Bachelorarbeit

Maximum likelihood estimation for data missing at random via the EM algorithm

Verfasser

Vincent Baumgartner

angestrebter akademischer Grad

Bachelor of Science (BSc)

Wien, 2025

Studienkennzahl lt. Studienblatt:	A 033 551
Fachrichtung:	Statistik
BetreuerIn:	Dr. Solomiia Dmytriv, M.Sc.

Contents

Abstract

Classical statistical methods often assume that data are complete, an assumption that is frequently not met in practice. Missing data can arise through various mechanisms, which are categorized into three main types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Not Missing at Random (NMAR) (cf. Little and Rubin, 2002, Chapter 1). When classical techniques such as listwise deletion or simple imputation are used, they can be inefficient under certain conditions and lead to biased estimates. To address this issue, specialized statistical methods are needed that can handle incomplete data. This article focuses on the EM algorithm for Maximum Likelihood Estimation under MAR, providing an efficient iterative framework for maximizing the observed-data likelihood when direct optimization is infeasible.

1 Introduction

Incomplete observations are a recurring challenge in applied statistics, stemming from various factors such as item nonresponse in surveys, participant dropout in clinical trials, or data loss due to mechanical failure. When done without caution methods such as listwise deletion or single imputation, often result in the inefficient use of information and, worse still, can introduce bias whenever the propensity for missing data is coupled to the data-generating process. Robust statistical inference therefore demands methods that exploit all observed information while respecting the stochastic mechanism. When considering such mechanisms, a principled approach is to use the tripartite classification system developed by Rubin (1976): missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). In accordance with the MAR principle, the probability that a value is missing depends only on the observed data and not on the unobserved (missing) information. Consequently, the data model and the missingness model are distinct, as per Rubin (1976) ignorability theorem. Ignorability legitimises basing likelihood inference solely on the marginal data likelihood, yet maximising it is typically intractable because it involves high-dimensional integration over the unobserved components. The expectation-maximisation (EM) algorithm of Dempster et al. (1977) provides a satisfactory solution by alternating between computing the conditional expectation of the complete-data log-likelihood and maximising this expectation with respect to the parameters. This ensures that the observed-data likelihood increases in a predictable and consistent way. When compared with brute-force numerical optimisation, EM transforms an ill-conditioned problem into a series of manageable ones, often exploiting closed-form expressions that arise in exponential-family models (cf. Schafer, 1997, p. 40). This article revisits maximum likelihood estimation under MAR through the lens of EM. By recasting EM as coordinate ascent on the evidence lower bound (ELBO), we unify classical likelihood theory with contemporary ideas from variational inference (cf. Blei et al., 2017). The remainder of the paper is structured as follows. Section 2 formalises the complete and incomplete data models and reviews the missing

data taxonomy. Section 3 states Rubin (1976) ignorability conditions and clarifies when the missingness process can be ignored. We then derive EM from first principles and establish its convergence properties by demonstrating monotonic likelihood ascent and stationarity. Afterward, we turn to the problem of obtaining standard errors in the EM framework, showing how Louis Method yields an exact observed-information matrix. Finally, we illustrate the Expectation Maximization Algorithm in a small Monte Carlo study under a logistic-MAR mechanism, comparing EM's efficiency against the bias of complete case analysis.

2 Definitions

Definition 2.1 (Complete Data Model)**.** Let $Y = (y_{ij}) \in \mathbb{R}^{n \times p}$ be the complete data matrix and denote its i-th row by $y_i = (y_{i1}, \ldots, y_{ip})^\top$. Further assume that $y_i \overset{\text{iid}}{\sim} \mathbb{P}_\theta$, and let f_θ denote the (discrete or continuous) density of $\mathbb{P}_\theta$ with respect to an appropriate dominating measure. Under these assumptions, the joint density of the entire data matrix Y factorises as

$$L(\theta; Y) \;=\; \prod_{i=1}^{n} f_\theta(y_i)$$

Definition 2.2 (Incomplete Data Model)**.** Let X and Z be the observed and missing parts of the complete data matrix $Y = (X, Z)$. Define the indicator matrix $M = (m_{ij})$ by $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ otherwise, and assume $M \sim \mathbb{P}_\psi$ with density $f_\psi(M)$. The resulting observed-data likelihood is then

$$L_{\text{obs}}(\theta, \psi; X, M) = f_{\theta,\psi}(X, M) = \int f_{\theta,\psi}(X, Z, M)\, dZ$$

Where the integral is taken over the entire support of Z. Moreover, in the discrete case, the term is indeed a summation.

Definition 2.3 (Likelihood Ignoring the Missing-Data Mechanism)**.** A likelihood is said to ignore the missing-data mechanism if it is any function of θ proportional to the marginal density of the observed data $f_\theta(X) = \int f_\theta(X, Z)\, dZ$ that is

$$L_{\text{ign}}(\theta; X) \;\propto\; f_\theta(X)$$

Definition 2.4 (Missing Completely at Random (MCAR))**.** The missingness pattern M is independent of the data

$$f_\psi(M \mid X, Z) = f_\psi(M) \quad \text{for all } X, Z, \psi$$

Definition 2.5 (Missing at Random (MAR))**.** Missingness may depend on the observed entries X but not on the unobserved entries Z

$$f_\psi(M \mid X, Z) = f_\psi(M \mid X) \quad \text{for all } Z, \psi$$

Definition 2.6 (Not Missing at Random (NMAR))**.** The probability of missingness depends on the unobserved values themselves, so

$$f_\psi(M \mid X, Z) \quad \text{for all } \psi$$

3 Ignorability

Theorem 3.1. *A missing-data mechanism is said to be ignorable if it satisfies the* MAR *condition and the parameter vectors θ (for the data model) and ψ (for the missingness mechanism) are distinct, i.e. $(\theta, \psi) \in \Theta \times \Psi$.*

While a precise and detailed proof is available in Rubin (1976) below we present the proof as given by (cf. Little and Rubin, 2002, p. 119).

Proof. Note that under MAR we have that $f_\psi(M \mid X, Z) = f_\psi(M \mid X)$ and thus the observed likelihood simplifies to,

$$\begin{aligned}
L_{\text{obs}}(\theta, \psi; X, M) &= \int f_{\theta,\psi}(X, Z, M)\, dZ \\
&= \int f_\psi(M \mid X, Z) f_\theta(X, Z)\, dZ \\
&= \int f_\psi(M \mid X) f_\theta(X, Z)\, dZ \\
&= f_\psi(M \mid X) \int f_\theta(X, Z)\, dZ \\
&= f_\psi(M \mid X) f_\theta(X) \\
&\propto f_\theta(X)
\end{aligned}$$

Since the observed-data likelihood is proportional to the marginal density of X, the missing-data mechanism is ignorable. Moreover, under the distinctness condition, the missingness process contains no information about θ, so inference for θ based on L_{ign} coincides exactly with that based on L_{obs}. □

This shows that, under MCAR and MAR, inference for θ can be based entirely on the marginal likelihood $f_\theta(X)$. In the event of the missingness mechanism being NMAR, the ignorability assumption is no longer valid. Consequently, it is necessary to specify a joint model for the variables (X, Z, M). The Heckman Selection Model Heckman (1979) for example is a renowned approach to the NMAR case that will not be further examined here. Henceforth, it is assumed that the data is MAR.

4 EM Algorithm

4.1 Maximizing the Marginal Likelihood

It has been established that, under the Missing-at-Random assumption, valid maximum-likelihood estimation requires the maximisation of the marginal likelihood. However, the question remains of how this maximisation should be carried out in practice. Moreover, the resulting optimization problem can be computationally challenging, or even intractable, because it takes the form

$$\hat{\theta}_{\text{MLE}} \;=\; \arg\max_{\theta\in\Theta} \; \log\!\left(\int f_\theta(X, Z)\, dZ\right)$$

where the integral over the missing data typically admits no closed-form solution, and its evaluation depends critically on both the specific pattern of missingness (see Appendix A.1) and the assumed data distribution. In Appendix B.1, we present an example from Anderson (1957) in which direct maximization of the marginal likelihood is feasible. However, in most practical applications, the model structure and the missingness pattern are far more complex, so one must resort to numerical methods to obtain maximum-likelihood estimates.

4.2 Motivation

The Expectation Maximization algorithm is motivated by the following two lemmas.

Lemma 4.1. *Let X and Z be random variables, and let $\mathbb{Q}$ be any distribution on Z with density $q(Z)$ such that $\mathbb{P}_\theta(X, Z) \ll \mathbb{Q}$ and $\mathbb{P}_\theta(Z \mid X) \ll \mathbb{Q}$. Then the marginal log-likelihood decomposes as*

$$\ell(\theta; X) = \mathcal{F}(q, \theta) + D_{KL}\left(q \,||\, f_\theta(Z \mid X)\right)$$

where

- $\mathcal{F}(q, \theta) := \mathbb{E}_{Z\sim\mathbb{Q}}\left[\log\left(\dfrac{f_\theta(X, Z)}{q(Z)}\right)\right]$ *(the Evidence Lower Bound, ELBO)*
- $D_{KL}\left(q \,||\, f_\theta(Z \mid X)\right) := -\mathbb{E}_{Z\sim\mathbb{Q}}\left[\log\left(\dfrac{f_\theta(Z \mid X)}{q(Z)}\right)\right]$ *(the Kullback-Leibler Divergence)*

Proof. First note that $f_\theta(X) = \frac{f_\theta(X,Z)}{f_\theta(Z|X)}$ and X is fixed when taking the expectation over any $\mathbb{Q}$, it follows that $\ell(\theta; X) = \mathbb{E}_{Z\sim\mathbb{Q}}\left[\ell(\theta; X)\right]$. Thus we can make the following observation,

$$
\begin{aligned}
\ell(\theta; X) &= \mathbb{E}_{Z\sim\mathbb{Q}}\left[\log\left(\frac{f_\theta(X,Z)}{f_\theta(Z \mid X)}\right)\right] \\
&= \mathbb{E}_{Z\sim\mathbb{Q}}\left[\log\left(\frac{f_\theta(X,Z)/q(Z)}{f_\theta(Z \mid X)/q(Z)}\right)\right] \\
&= \mathbb{E}_{Z\sim\mathbb{Q}}\left[\log\left(\frac{f_\theta(X,Z)}{q(Z)}\right)\right] - \mathbb{E}_{Z\sim\mathbb{Q}}\left[\log\left(\frac{f_\theta(Z \mid X)}{q(Z)}\right)\right] \\
&= \mathcal{F}(q,\theta) + D_{\mathrm{KL}}\left(q \,||\, f_\theta(Z \mid X)\right)
\end{aligned}
$$

□

Lemma 4.2. *For every distribution $\mathbb{Q}$ on Z that satisfies the assumptions of Lemma 4.1,*

$$\mathcal{F}(q,\theta) \;\leq\; \ell(\theta; X)$$

with equality iff $q(Z) = f_\theta(Z \mid X)$.

Proof. Since the logarithm is a concave function, we know the following by Jensen's Inequality,

$$
\begin{aligned}
\underbrace{\mathbb{E}_{Z\sim\mathbb{Q}}\left[\log\left(\frac{f_\theta(X,Z)}{q(Z)}\right)\right]}_{=\mathcal{F}(q,\theta)} &\leq \log\left(\mathbb{E}_{Z\sim\mathbb{Q}}\left[\frac{f_\theta(X,Z)}{q(Z)}\right]\right) \\
&= \log\left(\int \frac{f_\theta(Z,X)}{q(Z)} q(Z)\, dZ\right) \\
&= \underbrace{\log(f_\theta(X))}_{=\ell(\theta;X)}
\end{aligned}
$$

When setting $q = f_\theta(Z \mid X)$ we make the following observation,

$$
\begin{aligned}
\mathcal{F}(q,\theta) &= \mathbb{E}_{Z\sim\mathbb{Q}}\left[\log\left(\frac{f_\theta(X,Z)}{q(Z)}\right)\right] \\
&= \mathbb{E}_{Z\sim\mathbb{P}_\theta(\cdot|X)}\left[\log\left(\frac{f_\theta(X,Z)}{f_\theta(Z \mid X)}\right)\right] \\
&= \mathbb{E}_{Z\sim\mathbb{P}_\theta(\cdot|X)}\left[\log\left(f_\theta(X,Z)\right)\right] \\
&= \mathbb{E}_{Z\sim\mathbb{P}_\theta(\cdot|X)}\left[\ell(\theta; X)\right] \\
&= \ell(\theta; X)
\end{aligned}
$$

Thus we have also shown that lower bound is tight, when setting $q = f_\theta(Z \mid X)$

□

4.3 The Algorithm

Consequently, $\mathcal{F}(q, \theta)$ provides a suitable surrogate objective (given the posterior $f_\theta(Z \mid X)$ is tractable) and maximising $\mathcal{F}(q, \theta)$ in an alternating (coordinate-ascent) manner with respect to q and θ, one recovers the famous expectation–maximisation (EM) algorithm of Dempster et al. (1977).

$$\theta^{(i+1)} = \arg\max_\theta \left(\max_q \mathcal{F}(q, \theta) \right)$$

Fix an arbitrary parameter value $\theta^{(i)} \in \Theta$. The optimisation of the auxiliary distribution q proceeds by observing that

$$\mathcal{F}(q, \theta^{(i)}) = \ell(\theta^{(i)}; X) \; - \; D_{\mathrm{KL}}\big(q \mid\mid f_{\theta^{(i)}}(Z \mid X)\big).$$

Because $\ell(\theta^{(i)}; X)$ is independent of q, maximising $\mathcal{F}(q, \theta^{(i)})$ with respect to q is equivalent to minimising $D_{\mathrm{KL}}\big(q \mid\mid f_{\theta^{(i)}}(Z \mid X)\big)$. The KL divergence is non-negative, a fact that follows immediately from Jensen's inequality

Proof.

$$\begin{aligned}
D_{\mathrm{KL}}(q \mid\mid f_{\theta^{(i)}}(Z \mid X)) &= -\mathbb{E}_{Z\sim\mathbb{Q}}\left[\log\left(\frac{f_{\theta^{(i)}}(Z \mid X)}{q(Z)}\right)\right] \\
&\geq -\log\left(\mathbb{E}_{Z\sim\mathbb{Q}}\left[\frac{f_{\theta^{(i)}}(Z \mid X)}{q(Z)}\right]\right) \\
&= -\log\left(\int \frac{f_{\theta^{(i)}}(Z \mid X)}{q(Z)} q(Z)\, dZ\right) \\
&= -\log(1) \\
&= 0
\end{aligned}$$

□

Equality is achieved if and only if $q(Z) = f_{\theta^{(i)}}(Z \mid X)$, in which case the objective attains the exact log-likelihood: $\mathcal{F}(q, \theta^{(i)}) = \ell(\theta^{(i)}; X)$. Consequently

$$f_{\theta^{(i)}}(Z \mid X) \; = \; \arg\min_q \; D_{\mathrm{KL}}\big(q \mid\mid f_{\theta^{(i)}}(Z \mid X)\big)$$

With the optimal auxiliary distribution fixed at $q_i(Z) = f_{\theta^{(i)}}(Z \mid X)$, the second step of the procedure aims to update the parameter θ. We do so by maximising the functional

$$\mathcal{F}(q_i, \theta) = \mathbb{E}_{Z\sim\mathbb{Q}_i}\left[\log\left(\frac{f_\theta(X, Z)}{q_i(Z)}\right)\right]$$

with respect to θ while keeping $q_i(Z)$ constant. Substituting $q_i(Z) = f_{\theta^{(i)}}(Z \mid X)$ we get the following objective:

$$\max_\theta \mathbb{E}_{Z\sim\mathbb{P}_{\theta^{(i)}}(\cdot\mid X)}\left[\log\left(\frac{f_\theta(X, Z)}{f_{\theta^{(i)}}(Z \mid X)}\right)\right]$$

The logarithm in the integrand can be split as $\log(f_\theta(X,Z)) - \log(f_{\theta^{(i)}}(Z \mid X))$. The second term is constant with respect to θ, so it does not influence the maximisation. Thus, the θ estimate reduces to

$$\theta^{(i+1)} = \arg\max_\theta \mathbb{E}_{Z\sim\mathbb{P}_{\theta^{(i)}}(\cdot|X)}\big[\log(f_\theta(X,Z))\big]$$

Algorithm 1 Expectation Maxmization

Initialize parameters $\theta^{(0)}$
$i \leftarrow 0$
while not converged **do**
  Step 1: $q_i \leftarrow f_{\theta^{(i)}}\big(Z \mid X\big)$
  Step 2: $\theta^{(i+1)} \leftarrow \arg\max_\theta \mathbb{E}_{Z\sim q_i}\big[\log(f_\theta(X,Z))\big]$
  $i \leftarrow i+1$
end while
return $\theta^{(i)}$

In other words, we maximise the expected full-data log-likelihood, where the expectation is taken under the posterior of the latent variables computed with the current parameter estimates θ_i. Thus we have identified an iterative approach to the hard optimisation problem, which is sometimes easier to implement under specific conditions (e.g. Exponential Families (see Ng et al., 2004)) than direct optimisation of the marginal likelihood.

5 Analysis of the EM Algorithm

5.1 Guaranteed Likelihood Monotonicity

By now we have derived the EM algorithm and shown how it works by iteratively maximising a lower bound on the observed data likelihood. However, critical questions remain. Does this bound maximisation procedure actually drive the marginal likelihood upwards, and if so, does it converge to a global maximum or get stuck at a local optimum? Furthermore, what can we say about the speed of convergence? In what follows, we examine these fundamental properties of EM by establishing conditions for convergence, characterising the nature of the bounds, and deriving bounds on its asymptotic rate of ascent.

Theorem 5.1. *Let $\{\theta^{(i)}\}_{i=0}^\infty$ be the sequence of parameter estimates produced by the EM algorithm applied to the observed data X. Then for every iteration $i \geq 0$ we have that the observed data log-likelihood,*

$$\ell(\theta^{(i)}; X) \leq \ell(\theta^{(i+1)}; X)$$

ie. monotone increasing

Proof. Let $\theta^{(i)}$ be the current EM estimate and define $q_i = f_{\theta^{(i)}}(Z \mid X)$. By using Lemma 4.2, and we can make the following observation,

$$\begin{aligned}
\ell(\theta^{(i)}; X) &= \mathcal{F}(q_i, \theta^{(i)}) \\
&\leq \mathcal{F}(q_i, \theta^{(i+1)}) \quad \text{by second EM step} \\
&\leq \mathcal{F}(q_{i+1}, \theta^{(i+1)}) \quad \text{by first EM step} \\
&= \ell(\theta^{(i+1)}; X)
\end{aligned}$$

Thus we have shown that,

$$\ell(\theta^{(i)}; X) \leq \ell(\theta^{(i+1)}; X)$$

□

5.2 Convergence of the Likelihood Sequence

Theorem 5.2. *Let $\{\theta^{(i)}\}_{i=0}^{\infty}$ be the sequence of parameter estimates produced by the EM algorithm on the observed data X, and let*

$$\{\ell(\theta^{(i)}; X)\}_{i=0}^{\infty}$$

be the corresponding observed-data log-likelihoods. If this sequence is bounded above, then it converges.

Proof. Since, by Theorem 5.1, the sequence $\{\ell(\theta^{(i)}; X)\}_{i=0}^{\infty}$ is monotone increasing and by assumption bounded above. Thus the Monotone Convergence Theorem implies that $\{\ell(\theta^{(i)}; X)\}_{i=0}^{\infty}$ converges. □

As a direct consequence of Theorem 5.2 , we see that the only remaining requirement for convergence of the EM algorithm's log-likelihood sequence is boundedness. As long as $\ell(\theta; X)$ cannot grow without bound, the monotonic increase established in 5.1 guarantees that the sequence converges to some finite limit.

5.3 Convergence of Parameter Estimates

Next we ask whether the EM iterates $\{\theta^{(i)}\}$ itself converge to a point with vanishing score, i.e. $\nabla_\theta \ell(\theta; X) = 0$. Following the work by Wu (1983) and Vaida (2005) we will state the central convergence results. To this end we introduce the EM transition function $M : \Theta \to \Theta$ for which we know for every iteration that $M(\theta^{(i)}) = \theta^{(i+1)}$, and the interior stationary set $S = \{\theta \in \operatorname{int}\Theta : \nabla_\theta \ell(\theta; X) = 0\}$. Further we say that a point $\theta^* \in \Theta$ is a *limit point* of a sequence $\{\theta_k\}$ if there exists a subsequence $\{\theta_{k_j}\}$ with $\theta_{k_j} \to \theta^*$. If in fact the full sequence converges, $\theta_k \to \theta^*$, then θ^* is called the *convergence point* of $\{\theta_k\}$. Therefore we call the EM algorithm *convergent* if, for any initialization $\theta^{(0)}$, the iterate sequence $\{\theta^{(i)}\}$ satisfies $\theta^{(i)} \to \theta^*$ for some $\theta^* \in \Theta$. Throughout the convergence theorems we assume the following regularity conditions:

Definition 5.1 (Regularity Conditions)**.**

(a) Θ is an open subset of $\mathbb{R}^p$.

(b) The observed-data log-likelihood $\ell(\theta; X)$ is continuously differentiable on Θ, so that $\nabla_\theta \ell(\theta; X)$ exists everywhere in Θ.

(c) The level set $\Theta_\theta := \{\theta^* \in \Theta : \ \ell(\theta^*; X) \geq \ell(\theta; X)\}$ is compact for any $\ell(\theta; X) > -\infty$.

(d) The distribution $f_\theta(Z \mid X)$ has the same support for all $\theta \in \Theta$.

(e) The function $Q(\theta^*|\theta) := \mathbb{E}_{Z\sim\mathbb{P}_\theta(\cdot|X)}\big[\log f_{\theta^*}(X, Z)\big]$ is continuous in both θ^* and θ, and is continuously differentiable in θ^*.

(f) All stationary points in S are isolated.

Lemma 5.3. *Let $\theta \in \Theta$ be arbitrary. If the mapping $\theta' \mapsto Q(\theta'|\theta)$ admits a unique global maximizer in Θ, then the EM map*

$$M : \Theta \to \Theta, \quad M(\theta) = \arg\max_{\theta' \in \Theta} Q(\theta'|\theta)$$

is continuous at θ.

Proof. First, recall that $M(\cdot)$ is continuous at θ if for a sequence $\theta_k \to \theta \implies M(\theta_k) \to M(\theta)$, by EM we also have that

$$Q(M(\theta_k)|\theta_k) \geq Q(\theta'|\theta_k), \ \ \forall \theta' \in \Theta$$

Taking the $\liminf$ as $k \to \infty$ and using continuity of Q in both arguments Condition 5.1 (e) together with $\theta_k \to \theta$ gives

$$\liminf_{k\to\infty} Q(M(\theta_k)|\theta_k) = \liminf_{k\to\infty} Q(M(\theta_k)|\theta) \geq \liminf_{k\to\infty} Q(\theta'|\theta_k) = Q(\theta'|\theta)$$

By Theorem 5.1, each iterate $M(\theta_k)$ lies in the level set Θ_{θ_0}. Since Θ_{θ_0} is compact, it is bounded, and hence the sequence $\{M(\theta_k)\} \subset \Theta_{\theta_0}$ is bounded as well. Furthermore, now that $\{M(\theta_k)\}$ is bounded, the Bolzano–Weierstrass theorem guarantees that there exists a subsequence $\{M(\theta_{k_j})\}$ which converges to some point $\tilde{M}$. Thus, for every $\theta' \in \Theta$ we have

$$Q(\tilde{M}|\theta) \geq \liminf_{k\to\infty} Q(M(\theta_k)|\theta) \geq Q(\theta'|\theta)$$

Because $Q(\cdot|\theta)$ admits a unique global maximizer, this forces $\tilde{M} = M(\theta)$. Hence every convergent subsequence $\{M(\theta_{k_j})\}$ must tend to $M(\theta)$, and by combining this with the boundedness of $\{M(\theta_k)\}$ it is verifiable B.2 that this implies

$$M(\theta_k) \to M(\theta)$$

□

Lemma 5.4. *Fix any $l^* \in \mathbb{R}$ there exist at most a finite number of stationary points $\theta^* \in S$ which satisfy $\ell(\theta^*; X) = l^*$*

Proof. Fix an arbitrary level $l^* \in \mathbb{R}$, then define the set $S(l^*) := \{\theta^* \in S : \ell(\theta^*; X) = l^*\} \subset \Theta_{\theta^*}$. By the regularity assumptions we know that $S(l^*)$ is bounded. Suppose, for contradiction, that $S(l^*)$ is infinite. Then there exists a sequence of *distinct* points $\{\theta_k^*\} \in S(l^*)$ and by boundedness we are guaranteed that there exists some subsequence which converges to some $\theta_0 \in \Theta_{\theta^*}$. Regularity condition (b) states that $\nabla_\theta \ell(\theta; X)$ is continuous. Hence

$$0 = \lim_{k \to \infty} \nabla_\theta \ell(\theta_k^*; X) = \nabla_\theta \ell(\theta_0; X),$$

Hence θ_0 is a stationary point, i.e. $\theta_0 \in S$. Because θ_0 is simultaneously a limit point, this violates regularity condition (f). We therefore conclude that $S(l^*)$ must be finite. □

Other statements, which we omit proving, can be verified in (Mclachlan and Krishnan, 2007, Chapter 3) and (Vaida, 2005):

Lemma 5.5. *Let $\theta' \in \Theta$ satisfy $\ell(M(\theta'); X) = \ell(\theta'; X)$. Then $\theta' \in S$, and θ' is a maximiser of the auxiliary function $Q(\cdot \mid \theta')$.*

Theorem 5.6 (Wu (1983))**.** *Assume (i) M is closed over the complement $\Theta \backslash S$; and (ii) the likelihood is strictly increased outside S, namely $\ell(\theta^{(i+1)}; X) > \ell(\theta^{(i)}; X)$ whenever $\theta^{(i)} \notin S$. Then every limit point of the sequence $\{\theta^{(i)}\}$ lies in S, and the likelihood values increase monotonically to a finite limit $\ell(\theta^*; X)$ for some stationary point $\theta^* \in S$.*

Thus, the EM iteration possesses the following qualitative behaviour. It either converges to a single stationary point, or else every subsequential limit lies in S and all such limits attain the same likelihood value. Observe that assumption (i) is automatically met under the regularity conditions 5.1. However, Theorem 5.6 does not, by itself, guarantee convergence of the entire sequence $\{\theta^{(i)}\}$. Consequently, the objective is to ascertain whether it is possible to achieve a stronger result.

Theorem 5.7 (Vaida (2005))**.** *Assume that the EM map $M(\cdot)$ is continuous. Then the EM sequence $\{\theta^{(i)}\}$ exhibits exactly one of the following behaviours:*

(i) Convergence. *The sequence converges to a stationary point $\theta^* \in \Theta$ with $M(\theta^*) = \theta^*$.*

(ii) Finite cycle. *There exists a finite set $\mathcal{C} = \{\theta_1^*, \ldots, \theta_m^*\} \subset S$ such that*

- *(a) each θ_j^* attains the same likelihood value, $\ell(\theta_1^*; X) = \cdots = \ell(\theta_m^*; X)$.*
- *(b) the EM map cycles through the points in $\mathcal{C}$: $M(\theta_1^*) = \theta_2^*$, $M(\theta_2^*) = \theta_3^*$, $\ldots$, $M(\theta_m^*) = \theta_1^*$.*

Theorem 5.8 (Main Result Vaida (2005))**.** *Assuming Lemma 5.3 holds. Then, for any initial value $\theta^{(0)} \in \Theta$, the EM iterates satisfy*

$$\theta^{(i)} \;\to\; \theta^* \quad \text{as } i \to \infty,$$

where $\theta^* \in S$ is a stationary point with $M(\theta^*) = \theta^*$.

Proof. Lemma 5.3 guarantees that the assumptions of Theorem 5.7 are met, so the EM sequence $\{\theta^{(i)}\}$ either converges or enters a finite cycle of limit points. Assume, for contradiction, that the latter occurs and that the cycle contains at least two distinct $\theta_1^* \neq \theta_2^*$ with $M(\theta_1^*) = \theta_2^*$. Because θ_1^* and θ_2^* are limit points, Theorem 5.6 implies they lie in S, with $\ell(\theta_1^*; X) = \ell(\theta_2^*; X)$. Moreover, the cycle relation gives $\ell(\theta_1^*; X) = \ell(M(\theta_1^*); X)$. Lemma 5.5 now yields that θ_1^* maximises $Q(\,\cdot\,|\theta_1^*)$. Since $M(\theta_1^*) = \theta_2^*$, the point θ_2^* by definition of $M(\cdot)$ is also a maximiser of the same function. Uniqueness of the maximiser then forces $\theta_1^* = \theta_2^*$, contradicting our assumption of two distinct points. Hence a non-trivial cycle cannot occur, and the EM sequence must converge to a single stationary point $\theta^* \in S$. □

In the end a large collection of EM-convergence results is available each relying on some technical assumptions (see Mclachlan and Krishnan, 2007, Chapter 3). In practice, though, well-behaved models lead the algorithm to settle at a likelihood maximum, usually local and occasionally global. True saddle-point convergence is possible but seldom observed in empirical work, and the tendency to stop at a local optimum can also be encountered in other optimization techniques for non-convex landscapes (cf. Schafer, 1997, pp. 51–52). This means that, in practice, one should run EM from several different initial values $\theta^{(0)}$, compare the final likelihoods, and choose θ^* as the estimate for which the solution yields the highest observed-data log-likelihood.

5.4 Rate of Convergence

Next, we will briefly examine the convergence behaviour of the EM algorithm. Since it generally converges only at a linear (and thus relatively slow) rate, unlike the quadratic convergence of Newton's method, various accelerated or modified EM methods have been proposed (cf. Mclachlan and Krishnan, 2007).

Proof. Again define the EM self-mapping by $\theta^{(i+1)} = M(\theta^{(i)})$. Now suppose $\theta^{(i)}$ converges to some stationary point θ^* (ie. $\theta^* = M(\theta^*)$) and M is continuous, then a first-order Taylor expansion of M about θ^* gives

$$M(\theta^{(i)}) \approx M(\theta^*) + M'(\theta^*)(\theta^{(i)} - \theta^*)$$

First note that $M(\theta^{(i)}) = \theta^{(i+1)}$ and $M(\theta^*) = \theta^*$ then we can rewrite the equation the following way,

$$\theta^{(i+1)} - \theta^* \approx M'(\theta^*)(\theta^{(i)} - \theta^*)$$

In particular, the quantity $\epsilon^{(i+1)} := \theta^{(i+1)} - \theta^*$ is the error after $i+1$ steps, while $\epsilon^{(i)} := \theta^{(i)} - \theta^*$ is the error after i steps. The relation

$$\epsilon^{(i+1)} \approx M'(\theta^*)\epsilon^{(i)}$$

shows that each new error is obtained by applying the linear map $M'(\theta^*)$ to the previous one. Consequently the errors shrink by a roughly constant factor each iteration (i.e. the method converges at an approximately linear rate). □

6 Standard Errors

6.1 Review of MLE Asymptotics

A major drawback of the EM algorithm is that, unlike gradient methods, it does not provide immediate estimates of the standard errors. Consequently, multiple approaches have been developed to find them. We will focus on the work of Louis (1982). First, however, we will briefly recall a classical statistical result (cf. Casella and Berger, 2002, pp. 472-474).

Theorem 6.1 (Asymptotic normality of the MLE)**.** *Under some regularity conditions,*

$$\hat{\theta}_{\text{MLE}} \xrightarrow{d} \mathcal{N}\big(\theta_0,\, n^{-1}\mathcal{I}(\theta_0)^{-1}\big),$$

where the Fisher information matrix is defined by

$$\mathcal{I}(\theta_0) = -\,\mathbb{E}_{\theta_0}\big[\nabla^2_\theta\, \ell(\theta; X)\big].$$

Because the expectation in $\mathcal{I}(\theta_0)$ depends on the unknown model, practitioners replace it with the observed information, the negative second derivative of the log-likelihood at the data (cf. Schafer, 1997, p.57). With complete data this is simply $-\nabla^2_\theta \ell(\theta; X, Z)$. In the presence of missing values we should take $-\nabla^2_\theta \ell(\theta; X)$, but the required Hessian of the marginal log-likelihood is rarely available in closed form, the very obstacle that motivates the EM algorithm.

6.2 The Louis Method

The Louis (1982) Method bypasses this problem by rewriting the observed information through complete-data quantities:

Theorem 6.2. *Assuming that differentiation and integration are interchangeable.*

$$-\nabla^2_\theta \ell(\theta; X) = \mathbb{E}_{Z\sim\mathbb{P}_\theta(\cdot|X)}\big[-\nabla^2_\theta \ell(\theta; X, Z)\big] - \operatorname{Var}_{Z\sim\mathbb{P}_\theta(\cdot|X)}\big[\nabla_\theta \ell(\theta; X, Z)\big]$$

The expectations are taken with respect to the conditional distribution of the missing part Z given the observed part X. Both terms rely solely on the complete-data log-likelihood, whose derivatives are available from the model specification. In what follows we assume that differentiation under the integral sign is permissible. Dominated convergence provides a sufficient justification, something we know holds for regular exponential-family models.

Proof. By differentiating twice and using the chain rule one can verify that,

$$-\nabla_\theta^2 \ell(\theta; X) = -\frac{\nabla_\theta^2 f_\theta(X)}{f_\theta(X)} + \frac{\nabla_\theta f_\theta(X) \left(\nabla_\theta f_\theta(X)\right)^\top}{f_\theta(X)^2} \tag{6.1}$$

$$= -\frac{\nabla_\theta^2 f_\theta(X)}{f_\theta(X)} + \nabla_\theta \ell(\theta; X) \left(\nabla_\theta \ell(\theta; X)\right)^\top \tag{6.2}$$

Now, let's take a look at both summands.

$$\begin{aligned}
-\frac{\nabla_\theta^2 f_\theta(X)}{f_\theta(X)} &= -\frac{\nabla_\theta^2 \int f_\theta(X, Z) dZ}{f_\theta(X)} \\
&= -\int \frac{\nabla_\theta^2 f_\theta(X, Z)}{f_\theta(X)} dZ \\
&= -\int \frac{\nabla_\theta^2 f_\theta(X, Z)}{f_\theta(X)} \frac{f_\theta(X, Z)}{f_\theta(X, Z)} dZ \\
&= -\int \frac{\nabla_\theta^2 f_\theta(X, Z)}{f_\theta(X, Z)} f_\theta(Z \mid X) dZ
\end{aligned}$$

Observe that the expression $-\frac{\nabla_\theta^2 f_\theta(X)}{f_\theta(X)}$ matches $-\frac{\nabla_\theta^2 f_\theta(X,Z)}{f_\theta(X,Z)}$, the former is based on the marginal likelihood, while the latter comes from the complete likelihood. Consequently, applying the original (6.2), we can write

$$\begin{aligned}
-\frac{\nabla_\theta^2 f_\theta(X)}{f_\theta(X)} &= \int \left(-\nabla_\theta^2 \ell(\theta; X, Z) - \nabla_\theta \ell(\theta; X, Z) \left(\nabla_\theta \ell(\theta; X, Z)\right)^\top\right) f_\theta(Z \mid X) dZ \\
&= \mathbb{E}_{Z \sim \mathbb{P}_\theta(\cdot|X)} \left[-\nabla_\theta^2 \ell(\theta; X, Z)\right] - \mathbb{E}_{Z \sim \mathbb{P}_\theta(\cdot|X)} \left[\nabla_\theta \ell(\theta; X, Z) \left(\nabla_\theta \ell(\theta; X, Z)\right)^\top\right]
\end{aligned}$$

Next we take a closer look at $\nabla_\theta \ell(\theta; X)$ and essentially apply the same trick,

$$\begin{aligned}
\nabla_\theta \ell(\theta; X) &= \frac{\nabla_\theta \int f_\theta(X, Z) dZ}{f_\theta(X)} \\
&= \int \frac{\nabla_\theta f_\theta(X, Z)}{f_\theta(X)} dZ \\
&= \int \frac{\nabla_\theta f_\theta(X, Z)}{f_\theta(X)} \frac{f_\theta(X, Z)}{f_\theta(X, Z)} dZ \\
&= \int \frac{\nabla_\theta f_\theta(X, Z)}{f_\theta(X, Z)} f_\theta(Z \mid X) dZ \\
&= \int \nabla_\theta \ell(\theta; X, Z) f_\theta(Z \mid X) dZ \\
&= \mathbb{E}_{Z \sim \mathbb{P}_\theta(\cdot|X)} \left[\nabla_\theta \ell(\theta; X, Z)\right]
\end{aligned}$$

Putting everything back together we have,

$$\begin{aligned}
-\nabla_\theta^2 \ell(\theta; X) =& \mathbb{E}_{Z \sim \mathbb{P}_\theta(\cdot|X)} \left[-\nabla_\theta^2 \ell(\theta; X, Z)\right] - \mathbb{E}_{Z \sim \mathbb{P}_\theta(\cdot|X)} \left[\nabla_\theta \ell(\theta; X, Z) \left(\nabla_\theta \ell(\theta; X, Z)\right)^\top\right] + \\
& \mathbb{E}_{Z \sim \mathbb{P}_\theta(\cdot|X)} \left[\nabla_\theta \ell(\theta; X, Z)\right] \mathbb{E}_{Z \sim \mathbb{P}_\theta(\cdot|X)} \left[\nabla_\theta \ell(\theta; X, Z)\right]^\top \\
=& \mathbb{E}_{Z \sim \mathbb{P}_\theta(\cdot|X)} \left[-\nabla_\theta^2 \ell(\theta; X, Z)\right] - \mathrm{Var}_{Z \sim \mathbb{P}_\theta(\cdot|X)} \left[\nabla_\theta \ell(\theta; X, Z)\right]
\end{aligned}$$

□

To estimate the covariance matrix of the EM estimator, insert the final iterate $\hat{\theta}_{\mathrm{EM}}$ into Louis' identity 6.2:

$$\mathcal{I}_{\mathrm{obs}}\big(\hat{\theta}_{\mathrm{EM}}\big) \;=\; \mathbb{E}_{Z\sim\mathbb{P}_{\hat{\theta}_{\mathrm{EM}}}(\cdot|X)}\Big[-\nabla_\theta^2\,\ell\big(\hat{\theta}_{\mathrm{EM}};X,Z\big)\Big] \;-\; \mathrm{Var}_{Z\sim\mathbb{P}_{\hat{\theta}_{\mathrm{EM}}}(\cdot|X)}\big[\nabla_\theta\,\ell\big(\hat{\theta}_{\mathrm{EM}};X,Z\big)\big].$$

An approximate covariance matrix for the EM estimate is then given by the inverse observed Fisher information:

$$\mathrm{Cov}\big(\hat{\theta}_{\mathrm{EM}}\big) \;\approx\; \mathcal{I}_{\mathrm{obs}}^{-1}\big(\hat{\theta}_{\mathrm{EM}}\big).$$

Further note that at the Maximum-Likelihood and or EM solution, the expected complete-data score vanishes, $\mathbb{E}_{Z\sim\mathbb{P}_\theta(\cdot|X)}\left[\nabla_\theta\ell(\theta;X,Z)\right] = \nabla_\theta\ell(\theta;X) = 0$. Hence Louis' identity for the observed Fisher information simplifies to

$$\begin{aligned}\mathcal{I}_{\mathrm{obs}}\big(\hat{\theta}_{\mathrm{EM}}\big) \;&= \mathbb{E}_{Z\sim\mathbb{P}_{\hat{\theta}_{\mathrm{EM}}}(\cdot|X)}\Big[-\nabla_\theta^2\,\ell\big(\hat{\theta}_{\mathrm{EM}};X,Z\big)\Big] \;-\\ &\qquad \mathbb{E}_{Z\sim\mathbb{P}_{\hat{\theta}_{\mathrm{EM}}}(\cdot|X)}\left[\nabla_\theta\ell(\hat{\theta}_{\mathrm{EM}};X,Z)\left(\nabla_\theta\ell(\hat{\theta}_{\mathrm{EM}};X,Z)\right)^\top\right]\end{aligned}$$

In the end several techniques are available for obtaining standard-error estimates after fitting a model with the EM algorithm. Because the usual derivation of Louis' identity relies on interchanging integration and differentiation, its validity is not guaranteed in every setting. But nevertheless, for exponential-family models it furnishes a *consistent large-sample estimate* of the covariance matrix. Other approaches such as bootstrap are conceptually straightforward, but computationally expensive, because each replication requires a full rerun of the EM routine, whose convergence is typically slow. Alternatives that avoid repeated maximisation include Supplemented EM of Meng and Rubin (1991) and Oakes' Method Oakes (1999).

7 Example and Simulation

7.1 Linear Regression

We assume the following linear-regression setup: for each $i = 1,\dots,n$,

$$y_i = x_i^\top\beta + \epsilon_i, \quad x_i \perp\!\!\!\perp \epsilon_i, \quad x_i \sim \mathcal{N}_p(\mu_x,\Sigma_x), \;\; \epsilon_i \sim \mathcal{N}(0,\sigma^2)$$

Here $\beta \in \mathbb{R}^p$ is the unknown coefficient vector, x_i is a p-variate Gaussian predictor, and ϵ_i is independent Gaussian noise. Thus it also follows that

$$(x_i,\,\epsilon_i)^\top \;\sim\; \mathcal{N}_{p+1}\left(\begin{pmatrix}\mu_x\\0\end{pmatrix},\begin{pmatrix}\Sigma_x & 0\\0 & \sigma^2\end{pmatrix}\right)$$

Define

$$c := \begin{pmatrix}0\\\alpha\end{pmatrix}, \quad M := \begin{pmatrix}I_p & 0\\\beta^\top & 1\end{pmatrix}.$$

Then one can write

$$(x_i, y_i)^\top = c + M \begin{pmatrix} x_i \\ \epsilon_i \end{pmatrix},$$

and hence *by linearity*

$$(x_i, y_i)^\top \sim \mathcal{N}_{p+1} \left(\underbrace{\begin{pmatrix} \mu_x \\ \alpha + \beta^\top \mu_x \end{pmatrix}}_{(p+1)\times 1}, \begin{pmatrix} \underbrace{\Sigma_x}_{p\times p} & \underbrace{\Sigma_x \beta}_{p\times 1} \\ \underbrace{\beta^\top \Sigma_x}_{1\times p} & \underbrace{\beta^\top \Sigma_x \beta + \sigma^2}_{1\times 1} \end{pmatrix} \right) \tag{7.1}$$

For easier computation, note that the conditional distribution of y_i given x_i is

$$y_i \mid x_i \ \sim \ \mathcal{N}\big(\mathbb{E}[y_i \mid x_i],\, \mathbb{V}[y_i \mid x_i]\big),$$

where

$$\mathbb{E}[y_i \mid x_i] \ = \ \alpha \ + \ x_i^\top \beta, \quad \mathbb{V}[y_i \mid x_i] \ = \ \sigma^2.$$

Thus the joint density factorizes as $f(x_i, y_i) = f(x_i) f(y_i \mid x_i)$ and in case of no missing covariates the inference for $(\alpha, \beta, \sigma^2)^\top$ can be solely based on the conditional part.

7.1.1 EM in case of Missing Covariates

Now, we assume that missingness occurs exclusively in the covariates x_i. However, we do not impose or assume any particular missingness pattern or structure (e.g. the *Random Pattern* in A.1). Thus, we introduce the following notation to succinctly represent the missingness structure:

$$\mathcal{M}(i) = \{j \ : \ x_{ij} \text{ is not observed}\}, \quad \mathcal{O}(i) = \{j \ : \ x_{ij} \text{ is observed}\}.$$

Here, $\mathcal{M}(i)$ denotes the set of indices corresponding to missing covariates for observation i, while $\mathcal{O}(i)$ represents the indices of the observed covariates. Further, since the EM algorithm involves repeatedly computing conditional expectations, we leverage the well-known conditional distribution property of the multivariate normal (cf. Johnson and Wichern, 2007, p. 163). To conveniently express this, we define a combined vector w_i consisting of the observed outcome y_i and the observed covariates $x_{i,\mathcal{O}(i)}$:

$$w_i = \begin{pmatrix} y_i \\ x_{i,\mathcal{O}(i)} \end{pmatrix}.$$

Then, the conditional distribution of the missing covariates $x_{i,\mathcal{M}(i)}$, given the observed data $w_i = (y_i, x_{i,\mathcal{O}(i)}^\top)^\top$, is

$$x_{i,\mathcal{M}(i)} \mid w_i \sim \mathcal{N}\left(\mu_{x,\mathcal{M}(i)|w_i},\ \Sigma_{x,\mathcal{M}(i)|w_i}\right)$$

$$\mu_{x,\mathcal{M}(i)|w_i} = \mu_{\mathcal{M}(i)} + \Sigma_{\mathcal{M}(i),\mathcal{W}(i)} \Sigma^{-1}_{\mathcal{W}(i),\mathcal{W}(i)} (w_i - \mu_{\mathcal{W}(i)}) \tag{7.2}$$

$$\Sigma_{x,\mathcal{M}(i)|w_i} = \Sigma_{\mathcal{M}(i),\mathcal{M}(i)} - \Sigma_{\mathcal{M}(i),\mathcal{W}(i)}\Sigma^{-1}_{\mathcal{W}(i),\mathcal{W}(i)}\Sigma_{\mathcal{W}(i),\mathcal{M}(i)} \tag{7.3}$$

Note that Σ and μ refer to the joint parameters of the vector (x_i, y_i) (7.1), where we define $\mathcal{W}(i) := \mathcal{O}(i) \cup \{p+1\}$, the set of indices corresponding to the observed covariates and the response. The vector $x_{i,\mathcal{M}(i)}$ contains the missing components of x_i, and $\Sigma_{\mathcal{M}(i),\mathcal{W}(i)}$ is the submatrix of the joint covariance matrix of (x_i, y_i), taken over rows in $\mathcal{M}(i)$ and columns in $\mathcal{W}(i)$. To simplify notation, we adopt the convention that all subvectors and submatrices such as $x_{i,\mathcal{M}(i)}$, $\mu_{\mathcal{M}(i)}$, $\mu_{\mathcal{W}(i)}$ $\Sigma_{\mathcal{M}(i),\mathcal{W}(i)}$, etc. retain the global indices from the full vector or matrix they are sliced from. That is, for example,

$$(\Sigma_{\mathcal{M}(i),\mathcal{W}(i)})_{j,k} = \Sigma_{j,k}, \quad \text{for all } j \in \mathcal{M}(i),\, k \in \mathcal{W}(i),$$

where $\Sigma_{j,k}$ denotes the entry in the full covariance Matrix we derived above for $(x_i, y_i)^\top$ and similarly for other such quantities. This indexing convention enables concise expressions while maintaining consistency with the full joint structure.

7.1.2 Taking the Expectation

Since the joint full-data likelihood belongs to an exponential family, it's natural to express it using just five aggregates:

$$S_x = \sum_{i=1}^{n} x_i, \quad S_{xx} = \sum_{i=1}^{n} x_i x_i^\top, \quad S_{xy} = \sum_{i=1}^{n} x_i y_i, \quad S_y = \sum_{i=1}^{n} y_i, \quad S_{yy} = \sum_{i=1}^{n} y_i^2.$$

and those aggregates are then used directly in the maximization. Through straightforward manipulation using linear algebra, the joint log-likelihood can be compactly written as (see Appendix B.3)

$$\begin{aligned}\ell(\theta; X, y) \propto &- n\log(\det|\Sigma_x|) - \operatorname{Tr}\left(\Sigma_x^{-1}\left(S_{xx} - S_x\mu_x^\top - \mu_x S_x^\top + n\mu_x\mu_x^\top\right)\right) \\ &- n\log(\sigma^2) - \frac{1}{\sigma^2}\left(S_{yy} - 2\alpha S_y - 2S_{xy}^\top\beta + n\alpha^2 + 2\alpha S_x^\top\beta + \beta^\top S_{xx}\beta\right)\end{aligned}$$

Since the complete-data log-likelihood is linear in its sufficient statistics, taking its conditional expectation at a given θ and observed data boils down to computing the conditional expectations of those statistics.

$$\mathbb{E}_{x_\mathcal{M}|x_\mathcal{O},y}[S_x],\ \mathbb{E}_{x_\mathcal{M}|x_\mathcal{O},y}[S_{xx}],\ \mathbb{E}_{x_\mathcal{M}|x_\mathcal{O},y}[S_{xy}],\ \mathbb{E}_{x_\mathcal{M}|x_\mathcal{O},y}[S_y],\ \mathbb{E}_{x_\mathcal{M}|x_\mathcal{O},y}[S_{yy}],$$

In what follows we abbreviate these conditional expectations by $\tilde{S}_x, \tilde{S}_{xx}, \ldots$ and compute each:

Computation of $\tilde{S}_x$

$$\tilde{S}_x = \sum_{i=1}^{n} \mathbb{E}_{x_{i,\mathcal{M}(i)}|w_i}[x_i]$$

where the jth component of the inner expectation is given by (7.2):

$$\left(\mathbb{E}_{x_{i,\mathcal{M}(i)}|w_i}[x_i]\right)_j = \begin{cases} x_{ij}, & j \in \mathcal{O}(i), \\ (\mu_{x,\mathcal{M}(i)|w_i})_j, & j \in \mathcal{M}(i). \end{cases}$$

Computation of $\tilde{S}_{xy}$

$$\tilde{S}_{xy} = \sum_{i=1}^{n} y_i \mathbb{E}_{x_{i,\mathcal{M}(i)}|w_i}[x_i]$$

where again the jth component of the inner expectation is similarly given by (7.2):

$$\left(y_i \mathbb{E}_{x_{i,\mathcal{M}(i)}|w_i}[x_i]\right)_j = \begin{cases} y_i x_{ij}, & j \in \mathcal{O}(i), \\ y_i (\mu_{x,\mathcal{M}(i)|w_i})_j, & j \in \mathcal{M}(i). \end{cases}$$

Computation of $\tilde{S}_{xx}$

$$\tilde{S}_{xx} = \sum_{i=1}^{n} \mathbb{E}_{x_{i,\mathcal{M}(i)}|w_i}[x_i x_i^\top]$$

where the (j,k) component of the inner expectation is given by (7.2) and (7.3):

$$\left(\mathbb{E}_{x_{i,\mathcal{M}(i)}|w_i}[x_i x_i^\top]\right)_{jk} = \begin{cases} x_{ij}\, x_{ik}, & j,k \in \mathcal{O}(i) \\ x_{ij}\, (\mu_{x,\mathcal{M}(i)|w_i})_k, & j \in \mathcal{O}(i),\ k \in \mathcal{M}(i) \\ (\mu_{x,\mathcal{M}(i)|w_i})_j\, x_{ik}, & j \in \mathcal{M}(i),\ k \in \mathcal{O}(i) \\ (\Sigma_{x,\mathcal{M}(i)|w_i})_{jk} \ + \ (\mu_{x,\mathcal{M}(i)|w_i})_j\, (\mu_{x,\mathcal{M}(i)|w_i})_k, & j,k \in \mathcal{M}(i) \end{cases}$$

Computation of $\tilde{S}_y$ and $\tilde{S}_{yy}$ Given the complete observation of both quantities we can write

$$\tilde{S}_y = S_y, \quad \tilde{S}_{yy} = S_{yy}$$

By taking a closer look, we see that EM implicitly imputes each missing x_{ij}, its square x_{ij}^2, and any missing cross-product $x_{ij}x_{ik}$ by replacing them with their conditional expectations under the current parameter estimates.

7.1.3 The Algorithm

Now that we've determined how to compute the conditional expectations of the sufficient-statistics, we can turn to the closed-form parameter updates as shown in B.4.

$$\mu_x^{(i+1)} = \frac{1}{n}\,\tilde{S}_x^{(i)},$$

$$\Sigma_x^{(i+1)} = \frac{1}{n}\Big(\tilde{S}_{xx}^{(i)} - \frac{1}{n}\,\tilde{S}_x^{(i)}\tilde{S}_x^{(i)\top}\Big),$$

$$\beta^{(i+1)} = \Big(\tilde{S}_{xx}^{(i)} - \frac{1}{n}\,\tilde{S}_x^{(i)}\tilde{S}_x^{(i)\top}\Big)^{-1}\Big(\tilde{S}_{xy}^{(i)} - \frac{1}{n}\,\tilde{S}_y^{(i)}\,\tilde{S}_x^{(i)}\Big),$$

$$\alpha^{(i+1)} = \frac{1}{n}\,\tilde{S}_y^{(i)} \;-\; \frac{1}{n}\,\tilde{S}_x^{(i)\top}\beta^{(i+1)},$$

$$\sigma^{2\,(i+1)} = \frac{1}{n}\Big(\tilde{S}_{yy}^{(i)} - 2\alpha^{(i+1)}\tilde{S}_y^{(i)} - 2\tilde{S}_{xy}^{(i)\top}\beta^{(i+1)} + n\big(\alpha^{(i+1)}\big)^2 + 2\alpha^{(i+1)}\tilde{S}_x^{(i)\top}\beta^{(i+1)} + \beta^{(i+1)\top}\tilde{S}_{xx}^{(i)}\beta^{(i+1)}\Big).$$

Where the subscript i is used to indicate the current iteration. Although the algorithm reproduces the exact EM updates, its naive implementation is computationally costly. It loops over all n observations and, in the worst case, requires up to n separate inversions of sub–covariance matrices. Consequently, this pseudo code should be viewed as a pedagogical reference rather than applicable in real world scenarios. For practical and numerically stable implementations (see Schafer, 1997, Chapter 5).

Algorithm 2 EM for linear regression with covariate missingness

Require: observed pairs $(y_i, x_{i,\mathcal{O}(i)})_{i=1}^{n}$, tolerance ε

Notation: $\mathcal{O}(i)$ observed indices, $\mathcal{M}(i)$ missing indices, $w_i = (y_i, x_{i,\mathcal{O}(i)}^{\top})^{\top}$

initialise $\theta^{(0)} = (\mu^{(0)}, \Sigma_x^{(0)}, \alpha^{(0)}, \beta^{(0)}, \sigma^{2\,(0)})$ (mean imputation + OLS)
$t \leftarrow 0$
repeat
  reset $S_x, S_{xx}, S_{xy} \leftarrow 0$
  for $i = 1, \dots, n$ **do**
    if $\mathcal{M}(i) = \varnothing$ **then**
      $\hat{x}_i \leftarrow x_i, \quad C_i \leftarrow 0$
    else
      compute $\mu_{x,\mathcal{M}(i)|w_i}$, $\Sigma_{x,\mathcal{M}(i)|w_i}$ via (7.2)–(7.3)
      $\hat{x}_{i,\mathcal{M}(i)} \leftarrow \mu_{x,\mathcal{M}(i)|w_i}, \;\; \hat{x}_{i,\mathcal{O}(i)} \leftarrow x_{i,\mathcal{O}(i)}$
      $C_i \leftarrow \Sigma_{x,\mathcal{M}(i)|w_i}$
    end if
    $S_x \leftarrow S_x + \hat{x}_i$
    $S_{xx} \leftarrow S_{xx} + \hat{x}_i \hat{x}_i^{\top} + C_i$
    $S_{xy} \leftarrow S_{xy} + y_i \hat{x}_i$
  end for
  $\mu^{(t+1)} \leftarrow \dfrac{S_x}{n}$
  $\Sigma_x^{(t+1)} \leftarrow \dfrac{1}{n}\left(S_{xx} - \dfrac{1}{n} S_x S_x^{\top}\right)$
  $\beta^{(t+1)} \leftarrow \left(S_{xx} - \dfrac{1}{n} S_x S_x^{\top}\right)^{-1} \left(S_{xy} - \dfrac{S_y}{n} S_x\right)$
  $\alpha^{(t+1)} \leftarrow \dfrac{S_y - \beta^{(t+1)\top} S_x}{n}$
  $\sigma^{2\,(t+1)} \leftarrow \dfrac{1}{n}\Big(S_{yy} - 2\alpha^{(t+1)} S_y - 2\beta^{(t+1)\top} S_{xy} + n\big(\alpha^{(t+1)}\big)^2 + 2\alpha^{(t+1)} \beta^{(t+1)\top} S_x + \beta^{(t+1)\top} S_{xx} \beta^{(t+1)}\Big)$
  $\delta \leftarrow \max |\theta^{(t+1)} - \theta^{(t)}|$
  $t \leftarrow t + 1$
until $\delta < \varepsilon$
return $\theta^{(t)}$

7.2 Simulated Simple Linear Regression

To provide an intuitive visual example, we use the most elementary linear-regression setting. Assume the data arise from

$$y_i \;=\; \alpha \;+\; \beta x_i \;+\; \varepsilon_i,$$

where α is the intercept, β the slope, and ε_i denotes an independent error term for each observation.

Missingness mechanism. After generating (x_i, y_i), each covariate value x_i is missing according to a logistic model

$$\mathbb{P}(x_i \text{ missing } \mid y_i) \;=\; \text{logit}^{-1}\big(a_y + b_y\, y_i\big) \;=\; \frac{1}{1+\exp\big[-\big(a_y + b_y\, y_i\big)\big]}$$

So that the intercept a_y fixes the overall proportion of missing values, while the slope b_y determines how strongly the deletion probability increases ($b_y > 0$) or decreases ($b_y < 0$) with the response y_i. Because the missingness depends only on the observed variable y_i, the mechanism is *missing at random* (MAR).

Simulation setup. Throughout this section we draw observations with

$$\alpha = 0, \qquad \beta = 0.25, \qquad \sigma^2 = 50, \qquad x_i \sim \mathcal{N}\big(\mu_x = 0,\ \sigma_x^2 = 500\big)$$

For the logistic deletion model we fix $b_y = 1$ and set the intercept to $a_y = -\alpha - \beta\mu_x = 0$. Hence

$$\mathbb{P}(x_i \text{ missing } \mid y_i) \;=\; \frac{1}{1+\exp(-y_i)}.$$

the mechanism deletes, on average, 50 % of the covariate values x_i.

Results (one replication, $n = 1000$). The plot contrasts three regression lines estimated from the same simulated sample.

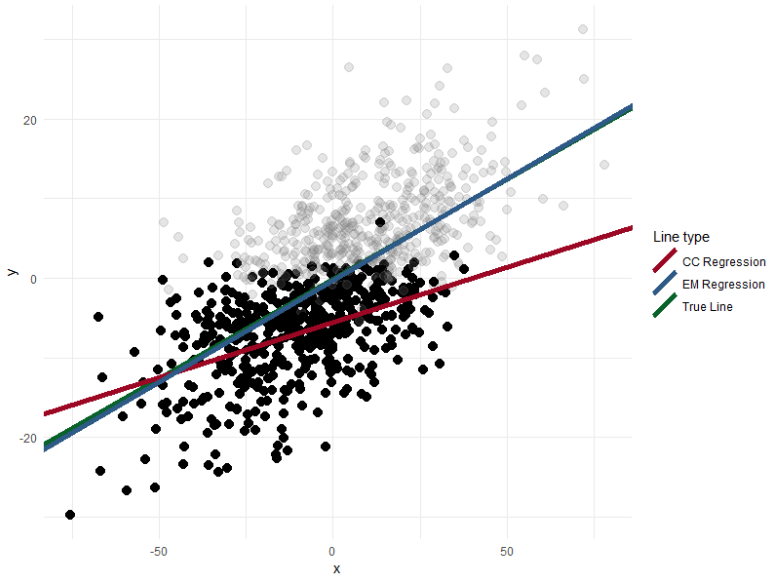

Figure 1: Scatter of the simulated sample with missing-data pattern (black = observed x_i, grey = missing x_i) and the three lines: *green* = true data-generating relationship $\big(\alpha = 0,\ \beta = 0.25\big)$, *red* = complete-case (CC) least squares fit, biased because high-y points tend to lose their x, *blue* = EM fit, which essentially recovers the true line.

We repeated the simulation 5 000 times with a smaller sample size of $n = 100$ (to reduce computation time). Table 7.2 reports the resulting bias and root-mean-squared error (RMSE). The EM estimator remains essentially unbiased for both the intercept and the slope and achieves markedly lower RMSEs. In contrast, the complete-case (CC) estimator shows a pronounced negative bias (about -5.66 for the intercept and -0.12 for the slope) and correspondingly larger RMSEs, so it fails to recover the true regression line.

Method	Parameter	Bias	RMSE
CC	Intercept	-5.6643	5.7116
CC	Slope	-0.1235	0.1296
EM	Intercept	-0.0560	1.1244
EM	Slope	-0.0052	0.0478

Table 1: Monte-Carlo simulation (true parameters: intercept = 0, slope = 0.25). Comparison of bias and RMSE for complete-case (CC) and EM estimators.

8 Conclusion

It has been demonstrated that when data are missing at random (MAR) or missing completely at random (MCAR), directly maximising the observed-data likelihood often becomes infeasible because the required integrals are intractable. The EM framework by Dempster et al. (1977) offers a practical workaround by iteratively solving a succession of simpler optimisation problems instead of one large, unwieldy problem. While the primary focus has been on the theoretical underpinnings, the practical success of the method hinges on the tractability of the conditional distribution of the missing values, the ease with which the corresponding ELBO can be optimised, and the availability of reliable standard errors, something EM does not deliver automatically. Note, however, that although Louis' method yields an exact expression for the observed-information matrix, it remains a large-sample estimate and as, (Schafer, 1997, pp. 63-64) cautions, may perform poorly or be misleading in smaller samples. In fact Schafer (1997), recommends obtaining uncertainty measures via the bootstrap, but this can be numerically burdensome because each bootstrap replicate requires a full EM run and the algorithm is inherently slow. Each sub-problem may present its own challenges, yet the heuristic frequently yields workable solutions across a broad spectrum of applications, including latent-variable models, censored-data analyses, and beyond.

A Figures

A.1 Missingness Patterns

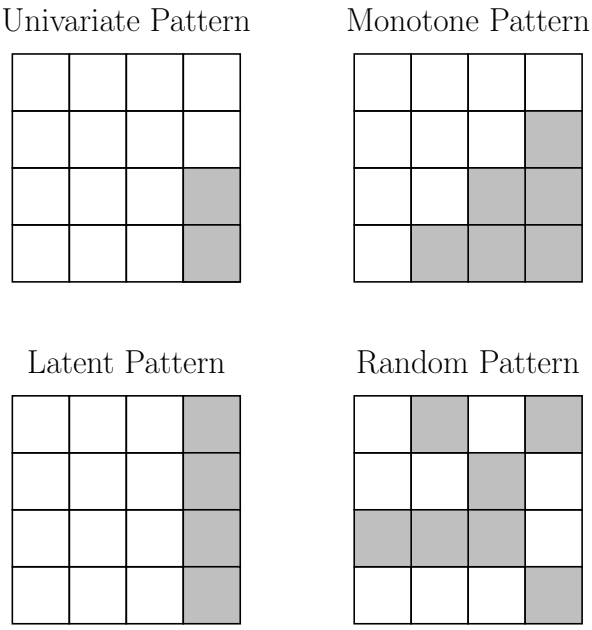

Figure 2: Illustration of four common missing-data patterns. Each grid represents a data matrix with rows as observations and columns as variables: white cells are observed values and grey cells are missing. *Univariate pattern* all missingness is confined to a single variable. *Monotone pattern* once a value is missing for a variable, all variables to its right are also missing in that row. *Latent pattern* entire blocks of variables are missing together, often reflecting an unmeasured latent factor. *Random pattern* missingness is scattered without an obvious structural order.

B Supplementary Examples and Derivations

B.1 Bivariate Normal with Univariate Missing Pattern

Let (x_i, y_i), $i = 1, \ldots, n$, be independently drawn from the bivariate normal distribution

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy} & \Sigma_y \end{pmatrix} \right)$$

where missingness occurs only in the y_i. The observed data can be arranged as

$$\begin{array}{ccccc} x_1 & \cdots & x_{n_1} & \cdots & x_n \\ y_1 & \cdots & y_{n_1} & & \end{array}$$

where n_1 is the number of non-missing y_i and $n - n_1$ the number of missing responses. Having this univariate missingness pattern we can make the following observation for the marginal likelihood.

$$\begin{aligned} L_{\text{ign}}(\theta; X, Y) &= \int \prod_{i=1}^{n} f_\theta(x_i, y_i)\, dy_{n_1+1} \dots dy_n \\ &= \prod_{i=1}^{n_1} f_\theta(x_i, y_i) \int \prod_{j=n_1+1}^{n} f_\theta(x_j, y_j)\, dy_{n_1+1} \dots dy_n \\ &= \prod_{i=1}^{n_1} f_\theta(x_i, y_i) \prod_{j=n_1+1}^{n} f_\theta(x_j) \\ &= \prod_{i=1}^{n_1} f_\theta(y_i \mid x_i) f_\theta(x_i) \prod_{j=n_1+1}^{n} f_\theta(x_j) \\ &= \prod_{i=1}^{n_1} f_\theta(y_i \mid x_i) \prod_{i=1}^{n} f_\theta(x_i) \end{aligned}$$

Now we have found the closed form solution for the problem because for any bivariate normal the corresponding conditional distribution (Johnson and Wichern, 2007, pp. 161-162) is given by

$$y_i \mid x_i \sim \mathcal{N}\left(\mu_y + \frac{\Sigma_{xy}}{\Sigma_x}(x_i - \mu_x),\, \Sigma_y - \frac{\Sigma_{xy}^2}{\Sigma_x}\right)$$

To simplify estimation, Anderson (1957) proposed an alternative parameterization that makes for a nice interpretation. Let $\beta_1 := \frac{\Sigma_{xy}}{\Sigma_x}$, $\beta_0 := \mu_y - \beta_1 \mu_x$ and $\nu := \Sigma_y - \frac{\Sigma_{xy}^2}{\Sigma_x}$ then we have that,

$$y_i \mid x_i \sim \mathcal{N}\left(\beta_0 + \beta_1 x_i,\, \nu\right)$$

and the corresponding log likelihood of the model for $\phi = (\mu_x.\Sigma_x, \beta_0, \beta_1, \nu)$ is given, by

$$\ell_{\text{ign}}(\phi) \propto -\frac{n}{2}\log(\Sigma_x) - \frac{1}{2\Sigma_x}\sum_{i=1}^{n}(x_i - \mu_x)^2 - \frac{n_1}{2}\log(\nu) - \frac{1}{2\nu}\sum_{i=1}^{n_1}(y_i - \beta_0 - \beta_1 x_i)^2$$

Maximizing the marginal likelihood with respect to μ_x and Σ_x yields the usual univariate MLEs for X, while the remaining parameters coincide with the

ordinary-least-squares estimates from regressing Y on X. Since this reparameterization is one-to-one (see Casella and Berger, 2002, Theorem 7.2.11), the corresponding MLEs follow immediately as

$$\hat{\mu}_y \;=\; \hat{\beta}_0 + \hat{\beta}_1\,\hat{\mu}_x, \quad \hat{\Sigma}_{xy} \;=\; \hat{\beta}_1\,\hat{\Sigma}_x, \quad \hat{\Sigma}_y \;=\; \hat{\nu} + \frac{\hat{\Sigma}_{xy}^2}{\hat{\Sigma}_x}\,.$$

B.2 Bounded Sequences whose every Subsequence shares the same Limit

Corollary B.1. *Let $\{a_n\}_{n\in\mathbb{N}}$ be a bounded sequence. Suppose that every subsequence $\{a_{n_k}\}_{k\in\mathbb{N}}$ converges to the same value $L \in \mathbb{R}$. Then the full sequence $\{a_n\}$ converges to L as well.*

Proof. Assume, toward a contradiction, that the sequence $\{a_n\}$ does *not* converge to L. By negating the definition of convergence we obtain

$$\exists\,\varepsilon > 0 \quad \forall\, N \in \mathbb{N} \quad \exists\, n \geq N \quad \text{such that} \quad |a_n - L| \geq \varepsilon.$$

Consequently one can extract a subsequence $\{a_{n_k}\}$ satisfying

$$|a_{n_k} - L| \;\geq\; \varepsilon \quad \forall k \in \mathbb{N}.$$

Because the original sequence is bounded, every subsequence is bounded as well. Hence, by the Bolzano–Weierstrass theorem, $\{a_{n_k}\}$ possesses a convergent subsequence $\{a_{n_{k_j}}\}$ with limit, say, L'. From $|a_{n_{k_j}} - L| \geq \varepsilon$ for all j we deduce $L' \neq L$. However, $\{a_{n_{k_j}}\}$ is itself a subsequence of the original sequence, so by hypothesis it must converge to L. This contradiction establishes that the assumption was false. Therefore the full sequence $\{a_n\}$ indeed converges to L. □

B.3 Expressing the Likelihood in terms of the Sufficient Statistics

$x_i \overset{\text{iid}}{\sim} \mathcal{N}_p(\mu_x, \Sigma_x)$ and $y_i \mid x_i \overset{\text{iid}}{\sim} \mathcal{N}(\alpha + x_i^\top \beta,\ \sigma^2)$ with $\theta = (\mu_x, \Sigma_x, \alpha, \beta, \sigma^2) \implies \ell(\theta; X, y) = \sum_{i=1}^n \ell(\mu_x, \Sigma_x; x_i) + \sum_{i=1}^n \ell(\alpha, \beta, \sigma^2; y_i \mid x_i)$

1.

$$\begin{aligned}
\sum_{i=1}^{n} \ell(\mu_x, \Sigma_x; x_i) &\propto \sum_{i=1}^{n} \left(-\frac{1}{2} \log(\det|\Sigma_x|) - \frac{1}{2}(x_i - \mu_x)^\top \Sigma_x^{-1}(x_i - \mu_x) \right) \\
&\propto -n \log(\det|\Sigma_x|) - \sum_{i=1}^{n} (x_i - \mu_x)^\top \Sigma_x^{-1}(x_i - \mu_x) \\
&\propto -n \log(\det|\Sigma_x|) - \sum_{i=1}^{n} \operatorname{Tr}\left(\Sigma_x^{-1}(x_i - \mu_x)(x_i - \mu_x)^\top\right) \\
&\propto -n \log(\det|\Sigma_x|) - \operatorname{Tr}\left(\Sigma_x^{-1}\left(\sum_{i=1}^{n}(x_i - \mu_x)(x_i - \mu_x)^\top\right)\right) \\
&\propto -n \log(\det|\Sigma_x|) - \operatorname{Tr}\left(\Sigma_x^{-1}\left(\sum_{i=1}^{n}(x_i x_i^\top - x_i \mu_x^\top - \mu_x x_i^\top + \mu_x \mu_x^\top)\right)\right) \\
&\propto -n \log(\det|\Sigma_x|) - \operatorname{Tr}\left(\Sigma_x^{-1}\left(S_{xx} - S_x \mu_x^\top - \mu_x S_x^\top + n\mu_x \mu_x^\top\right)\right)
\end{aligned}$$

2.

$$\begin{aligned}
\sum_{i=1}^{n} \ell(\alpha, \beta, \sigma^2; y_i \mid x_i) &\propto -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - (\alpha + x_i^\top \beta)\right)^2 \\
&\propto -n \log(\sigma^2) - \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(y_i^2 - 2\alpha y_i - 2 y_i x_i^\top \beta + \alpha^2 + 2\alpha x_i^\top \beta + (x_i^\top \beta)^2\right) \\
&\propto -n \log(\sigma^2) - \frac{1}{\sigma^2} \left(S_{yy} - 2\alpha S_y - 2 S_{xy}^\top \beta + n\alpha^2 + 2\alpha S_x^\top \beta + \sum_{i=1}^{n} \operatorname{Tr}(x_i^\top \beta) x_i^\top \beta \right) \\
&\propto -n \log(\sigma^2) - \frac{1}{\sigma^2} \left(S_{yy} - 2\alpha S_y - 2 S_{xy}^\top \beta + n\alpha^2 + 2\alpha S_x^\top \beta + \sum_{i=1}^{n} \beta^\top x_i x_i^\top \beta \right) \\
&\propto -n \log(\sigma^2) - \frac{1}{\sigma^2} \left(S_{yy} - 2\alpha S_y - 2 S_{xy}^\top \beta + n\alpha^2 + 2\alpha S_x^\top \beta + \beta^\top S_{xx} \beta \right)
\end{aligned}$$

B.4 Maximizing the Conditional Likelihood

Deriving $\mu_x^{(i+1)}$

$$\begin{aligned}
\frac{\partial}{\partial \mu_x} \ell(\theta; X, y) &\propto \frac{\partial}{\partial \mu_x} - \operatorname{Tr}\left(\Sigma_x^{-1}\left(S_{xx} - S_x \mu_x^\top - \mu_x S_x^\top + n\mu_x \mu_x^\top\right)\right) \\
&\propto \frac{\partial}{\partial \mu_x} \left(\operatorname{Tr}\left(\Sigma_x^{-1} S_x \mu_x^\top\right) + \operatorname{Tr}\left(\Sigma_x^{-1} \mu_x S_x^\top\right) - n \operatorname{Tr}(\Sigma_x^{-1} \mu_x \mu_x^\top) \right) \\
&\propto \frac{\partial}{\partial \mu_x} \left(S_x^\top \Sigma_x^{-1} \mu_x + \mu_x^\top \Sigma_x^{-1} S_x - n \mu_x^\top \Sigma_x^{-1} \mu_x \right) \\
&\propto \Sigma_x^{-1} S_x + \Sigma_x^{-1} S_x - 2n \Sigma_x^{-1} \mu_x \\
&\propto 2\Sigma_x^{-1} S_x - 2n \Sigma_x^{-1} \mu_x
\end{aligned}$$

Setting this expression to zero results in:

$$\frac{\partial}{\partial \mu_x}\ell(\theta; X, y) = 0 \iff \mu_x = \frac{1}{n}S_x$$

Thus we have,

$$\mu_x^{(i+1)} = \frac{1}{n}\tilde{S}_x^{(i)}$$

Deriving $\Sigma_x^{(i+1)}$ We will use the following matrix-derivative identities for the calculation (cf. Petersen and Pedersen, 2012),

$$\frac{\partial}{\partial X}\log(\det|X|) = (X^\top)^{-1}, \quad \frac{\partial}{\partial X}\operatorname{Tr}(X^{-1}B) = -(X^{-1}BX^{-1})^\top$$

$$\begin{aligned}\frac{\partial}{\partial \Sigma_x}\ell(\theta; X, y) &\propto \frac{\partial}{\partial \Sigma_x}\left(-n\log(\det|\Sigma_x|) - \operatorname{Tr}\left(\Sigma_x^{-1}(\underbrace{S_{xx} - S_x\mu_x^\top - \mu_x S_x^\top + n\mu_x\mu_x^\top}_{=B})\right)\right) \\ &\propto -n\Sigma_x^{-1} + (\Sigma_x^{-1}B\Sigma_x^{-1})^\top \\ &\propto -n\Sigma_x^{-1} + (\Sigma_x^{-1}B\Sigma_x^{-1})\end{aligned}$$

Setting this expression to zero results in:

$$\Sigma_x = \frac{1}{n}B = \frac{1}{n}\left(S_{xx} - S_x\mu_x^\top - \mu_x S_x^\top + n\mu_x\mu_x^\top\right)$$

Furthermore, when the μ_x found above is plugged in, the following is obtained:

$$\Sigma_x = \frac{1}{n}\left(S_{xx} - \frac{1}{n}S_xS_x^\top - \frac{1}{n}S_xS_x^\top + \frac{1}{n}S_xS_x^\top\right) = \frac{1}{n}\left(S_{xx} - \frac{1}{n}S_xS_x^\top\right)$$

Thus,

$$\Sigma_x^{(i+1)} = \frac{1}{n}\left(\tilde{S}_{xx}^{(i)} - \frac{1}{n}\tilde{S}_x^{(i)}\tilde{S}_x^{(i)^\top}\right)$$

Deriving $\alpha^{(i+1)}$

$$\begin{aligned}\frac{\partial}{\partial \alpha}\ell(\theta; X, y) &\propto -\frac{\partial}{\partial \alpha}\left(S_{yy} - 2\alpha S_y - 2S_{xy}^\top\beta + n\alpha^2 + 2\alpha S_x^\top\beta + \beta^\top S_{xx}\beta\right) \\ &\propto -2S_y - 2n\alpha + 2S_x^\top\beta\end{aligned}$$

Setting the expression to zero yields:

$$\alpha = \frac{1}{n}S_y - \frac{1}{n}S_x^\top\beta$$

Thus for the updating we have,

$$\alpha^{(i+1)} = \frac{1}{n}\tilde{S}_y^{(i)} - \frac{1}{n}\tilde{S}_x^{(i)^\top}\beta^{(i+1)}$$

Deriving $\beta^{(i+1)}$

$$\begin{aligned}\frac{\partial}{\partial\beta}\ell(\theta;X,y) &\propto -\frac{\partial}{\partial\beta}\left(S_{yy}-2\alpha S_y-2S_{xy}^\top\beta+n\alpha^2+2\alpha S_x^\top\beta+\beta^\top S_{xx}\beta\right)\\ &\propto -2S_{xy}-2\alpha S_x+2S_{xx}\beta\end{aligned}$$

Setting the expression to zero and plugging in the α from above yields,

$$\begin{aligned}&\left(\frac{1}{n}S_y-\frac{1}{n}S_x^\top\beta\right)S_x+S_{xx}\beta=S_{xy}\\ \iff &-\underbrace{(\frac{1}{n}S_x^\top\beta)}_{1\times 1}S_x=S_{xy}-\frac{1}{n}S_yS_x\\ \iff &-\frac{1}{n}\underbrace{(S_xS_x^\top)}_{p\times p}\beta+S_{xx}\beta=S_{xy}-\frac{1}{n}S_yS_x\\ \iff &\beta=\left(S_{xx}-\frac{1}{n}S_xS_x^\top\right)^{-1}\left(S_{xy}-\frac{1}{n}S_yS_x\right)\end{aligned}$$

Thus,

$$\beta^{(i+1)}=\left(\tilde{S}_{xx}^{(i)}-\frac{1}{n}\tilde{S}_x^{(i)}\tilde{S}_x^{(i)^\top}\right)^{-1}\left(\tilde{S}_{xy}^{(i)}-\frac{1}{n}\tilde{S}_y^{(i)}\tilde{S}_x^{(i)}\right)$$

Deriving $\sigma^{2^{(i+1)}}$

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\ell(\theta;X,y) &\propto \frac{\partial}{\partial\sigma^2}\left(-n\log(\sigma^2)-\frac{1}{\sigma^2}\underbrace{\left(S_{yy}-2\alpha S_y-2S_{xy}^\top\beta+n\alpha^2+2\alpha S_x^\top\beta+\beta^\top S_{xx}\beta\right)}_{R(\alpha,\beta)}\right)\\ &\propto -\frac{n}{\sigma^2}+\frac{1}{\sigma^4}R(\alpha,\beta)\end{aligned}$$

Setting the above to zero results in:

$$\sigma^2=\frac{1}{n}R(\alpha,\beta)$$

Thus we update by,

$$\sigma^{2^{(i+1)}}=\frac{1}{n}\left(\tilde{S}_{yy}^{(i)}-2\alpha^{(i+1)}\tilde{S}_y^{(i)}-2\tilde{S}_{xy}^{(i)^\top}\beta^{(i+1)}+n(\alpha^{(i+1)})^2+2\alpha\tilde{S}_x^{(i)^\top}\beta^{(i+1)}+\beta^{(i+1)^\top}\tilde{S}_{xx}^{(i)}\beta^{(i+1)}\right)$$

C Github link for corresponding R-code

https://github.com/Vincent-Baumgartner/SeminarStatistics.git

References

T. W. Anderson. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52(278):200–203, 1957. ISSN 01621459. URL http://www.jstor.org/stable/2280845.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe and. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury Press, Pacific Grove, CA, 2 edition, 2002.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912352.

Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Prentice Hall, Upper Saddle River, NJ, 6 edition, 2007.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 2 edition, 2002. ISBN 978-0-471-18386-0. doi: 10.1002/9781119013563.

Thomas A. Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233, 1982. ISSN 00359246. URL http://www.jstor.org/stable/2345828.

G. Mclachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*. 01 2007. ISBN 0471201707. doi: 10.1002/9780470191613.

Xiao-Li Meng and Donald B. Rubin. Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, 86(416):899–909, 1991. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2290503.

See Ng, Thriyambakam Krishnan, and G. Mclachlan. The em algorithm. *Handbook of Computational Statistics: Concepts and Methods*, 01 2004. doi: 10.1007/978-3-642-21551-3_6.

David Oakes. Direct calculation of the information matrix via the em algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61 (2):479–482, 1999. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/2680653.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, nov 2012. URL http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html. Version 20121115.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2335739.

Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

Florin Vaida. Parameter convergence for em and mm algorithms. *Statistica Sinica*, 15(3):831–840, 2005. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24307335.

CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.