Factor Analysis

COS 424/524, SML 302: Fundamentals of Machine Learning

Professor Engelhardt

COS 424/524, SML 302

Lecture 17

Dimension reduction: Factor analysis

Last lecture, we discussed principal component analysis (PCA) from three perspectives:

- greedily finding the directions of maximal variance in the data;
- finding a linear projection onto an orthogonal subspace minimizing reconstruction error;
- latent variable model for matrix factorization.

Today, we will discuss *factor analysis*, which generalizes the latent variable model of PCA.

Extensions to PCA and related methods

- Factor analysis: this lecture
- Bayesian PCA: Regularize with appropriate Bayesian priors
- Independent component analysis (ICA): non-Gaussian Z
- Canonical correlation analysis (CCA): multiple observations
- Latent Dirichlet allocation – next lecture
- Non-negative matrix factorization (NMF)
- Kernelized PCA: project observations to higher dimension
- Linear discriminant analysis (Fisher): PCA but includes class labels
- Sparse PCA: add sparsity in the weight matrix
- Nonlinear PCA: nonlinear projection to latent dimensions
- Many more...

Factor analysis introduction

Although factor analysis and PCA have almost identical latent variable models, they were developed for *different types of analyses*.

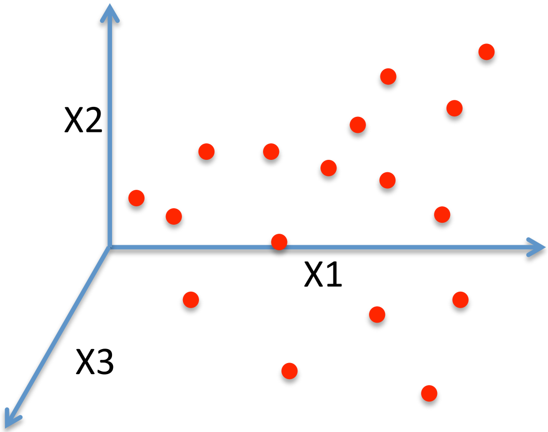
Factor analysis applications

- An exam question may test multiple topics, such as calculus, geometry, topology, or probability. Students' performance on each question is the matrix of observations (n students; p questions)
- A images has objects: sunset, tree, mouse, cat, etc. Pixels may be included in multiple features; (n images, p pixel features)
- A document may have multiple topics: economy, government, education, or sports. The counts of words in each document are the observed data; (n documents; p vocabulary)
- Gene expression levels may be a function of many underlying variables: sample age, sample cell type, sample batch (n samples, p genes).

Factor analysis (FA) as dimension reduction

Factor analysis is a method for *dimension reduction*

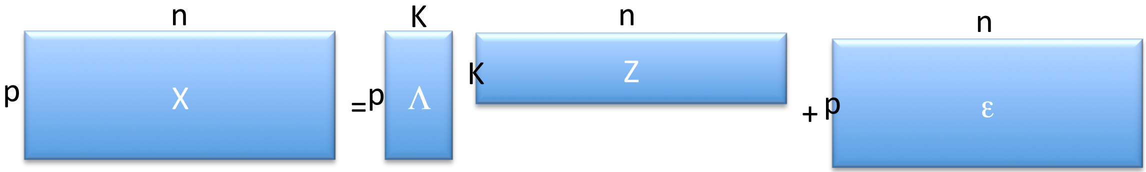
FA projects high dimensional data onto low dimensional linear subspace, assuming independent Gaussian noise.



Statistical model for FA

Let $X \in \Re^{p \times n}$ be the observed variables:

$$X = \Lambda Z + \epsilon$$

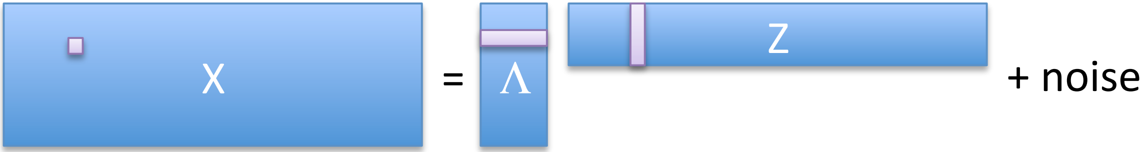


- $X \in \Re^{p \times n}$ is the observed data: n observations of p *features*;
- $\Lambda \in \Re^{p \times K}$ is *loadings matrix*: K-dim latent space within features;
- $Z \in \mathbb{R}^{K \times n}$ is *factor matrix*: projects n samples to K space;
- ϵ is Gaussian noise, $\epsilon_i \sim \mathcal{N}_p(0, \Psi)$ for $\Psi = diag(\psi_1, \ldots, \psi_p)$.
- Noise is in p space, not K space.

Statistical model for FA

We can rewrite this equation differently:

$$X_{j,i} = \sum_{k=1}^{K} \Lambda_{j,k} Z_{k,i} + \epsilon_{i,j}.$$

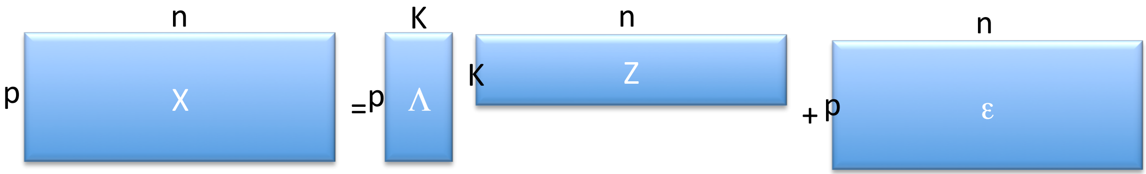


- Each observation $X_{j,i}$ is represented as the inner product of the loadings for feature j and the factors for sample i.
- Each feature j has its own variance term ψ_j. This is the only model difference between PPCA and FA.

Statistical model for probabilistic PCA

Let $X \in \Re^{p \times n}$ be the observed variables:

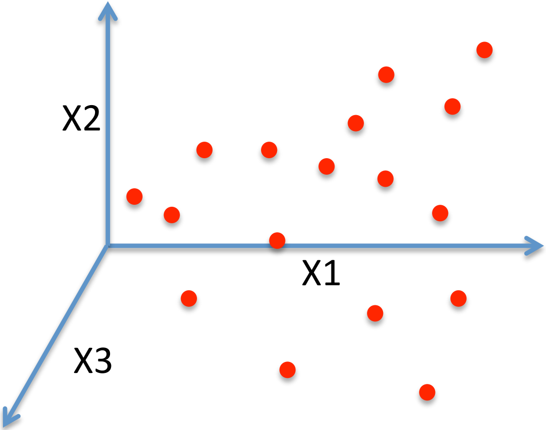
$$X = \Lambda X + \epsilon$$



- $X \in \Re^{p \times n}$ is the observed data: n observations of p *features*;
- $\Lambda \in \Re^{p \times K}$ is *loadings matrix*, weighting feature map to latent space
- $Z \in \mathbb{R}^{K \times n}$ is *factor matrix* projecting n observations to K space;
- ϵ is Gaussian noise, $\epsilon_i \sim \mathcal{N}_p(0, \Psi)$ for $\Psi = diag(\psi, \ldots, \psi)$.
- Noise is independent and identical across features.
- Principal components Z are an MLE solution to this model.

Factor analysis interpretation

Start with a set of $n = 17$ observations of $p = 3$ features.

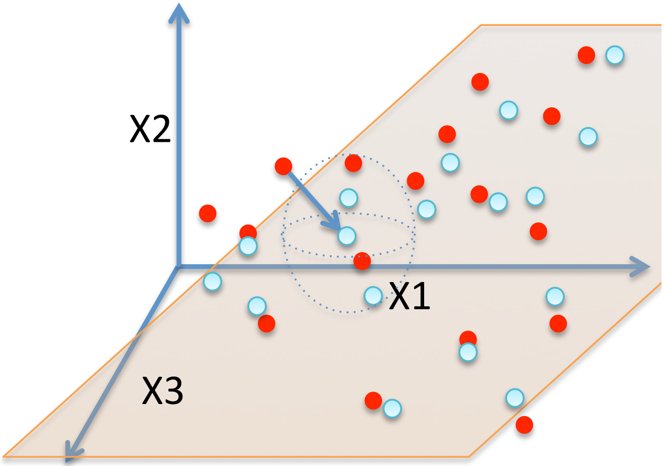


Factor analysis interpretation

Find a 2 dimensional hyperplane in this p space.

Project each observation onto $K = 2$ latent space.

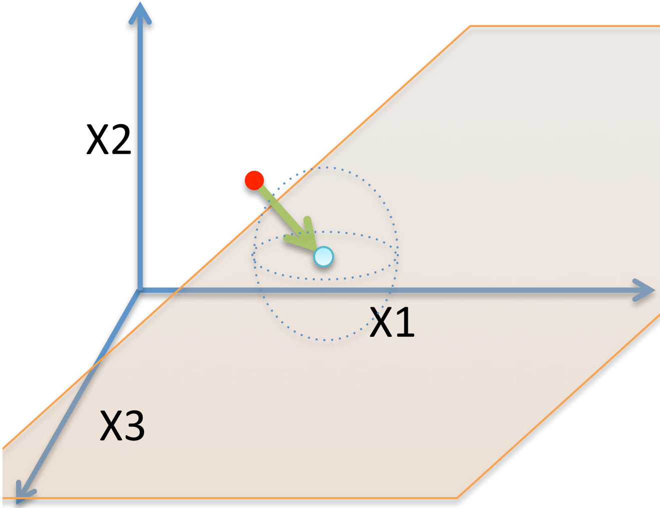
Note that: $X - \Lambda Z \sim \mathcal{N}(0, \Psi)$



Factor analysis interpretation

$$X = Z\Lambda + \epsilon$$

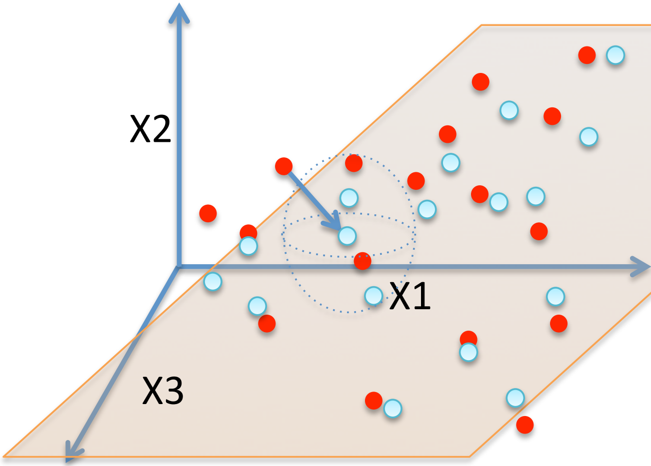
- Tan hyperplane represents Λ loadings
- Red point is observation x_i
- Blue point is $\hat{x}_i = \Lambda z_i$
- Green arrow is residual error ϵ_i when representing x_i as $\hat{x}_i$
- Dotted blue lines are Gaussian noise of projection



Factor analysis interpretation

- Latent hyperplane: feature dimensions that covary have correlated values in Λ_k.

- Latent hyperplane: feature dimensions with larger magnitudes in Λ_k contribute more to variance explained by that factor



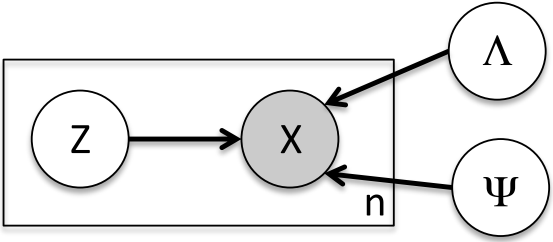
FA: statistical framework

Let's look at another interpretation: covariance estimation

First, let's write out the generative model for FA:

$$\begin{aligned} z_i &\sim \mathcal{N}_K(0, \Sigma) \\ x_i &\sim \mathcal{N}_p(\Lambda z_i, \Psi) \end{aligned}$$

(We may put prior distributions on Ψ, Λ)



FA: simplifying the framework

When $\Sigma = I$, we can integrate out z_i from $p(x_i, z_i \mid \Lambda, \Psi)$:

$$\int_{\mathcal{Z}} p(x_i|z_i, \Lambda, \Psi) p(z_i|\Sigma) dz_i = p(x_i|\Lambda, \Psi) = \mathcal{N}_p(x_i|0, \Lambda\Lambda^T + \Psi)$$

Note that the marginal distribution of x_i is Gaussian.

FA: low dimensional estimation of covariance

Implication of marginal Gaussian distribution for x_i is interesting:

$$cov[X|\Lambda, \Psi] = \Lambda\Lambda^T + \Psi$$

where:

- $\Lambda\Lambda^T$ models covariance structure of matrix X in dimension p
- Ψ models the variance in dimension p
- Ψ is not required to be diagonal
- when Ψ is diagonal, Λ recovers the covariance of the data matrix X
- What happens when Ψ is not diagonal?

Factor analysis is a low-dimensional estimate of covariance of X.

FA versus PCA

A note on the relationship between FA and PCA

- In FA, no need to mean-center original features: mean is absorbed in latent factors (with no prior on Λ)
- In FA, no need to standardize original features: differences in variance of features are explicitly modeled in Ψ
- PCA: orthogonal latent dimensions capture a disjoint proportion of variance in data; FA: PVE is not disjoint across factors
- PCA is often thought of as finding the highest variance dimensions; FA instead explicitly models covariance structure in features.

Key point: small changes in model can have profound changes on model interpretation.

Fitting FA with EM

We can fit FA model with expectation-maximization (EM)

- In the E-step we compute the posterior weights, $p(z_i \mid \mathbf{x_i}, \boldsymbol{\Lambda}, \boldsymbol{\Psi})$.
- In the M-step we re-estimate the parameters Λ, Ψ.

For details, see *[Ghahramani & Hinton 1998]*.

FA Expectation-Maximization

Key insight for deriving EM: The joint distribution of our observed and latent variables (X, Z) is a $p + K$ dimensional Gaussian. Why?

$$\begin{pmatrix} x_i \\ z_i \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Lambda\Lambda^T + \Psi & \Lambda \\ \Lambda^T & I \end{pmatrix}\right]$$

- I: $K \times K$ identity matrix
- Λ: $p \times K$ matrix
- Ψ is diagonal $p \times p$ matrix

Does diagonal Ψ enforce assumption that features are independent?

FA M-step

In the M-step, we re-estimate parameters Λ, Ψ, assuming expected values of latent variables $\langle z_i \rangle$ exist.

For the M-step, notice that

$$x_i = \Lambda \langle z_i \rangle + \epsilon_i,$$

where $\epsilon \sim \mathcal{N}_p(0, \Psi)$.

This is **linear regression**, where

- The response is x_i
- The covariates are $\langle z_i \rangle$ (the conditional expectations of z_i)
- The coefficients are Λ
- The residual covariance is Ψ

FA expected complete log likelihood

Expected complete log likelihood

$$
\begin{aligned}
\mathrm{E}[\log p(X, Z \mid \Psi, \Lambda)] &= \mathrm{E}\left[\log \prod_{i=1}^{n} (2\pi)^{p/2} |\Psi|^{-1/2} \right. \\
&\quad \left. \exp\left\{ -\frac{1}{2}[x_i - \Lambda z_i]^T \Psi^{-1} [x_i - \Lambda z_i] \right\} \right] \\
&= \mathrm{E}\left[\sum_{i=1}^{n} p/2 \log(2\pi) - 1/2 \log |\Psi| - \right. \\
&\quad \left. \left\{ \frac{1}{2} x_i^T \Psi^{-1} x_i - x_i^T \Psi^{-1} \Lambda z_i + \frac{1}{2} z_i^T \Lambda^T \Psi^{-1} \Lambda z_i \right\} \right] \\
&= \sum_{i=1}^{n} p/2 \log(2\pi) - 1/2 \log |\Psi| - \\
&\quad \left\{ \frac{1}{2} x_i^T \Psi^{-1} x_i - x_i^T \Psi^{-1} \Lambda E[z_i] + \frac{1}{2} \Lambda^T \Psi^{-1} \Lambda E[z_i z_i^T] \right\}
\end{aligned}
$$

FA M-step

MLE for Λ, we see that the M-step has the form of a posterior expectation solution to linear regression

$$\Lambda^{(t+1)} = \left(\sum_{i=1}^{n} \mathrm{E}[z_i z_i^\top \mid x_i] \right)^{-1} \left(\sum_{i=1}^{n} \mathrm{E}[z_i \mid x_i]^\top x_i \right).$$

Does this equation look familiar?

FA M-step

MLE for Λ, we see that the M-step has the form of a posterior expectation solution to linear regression

$$\Lambda^{(t+1)} = \left(\sum_{i=1}^{n} \mathrm{E}[z_i z_i^\top \mid x_i] \right)^{-1} \left(\sum_{i=1}^{n} \mathrm{E}[z_i \mid x_i]^\top x_i \right).$$

These are the normal equations substituting the expected sufficient statistics from the expected complete log likelihood.

$$\Psi^{(t+1)} = \frac{1}{n} diag \left(\sum_{i=1}^{n} x_i x_i^T - \Lambda \mathrm{E}[z_i \mid x_i]^\top x_i^T \right).$$

This equation is the empirical residual variance (with expectations).

Exercise: derive these updates.

FA E-step

In the E-step, we compute conditional expectations, $p(z_i \mid x_i, \Lambda, \Psi)$ to get the expected sufficient statistics.

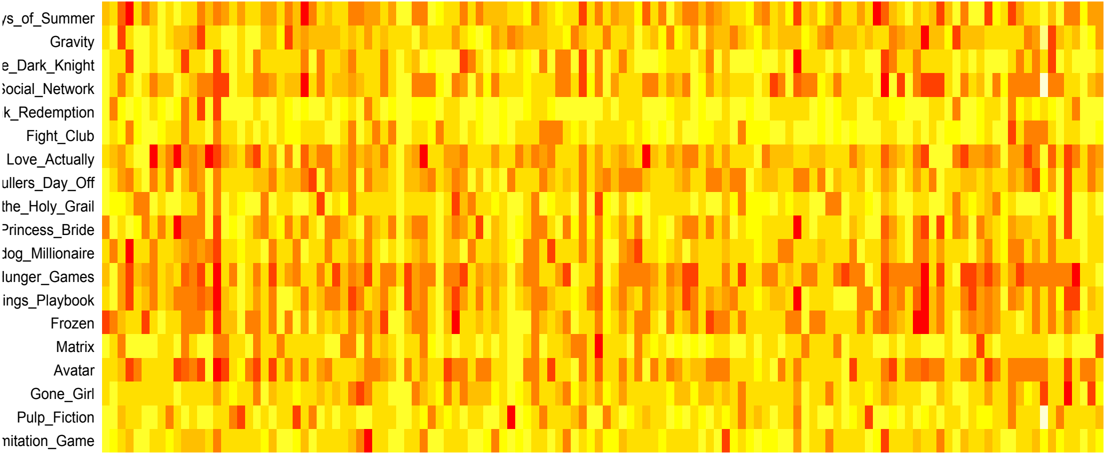
We know from last lecture that, because x_i and z_i are jointly Gaussian, we can compute these conditional expectations trivially:

$$\begin{aligned}
\mathrm{E}[z_i \mid x_i] &= \Lambda^T(\Lambda^T\Lambda + \Psi)^{-1}x_i \\
\mathrm{E}[z_i z_i^T \mid x_i] &= Var[z_i \mid x_i] + \mathrm{E}[z_i \mid x_i]\mathrm{E}[z_i \mid x_i]^T \\
&= I - \Lambda^T(\Lambda^T\Lambda + \Psi)^{-1}\Lambda + \mathrm{E}[z_i \mid x_i]\mathrm{E}[z_i \mid x_i]^T
\end{aligned}$$

What is the computational complexity of these computations?

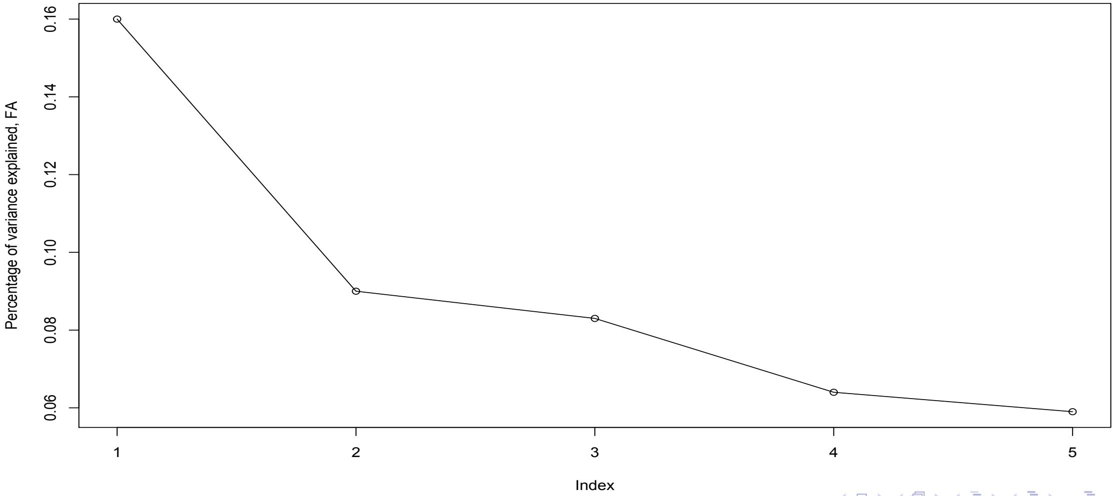
Example: Movie rating data

X represents $n = 127$ respondents' ratings for $p = 19$ movies. Let $K = 5$.



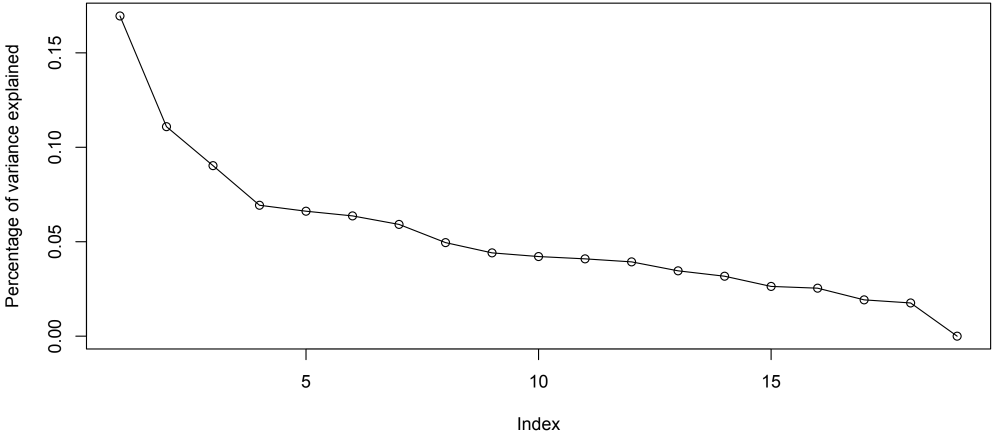
Example: Movie rating data

For $K = 5$ look at percentage of variance explained by each factor.



Example: Movie rating data

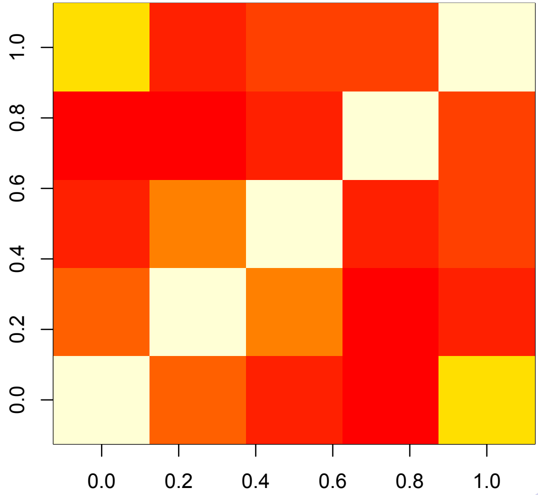
For $K = 5$ compare with PVE for PCA



Top PC/factor explain about the same variance; in FA the drop off is much steeper.

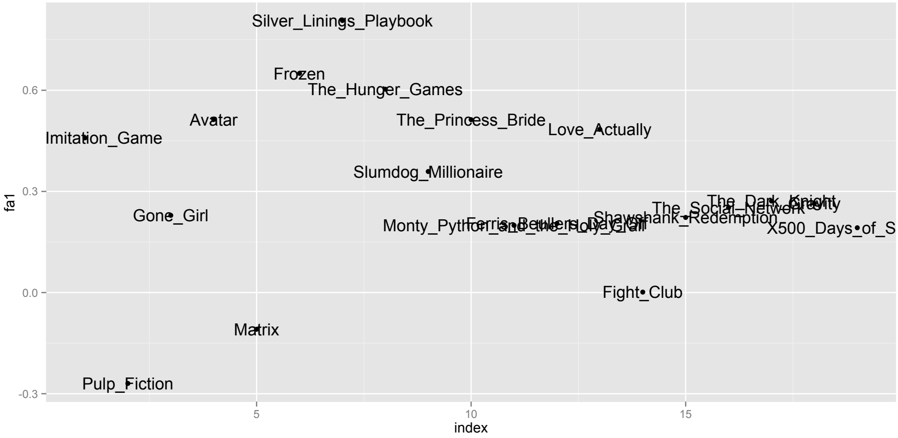
Example: Movie rating data

For $K = 5$, look at correlation across loadings



Example: Movie rating data

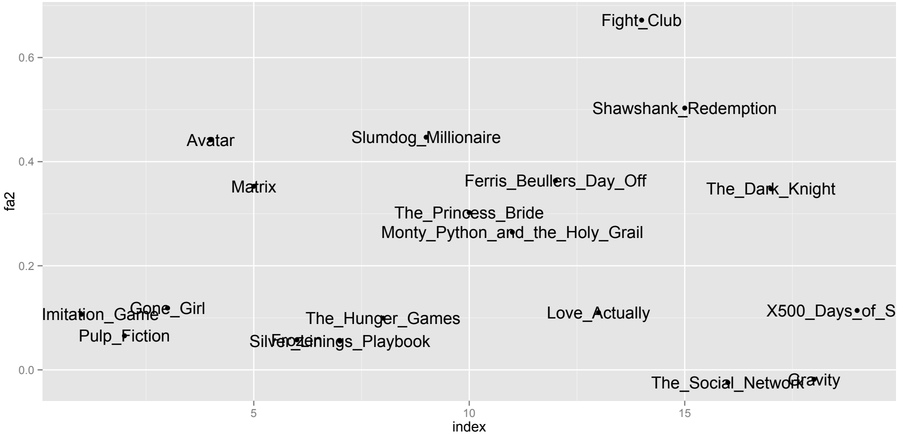
For $K = 5$, look each factor loading separately.



Magnitude of feature loading proportional to representation in the component of the subspace.

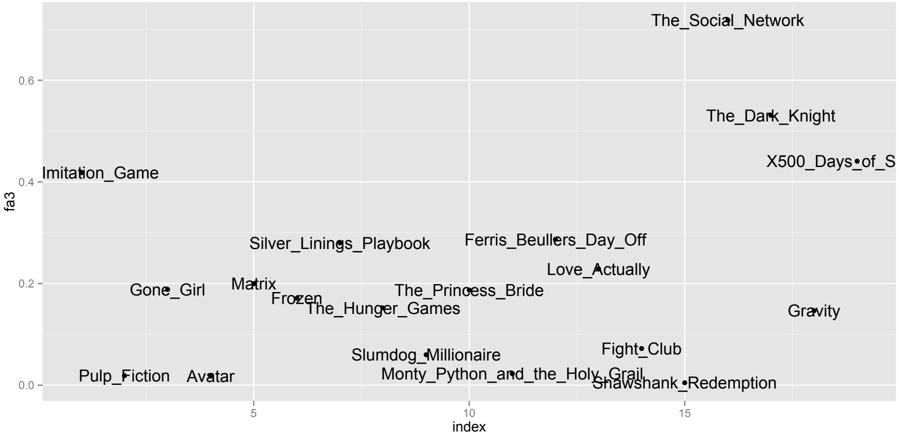
Example: Movie rating data

For $K = 5$, look each factor loading separately.



Example: Movie rating data

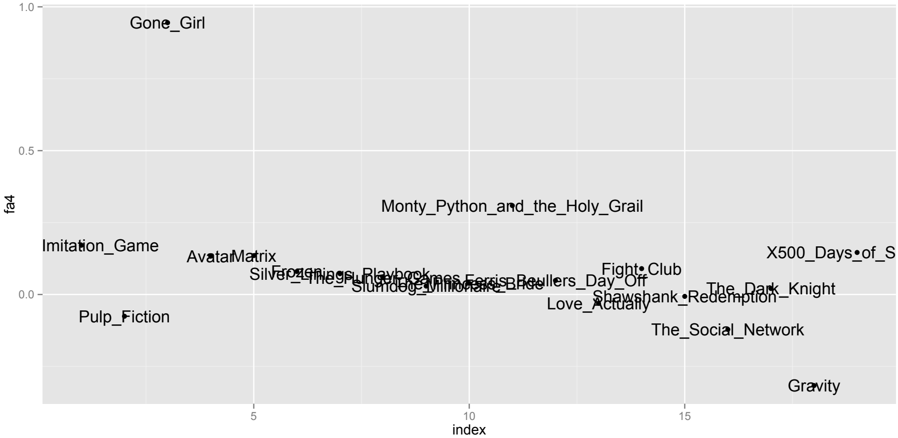
For $K = 5$, look each factor loading separately.



Somewhat correlated with factor 2.

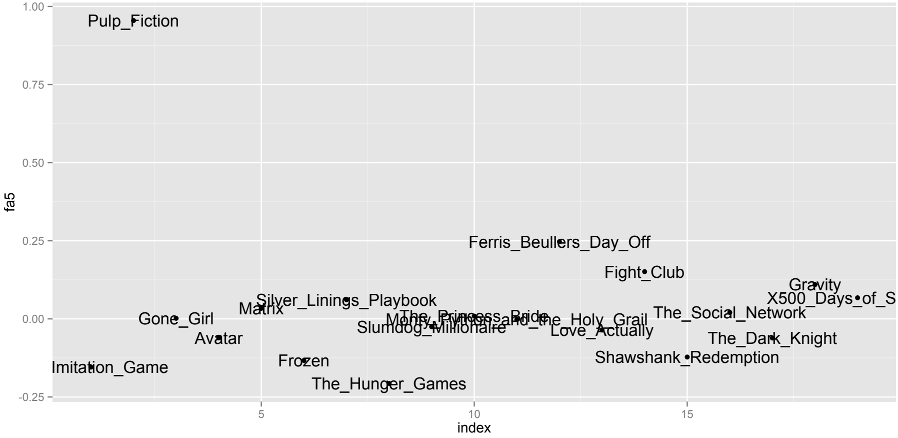
Example: Movie rating data

For $K = 5$, look each factor loading separately.



Example: Movie rating data

For $K = 5$, look each factor loading separately.



Well correlated with factor 1. Note: high-magnitude features help design factor "label."

Example: Interpreting movie ratings

Magnitude of $\Lambda_{j,k}$ is now relevance of movie j to factor k (use to give the factors "labels"):

Λ: captures linear structure that explains the deviation of movie ratings from the empirical mean rating

$z_{k,i}$ has the interpretation of the magnitude of the respondent i's ratings specific to factor k.

A large value in $z_{k,i}$ means subject i rated movies that contribute to this factor differently than the mean movie rating.

Non-identifiability and FA

Non-identifiability

Non-identifiability refers to the statistical situation where multiple different parameteriziations of a model produce identical data likelihoods:

$$p(x|\theta) = p(x \mid \theta').$$

The factor analysis model has three types of non-identifiability:

- orthogonal matrix rotation,
- scale,
- label switching.

Let's understand each identifiability problem, and consider solutions

Identifiability: Orthogonal rotation

Suppose that $\mathbf{R}$ is an arbitrary $k \times k$ orthogonal rotation matrix satisfying $\mathbf{R}\mathbf{R}^T = \mathbf{I}$, then define $\tilde{\boldsymbol{\Lambda}} = \mathbf{R}\boldsymbol{\Lambda}$, and also rotate the latent factors as:

$$\tilde{\mathbf{Z}} = \mathbf{Z}\mathbf{R}^T.$$

Then we have:

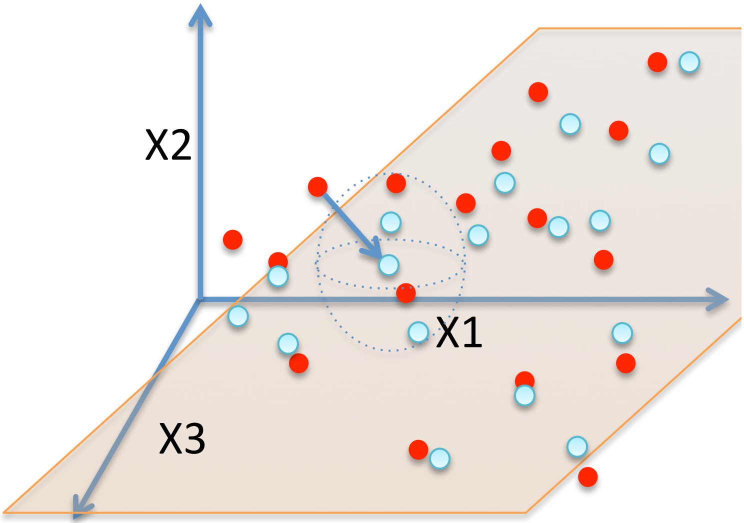
$$\tilde{\mathbf{Z}}\tilde{\boldsymbol{\Lambda}} = \mathbf{Z}\mathbf{R}^T\mathbf{R}\boldsymbol{\Lambda} = \mathbf{Z}\boldsymbol{\Lambda}.$$

The mean of x_i conditioned on latent factors z_i does not change under rotation of the Z and Λ matrices.

We cannot uniquely identify loadings and corresponding latent factors up to orthogonal rotation.

What does invariance mean with respect to specification of latent space? Interpretability?

Factor analysis: orthogonal rotation



Orthogonal rotation: effect on covariance estimation

Orthogonal rotation does not affect the covariance matrix estimate for $\mathbf{x}$:

$$\mathrm{cov}[\mathbf{x}] = \tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}^T + \Psi = \mathbf{\Lambda}\mathbf{R}\mathbf{R}^T\mathbf{\Lambda}^T + \Psi = \mathbf{\Lambda}\mathbf{\Lambda}^T + \Psi$$

After rotation, the covariance matrix of $\mathbf{x}$ does not change.

We are not able to identify the latent space $\mathbf{\Lambda}$ up to orthogonal rotation.

Solutions to identifiability to orthogonal rotation

- *Force* $\mathbf{\Lambda}$ *to be orthonormal.* PCA enforces a unique $\mathbf{\Lambda}$ and $\mathbf{Z}$. Comes at the possible cost of interpretability.
- *Force* $\mathbf{\Lambda}$ *to be lower triangular.* Force specific directions and, implicitly, orthogonality in the loadings matrix. But forced zeros may not recover "true" latent structure.
- *Choose an informative rotation matrix* $\mathbf{R}$. One approach (**varimax**) chooses rotation that forces the most loading elements to zero.
- *Put sparse priors on* $\mathbf{\Lambda}$. Sparse priors such as ℓ_1 regularization, automatic relevance determination, or a spike-and-slab prior improves interpretability and effectively clusters features in loadings.
- *Use non-Gaussian priors for the latent factors.* Choosing non-Gaussian priors on $\mathbf{z_i}$ help to uniquely identify $\mathbf{\Lambda}$ and $\mathbf{Z}$; this is called Independent Component Analysis (ICA).

Non-identifiability with respect to scale

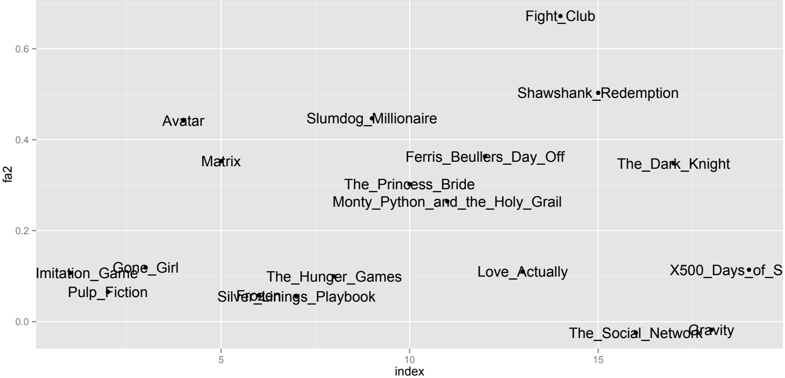
We could simultaneously scale the factor loadings and the latent factors by constant α.

$$\mathbf{z}_j \mathbf{\Lambda}_i = \left(\mathbf{z}_j \frac{1}{\alpha}\right) (\mathbf{\Lambda}_i \alpha)$$

This produces another non-identifiability.

Solutions to scale

- *Force* $\mathbf{\Lambda}$ *to be orthonormal (or normalize* $\mathbf{\Lambda}$*).* Orthonormality resolves the issue of scale, also explicit in PCA.
- *Avoid consideration of the loadings or factors in side-by-side comparison.* Think instead of within-factor magnitude.



Non-identifiability due to label switching

We may have two different orderings of the rows of $\boldsymbol{\Lambda}$ and their corresponding columns of $\mathbf{z}$:

$$\sum_{k=1}^{K} \mathbf{z}_{j,k} \boldsymbol{\Lambda}_{i,k} = \sum_{k'=1}^{K'} \mathbf{z}_{j,k'} \boldsymbol{\Lambda}_{i,k'}$$

$$K \in \{2,3,1\} \qquad K' \in \{1,2,3\}$$

But their inner product is the same.

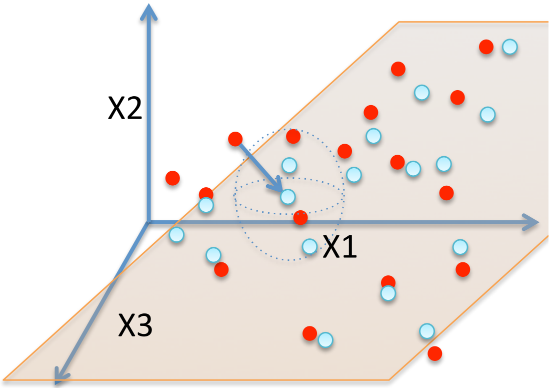
This is known as the *label-switching problem*

Solutions to label-switching non-identifiability

- *Put a prior on percentage of variance explained by each factor.* Methods will produce a fairly robust ordering of latent factors.
- *Avoid matching factors and loadings directly across runs.* Compare instead based, for example, on covariance matrix estimates between the two sets of latent dimensions.
- *Explicitly match factors and loadings across runs.* Find factors and loadings that are best correlated across runs to match them.

Sparse factor analysis

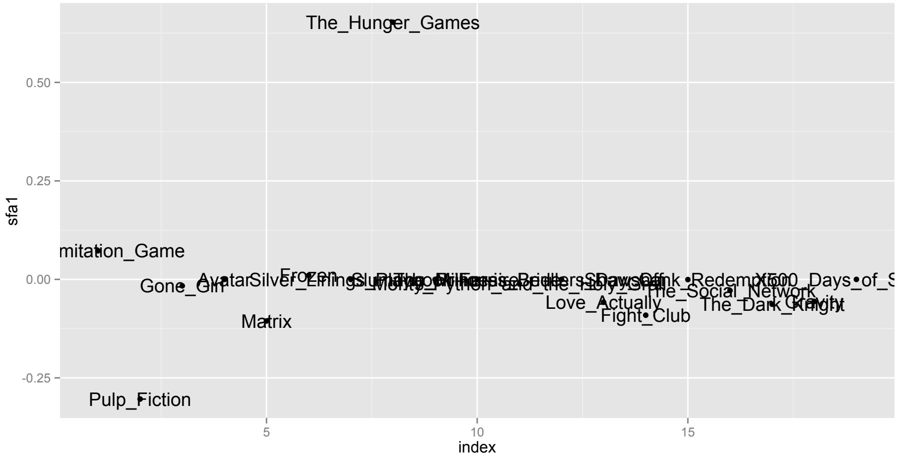
- We can include a sparse prior on the loadings matrix Λ
- Sparsity is often a solution to the problem of rotational invariance
- Sparsity also adds another level of interpretability to the lower dimensional space.



What is the effect on the latent space of sparse Λ?

Example: Movie rating data

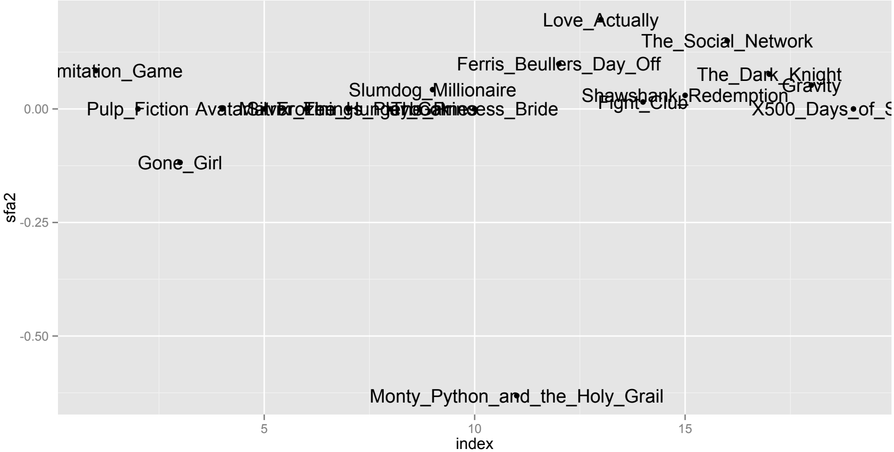
For $K = 18$, look each sparse loading separately.



This factor captures the anti-correlation in ratings about *Pulp Fiction* and *The Hunger Games*.

Example: Movie rating data

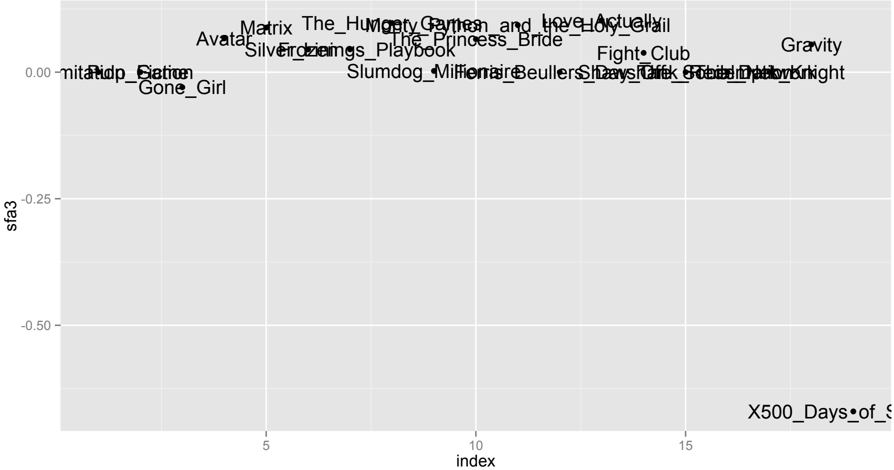
For $K = 18$, look each sparse loading separately.



Factor captures the variance due to *Monty Python and the Holy Grail* ratings.

Example: Movie rating data

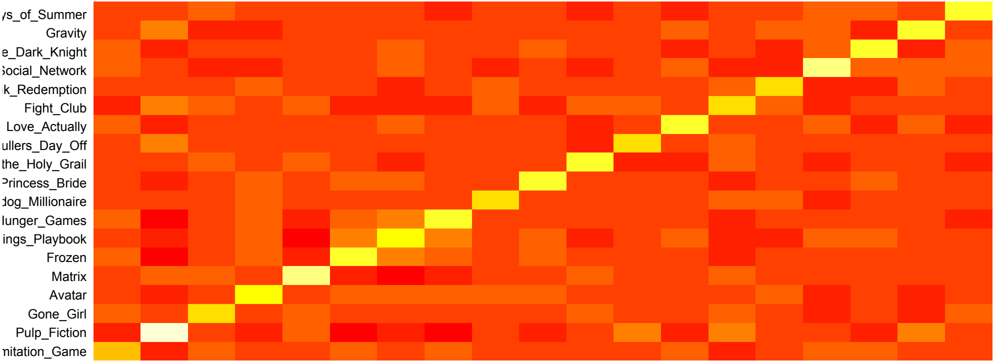
For $K = 18$, look each sparse loading separately.



This factor captures the variance due to *500 Days of Summer* ratings.

Example: Movie rating data

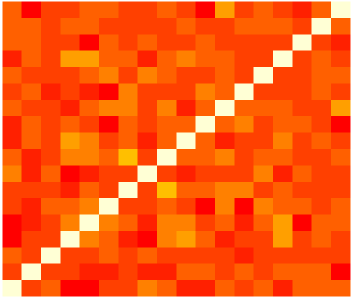
For $K = 18$, look at $\Lambda^T \Lambda$ for an estimate of the covariance matrix (diagonals incorrect).



Example: Movie rating data

How is SFA different than FA?

- non-disjoint movie clusters: movies with non-zero values on loading k.
- substantial sparsity in underlying data: loadings matrix will not rotate with random EM restarts.
- need many more factors to explain same variance



Factor analysis: summary

Main assumptions of FA:

- Assumes data are jointly Gaussian
- Assumes residuals are uncorrelated
- Assumes latent space is low-dimensional, linear

Limitations of FA:

- Likelihood invariant to scale, rotation, label switching.
- Interpretation is manual and weak.
- Sensitive to local optima (EM) and choice of K.
- No analysis or interpretation of causality

When to use PCA vs FA *[Brown 2009]*:

- use PCA when the goal of the analysis is to explore patterns in data
- use factor analysis when relationships between features exist

Factor analysis: extensions

- Many different types of sparse FA
- exploratory FA vs confirmatory FA
- non-linear latent space
- non-parametric priors on number of factors
- different assumptions of distribution of data
- imposing orthogonality on factor loadings
- many more...

Additional Resources

- MLAPA: Chapter 12
- MacKay: Chapter 34 (Independent Component Analysis and Latent Variable Modeling)
- *[Ghahramani & Hinton 1996]*. Derivation of EM for factor analysis.
- *[Roweis & Ghahramani 1999] A Unifying Review of Linear Gaussian Models*
- *[Cunningham & Ghahramani 2014] Unifying linear dimensionality reduction*
- *[Brown 2009] Principal components analysis and exploratory factor analysis—Definitions, differences and choices*

- *Metacademy*: Factor Analysis and Principal Components Analysis