Chapter 3

Analytical and iterative factor modeling

Modeling a process which generated a given multivariate time series dataset for the purpose of learning motivates this thesis. In Chapter 2, the essential time series analysis tools that are needed in the modeling were presented; whereas in this chapter the elements of building the model itself will be discussed. Models with parameters that could be tuned to fit the statistical characteristics of the dataset at hand will be chosen; this tuning is called **parametric estimation**. It may be seen as a limitation because it warrants assumptions on the type of the data generation process involved. However, essential precautions will be taken by modeling on a dataset that is representative enough of the process. Moreover, in order to avoid any overfitting, learning methods which caution when wide deviations from the assumptions on the model and data are detected will be used.

The treatment of this chapter from the rest of the thesis has one main difference; here, any temporal correlation of the data samples in the given dataset is ignored. Yet, later on in the thesis, the parametric modeling techniques presented herein with the time series techniques of Chapter 2 will be utilized to achieve the thesis objectives.

In Section 3.1, a well-founded modeling strategy based on the **principle of maximum likelihood** will be introduced [96]. The principle assumes that the data has been generated by a known class of probability distributions whose parameters are to be estimated such that the **likelihood** of observing the data is maximized.

In Section 3.2, the concept of a **linear model** whose parameters are linear combinations of the data samples will be introduced. The derivation of the optimal parameters will be summarized by the **Gauss-Markov theorem**. The ideas of **unbiased** and **efficient** parameters defined there are desirable properties for any parametric model.

In Section 3.3, the **factor model** is presented. While the linear model of Section 3.2 utilizes some **measured variables** of the dataset to explain themselves or other measured variables, the factor model is remarkably different. The latter assumes existence of a fewer number of unmeasured **latent variables** responsible for generating all measured variables of a given dataset. The transformation from latent variables to measured variables is assumed to be non-random but unknown; this transformation will account for the covariations in the data. However, in the measured data, there will be deviations unexplained by such a generative model. Those deviations will be assumed unique to each of the measured variables and the variables that absorb these unique

deviations will be called **unique factors**. This characteristic of the factor model is actually facilitated by imposing a diagonal structure on the covariance matrix of the unique factors; whereas the latent variables are transformed such that they absorb the common variation of the measured variables. The transformed latent variables are, hence, called **common factors**.

Note that the parameters of the factor model are (i) **transformation matrix** of the latent variables to the measured variables and (ii) **variances of the unique factors**. The principle of maximum likelihood cannot yield these two sets of parameters independently. By assuming knowledge or guessing one of them, an estimate for the other parameter could be found.

In Section 3.4.1, the **principal factor model** first estimates the covariance matrix of the unique factors in order to estimate the transformation matrix; the reverse procedure is followed in **principal component factor model** of Section 3.4.2.

In Section 3.5, the **Expectation - Maximization** (EM) algorithm for maximum likelihood estimation of the factor model will be first narrated in an original manner. It is an iterative scheme established by [33]. The expression for **complete log-likelihood** of the measured variables as well as the latent variables are written out. However, its analytical tediousness in direct maximization is realized. This is overcome by probing its lower bound. It turns out that the local maximum of the lower bound is attained whenever the complete log-likelihood converges to the log-likelihood of the measured variables. Hence, starting with a set of guessed parameters, iteratively maximizing the complete log-likelihood converges towards the standard log-likelihood. Writing the lower bound of the complete log-likelihood scheme in an *a posteriori* expectation format and maximizing it for the optimal parameters is the crux of the EM algorithm.

In Sections 3.6 and 3.7, the scheme for using the EM algorithm for iteratively estimating factor model parameters is presented; it is partly along the lines of [14]. In doing so, the expression for the log-likelihood in the complete log-likelihood form is first written out; the latter is conducive for use with the algorithm. In the E-step of the algorithm presented in Section 3.7.1, the *a posteriori* mean and covariance of the latent variables are derived. In the M-step of Section 3.7.2, *a posteriori* expectation format of the log-likelihood is maximized; the parameters of the factor model, viz., transformation matrix of the common factors and covariance matrix of the unique factors, are thereby estimated.

3.1 Maximum likelihood model

This section starts by presenting some notions and usages that will help in explaining the characteristics of the data to be modeled. The primary assumption is that the data is a collection of samples of some relevant variables measured in an experiment; this collection will be denoted by $\mathcal{D}$ and will be called simply the **dataset**. Let $\mathcal{D}$ be constituted by n data samples written in the sequence $\mathcal{D} = \{y_l\}$, $l = 1, \ldots, n$ and y_l be called the l-th **sample**.

In this chapter, unlike in Chapter 2, any sequential dependence of the value or occurrence of a data sample on next or any other sample is ignored. So there is no need to sort the data samples based on the time of data acquisition or any other criteria. But an index to identify the samples individually may be used.

Let the samples in $\mathcal{D}$ be realizations of the sequence y_l of random variables. Let $\mathrm{p}^{\mathrm{y}_l}$ denote density functions of their respective probability distributions. Throughout this

chapter, it is assumed that the samples encompassing $\mathcal{D}$ adhere to the characteristic defined below [108]:

Definition 3.1. *A dataset $\{y_l\}$, $l = 1, \ldots, n$ due to its respective random variables y_l is said to be **independently and identically distributed or iid** if its joint probability density function is $\prod_{l=1}^{n} \mathsf{p}^{\mathsf{y}}(y_l)$, where y is a random variable such that $\mathsf{p}^{\mathsf{y}}(y_l) = \mathsf{p}^{\mathsf{y}_l}(y_l)\ \forall l = 1, \ldots, n$.*

Based on Definition 3.1, $\mathcal{D}$ is interpreted as n realizations of a random variable y whose probability density function is p^{y}. Therefore, any realization y of y has a corresponding probability density $\mathsf{p}^{\mathsf{y}}(y)$. Extending $\mathsf{p}^{\mathsf{y}}(y)$ based on Definition 3.1 to $\mathcal{D}$ leads to the joint probability density of the dataset, which is simply denoted by $\mathsf{p}^{\mathsf{y}}(\mathcal{D})$, and is given by

$$\mathsf{p}^{\mathsf{y}}(\mathcal{D}) \triangleq \prod_{l=1}^{n} \mathsf{p}^{\mathsf{y}}(y_l). \tag{3.1}$$

In order to maintain simplicity for the model which generated $\mathcal{D}$, a set of parameters θ will be introduced for the model. In the context of Definition 3.1, θ refers to the set of parameters of the probability density function of y. However, θ is unknown and has to be estimated. In this setup, $\mathsf{p}^{\mathsf{y}|\theta}(\mathcal{D} \mid \theta)$ will be termed as the likelihood of the dataset whilst the parameters θ are available. Note that the distribution is over y and the likelihood is a function of non-random θ.

At a set of parameters θ, due to (3.1), the likelihood of the dataset $\mathcal{D}$ consisting of iid samples y_l, $l = 1, \ldots, n$ factorizes as

$$\mathsf{p}^{\mathsf{y}|\theta}(\mathcal{D} \mid \theta) = \prod_{l=1}^{n} \mathsf{p}^{\mathsf{y}|\theta}(y_l \mid \theta). \tag{3.2}$$

In order to find an appropriate model for a given dataset, the intention is to utilize the following statistical methodology [96].

Definition 3.2. *According to the **principle of maximum likelihood**, an optimal set of parameters $\hat{\theta}$ for the model corresponding to a dataset $\mathcal{D}$ is the set of parameters θ for which the likelihood of $\mathcal{D}$ is maximized, i.e.,*

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \mathsf{p}^{\mathsf{y}|\theta}(\mathcal{D} \mid \theta). \tag{3.3}$$

A modification to (3.3) is made now: Since probability density is a non-negative function, the logarithm of the likelihood is maximized to arrive at the same solution for the optimal parameters as per Definition 3.2. Such an analysis of the exponential family of probability density functions will lead to substantial simplification [96]. Hence, (3.3) may be rewritten equivalently as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \mathsf{L}(\theta), \tag{3.4}$$

where

$$\mathsf{L}(\theta) = \log_e \mathsf{p}^{\mathsf{y}|\theta}(\mathcal{D} \mid \theta). \tag{3.5}$$

3.2 Linear model

In Section 3.1, an appropriate estimate $\hat{\theta}$ of θ is considered as parameter for a model given the dataset $\mathcal{D}$. It is assumed that θ is a non-random quantity. Now consider the estimator Θ of θ, i.e., Θ is a random variable. It is hoped that Θ gives a reasonably good estimate of the true set of parameters θ given $\mathcal{D}$. Denoting mean of Θ by μ^{Θ} and variance by γ^{Θ}, the following properties indicate the quality of Θ; refer §4.4 of [70] and §10.3 of [94]:

Property 3.1. *An estimator Θ of θ is **unbiased** if $\mu^{\Theta} = \theta$.*

Property 3.2. *An estimator Θ of θ is **efficient** if $\Theta = \underset{\widetilde{\Theta} \in \mathcal{C}}{\operatorname{argmin}} \gamma^{\widetilde{\Theta}}$, where $\mathcal{C}$ is the class of all unbiased estimators of θ.*

A modeling strategy in Section 3.1 with a dataset constituted by iid samples was considered. Another popular parametric modeling paradigm called **linear model** involves treating a set of r-variate iid samples $y_1, \ldots, y_n$ as dependent on a set of q-variate iid samples $x_1, \ldots, x_n$, where $n > q$. The simplest of linear models regresses x_l towards y_l $\forall l = 1, \ldots, n$ through the relation

$$y_l = W x_l + z_l, \tag{3.6}$$

where $W \in \mathbb{R}^{r \times q}$ is a linear function of $x_l \in \mathbb{R}^q$ and $z_l \in \mathbb{R}^r$ is the error in the regression [119, 109]. The model parameters θ discussed above refer to W here. Suppose the modeling errors z_l are realizations of the vector random variable z, then the measurements y_l may also be treated as realizations of the vector random variable y. The linear model may then be effectively written as

$$\mathsf{y} = W x + \mathsf{z}, \tag{3.7}$$

for any $x \in \mathbb{R}^q$. By restricting the quality of the regression error z, the following theorem defines a popular linear model; for details one may refer to §6.2.1 of [62], §7.1 of [38] or §8.1 of [51] among plenty of references in the literature: Suppose each realization $y_l \in \mathbb{R}^r$ of the vector random variable y is related to $x_l \in \mathbb{R}^q$, $l = 1, \ldots, n$ through (3.6) or (3.7) where z_l are due to zero mean uncorrelated Gaussian vector random variable z. According to the Gauss-Markov theorem, an efficient estimator of W is given by

$$\widehat{W} = [y_1 \cdots y_n] X' \left(XX'\right)^{-1}, \tag{3.8}$$

where $X = (x_1 \cdots x_n) \in \mathbb{R}^{q \times n}$ has $\operatorname{rank}(X) = q$. Then, the error estimate for the l-th sample becomes $\hat{z}_l = (\hat{z}_{1_l} \cdots \hat{z}_{r_l})' = y_l - \widehat{W} x_l \; \forall l = 1, \ldots, n$. The unbiased estimator of the covariance matrix $\Gamma^{\mathsf{z}} = \operatorname{diag}(\gamma^{\mathsf{z}_1}, \ldots, \gamma^{\mathsf{z}_r})$ of z is given by

$$\gamma^{\mathsf{z}_k} = \frac{1}{n-q} \sum_{l=1}^{n} \hat{z}_{k_l}^2 \quad \forall k = 1, \ldots, r. \tag{3.9}$$

3.3 Factor model

The linear model reviewed in Section 3.2 could be seen as the q-variables of the iid samples x_l, $l = 1, \ldots, n$ together explaining the r-variables of each and every iid sample y_l, where both these sets of measured variables are available as part of the dataset $\mathcal{D}$. Suppose y_l and x_l , $l = 1, \ldots, n$ are treated to be due to vector random variables y and x, respectively, so that y is the result of the transformation Wx, where $W \in \mathbb{R}^{r \times q}$. Then, the challenge is to explain y when x is unavailable or inaccessible in $\mathcal{D}$. One way to proceed is by assuming latent existence of the q-dimensional vector random variable x in generating the r-dimensional vector random variable y. In that context, y is named the set of **measured variables** and x the **latent variables**.
It is wished to pursue here a parametric model by involving the probability density function; this will help extract the statistical characteristics of the dataset in a finite number of parameters. Hence, if the probability density function of x is assumed known in the transformation $\mathsf{y} = W\mathsf{x}$, then W serves as the parameter that needs to be estimated from the measured y.
However, the model $\mathsf{y} = W\mathsf{x}$ is very restrictive because it assumes that any randomness in y is due to x whose characteristics are assumed. The model is relaxed by introducing an r-dimensional random variable z uncorrelated with x and designated to absorb all deviations in y that cannot be retained by

$$\mathsf{v} \triangleq W\mathsf{x}. \tag{3.10}$$

Thus, the measured variables y are split into the common factors $\mathsf{v} = W\mathsf{x}$ and unique factors z; the following is such a model [80]:

Definition 3.3. *A **factor model** is defined as*

$$\mathsf{y} = W\mathsf{x} + \mathsf{z}, \tag{3.11}$$

where y *and* z *are* r*-dimensional vector random variables,* x *is a* q*-dimensional vector random variable,* $W \in \mathbb{R}^{r \times q}$ *is a non-random transformation matrix, and*

$$\mu^{\mathsf{x}} = 0,\ \mu^{\mathsf{y}} = 0,\ \mu^{\mathsf{z}} = 0, \tag{3.12}$$

$$\Gamma^{\mathsf{x}} = I_q, \tag{3.13}$$

Γ^{z} *is diagonal, and* (3.14)

$$\Gamma^{\mathsf{x},\mathsf{z}} = 0. \tag{3.15}$$

Given a dataset of realizations of y, the parameters of the factor model that need to be estimated are W and the covariance matrix Γ^{z} of z. The factor model, in contrast to the linear model of (3.6), does not observe any realizations of x.
The following essential result for the moments of a function of a vector random variable is summarily provided; refer Chapter 6 of [35]: For $\mathsf{v} = W\mathsf{x}$, $\mu^{\mathsf{v}} = W\mu^{\mathsf{x}}$ and $\Gamma^{\mathsf{v}} = W\Gamma^{\mathsf{x}}W'$. Therefore, applying (3.13) gives

$$\Gamma^{\mathsf{v}} = WW'. \tag{3.16}$$

In the factor model, the condition (3.15) of zero correlation between x and z is crucial. Naturally, it leads to $\Gamma^{\mathsf{v},\mathsf{z}} = 0$. Therefore, taking the second-order moments on both sides of (3.11) gives

$$\Gamma^{\mathsf{y}} = \Gamma^{\mathsf{v}} + \Gamma^{\mathsf{z}} = WW' + \Gamma^{\mathsf{z}}. \tag{3.17}$$

Due to (3.14), the r components of z are uncorrelated, or, all cross-covariances between the r components of y are inherited by only the covariance matrix $\Gamma^{\mathsf{v}} = WW'$ of $\mathsf{v} = W\mathsf{x}$ and not by Γ^{z}. This can be seen as each component of z inheriting only a part of the variance unique to its corresponding component in y. Hence, the components of z are called the **unique factors**. Since no part of the covariance of y common to all its components are held by z but instead by the transformation $\mathsf{v} = W\mathsf{x}$, v is called the **common factors**.
Note that in the factor model, the rq elements of W and r diagonal elements of Γ^{z} are to be estimated. Since (3.17) is just one equation with two unknowns, i.e., W and Γ^{z}, it cannot be solved uniquely; more conditions and assumptions may be placed to restrict possible solutions.

3.4 Maximum likelihood factor model

Well-known is the following assumption towards a proper solution of the factor model parameters, e.g., refer §3.5 of [62]: The measured variables follow a Gaussian distribution with parameters $\theta = \{\mu^{\mathsf{y}}, \Gamma^{\mathsf{y}}\}$, i.e.,

$$\mathsf{p}^{\mathsf{y}|\theta}(y \mid \theta) = \mathcal{N}(y \mid \mu^{\mathsf{y}}, \Gamma^{\mathsf{y}}), \tag{3.18}$$

as defined in (2.26).

Given samples y_l , $l = 1, \ldots, n$ of the measured variables y, the principle of maximum likelihood as per Definition 3.2 could be used to estimate an optimal set of parameters according to (3.4). The maximum likelihood parameters $\widehat{\mu}^{\mathsf{y}}$ and $\widehat{\Gamma}^{\mathsf{y}}$ of the mean and covariance matrix of the Gaussian distribution in (3.18) are the sample mean and sample covariance matrix, respectively, i.e.,

$$\widehat{\mu}^{\mathsf{y}} = \frac{1}{n}\sum_{l=1}^{n} y_l, \tag{3.19}$$

$$\widehat{\Gamma}^{\mathsf{y}} = \frac{1}{n}\sum_{l=1}^{n} (y_l - \widehat{\mu}^{\mathsf{y}})(y_l - \widehat{\mu}^{\mathsf{y}})'. \tag{3.20}$$

Then (3.20) may be substituted in (3.17) to get

$$\widehat{\Gamma}^{\mathsf{y}} = WW' + \Gamma^{\mathsf{z}}. \tag{3.21}$$

However, it gives no clue regarding the maximum likelihood W and Γ^{z}, which are the parameters of interest to the factor model. In what follows, two relevant methods which derive appropriate solutions on the basis of the general maximum likelihood solution are briefly presented.

3.4.1 Principal factor model

One of the approaches to finding possible solutions to the maximum likelihood factor model of the r-dimensional measured variables y using q-dimensional common factors x and r-dimensional unique factors z as per Definition 3.3 starts with a good guess $\widehat{\Gamma}^{\mathsf{z}}$ of Γ^{z}. The approach is called the principal factor model. One may refer to §10.2 of [54] or §6.3 of [46] to know how this guess could be made reliable; the details which are unnecessary for the objective of the present discussion are skipped. Substituting $\widehat{\Gamma}^{\mathsf{z}}$ in (3.21) gives

$$\widehat{\Gamma}^{\mathsf{y}} = WW' + \widehat{\Gamma}^{\mathsf{z}}. \tag{3.22}$$

The problem then is to estimate a W such that $WW' = \widehat{\Gamma}^{\mathsf{y}} - \widehat{\Gamma}^{\mathsf{z}}$ subject to some quality criterion. Suppose columns $u_1, \ldots, u_r$ of $U \in \mathbb{R}^{r\times r}$ are the eigenvectors of $\widehat{\Gamma}^{\mathsf{y}} - \widehat{\Gamma}^{\mathsf{z}}$ whose corresponding eigenvalues $d_1^2 \geq \cdots \geq d_r^2 > 0$ constitute the diagonal elements of a diagonal matrix D^2 from top-left to bottom-right. If the eigenvalue-eigenvector decomposition of $WW' = UD^2U'$ and the subscript $1:q$ is used to refer the first q column indices of a matrix, then the optimal transformation matrix of the principal factor model is

$$\widehat{W} = U_{1:q}D_{1:q}. \tag{3.23}$$

If necessary, the estimation between $\widehat{\Gamma}^{\mathsf{z}}$ and $\widehat{W}$ may alternate iteratively.

3.4.2 Principal component factor model

Another approach, for which [103] is referred to, involves first estimating W and then the covariance matrix Γ^{z} of the unique factors. In order to estimate Γ^{v} and W, first, it has to be reminded that WW' is of rank q. Second, note that the relation (3.21) may be thought as WW' approximating the variance-covariance of the measured variables y as contained in Γ^{y}. There could be infinitely many ways WW' could approximate Γ^{y} and an approximation with respect to the Frobenius norm $\|\widehat{\Gamma}^{\mathsf{y}} - \Gamma^{\mathsf{v}}\|_F$ seems reasonable and standard practice; refer §10.2 of [54] and and §2.12 of [103]. In that context, the following theorem is used; refer Lecture 5 of [120]:

Theorem 3.1. *For full rank matrix $A \in \mathbb{C}^{r\times r}$ with eigenvectors $u_1, \ldots, u_r$ whose corresponding eigenvalues are $\alpha_1 \geq \cdots \geq \alpha_r$, matrix $\widetilde{A} \in \mathbb{C}^{r\times r}$ with $\mathsf{rank}(\widetilde{A}) = q < r$ defined as*

$$\widetilde{A} = [u_1 \ldots u_q]\,\mathsf{diag}(\alpha_1, \ldots, \alpha_q)\,[u_1 \ldots u_q]^* \tag{3.24}$$

is such that

$$\|A - \widetilde{A}\|_F = \inf_{\substack{B\in\mathbb{C}^{r\times r} \\ \mathsf{rank}(B)=q}} \|A - B\|_F = \alpha_{q+1}. \tag{3.25}$$

Due to Theorem 3.1, the optimal approximation of Γ^{y} using WW' in the Frobenius norm sense is declared as $\widehat{W}\widehat{W}' = E_{1:q}\Lambda^2_{1:q}E'_{1:q}$, where columns of E are eigenvectors of $\widehat{\Gamma}^{\mathsf{y}}$ whose corresponding eigenvalues in decreasing order form the diagonal of Λ^2.

Therefore, if the subscript $1:q$ to refer to the first q column indices of a matrix, the estimate of W sought is given by

$$\widehat{W} = E_{1:q}\Lambda_{1:q}. \tag{3.26}$$

Since Γ^{z} ought to be diagonal due to uncorrelated z, taking into account (3.21), an approximate solution for Γ^{z} is

$$\widehat{\Gamma}^{\mathsf{z}} \approx \widetilde{\mathrm{diag}}(\widehat{\Gamma}^{\mathsf{y}} - \widehat{W}\widehat{W}'), \tag{3.27}$$

where $\widetilde{\mathrm{diag}}$ refers to setting the off-diagonal elements to zero.

3.5 EM algorithm

Now a presentation of the Expectation-Maximization (EM) algorithm is attempted; as stated in the introduction of this chapter, it is a popular iterative method for maximum likelihood estimation.

Note 3.1. *In this section,* x *is assumed a discrete and univariate random variable; this is to avoid any unnecessary analytical complications otherwise leading to equivalent conclusions. E.g. the summation over* x *has to be replaced by an integration for continuous* x*. And, a summation or integration across all dimensions of* x *is to be applied had* x *been a vector random variable.*

Note the following lemma (refer §4.5 of [49]):

Lemma 3.1. *If a random variable* x *is* ***marginalized*** *from its joint distribution* $\mathsf{p}^{\mathsf{x},\mathsf{y}}$ *with the random variable* y*, the result is the distribution of* y*, i.e.,*

$$\mathsf{p}^{\mathsf{y}}(y) = \sum_{x} \mathsf{p}^{\mathsf{x},\mathsf{y}}(x, y). \tag{3.28}$$

The definition of log-likelihood in (3.5) may be rewritten through Lemma 3.1 as

$$\mathsf{L}(\theta) = \log_e \sum_{x} \mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta); \tag{3.29}$$

the maximization of this expression for the log-likelihood is intractable due to the summation inside the logarithm. In order to evade this situation, a dummy function $\eta(\mathsf{x})$ such that

$$\sum_{x} \eta(x) = 1; \quad \eta(\mathsf{x}) > 0 \tag{3.30}$$

is introduced and the complete log-likelihood

$$\mathsf{L}(\theta, \eta) = \log_e \sum_{x} \eta(x) \frac{\mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta)}{\eta(x)} \tag{3.31}$$

is formed. The purpose in introducing $\eta(x)$ is to seek possibilities to maximize $\mathsf{L}(\theta, \eta)$ in lieu of $\mathsf{L}(\theta)$. In that pursuit, as shown in Section B.1.1, the logarithm may be brought inside the summation, i.e.,

$$\mathsf{L}(\theta, \eta) \geq \sum_x \eta(x) \log_e \frac{\mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta)}{\eta(x)}. \tag{3.32}$$

Referring to Section B.1.2, it is possible to decompose the complete log-likelihood as

$$\mathsf{L}(\theta, \eta) \geq \mathsf{L}(\theta) + \mathsf{K}(\theta, \eta), \tag{3.33}$$

where

$$\mathsf{K}(\theta, \eta) = \sum_x \eta(x) \log_e \frac{\mathsf{p}^{\mathsf{x}|\mathsf{y},\Theta}(x \mid \mathcal{D}, \theta)}{\eta(x)}. \tag{3.34}$$

Now think of two iterative steps:

Step 1 – Find optimal η for a fixed θ: For a particular $\theta = \theta_i$, let $\hat{\eta}_i = \underset{\eta}{\text{argmax}}\ \mathsf{L}(\theta_i, \eta)$. Since local increase of $\mathsf{L}(\theta, \eta)$ is guaranteed by locally maximizing its global lower bound $\mathsf{L}(\theta) + \mathsf{K}(\theta, \eta)$, $\hat{\eta}_i = \underset{\eta}{\text{argmax}}\ \mathsf{K}(\theta_i, \eta)$; one may refer [87] for more details. By differentiating (3.34) with respect to $\eta(x)$, it may be found that

$$\hat{\eta}_i = \mathsf{p}^{\mathsf{x}|\mathsf{y},\Theta}(x \mid \mathcal{D}, \theta_i). \tag{3.35}$$

However, $\mathsf{K}(\theta_i, \hat{\eta}_i) = 0$ whereby

$$\mathsf{L}(\theta_i, \hat{\eta}_i) = \mathsf{L}(\theta_i). \tag{3.36}$$

Note that for a Gaussian density for y, the conditional probability for $\hat{\eta}_i$ in (3.35) is tractable.

Step 2 – Find optimal θ for a fixed η: Having found the locally optimal η for a fixed θ, the locally optimal θ for a fixed $\eta = \hat{\eta}_i$ is pursued. Based on (3.33) and (3.36), it may be written that

$$\theta_{i+1} = \underset{\theta}{\text{argmax}}\ \mathsf{L}(\theta, \hat{\eta}_i) \tag{3.37}$$

Note that (3.36) ensures that likelihood $\mathsf{L}(\theta_i)$ is approached in every i-th iteration whenever $\mathsf{L}(\theta_i, \hat{\eta}_i)$ is maximized to obtain the $i+1$-th estimate θ_{i+1}; in other words, the iterations converge to a local maximum of $\mathsf{L}(\theta_i)$.

E and M steps

The two steps arrived at above are now compiled. Suppose there is an initial guess θ_0 of θ. Then a local maximization of likelihood may be performed such that, in the i-th iteration, where $i = 1, 2, \ldots$, has the two steps:

Step 1: estimate $\hat{\eta}_i$ based on θ_i, and

Step 2: locally maximize the likelihood to obtain θ_{i+1}.

These steps of an iteration become explicit if we note as shown in Section B.1.3 that

$$\theta_{i+1} = \underset{\theta_i}{\text{argmax}}\ \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}\big[\log_e \mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta_i)\big]. \tag{3.38}$$

Hence, the i-th iteration involves:

Step 1 (Expectation): Evaluating the expectation $\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}\big[\log_e \mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta_i)\big]$, and

Step 2 (Maximization): Maximizing $\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}\big[\log_e \mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, x \mid \theta_i)\big]$ locally with respect to θ_i.

3.6 Basic setup of EM algorithm for factor modeling

The difference between the linear model and the factor model is obvious and wide. While access to the samples x_n and y_n in (3.6) of the linear model is available, x in (3.11) is assumed inaccessible in the factor model. Hence, for the factor model, the conditional distribution $\mathsf{p}^{\mathsf{y}|\mathsf{x}}$ of y given x is of interest.
Firstly, using the properties of the conditional distribution, e.g. refer §8 in Chapter 1 of [110], it could easily be shown for the factor model that $\mathsf{E}^{\mathsf{y}|\mathsf{x}}[\mathsf{y}|\mathsf{x}] = W\mathsf{x} + \mathsf{E}^{\mathsf{z}|\mathsf{x}}[\mathsf{z}]$ $= W\mathsf{x}$ because z is independent of x and has zero mean.
Secondly, the conditional variance $\Gamma^{\mathsf{y}|\mathsf{x}}$ is $\mathsf{E}^{\mathsf{y}|\mathsf{x}}[\mathsf{y}\mathsf{y}'|\mathsf{x}] - (\mathsf{E}^{\mathsf{y}|\mathsf{x}}[\mathsf{y}|\mathsf{x}])\ (\mathsf{E}^{\mathsf{y}|\mathsf{x}}[\mathsf{y}|\mathsf{x}])'$. On expansion based on (3.11), $\Gamma^{\mathsf{y}|\mathsf{x}} = \mathsf{E}^{\mathsf{y}|\mathsf{x}}\ [(W\mathsf{x} + \mathsf{z})(W\mathsf{x} + \mathsf{z})'|\mathsf{x}] - (W\mathsf{x})(W\mathsf{x})'$. Upon term-by-term expansion, due to the independence of z and x and since $\mathsf{E}^{\mathsf{z}}[\mathsf{z}] = 0$, the only surviving term will be $\mathsf{E}^{\mathsf{z}|\mathsf{x}}[\mathsf{z}\mathsf{z}' \mid \mathsf{x}]$, which becomes $\mathsf{E}^{\mathsf{z}}[\mathsf{z}\mathsf{z}'] = \Gamma^{\mathsf{z}}$.
It is also well-known that the distribution of a Gaussian random vector conditioned on another is itself Gaussian; one may refer of §4.8 of [48] or Theorem 3.10.1 of [15] among many methods to verify it. Therefore, the Gaussian probability density of y | x with parameters $\theta = \{W, \Gamma^{\mathsf{z}}\}$ for the factor model may be written as

$$\mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(y \mid x, \theta) = \mathcal{N}(y \mid W\mathsf{x}, \Gamma^{\mathsf{z}}). \tag{3.39}$$

Note that θ in $\mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}$ refers to the availability of the set of parameters; the distribution is conditioned only on x. Based on the discussions in Section 3.1, the conditional probability density $\mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}$ underpins the likelihood of the factor model. As with (3.1), the dataset $\mathcal{D}$ is considered to consist of the iid samples y_l, $l = 1, \ldots, n$ of y. The likelihood of the dataset is

$$\mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(\mathcal{D} \mid \mathsf{x}, \theta) = \prod_{l=1}^{n} \mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(y_l \mid \mathsf{x}, \theta). \tag{3.40}$$

Using Theorem B.1 known as Bayes theorem, $\mathsf{p}^{\mathsf{y},\mathsf{x}|\theta}(\mathcal{D}, \mathsf{x} \mid \theta) = \mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(\mathcal{D} \mid \mathsf{x}, \theta)\mathsf{p}^{\mathsf{x}}(\mathsf{x})$. If it is assumed that the distribution $\mathsf{p}^{\mathsf{x}}(x)$ to be independent of θ, then (3.38) of the EM-algorithm reduces to iteratively solving

$$\theta_{i+1} = \underset{\theta_i}{\text{argmax}}\ \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}\big[\log_e \mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(\mathcal{D} \mid \mathsf{x}, \theta_i)\big]. \tag{3.41}$$

From the above three equations, it may be written that

$$\theta_{i+1} = \underset{\theta_i}{\text{argmax}}\ \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})], \tag{3.42}$$

$$\begin{aligned} \mathsf{f}(\theta_i, \mathsf{x}) &= \log_e \mathsf{p}^{\mathsf{y}|\mathsf{x},\theta}(\mathcal{D} \mid \mathsf{x}, \theta_i) \\ &= \sum_{l=1}^{n} \log_e \mathcal{N}(y_l \mid W_i \mathsf{x}, \Gamma_i^{\mathsf{z}}), \end{aligned} \tag{3.43}$$

where the parameters

$$\theta_i \triangleq \{W_i, \Gamma_i^{\mathsf{z}}\} \tag{3.44}$$

correspond to the i^{th} iteration.

3.7 Two steps of EM algorithm for factor modeling

In light of the discussion in Section 3.5 and the parameter update equations of (3.42), it may be stated that the i-th iteration of the factor model estimation alternates between:

1. **Expectation-step** Evaluate the expectation $\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}\big[\mathsf{f}(\theta_i, \mathsf{x})\big]$, and
2. **Maximization-step** Update $\theta_{i+1} \leftarrow \theta_i$ by maximizing $\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}\big[\mathsf{f}(\theta_i, \mathsf{x})\big]$ with respect to θ_i.

To proceed note that in (3.43) that $\mathsf{f}(\theta_i, \mathsf{x}) = -n \log_e(\det(\Gamma_i^{\mathsf{z}})) - \sum_{l=1}^{n} \mathsf{M}(y_l, W_i \mathsf{x}, \Gamma_i^{\mathsf{z}})$, where for any compatible vectors a, b and matrix C

$$\mathsf{M}(a, b, C) = (a - b)' C^{-1} (a - b), \tag{3.45}$$

whose expansion gives $\mathsf{M}(y_l, W_i \mathsf{x}, \Gamma_i^{\mathsf{z}}) = y_l' (\Gamma_i^{\mathsf{z}})^{-1} y_l - 2 y_l' (\Gamma_i^{\mathsf{z}})^{-1} W_i \mathsf{x} + \mathsf{tr}((\Gamma_i^{\mathsf{z}})^{-1} W_i \mathsf{x}\mathsf{x}' W_i')$. Note the presence of terms with random variables x and $\mathsf{x}\mathsf{x}'$ in $\mathsf{M}(y_l, W_i \mathsf{x}, \Gamma_i^{\mathsf{z}})$. Therefore, the EM algorithm, as a result of this expansion of $\mathsf{M}(y_l, W_i \mathsf{x}, \Gamma_i^{\mathsf{z}})$, will involve alternating between:

1. **Expectation-step** Evaluate, for $l = 1, \ldots, n$,

$$\begin{aligned} \langle x \rangle_{i,l} &\triangleq \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{x} \mid y_l, \theta_i], \\ \langle x x' \rangle_{i,l} &\triangleq \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{x}\mathsf{x}' \mid y_l, \theta_i], \end{aligned} \tag{3.46}$$

   where $\langle x \rangle_{i,l} \in \mathbb{R}^q$ and $\langle x x' \rangle_{i,l} \in \mathbb{R}^{q \times q}$, and

2. **Maximization-step** Update $\theta_{i+1} \leftarrow \theta_i$ by maximizing $\mathsf{f}(\theta_i, \mathsf{x})$ with respect to θ_i, where x and $\mathsf{x}\mathsf{x}'$ are replaced by their corresponding *a posteriori* estimates, i.e., in (3.43)

$$\begin{aligned} \mathsf{x} &\Longleftarrow \langle x \rangle_{i,l} \\ \mathsf{x}\mathsf{x}' &\Longleftarrow \langle x x' \rangle_{i,l}. \end{aligned} \tag{3.47}$$

The following analysis between (3.48) and (3.52) is inspired by [14].

3.7.1 E-step

Note that $\langle x \rangle_{i,l}$ is the mean of the Gaussian distribution $\mathsf{p}^{\mathsf{x}|\mathsf{y},\theta}(\mathsf{x} \mid y_l, \theta_i)$, which is evaluated in Appendix B.2 to be

$$\begin{aligned} \langle x \rangle_{i,l} &= \Omega_i W_i' (\Gamma_i^{\mathsf{z}})^{-1} y_l, \\ \Omega_i &= (I_q + W_i' (\Gamma_i^{\mathsf{z}})^{-1} W_i)^{-1}, \end{aligned} \tag{3.48}$$

where $\Omega_i \in \mathbb{R}^{q \times q}$. From the classical relation of mean and covariance of any distribution, it is known that $\langle xx' \rangle_{i,l}$ is the sum of $\langle x \rangle_{i,l} \langle x' \rangle_{i,l}$ and covariance of $\mathsf{x} \mid y_l, \theta_i$, i.e.,

$$\langle xx' \rangle_{i,l} = \langle x \rangle_{i,l} \langle x' \rangle_{i,l} + \Omega_i. \tag{3.49}$$

This completes the E-step of the EM Algorithm.

3.7.2 M-step

Towards the M-step of the EM algorithm, the substitutions in (3.47) give

$$\begin{aligned} \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})] = -n \log_e(\det(\Gamma_i^{\mathsf{z}})) - \sum_{l=1}^{n} \mathsf{tr}((\Gamma_i^{\mathsf{z}})^{-1} W_i \langle xx' \rangle_{i,l} W_i') \\ - 2 y_l' (\Gamma_i^{\mathsf{z}})^{-1} W_i \langle x \rangle_{i,l} + y_l' (\Gamma_i^{\mathsf{z}})^{-1} y_l. \end{aligned} \tag{3.50}$$

Now $\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})]$ may be maximized to update the parameters W_i and Γ_i^{z}:

Update W_i : The problem that has to be solved is

$$\begin{aligned} W_{i+1} &= \underset{W_i}{\mathsf{argmax}}\ \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})] \\ &= \underset{W_i}{\mathsf{arg}} \left(\frac{\partial}{\partial W_i} \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})] = 0 \right). \end{aligned} \tag{3.51}$$

It is easy to see using matrix differentiation rules, e.g., refer [98], that

$$\frac{\partial}{\partial W_i} \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i, \mathsf{x})] = -\sum_{l=1}^{n} 2(\Gamma_i^{\mathsf{z}})^{-1} W_i \langle xx' \rangle_{i,l} - 2(\Gamma_i^{\mathsf{z}})^{-1} y_l \langle x' \rangle_{i,l},$$

which when equated to zero gives

$$W_{i+1} = \left(\sum_{l=1}^{n} y_l \langle x' \rangle_{i,l} \right) \left(\sum_{l=1}^{n} \langle xx' \rangle_{i,l} \right)^{-1}. \tag{3.52}$$

Update Γ_i^{z} : The access to the updated W_{i+1} is available and if

$$v_{i,l} = W_{i+1} \langle x \rangle_{i,l},$$

then $\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i,\mathsf{x})] = -n\log_e(\det(\Gamma_i^{\mathsf{z}})) - \sum_{l=1}^{n}\mathsf{M}(y_l, v_{i,l}, \Gamma_i^{\mathsf{z}})$. Now, consider the update

$$\Gamma_{i+1}^{\mathsf{z}} = \underset{W_i}{\arg}\left(\frac{\partial}{\partial W_i}\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i,\mathsf{x})] = 0\right). \tag{3.53}$$

For $\Gamma_i^{\mathsf{z}} = \mathsf{diag}(\gamma_i^{\mathsf{z}_1},\ldots,\gamma_i^{\mathsf{z}_r})$ it can be seen that

$$\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i,\mathsf{x})] = -n\sum_{k=1}^{r}[\log_e(\gamma_i^{\mathsf{z}_k}) + \frac{1}{\gamma_i^{\mathsf{z}_k}}a_i^{\mathsf{z}_k}]$$

where

$$a_i^{\mathsf{z}_k} = \frac{1}{n}\sum_{l=1}^{n}(y_l - v_{i,l})^2.$$

Then, $\partial\mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i,\mathsf{x})]/\partial\gamma_i^{\mathsf{z}_k} = 0$ at

$$\gamma_{i+1}^{\mathsf{z}_k} = a_i^{\mathsf{z}_k}. \tag{3.54}$$

Factor model estimation via EM algorithm

Given the dataset, in Algorithm 2, the results of the analysis of the iterative parametric estimation of the factor model using the EM algorithm are summarized.

Algorithm 2: EM algorithm for the factor model

Input: $\mathcal{D} = \{y_l\}, l = 1,\ldots n$
Output: $\widehat{W}, \widehat{\Gamma}^{\mathsf{z}} = \mathsf{diag}(\hat{\gamma}^{\mathsf{z}_1},\ldots,\hat{\gamma}^{\mathsf{z}_r})$
initialize $i = 0$;
initialize randomly $W_i, \Gamma_i^{\mathsf{z}}$;
do
  E-step:
  for $l = 1$ *to* n **do**
    compute
      $\langle x\rangle_{i,l}$ using (3.48);
      $\langle xx'\rangle_{i,l}$ using (3.49);
  end
  M-step: update
    W_{i+1} using (3.52);
    $\gamma_{i+1}^{\mathsf{z}_k}\ \forall k = 1,\ldots,r$ using (3.54);
  $i \longleftarrow i+1$;
  $\epsilon \longleftarrow \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_i,\mathsf{x})] - \mathsf{E}^{\mathsf{x}|\mathsf{y},\theta}[\mathsf{f}(\theta_{i-1},\mathsf{x})]$ using (3.50);
while $\epsilon > 10^{-8}$ **and** $i < 20$;
$\widehat{W} \longleftarrow W_i, \hat{\gamma}^{\mathsf{z}_k} \longleftarrow \gamma_i^{\mathsf{z}_k}\ \forall k = 1,\ldots,r$;

A major drawback of the EM algorithm is the possibility that the estimation might get trapped in a local maximum of the log-likelihood and hence might require random restarts or other heuristic measures to be more certain regarding the estimates.

3.8 Summary

Two possibilities of modeling an r-dimensional measured vector random variable y were considered, viz., (i) the linear model where a measured variable $x \in \mathbb{R}^q, q < r$ is transformed to y and (ii) the factor model where a latent q-variate random variable x is transformed to y. Essentially, a factor model transforms a latent vector random variable of known probability distribution to a measured vector random variable of higher dimensionality that is perturbed by independent and uncorrelated noise. For the linear model, an efficient estimator of the transformation matrix was presented; whereas for the factor model there is no unique transformation. However, by restricting the variances unique to each of the measured variable, it is possible to estimate meaningful transformations. Thus, from a parametric modeling perspective, the transformation matrix and the unique variances are the parameters of the factor model.
In order to estimate the factor model parameters, two approaches based on the principle of maximum likelihood were discussed: The analytical estimation approach involves approximating the covariance structure of the measured variables using that of the transformed variables. For the iterative approach based on the EM algorithm, the log-likelihood function being lower bound by the *a posteriori* expectation of the logarithm of the joint probability density of the measured variables and the latent variables was exploited. Starting from guesses of the parameters, the EM algorithm maximizes the complete log-likelihood function of the latent variables and the measured variables by iteratively converging to the log-likelihood with every update of the parameters.