# Inference and Representation: Factor Analysis and PCA

Rahul G. Krishnan

New York University

Lab 7, October 18, 2016

# Outline

- Technique to factorize a matrix
- Given a matrix $M \in \mathbb{R}^{m \times n}$, a singular value decomposition yields: $U \in \mathbb{R}^{m \times m}$ (orthogonal matrix), $D \in \mathbb{R}^{m \times n}$ (a rectangular diagonal matrix), $V \in \mathbb{R}^{n \times n}$ (orthogonal matrix) where $M = UDV^T$.
- We'll return back to this concept later when we talk about doing an eigendecomposition.

# PCA as transforming Gaussian Random Variables

- In lecture, you learned about one way to interpret PCA (Principal Component Analysis)
- If your data $X$ is Gaussian, then $\exists Y = AX$ is Gaussian and the covariance of $Y$ is given by: $A\Sigma_X A^T$
- We're interested in finding $A$ such that $A\Sigma_X A^T = \mathbb{I}$
- You can achieve this with an eigendecomposition of the covariance matrix.

# Another Interpretation

- What is the objective function for which eigendecomposition of the covariance matrix is the solution?
- If $\mathbf{u}_1 \ldots \mathbf{u}_M$ form an orthonormal basis ($\mathbf{u}_m^T \mathbf{u}_n = 1$, if $m = n$ and 0 otherwise)
- For any vector in $x_m \in \mathbb{R}^N$, we can write $x_i = \sum_{j=1}^{N} \alpha_{ij} \mathbf{u_j}$ and $\alpha_{ij} = \mathbf{u}_j^T x_i$
- Consider approximating a vector in that space with fewer than $M$ basis functions
- Use a $K$ term approximation to $x_i$ : $\hat{x}_i = \sum_{j=1}^{K} \alpha_{ij} \mathbf{u_j}$

# Minimizing Mean-Squared Error

Given data matrix $X \in \mathbb{R}^{M \times N}$, we are interested in the following optimization problem:

$$\min_{\mathbf{u}_1,\ldots,\mathbf{u}_K} \frac{1}{M} \sum_{i=1}^{M} ||x_i - \hat{x}_i||^2$$

$$\frac{1}{M} \sum_{i=1}^{M} ||x_i - \hat{x}_i||^2 = \sum_{i=1}^{M} || \sum_{j=k+1}^{N} \alpha_{ij} \mathbf{u}_j ||^2$$

$$= \frac{1}{M} \sum_{i=1}^{M} \sum_{j=k+1}^{N} \mathbf{u}_j^T x_i x_i^T \mathbf{u}_j$$

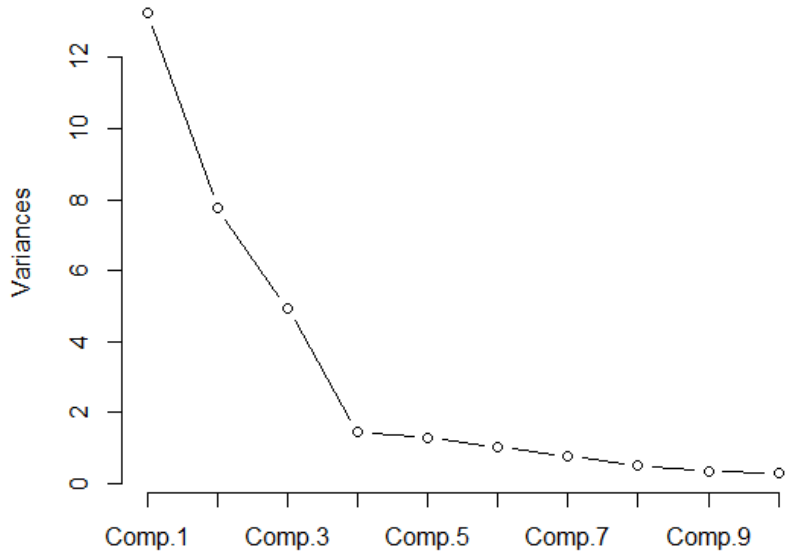
$$= \sum_{j=k+1}^{N} \mathbf{u}_j^T \hat{\Sigma} \mathbf{u}_j$$

# Implications

- Consider the eigenvalues of $\hat{\Sigma}$ (sorted by largest to smallest eigenvalues $\lambda_i$) as $\phi_1, \ldots, \phi_N$
- $\phi_N \hat{\Sigma} \phi_N = \lambda_N$ yields the smallest number among all orthonormal vectors
- To minimize the error the set $\{\mathbf{u}_i\}_{i=k+1}^{N} = \{\phi_{k+1} \ldots, \phi_N\}$
- Therefore: $\{\mathbf{u}_i\}_{i=1}^{K} = \{\phi_1, \ldots, \phi_K\}$

# Choosing Number of Components $K$

- Look at reconstruction error $\mathcal{E}(K) = \sum_{k=K+1}^{N} \lambda_k$
- Yet another interpretation: The eigenvalues correspond to the variance of the original data captured by the corresponding dimension
- Plot variance versus number of components
- Choose the smallest $K$ that captures most of the variance



# SVD instead of Eigendecomposition

- For PCA, we need to do an eigendecomposition of the emperical covariance matrix given data matrix $X \in \mathbb{R}^{M \times N}$
- We're interested in the eigenbasis of $\hat{\Sigma}_X = XX^T = Q\Lambda Q^{-1}$.
- SVD of $X = UDV^T$

$\hat{\Sigma} = XX^T$ (Substitute factorization for $X$)

$\hat{\Sigma} = UDV^T VDU^T$

($V^T V = \mathbb{I}$ (orthogonal matrix) and $DD = D^2$ (diagonal matrix))

$\hat{\Sigma} = UD^2U^T$

- The left singular vectors ($U$) of the data matrix $X$ are the eigenvectors

# Is this is probablistic model

- PCA is not a generative model. Why? We assume no noise from $Y$ to $X$
- Is there a probablistic variant of PCA

# Graphical Model

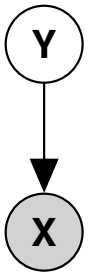



Figure: Factor Analysis and Probablistic PCA

# Probablistic PCA

- $y \sim \mathcal{N}(0, \mathbb{I})$
- $x \sim \mathcal{N}(Wy + \mu, \sigma^2 \mathbb{I})$
- The joint density $p(x, y)$, the marginal $p(x) = \int_y p(y)p(x|y)$, and the posterior $p(y|x)$ are all Gaussian
- $x \sim \mathcal{N}(\mu, WW^T + \sigma^2 \mathbb{I})$

# Relevance of PPCA

Why go through this trouble of defining a generative model that is similar to PCA when we can just use PCA?

- Understand other models as generalizations
- Can extend to mixtures of PPCA
- Handle missing data in a principled manner
- Reference: `https://www.microsoft.com/en-us/research/publication/probabilistic-principal-component-analysis/`

# Implications for PCA

- PCA = $\lim_{\sigma \to 0}$ PPCA
- From the view of the graphical model, it corresponds to having a model where the posterior covariance is zero.

# Factor Analysis

We can now view Factor Analysis as a generalization of PPCA.

- $y \sim \mathcal{N}(0, \mathbb{I})$
- $x \sim \mathcal{N}(Wy + \mu, \Psi)$
- Assume $\Psi$ is diagonal
- vs PPCA: Instead of a fixed variance for all dimensions, FA assumes that each dimension has its own variance.
- $W$ is known as the factor loading matrix

# Playing with PCA and FA

Ipython Notebook

# References


- http://pages.cs.wisc.edu/~jerryzhu/cs731/dim.pdf
- https://www.cs.toronto.edu/~hinton/csc2515/notes/lec7middle.pdf