix $\mathbf{G}$ and $\mathbf{D}$ contains the (non-negative) eigenvalues $d_1 \geq d_2, \ldots \geq d_p \geq 0$.
- The geometric interpretation of $\mathbf{U}$ is a rotation and $\sqrt{\mathbf{D}}$ is a rescaling.

Math behind PCA: obtaining PCs using eigen decomposition

- After the rotation $\mathbf{U}$ applied to $\mathbf{X}$, we get

$$(\mathbf{Z}_1, ..., \mathbf{Z}_p) = \mathbf{X}\mathbf{U} = (\mathbf{X}_1, ..., \mathbf{X}_p)(\mathbf{u}_1, ..., \mathbf{u}_p)$$

- $$\mathbf{Z}_j = (\mathbf{X}_1, ..., \mathbf{X}_p)\mathbf{u}_j = u_{1j}\mathbf{X}_1 + ... + u_{pj}\mathbf{X}_p,$$

  where $(u_{1j}, ..., u_{pj})^T = \mathbf{u}_j$ and $||\mathbf{u}_j||^2 = \sum_{k=1}^{p} u_{kj}^2 = 1$
- $\mathbf{Z}_1, ..., \mathbf{Z}_p$ are called Principal Components (PCs) and

$$\mathbf{Z}^T\mathbf{Z} = (\mathbf{X}\mathbf{U})^T(\mathbf{X}\mathbf{U}) = \mathbf{U}^T\mathbf{G}\mathbf{U} = \mathbf{D}$$

- Hence
  var$(\mathbf{Z}_j) = d_j$, cov$(\mathbf{Z}_i, \mathbf{Z}_j) = 0$ for $i \neq j$.
  That is, the PCs, or Z_j's, are orthogonal and their lengths are $\sqrt{d_j}$.

Math behind PCA

- First PC is the direction of maximum variance, let's derive it mathematically.
- Consider the set of vectors $\mathcal{S} = \mathbf{x}_1, \ldots, \mathbf{x}_n$ such that their mean $\overline{\mathbf{x}} = \mathbf{0}$ is zero.
- We need to find a direction $\mathbf{v}$ such that $\mathrm{Var}(\mathbf{v}^\top \mathbf{x})$ is maximized where $\mathbf{x}$ is selected uniformly at random from the set $\mathcal{S}$.
- Mathematically, we have to solve the problem:

$$\max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \sum_{i=1}^{n} (\mathbf{v}^\top \mathbf{x}_i - \mathbf{v}^\top \overline{\mathbf{x}})^2, \text{ or equivalently}$$

$$\max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \sum_{i=1}^{n} (\mathbf{v}^\top \mathbf{x}_i)^2$$

Math behind PCA

- Mathematically, we have

$$\begin{aligned}
\max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \sum_{i=1}^{n} (\mathbf{v}^\top \mathbf{x}_i)^2 &= \max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \sum_{i=1}^{n} \mathbf{v}^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v} \\
&= \max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \mathbf{v}^\top \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{v} \\
&= \max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \mathbf{v}^\top \underbrace{\left(\mathbf{X}^\top \mathbf{X} \right)}_{\mathbf{G}} \mathbf{v}
\end{aligned}$$

Math behind PCA: proof continues

- Let $d_1, \ldots, d_p$ denote the eigenvalues of the sample covariance matrix $\mathbf{G} = \mathbf{X}^T X$ with corresponding eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_p$.
- We have

$$\max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \mathbf{v}^\top \mathbf{G} \mathbf{v} = d_1, \quad \text{and}$$
$$\arg\max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \mathbf{v}^\top \mathbf{G} \mathbf{u} = \mathbf{v}_1$$

- Two ways to prove:
  1. Lagrange method of multipliers.
  2. Using SVD decomposition of the symmetric PSD matrix $\mathbf{G}$.

Math behind PCA: proof finishes

- It follows that we have

$$\max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \mathbf{v}^\top \mathbf{G}\mathbf{v} = \max_{\mathbf{v}:\|\mathbf{v}\|_2=1} \mathbf{v}^\top \mathbf{U}\mathbf{D}\mathbf{U}^\mathbf{T}\mathbf{v} = \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top \mathbf{D}\mathbf{w}$$

  where $\mathbf{w} = \mathbf{U}^T\mathbf{v}$, andbecause $\mathbf{U}$ is a rotation and L2 norm stays the same under rotation.

- Let $\mathbf{w} = (w_1, ..., w_p)^T$,
  $\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top \mathbf{D}\mathbf{w} = \max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \sum_{j=1}^p w_j^2 d_j$ which is maximized under the constraint $||w||_2 = 1$ when $w_1 = 1$ and $w_2 = ... = w_p = 0$.

- Hence
  $\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top \mathbf{D}\mathbf{w} = d_1$ and
  $\arg\max_{\mathbf{w}:\|\mathbf{w}\|_2=1} \mathbf{w}^\top \mathbf{D}\mathbf{w} = (1, 0, ..., 0)^T$ which corresponds to $\mathbf{u}_1$ the fist column of $\mathbf{U}$ since $w = \mathbf{U}^T v$ implying that the maximizing $v = \mathbf{U}(1, 0, ..., 0)^T = \mathbf{u}_1$.