Factor Analysis

Slides adapted from the course material on Unsupervised Machine Learning - University of Helsinki, Finland

Generative vs Discriminative Models

Generative vs Discriminative

- Generative
  - is a model of observable data values, given some parameters
  - Model joint distribution of the data
  - ‘Generative’, as sampling can generate synthetic data points
- Discriminative
  - is a model dependency of an unobserved variable on an observed one.
  - Directly estimate conditional posterior probabilities
  - Focus resources on given task

Popular Models

- Generative
  - Mixtures of Gaussians, Mixtures of experts
  - Sigmoidal belief networks, Bayesian networks
  - Markov random fields, Hidden Markov Models (HMM)
  - Factor Analysis, Generative Adversarial Networks,
- Discriminative
  - SVM, Logistic regression
  - Traditional neural networks, Nearest neighbor
  - Conditional Random Fields (CRF)

Generative Process

In several applications, the data measurements can be hypothesised to have been generated from an underlying process that may not be measurable itself.

The measurements;

- are typically high-dimensional; and
- may contain correlated variables,
- which are additionally corrupted with noise.

Latent Variables

Latent

- Latent
  - hidden, unobserved, unmeasured, hypothetical, not directly measurable.
- Latent Variable
  - A latent variable is a variable which is not directly observable and is assumed to affect the observed variables.

Latent Variables

One of the key features of machine learning is to learn latent (hidden) representations of the data:

- that summarise the data, and
- are low-dimensional.
- The learned data summaries can then be hypothesized as the descriptions of the phenomenon that have generated the data.

These summaries are then commonly used to understand the mechanisms of the data generation process, or then to predict the unseen outcomes.

Latent Variable Models

The distribution of high-dimensional (data) variables $\mathbf{x} \in \mathcal{R}^D$, can be expressed by small number of latent (hidden) variables $\mathbf{z} \in \mathcal{R}^K$.

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = p(\mathbf{z}) \prod_{d=1}^{D} p(\mathbf{x}_d|\mathbf{z})$$

where $K << D$, and each $\mathbf{z}_k$ follows a simple distribution.

Why to use?

- The complexity of the problem demands it
- Reveal underlying truth (e.g. ‘discover’ latent types)
- Represent effects of unobserved factors and unobserved heterogeneity between subjects.

Debate on Latent Variables

- Latent variable models
  - can handle measurement errors
  - provide parsimonious summarization
  - describe heterogeneity
- Open questions revolve around
  - Existence of latent variables
  - Uniqueness (identifiability)
  - Estimability (comparable fit of different models)

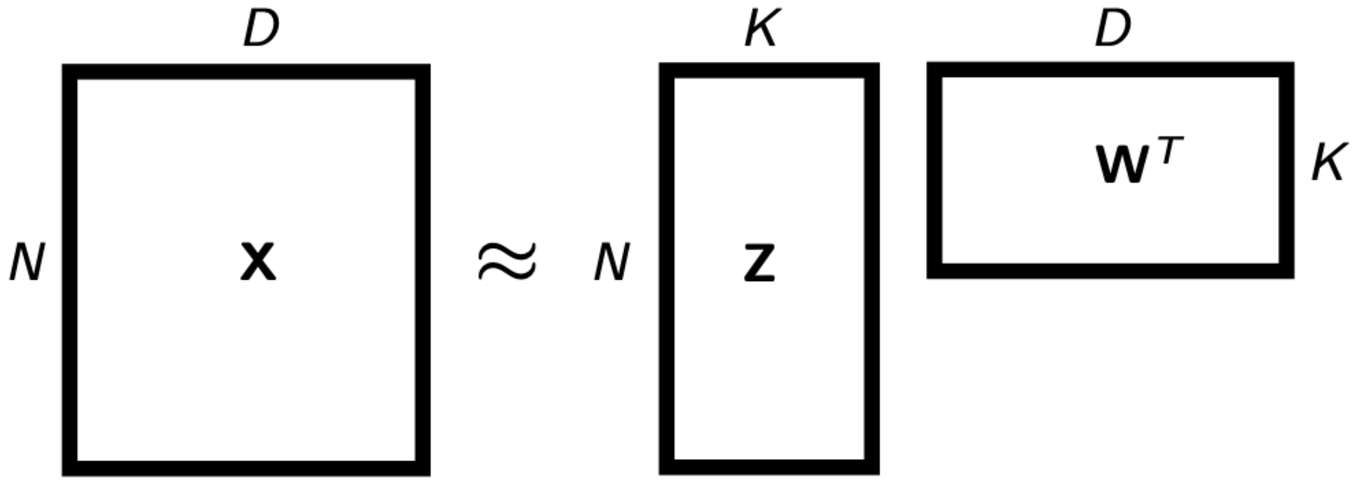
Matrix Factorisation

Matrix Factorisation

Matrix factorisation can be characterised as an unsupervised latent-variable approach for learning low-dimensional representation of a single matrix.

- Factor Analysis
- Principal Component Analysis
- Latent Dirichlet allocation (Topic Models)
- Independent Component Analysis
- and others..

Factor Analysis: Matrix Factorisation



$$\mathbf{X} \sim \mathbf{Z}\mathbf{W}^{\mathsf{T}}$$

Given a matrix $\mathbf{X} \in \mathcal{R}^{N \times D}$ we can factorise it into a product of two smaller matrices $\mathbf{Z} \in \mathcal{R}^{N \times K}$ and $\mathbf{W} \in \mathcal{R}^{D \times K}$

Linear Transformation

Matrix factorisation

- identifies a low-dimensional representation which can be linearly transformed to the original data: Z.
- identifies a new and simpler coordinate system through a set of linearly independent basis vectors: W.

Factor Analysis

Factor Analysis: Latent Variable Model

- FA attempts to explain the correlation between a large set of visible variables (x) in terms of a small number of hidden factors (z).

$$\mathbf{x}_n = \mathbf{W}\mathbf{z}_n + \boldsymbol{\epsilon}_n$$

- It is not possible to observe the factors directly. The variables depend on the factors but are also subject to random error.
- FA is one of the central tools in statistical analysis.

Factor Analysis: Latent Variable Model

Linear generative model:

$$\begin{aligned}\mathbf{x}_n =& \mathbf{W}\mathbf{z}_n + \boldsymbol{\epsilon}_n \\ \mathbf{x}_n \sim& N(\mathbf{W}\mathbf{z}_n, \Sigma) \\ \mathbf{z}_n \sim& N(0, \mathbf{I}) \ ,\end{aligned}$$

where $\mathbf{Z}$ are

- K-dimensional zero-mean unit-variance multivariate Gaussian vectors,
- independent latent variables aka *factors*, *components* or *scores*,
- form the low-dimensional representation of $\mathbf{X}$, i.e. $K < D$ and usually $K << D$.

Factor Analysis: Latent Variable Model

Linear generative model:

$$\begin{aligned}\mathbf{x}_n &= \mathbf{W}\mathbf{z}_n + \boldsymbol{\epsilon}_n \\ \mathbf{x}_n &\sim N(\mathbf{W}\mathbf{z}_n, \Sigma) \\ \mathbf{z}_n &\sim N(0, \mathbf{I}) \ ,\end{aligned}$$

- $w_{d,k}$ are projection weights aka factor loadings and indicate how strongly each factor contributes to each variable.
- Σ is a diagonal noise covariance matrix, with a separate term $\sigma_1^2, \ldots, \sigma_D^2$ for each of the D variables, and explain the part of the data not explained by the factors.

Factor Analysis: Covariance

By integrating out the factors one can see that

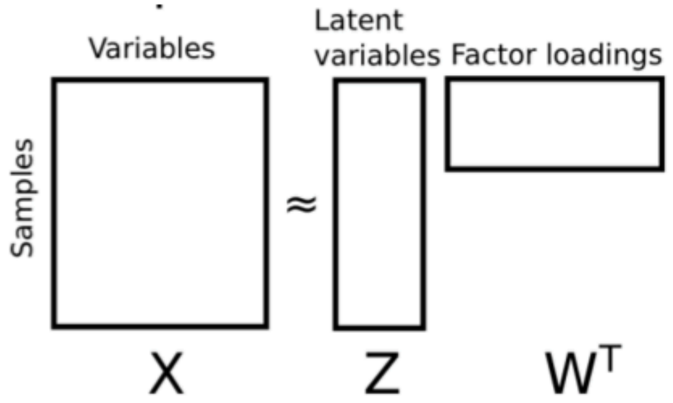
$$p(\mathbf{X}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})dz = N(0, \mathbf{W}\mathbf{W}^{\mathsf{T}} + \Sigma)$$

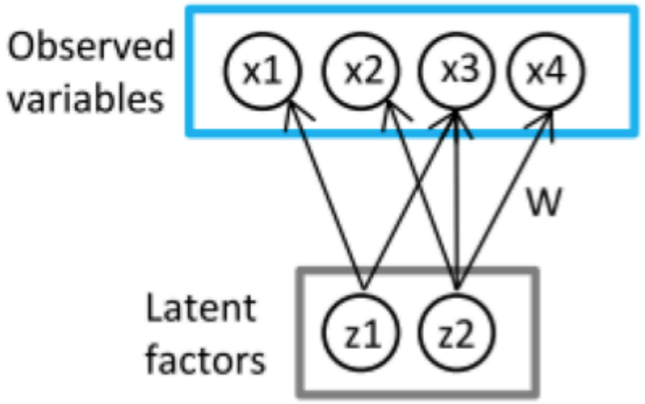
By choosing $K < D$, factor analysis makes it possible to model a Gaussian density for high dimensional data without requiring $O(D^2)$ parameters.

Means, K most significant correlations can be captured while still ensuring that the total number of parameters grows only linearly with D.

Factor Analysis

In factor analysis the observed variables are modeled using unobserved latent factors:





$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \qquad \mathbf{x} \sim \mathcal{N}(\mathbf{z}\mathbf{W}^T, \mathbf{\Sigma})$$

Factor Rotation Problem

Consider $\mathbf{R}$ to be any orthogonal matrix, then $\mathbf{R}\mathbf{R}^\intercal = \mathbf{I}$ and:

$$\begin{aligned}\mathbf{x}_n =& \mathbf{W}\mathbf{z}_n + \boldsymbol{\epsilon}_n \\ =& \mathbf{W}(\mathbf{R}\mathbf{R}^\intercal)\mathbf{z}_n + \boldsymbol{\epsilon}_n = (\mathbf{W}\mathbf{R})(\mathbf{R}^\intercal\mathbf{z}_n) + \boldsymbol{\epsilon}_n = (\hat{\mathbf{W}})(\hat{\mathbf{z}}_n) + \boldsymbol{\epsilon}_n \ ,\end{aligned}$$

where
$\mathbb{E}(\hat{\mathbf{Z}}) = \mathbf{R}.\mathbb{E}(\mathbf{Z}) = 0$, and
$cov(\hat{\mathbf{Z}}) = \mathbf{R}^\intercal cov(\mathbf{Z})\mathbf{R} = \mathbf{R}^\intercal\mathbf{R} = \mathbf{I}$,
i.e latent vectors in $\mathbf{Z}$ are still independent, as $\mathbf{R}$ was orthogonal.

Therefore, there are an infinite number of possibilities for $\hat{\mathbf{W}}$ and $\hat{\mathbf{z}}_n$, the solution is non-unique, and the exact values can not be interpreted directly. This is known as rotational ambiguity.

How to fix rotational ambiguity?

Simplify the structure

- Idea is to constrain the structure of the factor loading matrix W. Can be done as a post-processing step or built into the model.

Constrain the statistical properties of the factors

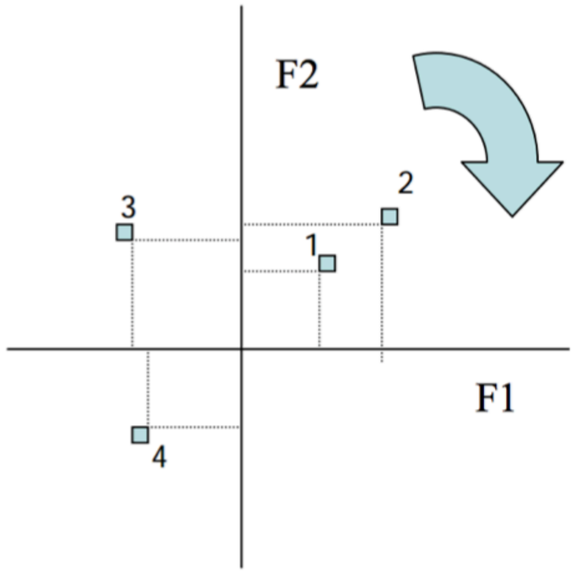
- For example, under certain conditions non-negative and non-gaussian factors can be uniquely identified. Key example approaches are NNMF and ICA.

Factor Rotations

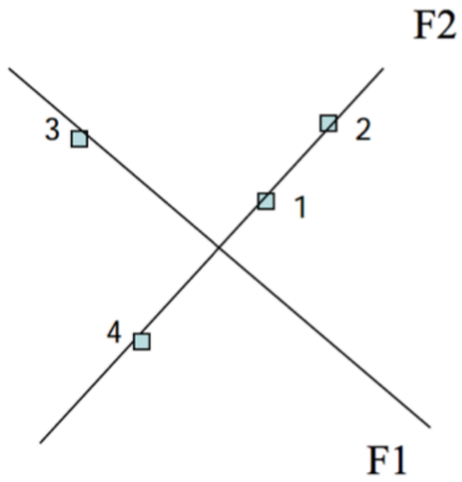
What's the idea?

- Identify a rotation that transforms the factor loadings into a simpler one, i.e. makes them easier to interpret.
- After rotation:
  - each factor should have nonzero, or significant loadings, for only some of the variables (but not too few!!).
  - each variable should have nonzero, or significant loadings, in only a few factors, if possible in only one.

Factor Rotations



	Factor 1	Factor 2
x1	0.5	0.5
x2	0.8	0.8
x3	-0.7	0.7
x4	-0.5	-0.5



	Factor 1	Factor 2
x1	0	0.6
x2	0	0.9
x3	-0.9	0
x4	0	-0.9

Factor Rotations

varimax

- Maximize squared loading variance across variables.
- Minimises the number of variables with high loadings in each factor.
- Tends to produce factors with few variables in each.

$$\arg\max_{\mathbf{R}} \sum \left((\mathbf{WR})^2_{d,k} - \overline{(\mathbf{WR})^2_{d,k}} \right)^2$$

where mean is over D.

Factor Rotations

quartimax

- Maximize squared loading variance across factors.
- Minimises the number of factors a variable loads on.
- Tends to produce a general factor and factors that have smaller groups of variables.

$$\arg\max_{\mathbf{R}} \sum \left((\mathbf{WR})^2_{d,k} - \overline{(\mathbf{WR})^2_{d,k}} \right)^2$$

where mean is over K.

Model Selection: Number of Parameters

- The number of factors or rank of the matrix depends on the application.
- $1 \leq K \leq min(N, D)$
- Ideally, we want $K \ll min(N, D)$, however, large enough to capture the interesting factors.
  - Do not want a too large K, difficult for interpretation.
  - Factors represented by 1-variable are meaningless.

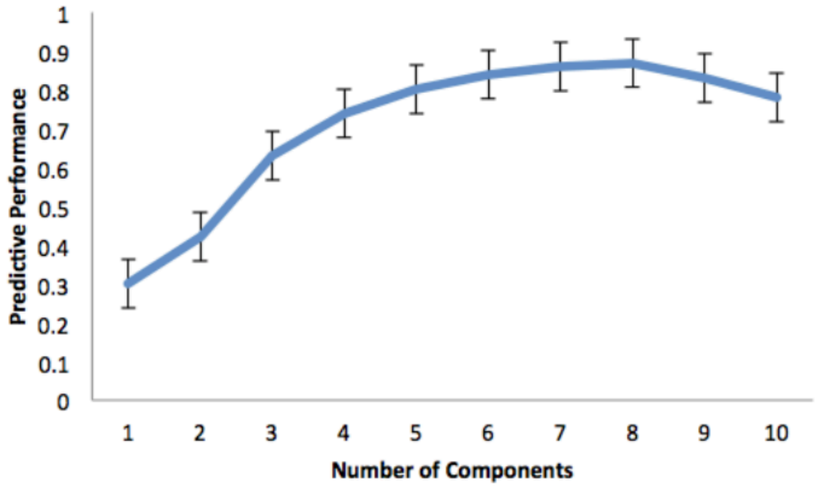
Model Selection: Number of Factors

- Iterative approaches (Cross-validation, Bayesian model selection, …
- Automatic relevance determination (ARD)
  - shrink unneeded aspects of the model such that they have no impact.
- Nonparametric methods
  - assume infinite number of dimensions with diminishing importance.
  - avoids selection of any fixed dimension (in principle)

Model Selection: Number of Factors

- Iterative approach
  - Iterate K at several positions between 1 and min(N,D).
  - Learn the factor analysis model (cross-validate/bootstrap)
  - Evaluate a performance metric (such as: Predictive Likelihood on test data, BIC, AIC).
  - Select the K with best performance/asymptote.

- Pros: works well in practice, especially for predictive applications.
- Cons: cost prohibitive for large N or D.



Factor Analysis: Interpretation

- **Goal**: describe the underlying processes that have generated the data. Assumption: Factors have been made interpretable.
- In FA, measurements are thought to have been generated from a combination of multiple latent processes (factors), each generating some parts of the data.
- The interpretation task then is to describe the underlying (hopefully meaningful) processes, based on the variables that load on them, and the latent variables of representative samples.

Factor Analysis: Interpretation

A factor can be interpreted in terms of the variables that load high on it, with respect to other variables as well as other factors.

- Select the non-zero variables.
- Significantly extreme than the prior distribution.
- Plot the variables, using the factor loadings as coordinates. Variables at the end of an axis are those that have high loadings on only that factor, and hence describe the factor.

FA: Identifying Dietary Patterns, Venkaiah et. al. 2011

Factors:

1. income-elastic (Milk, Sugar)
2. plant
3. traditional (pulses and spices)
4. micronutrient (leafy)
5. aquatic
6. protein

Food-group (g)	Adult men component						Adult women component					
	1	2	3	4	5	6	1	2	3	4	5	6
Cereals and millets	-0.32	0.36	0.35	0.35	-0.26	-0.04	-0.28	0.19	0.49	-0.16	0.38	0.04
Pulses and legumes	0.33	0.14	0.54	-0.11	-0.41	0.00	0.38	0.02	0.62	-0.23	-0.03	-0.08
Green-leafy vegetables	0.04	-0.24	-0.10	0.77	0.00	0.10	0.01	-0.22	-0.19	-0.03	0.78	0.06
Roots and tubers	0.05	0.72	-0.01	0.06	0.21	0.06	0.05	0.74	-0.12	0.21	0.06	0.05
Other vegetables	0.13	0.69	-0.06	-0.04	-0.22	-0.05	0.06	0.70	0.12	-0.29	-0.08	-0.06
Nuts and oilseeds	0.18	-0.15	0.30	0.04	0.11	-0.67	0.15	-0.11	0.28	0.17	-0.06	-0.53
Condiments and spices	-0.15	-0.09	0.69	-0.02	0.19	0.02	-0.17	-0.06	0.56	0.36	-0.03	0.08
Fruits	0.02	0.26	0.04	0.66	0.01	-0.09	0.07	0.17	0.18	0.05	0.57	-0.08
Fish and other sea-foods	-0.02	0.02	0.08	-0.04	0.81	-0.08	-0.02	-0.01	-0.05	0.80	-0.02	-0.10
Meat and poultry	0.15	-0.09	0.28	0.03	0.02	0.71	0.14	-0.06	0.20	0.07	-0.08	0.83
Milk and milk products	0.74	0.14	-0.09	-0.05	0.06	0.07	0.72	0.15	-0.11	0.03	-0.06	0.08
Fats and oils	0.32	0.43	0.34	0.07	0.39	0.21	0.32	0.45	0.23	0.46	0.11	0.14
Sugar and jaggery	0.80	0.03	0.05	0.10	-0.10	-0.08	0.80	-0.01	0.09	-0.02	0.10	-0.05
Eigen value	1.94	1.35	1.22	1.09	1.09	1.02	1.87	1.28	1.21	1.13	1.03	1.03
Variance explained	14.96	10.39	9.38	8.40	8.38	7.84	14.41	9.84	9.29	8.70	7.95	7.79

Factor Analysis: Applications

- Describing the generative process:
  - clusters samples and variables into homogeneous sets
  - factors allow gaining insight into the mechanisms
  - Ex: low-dimensional factors are used to represent the biological processes driving the mechanisms that generate different cellular response patterns
- Dimensionality reduction:
  - identifies groupings to allow us to select one variable to represent many
  - Ex: useful in visualisation or down stream analysis such as regression
- Missing value prediction
  - when some values of the matrix are unobserved the latent variables can be used to predict those. Ex: Netflix user-movie rating predictions, Drug-Target interactions.

Important Points

- Factor analysis model explains correlations between variables using latent variables (the factors) that affect several observed variables simultaneously, thus explaining the observed correlations
- FA model can be represented both with and without latent variables
- Factor loading matrix can be rotated without changing the likelihood, this must be kept in mind when interpreting the factors, but does not matter for prediction.

Summary

- Latent Variables
- Matrix Factorisation
- Factor Analysis
  - Latent Variable Model
  - Rotation Problem
  - Model Selection
- Applications