Introduction

http://www.analytictech.com/mb876/handouts/nb_eigenstructures.htm
(http://www.analytictech.com/mb876/handouts/nb_eigenstructures.htm)

Factor analysis is not a single technique but a family of methods for analyzing a set of observed variables. Two basic branches in family tree: defined factors (aka principal components) and inferred factors (aka common factor analysis or classical factor analysis).

In principal components, we define new variables (factors), which are linear combinations of our observed variables, that summarize our input data. The focus is on expressing the new variables (the principal components) as weighted averages of the observed variables. In common factor analysis, we infer the existence of latent variables that explain the pattern of correlations among our observed variables. The focus is on expressing the observed variables as a **linear combination** of underlying factors.

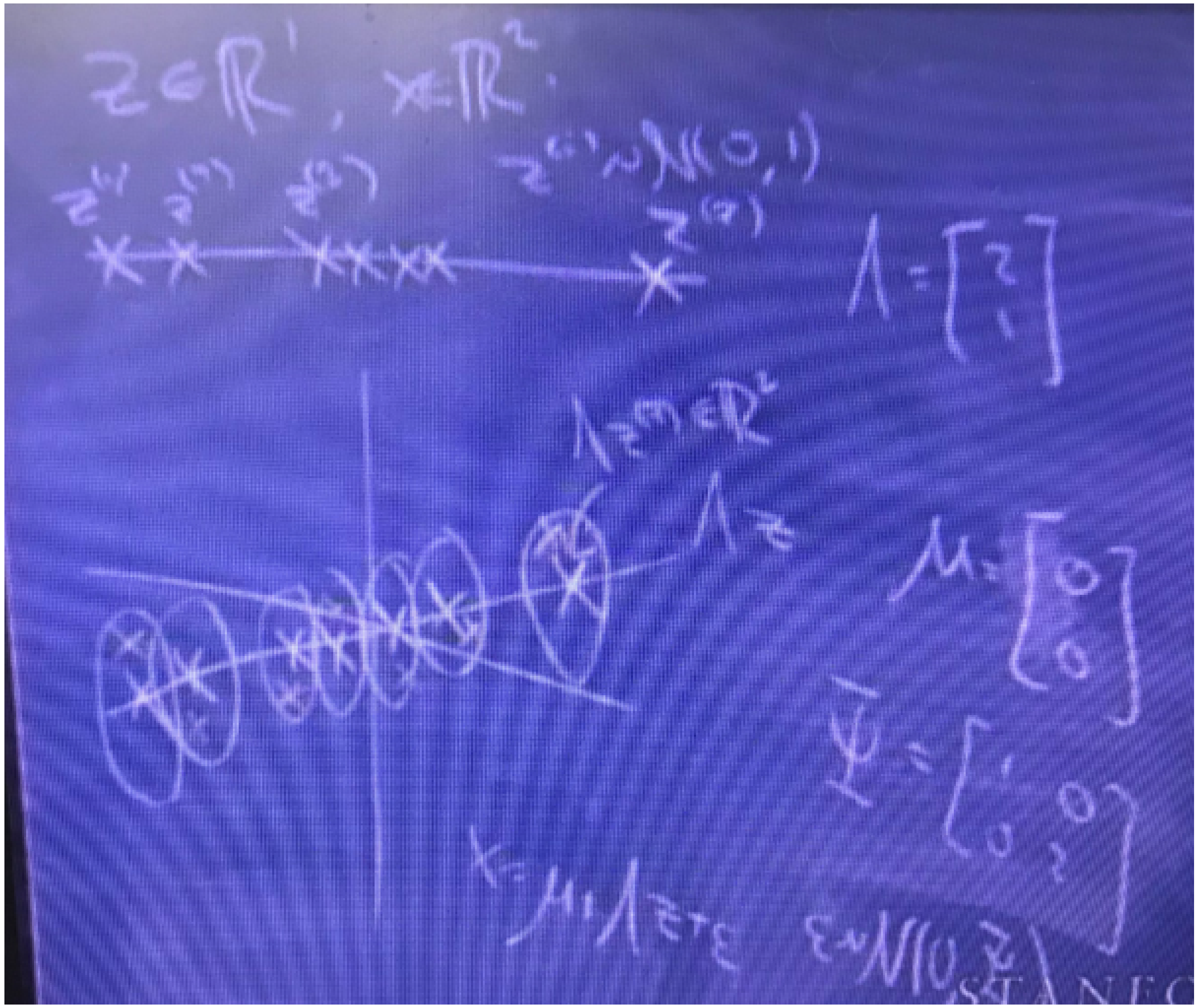
As above, the factor analysis is a more general concept as compared to PCA. However, factor analysis usually refers to the common factor analysis.

Derivation of factor analysis (FA)

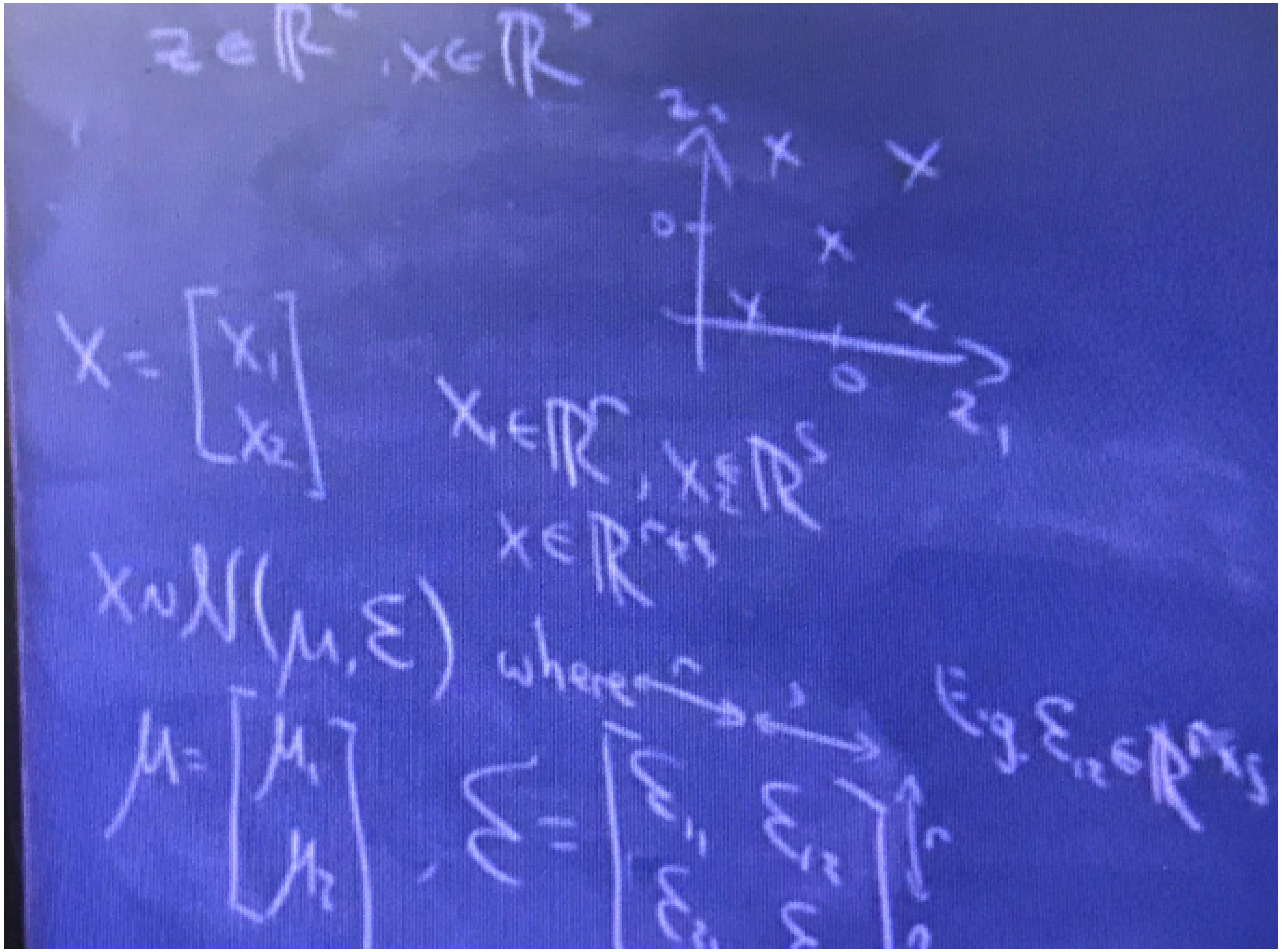
http://cs229.stanford.edu/notes/cs229-notes9.pdf (http://cs229.stanford.edu/notes/cs229-notes9.pdf)
The derivation of factor analysis is similar to other unsupervised algorithm such as mixture of Gaussians and mixture of naïve Bayes. In FA we just use a different data model from either a Gaussian or Bayes models, **but the general process of using expectation maximization (EM) is same.** FA is usually less popular than mixture of Gaussians and mixture of naïve Bayes. The example introduced in the link above handles a problem with continuous latent variable z. which different from the discrete random variables in the model of mixtures of Gaussians.

FA is useful to handle the case of $n >> m$ where the Σ in Gaussian mixture will be singular and we cannot write out the Gaussian densities.

Two cartoons show the factor analysis model as below. The figure below provides a 1D Gaussian continuous variable $x \mid z \sim N(\mu + \lambda z, \Psi)$. This is EQUIVALENT to $x = \mu + \lambda z + \epsilon$, where $\epsilon \sim N(0, \Psi)$. The generated data will be like in the second part of the figure. Here $\mu = 0$ is for simplicity.



The figure below assumes z is 2D continuous variable and first generate data on the (z_1, z_2) plane in the blackboard. The λz then tilts plane out of the board, and then add Gaussian noise. So factor analysis model generate data in the 3D space, but data are mainly on the 2D plane.



From the two figures above, we know FA models the data as coming from low-dimensional space with some noise. Here we assumes that the data is generated from a low-dimensional latent variable z belonging to d-dimension where $d < n$.

Key points in derivation of FA

To have a clear idea on how FA works, compare the derivation of mixtures of Gaussians by general EM theory [http://cs229.stanford.edu/notes/cs229-notes8.pdf (http://cs229.stanford.edu/notes/cs229-notes8.pdf)](http://cs229.stanford.edu/notes/cs229-notes8.pdf) and the derivation of FA in [http://cs229.stanford.edu/notes/cs229-notes9.pdf (http://cs229.stanford.edu/notes/cs229-notes9.pdf)](http://cs229.stanford.edu/notes/cs229-notes9.pdf). In fact, their main difference is the different form of Q_i.

In mixtures of Gaussians,

$$Q_i(z_i = j) = w_j^{(i)} = p(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)} \mid z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} p(x^{(i)} \mid z^{(i)} = l; \mu, \Sigma) p(z^{(i)} = l; \phi)}$$

where we don't know the explicit form of $Q_i(z_i = j)$. Instead we use Bayes rules to write $Q_i(z_i = j)$ in terms of likelihood and marginals. However, in FA, we assume a model that explicitly has the closed form of $Q_i(z_i)$. In the particular model we used, the random variable z_i is also continuous. Therefore FA is for the cases we already have an idea on the data.

In M-step, because z is continuous,we need maximize an integral quantity. In a step of integral over $Q(z)zdz$, it is just expectation value of z, and therefore we don't have to plug the detailed $Q(z)$ to it. See the lectures of photos.

In the generating formula $x = \mu + \lambda z + \epsilon$, we know that x and z are correlated. So it is better than the models by constraining the covariance matrix to be diagonal as in the beginning of the chapter, where correlation is all lost.

PCA vs factor analysis

Sometimes factor analysis is a general concept which include PCA.

https://mathoverflow.net/questions/40191/the-difference-between-principal-components-analysis-pca-and-factor-analysis (https://mathoverflow.net/questions/40191/the-difference-between-principal-components-analysis-pca-and-factor-analysis)

https://stats.stackexchange.com/questions/1576/what-are-the-differences-between-factor-analysis-and-principal-component-analysi (https://stats.stackexchange.com/questions/1576/what-are-the-differences-between-factor-analysis-and-principal-component-analysi)