Advanced probabilistic methods

Lecture 8: Factor analysis

Pekka Marttinen

Aalto University

March, 2023

Lecture 8 overview

- Factor analysis (FA)
- Probabilistic formulation of the FA model
- Intuition, usage
- Extensions
- Suggested reading: Ch. 21 of Barber

Two different views on classical multivariate analysis[1]

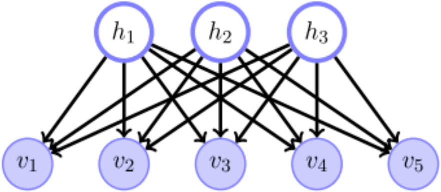
Given an $N \times D$ data matrix, we may be interested in comparing

1. rows of the data matrix (**individuals**)
    - starting point: similarities between individuals
    - techniques: **clustering**, multidimensional scaling, discriminant analysis
2. columns of the data matrix (**variables**)
    - starting point: correlation/covariance matrix between variables
    - techniques: **factor analysis**, principal component analysis, canonical correlation analysis

[1] From Mardia, K.V. (1980). Multivariate Analysis

Factor analysis - intuition

- Factor analysis attempts to explain correlation between a large set of visible variables (**v**) using a small number of hidden factors (**h**).
- It is not possible to observe the factors directly. The visible variables depend on the factors but are also subject to random error.
- A central tool in statistics, a simple example of **representation learning**, and a building block for more complex (deep) models.



Factor analysis, probabilistic description

- FA model generates a D-dimensional observation $\mathbf{v}$ from the H-dimensional vector $\mathbf{h}$ according to

$$\mathbf{v} = F\mathbf{h} + \mathbf{c} + \epsilon,$$

where

$$\epsilon \sim N(0, \Psi), \qquad \Psi = \mathrm{diag}(\psi_1, \ldots, \psi_D).$$

- The $D \times H$ *factor loading matrix* F tells how the factors affect the observations: f_{ij} is the effect of factor h_j on variable v_i.

Factor analysis (example) (1/3)

- Data matrix contains results of 5 exams for 120 students (see *factorandemo* in Matlab)
  - Exams 1 and 2 are about mathematics, exams 3 and 4 about literature, and exam 5 is comprehensive.
- Goal of analysis: to investigate if the results could be understood using a smaller number of characteristics (or, factors) of students, e.g., 'quantitative' and 'qualitative' skills.

$$\text{Data} = \begin{bmatrix} 65 & 77 & 69 & 75 & 69 \\ 61 & 74 & 70 & 66 & 68 \\ \dots & \dots & \dots & \dots & \dots \end{bmatrix}$$

- The n^{th} row of the data matrix is $v_n^T = (v_{n1}, \dots, v_{n5})$

Factor analysis (example) (2/3)

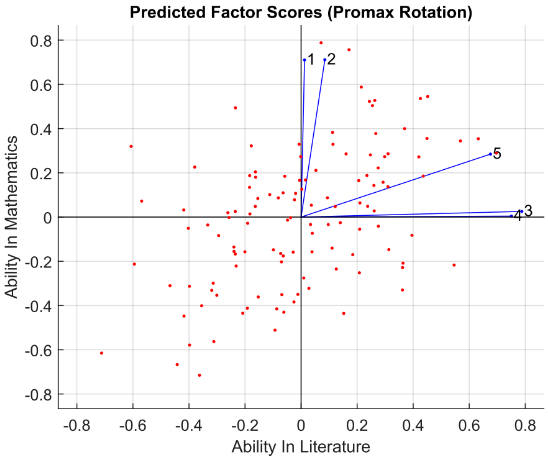
- Underlying model in detail

$$\begin{bmatrix} v_{n1} \\ v_{n2} \\ v_{n3} \\ v_{n4} \\ v_{n5} \end{bmatrix} = \begin{bmatrix} f_{11} & f_{12} \\ f_{21} & f_{22} \\ f_{31} & f_{32} \\ f_{41} & f_{42} \\ f_{51} & f_{52} \end{bmatrix} \times \begin{bmatrix} h_{n1} \\ h_{n2} \end{bmatrix} + \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix} + \epsilon_n$$

$$\epsilon_n \sim N_5(0, \Psi), \ \Psi = \begin{bmatrix} \psi_1 & 0 & 0 & 0 & 0 \\ 0 & \psi_2 & 0 & 0 & 0 \\ 0 & 0 & \psi_3 & 0 & 0 \\ 0 & 0 & 0 & \psi_4 & 0 \\ 0 & 0 & 0 & 0 & \psi_5 \end{bmatrix}$$

Factor analysis (example) (3/3)

- Results



$$\widehat{F} = \begin{bmatrix} 0.01 & 0.71 \\ 0.08 & 0.71 \\ 0.79 & 0.03 \\ 0.75 & 0.00 \\ 0.68 & 0.28 \end{bmatrix}$$

Equivalent model without latent factors

- Given

$$p(\mathbf{v}|\mathbf{h}) = N_D(\mathbf{v}|F\mathbf{h} + \mathbf{c}, \Psi)$$

and assuming a prior on $\mathbf{h}$:

$$p(\mathbf{h}) = N_H(\mathbf{h}|0, I),$$

integrating out $\mathbf{h}$ yields

$$p(\mathbf{v}) = \int p(\mathbf{v}|\mathbf{h})p(\mathbf{h})d\mathbf{h} = N(\mathbf{v}|\mathbf{c}, FF^T + \Psi)$$

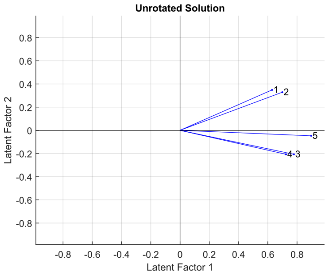
- The result follows from the *Linear transformation* of a Gaussian (see Lecture 3).

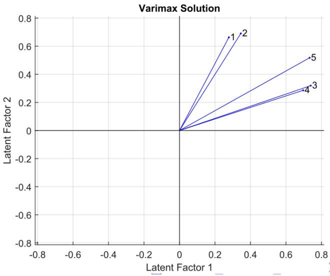
Rotation invariance

- The likelihood is unchanged if we rotate F using FR, with $RR^T = I$:

$$FR(FR)^T + \Psi = FRR^T F^T + \Psi = FF^T + \Psi.$$

- R is often selected to produce interpretable factors. *Varimax* rotation makes each column of F to have only a small number of large values.
- Note: rotation invariance does not matter if the goal is to fit the model in order to use it for prediction. For interpreting the factors, it does.





Probabilistic principal component analysis

- Probabilistic PCA has almost same the model as FA

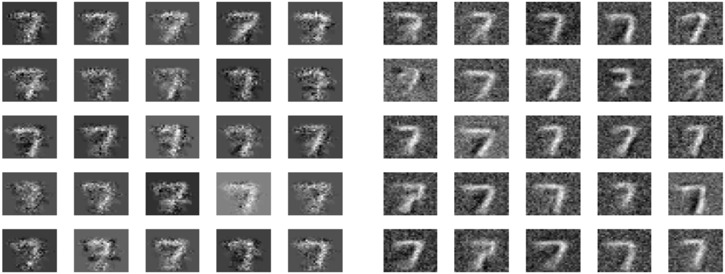
$$\mathbf{v} = F\mathbf{h} + \mathbf{c} + \epsilon,$$

$$\epsilon \sim N(0, \Psi), \qquad \Psi = \sigma^2 I.$$

- In FA

$$\Psi = \mathrm{diag}(\psi_1, \dots, \psi_D).$$

Example PPCA and FA, digit modeling

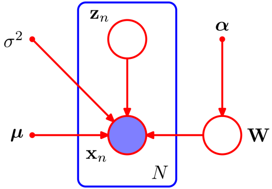


(a) Factor Analysis (b) PPCA

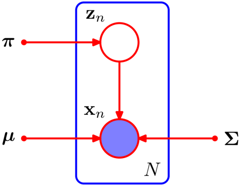
- Samples drawn from FA and PPCA models trained for digit 7.
- FA has different noise parameters for each pixel $\rightarrow$ reduced noise in boundary regions.

FA vs. GMM

- How are FA and GMM similar? How are they different?



Bayesian PCA (Bishop, Fig. 12.13)



GMM (Bishop, Fig. 9.6)

FA, geometric intuition (1/2)

- FA assumes that the data lies close to a low-dimensional linear manifold
- For example, if $H = 1$ and $D = 2$:

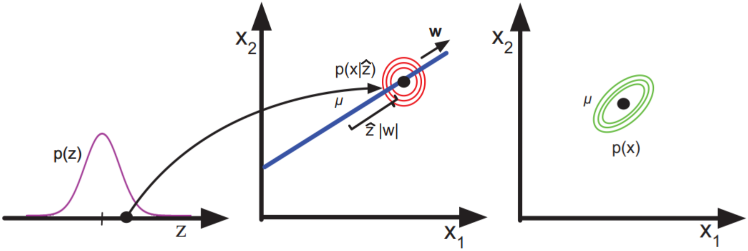


Figure: 12.1 in Murphy

FA, geometric intuition (2/2)

- If $H = 2$ and $D = 3$, the data points form a 'pancake'

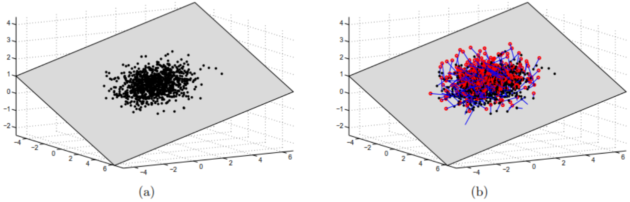


Figure: 21.2 in Barber

- Left: latent 2D points $\mathbf{h}_n$ sampled from $N(\mathbf{h}|\mathbf{0}, \mathbf{I})$ and mapped to the 3D plane by $\mathbf{v}_n^0 = F\mathbf{h}_n + \mathbf{c}$.
- Right: data points $\mathbf{v}_n$ are obtained by adding noise $\mathbf{v}_n = \mathbf{v}_n^0 + \epsilon_n$, where $\epsilon_n \sim N(\mathbf{0}, \Psi)$

Fitting the FA model

- EM algorithm
- Mean-field VB straightforward with conjugate priors (left as an exercise)
- Stochastic variational inference (next week)
- MCMC
- etc.

Determining the number of factors

- Same techniques as for determining the number of clusters in GMMs
  - Bayesian model selection
  - Cross-validation
  - ...
- Automated relevance determination (ARD)
  - **shrink unneeded aspects** of the model, such that they have no impact
  - empty clusters in GMM (corresponding to mixture weights driven to zero)
  - factors that don't have any effect (apply a shrinkate prior on the columns of the factor loading matrix)
- Nonparameteric methods
  - Assume infinite number of dimensions with diminishing importance
  - Avoids the selection of any fixed dimension (in principle)
  - Dirichlet process prior for clustering, Beta process prior for factor analysis

Remark

- FA-model is based on the Gaussian distribution, but often used with other data types as well.
- Pragmatic justification that FA often works well with other data types.
- Performance may not be good with highly non-Gaussian variables, for example binary 0-1 variables with a very small number of individuals with value 1.

Extension: a mixture of factor analysers*

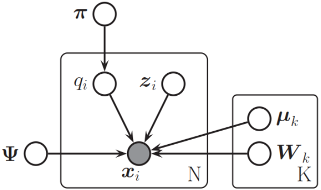


Figure: 12.3 in Murphy

Left: $K = 1$, right, $K = 10$:

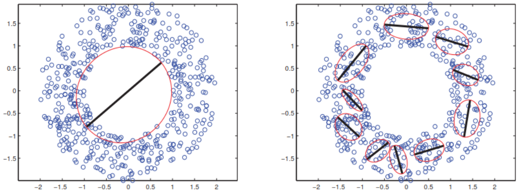
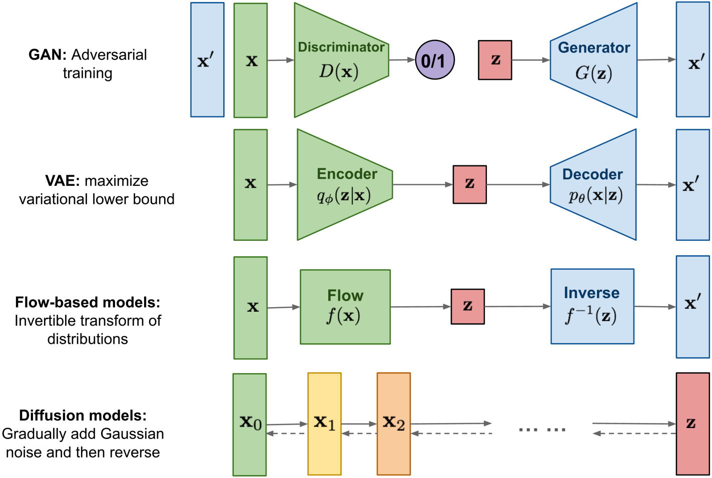


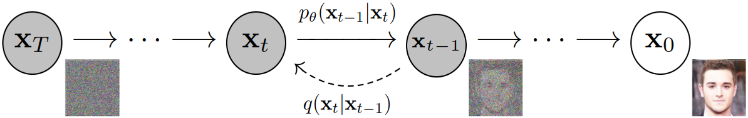
Figure: 12.4 in Murphy

Other latent variable models*



https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Other latent variable models: Diffusion*



Ho et al. (2020). https://arxiv.org/abs/2006.11239

- Incrementally add noise using a known $q(x_t|x_{t-1})$.
- Learn to remove noise by approximating $p_\theta(x_{t-1}|x_t)$ with variational inference.

Dall-E 2: Machine learning professor in front of a class, surrealist.

Important points

- Factor analysis model explains correlations between variables using latent variables (the factors) that affect several observed variables simultaneously.
- 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.
- FA model can be extended in many ways, and latent variable models are an important tool in modern ML.