缺失数据下的统计推断

邵智轩

1400012141 物理学院

汪谷

1600010755 数学科学学院

石少宏

1500011078 数学科学学院

(日期：2018 年 7 月 1 日)

本文在二元正态分布以及 MAR 假设下，对比了多种处理缺失数据的模型，包括完全案例分析，均值填补和回归填补，最大似然估计以及 EM 算法，并进行了随机模拟试验，比较相合性以及估计效率，最终总结了各模型的适用范围与优缺点。

关键词： Missing Data，MCAR，MAR，Complete Cases，Imputation，MLE，EM Algorithm

CONTENTS

I. 引言：问题背景

数据分析师们最喜欢看到这样的数据——一个规整的矩形数据框（data matrix）。通常行代表单元（units)/案例（cases）/观测（observations），列代表变量（variables）。有数不胜数的经典统计方法可以直接对矩形数据进行分析。

然而，实际的数据往往很少是完整的矩形。缺失数据（missing data）是非常常见的情况，即数据框中的一些元素没有被观测到。这种情况下我们能否继续分析数据，做出合理的统计推断？

本文对比了几种简单的，然而在实际中很常用的缺失数据处理模型。在二元正态分布下，对比它们的优缺点、适用条件，并通过随机模拟实验比较各模型优劣。

II. 缺失数据的基本术语与框架

A. 缺失模式（Missing Patterns）

记 $Y=(y_{ij})$ 为 $n\times K$ 的不存在缺失时的数据矩阵，第 i 行 $y_i=(y_{i1},\dots,y_{iK})$ 代表个体 i 的 K 个变量的取值。定义缺失示性矩阵（missing-data indicator matrix）$M=(m_{ij})$，$m_{ij}=1$ 如果 y_{ij} 缺失，$m_{ij}=0$ 如果 y_{ij} 被观测。那么矩阵 M 就定义了缺失模式。

简明而不失一般性，**本文仅在 $K=2$ 的二元正态分布下讨论，并探讨如下两种缺失模式：**

a. 单变量缺失，属于单调（Monotone）缺失，如图 (1a) 所示。

b. 两个变量都缺失，属于一般（General）缺失，如图 (1b) 所示。

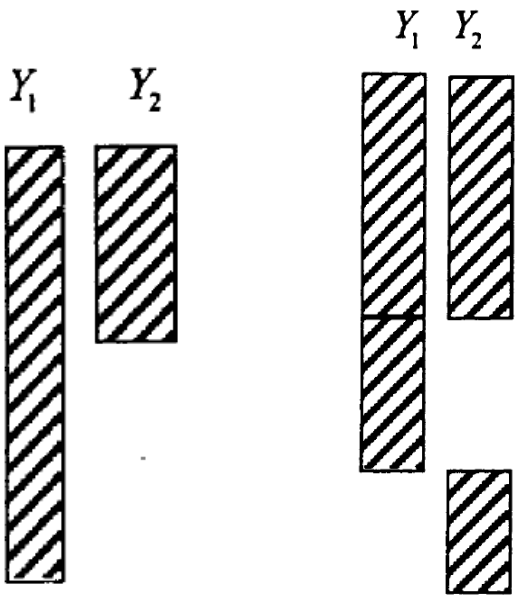
我们会看到，某些模型只适用于特定的缺失模式。

B. 缺失机制（Missing Mechanism）

缺失机制一个很关键的问题——缺失是否与观测或未观测到的数据有关？然而在 Rubin (1976a) 完整地提出这一概念以前，这一问题往往是被忽视的，这往往会导致严重的错误。

所谓缺失机制，说白了就是把缺失与否也当作随机变量来处理，即用给定 Y 时 M 的条件分布 $f(M|Y,\phi)$ 来对缺失机制建模，ψ 是一个未知参数。如果缺失与 Y 的值，无论观测还是缺失，都无关，即

$$f(M|Y,\psi)=f(M|\psi)\quad \text{for all}\, Y,\psi \tag{1}$$



(a) 单变量缺失，属于单调缺失　　**(b)** 两个变量都有缺失，属于一般缺失

图 1: 本文探讨的两种缺失模式

这类数据称为完全随机缺失（missing completely as random, MCAR）。这是一个很强的条件，由于缺失往往并不受采样人的控制，实际中满足完全随机缺失条件的数据并不多见

记 Y_{obs} 为观测到的 Y 的值，Y_{mis} 为缺失的 Y 的值。一个更弱一点的条件是缺失只与观测到的值 Y_{obs} 有关，而与缺失值 Y_{mis} 无关。即

$$f(M|Y,\phi) = f(M|Y_{obs},\phi) \quad \text{for all } Y_{mis},\phi \tag{2}$$

这类数据称为随机缺失（missing at random, MAR）。

如果缺失与未观测到的 Y_{mis} 有关，称为非随机缺失（NMAR）。可以想象，实际数据中 NMAR 应该是最常见的，但不幸的是，这种情况是最难处理的。虽然 MAR 假设通常看上去过于理想，但它相比 MCAR 给出了对真实情况 NMAR 更好的近似。在某些设定下，MAR 假设甚至能比基于 NMAR 的方法，对缺失数据做出更精确的预测。（Rubin, Stern and Vehovar, 1996）

本文仅限于研究 MCAR 或 MAR 假设下的数据。

C. 可忽略性（ignorable）条件

记观测值 Y_{obs} 和缺失值 Y_{mis} 的联合分布概率为 $f(Y|\theta) \equiv f(Y_{obs}, Y_{mis}|\theta)$。那么观测值 Y_{obs} 的边际分布可以通过对 Y_{mis} 积分得到：

$$f(Y_{obs}) = \int f(Y_{obs}, Y_{mis}|\theta)\mathrm{d}Y_{mis} \tag{3}$$

我们定义：θ 的基于 Y_{obs} 的似然函数称为**忽略缺失机制**的（ignoring the missing-data mechanism），如果这一似然函数是任意一个正比于 $f(Y_{obs}|\theta)$ 的函数：

$$L_{ign}(\theta|Y_{obs}) \propto f(Y_{obs}|\theta), \quad \theta \in \Omega_\theta \tag{4}$$

这样在可忽略性条件下，对于参数 θ 的推断就基于式 (4) 的似然函数，最大似然估计就是最大化式 (4)。

在一般情况下，如果我们将缺失 M 作为随机变量引入模型，Y 和 M 的联合分布可以写作 Y 的分布与给定 Y 时 M 的条件分布的乘积，后者代表缺失机制，包含一个未知参数 ψ：

$$f(Y, M|\theta, \psi) = f(Y|\theta)f(M|Y, \psi), \quad (\theta, \psi) \in \Omega_{\theta,\psi} \tag{5}$$

实际观测到的数据为 (Y_{obs}, M) 的值。通过对联合密度式 (5) 中的 Y_{mis} 积分，可得到观测数据的联合分布：

$$f(Y_{obs}, M|\theta, \psi) = \int f(Y_{obs}, Y_{mis}|\theta)f(M|Y_{obs}, Y_{mis}, \psi)\mathrm{d}Y_{mis} \tag{6}$$

关于参数 θ 和 ψ 的完整似然函数 (full likelihood) 是任一正比于式 (6) 的函数：

$$L_{full}(\theta, \psi|Y_{obs}, M) \propto f(Y_{obs}, M|\theta, \psi), \quad (\theta, \psi) \in \Omega_{\theta,\psi} \tag{7}$$

在最一般的情况下（NMAR），我们基于观测数据对 θ 的推断要基于式 (7) 的完整似然函数，即需要将缺失机制考虑在内。这使得问题处理起来非常复杂。什么情况下我们能忽略缺失机制呢？如果我们假设缺失是 MAR 的（式 (2)），式 (6) 能化简为：

$$\begin{aligned} f(Y_{obs}, M|\theta, \psi) &= f(M|Y_{obs}, \psi) \times \int f(Y_{obs}) \\ &= f(M|Y_{obs}, \psi)f(Y_{obs}|\theta) \end{aligned} \tag{8}$$

在大多数情况下，参数 θ 和 ψ 是 distinct 的，即 (θ, ψ) 的联合参数空间是 θ 的参数空间和 ψ 的参数空间的笛卡儿积。在 MAR 和 Distinctness 这两个条件下，缺失机制就是

可忽略的，因为从 $L_{full}(\theta,\psi|Y_{obs},M)$ 推断 θ 等价于从 $L_{ign}(\theta|Y_{obs})$ 推断 θ。

D. 完全数据下的最大似然估计

假设二变量数据 $Y=(Y_1,Y_2)\sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$，其中均值向量 $\boldsymbol{\mu}=(\mu_1,\mu_2)$，协方差矩阵 $\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_{11} & \sigma_{12}\\ & \sigma_{22}\end{pmatrix}$，则给定数据 Y，对数似然函数为：

$$l(\boldsymbol{\mu},\boldsymbol{\Sigma}|Y)=-\frac{1}{2}n\ln|\boldsymbol{\Sigma}|-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{y}_i-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu})' \tag{9}$$

对这一似然函数的最大似然估计（对 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 求导，具体细节可参见任一数理统计教材[3]），结果为：

$$\hat{\boldsymbol{\mu}}=\bar{\boldsymbol{y}},\quad \hat{\boldsymbol{\Sigma}}=S/n \tag{10}$$

其中 $\bar{\boldsymbol{y}}=(\bar{y}_1,\bar{y}_2)$ 为样本均值，$S=(s_{jk})$ 为 2×2 的样本的校正叉积矩阵，$s_{jk}=\sum_{i=1}^{n}(y_{ij}-\bar{y}_j)(y_{ik}-\bar{y}_k)$。如果将分母 n（自由度）修正为 $n-1$，则得到 $\boldsymbol{\Sigma}$ 的无偏估计。

III. 模型一：完全案例分析（COMPLETE-CASE ANALYSIS）

一个最简单的权宜之策，就是直接任何变量有缺失的案例删去。这一方法优点显而易见：简单易操作；且不同变量的统计量之间是可比较的，因为它们基于相同的案例。然而我们接下来会证明，CC 这一实际中常用方法的问题也很明显——信息的损失。这一信息损失会造成两方面的后果：其一是**效率的损失**，其二是当数据非 MCAR 时会**引入偏差**。

第二点是显然的，以 MAR 缺失为例，如果 y_2 与 y_1 正相关，y_1 较大时 y_2 大概率缺失，则 y_2 的均值估计会偏小。

第一点体现在样本量从 n 减小到 Complete cases 的个数 r 后，估计值的方差，相比于数据没有缺失时会增大：

$$\mathrm{Var}(\hat{\theta}_{CC})=\mathrm{Var}(\hat{\theta}_{NM})(1+\Delta^{*}_{CC}) \tag{11}$$

Δ^{*}_{CC} 就是由于缺失数据造成的方差增大的比例。由于这一方差的增大很大程度上来源于缺失本身，从而是不可避免的，一个更常用的效率损失的度量是相对效率损失 Δ_{CC}：

$$\mathrm{Var}(\hat{\theta}_{CC})=\mathrm{Var}(\hat{\theta}_{EFF})(1+\Delta_{CC}) \tag{12}$$

$\hat{\theta}_{EFF}$ 是 θ 的一个基于观测数据的 efficient 的估计，比如后面介绍的 MLE 估计。

以二元正态下，单变量 MCAR 缺失为例（模式如图 (1a)），分析 CC 方法对均值估计的效率的损失。假设原样本量为 n，Y_2MCAR 缺失，完全案例个数为 r，$n-r$ 个案例中 Y_1 观测到而 Y_2 缺失。CC 对 Y_j 均值的估计为 $\bar{y}_j^{CC}$

对于 μ_1 的估计，显然有

$$\Delta_{CC}^{*} = \Delta_{CC} = \frac{n-r}{r} \tag{13}$$

对于 μ_2 的估计，效率的损失不仅与完全数据的比例有关，还与 Y_1 和 Y_2 的相关系数 ρ^2 有关：

$$\Delta_{CC}^{*} = \frac{n-r}{r}, \quad \Delta_{CC} \approx \frac{(n-r)\rho^2}{n(1-\rho^2)+r\rho^2} \tag{14}$$

证明见 [4]。从式 (14) 我们看到当 Y_1 和 Y_2 不相关时没有效率损失，而 Y_1 和 Y_2 线性相关（$\rho^2 \to 1$）时，效率的损失和 μ_1 一样多。

由此可见，除非样本量很大与之相比缺失值很少，CC 这一过于 naive 的方法在计算力不成问题的今天是不值得推荐的。

这里顺带提一下所谓的 Available-case Analysis。不同于 Complete case analysis 只要一个案例中有一个变量缺失就将整个案例删去，它在估计感兴趣的参数时使用与该参数估计直接相关的，尽可能多的案例。Available-case Analysis 看起来比 Complete-case Analysis 利用更多的信息，从而应该有更好的 efficiency。在 MCAR 假设下，它能给出协方差和相关系数的相合估计。然而它在实际分析中会引入更严重的困难，例如绝对值大于 1 的相关系数，又如 $K > 3$ 时非正定的协方差矩阵。

IV. 模型二：（单一）填补法

既然 CC 方法的一大弊端在于丢弃了过多信息，一个自然的想法就是与其删不如补——尤其是如果我们有自信可以较好地预测缺失值，比如有缺失的变量与另一个完整的变量高度相关。

填补法是一个通用，灵活的处理缺失数据的方法。然而，这一方法有很多陷阱。Rubin and Dempster (1983) 这样评论填补方法：

The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases.

下面我们介绍两种比较简单的填补方法。

A. （非条件）均值填补（Unconditional Mean Imputation）

最简单的填补方法就是将某一变量地每个缺失值 y_{ij} 替换成该变量观测值的均值 $\bar{y}_j^{(j)}$。也就是说，这一方法在处理某个含缺失值的变量时，没有考虑其他变量。填补后对该变量均值的估计等价于 CC 法，所以在 MCAR 下是无偏的，MAR 下是有偏的。

然而不同于 MCAR 下 CC 法正确估计方差和协方差，均值填补后的变量的样本方差为 $s_{jj}^{(j)}(n^{(j)}-1)/(n-1)$，其中 $s_{jj}^{(j)}$ 是用 $n^{(j)}$ 个 Y_j 的观测值估计的。在 MCAR 下，$s_{jj}^{(j)}$ 是相合估计，所以均值填补后的样本方差以 $(n^{(j)}-1)/(n-1)$ 低估了方差。

同样,均值填补后 Y_j 与 Y_k 的协方差也被低估了。填补后的样本协方差为 $\bar{s}_{jk}^{(jk)}(n^{(jk)}-1)/(n-1)$，$n^{(jk)}$ 是 Y_j 和 Y_k 都被观测的案例数量，$\bar{s}_{jk}^{(jk)}$ 为从这些案例估计的样本协方差（Available-case Analysis）。在 MCAR 下 $\bar{s}_{jk}^{(jk)}$ 为相合估计，所以均值填补法以 $(n^{(jk)}-1)/(n-1)$ 低估了协方差。

虽然均值填补法和无缺失时一样，给出了对称正定的方差估计。但是估计存在系统偏差。如果我们矫正这一偏差，将方差放大 $(n-1)/(n^{(j)}-1)$ 倍，将协方差放大 $(n-1)/(n^{(jk)}-1)$ 倍，则会出现和 Available-case Analysis 相同的问题——协方差矩阵非正定，尤其当变量之间高度相关时。

B. 回归填补（Conditional Mean Imputation: Regression）

不同于非条件均值填补，条件均值填补考虑给定其他变量的值后当前变量的条件均值。在多元正态假设下，最常用的方法当属回归填补（Regression Imputation）。

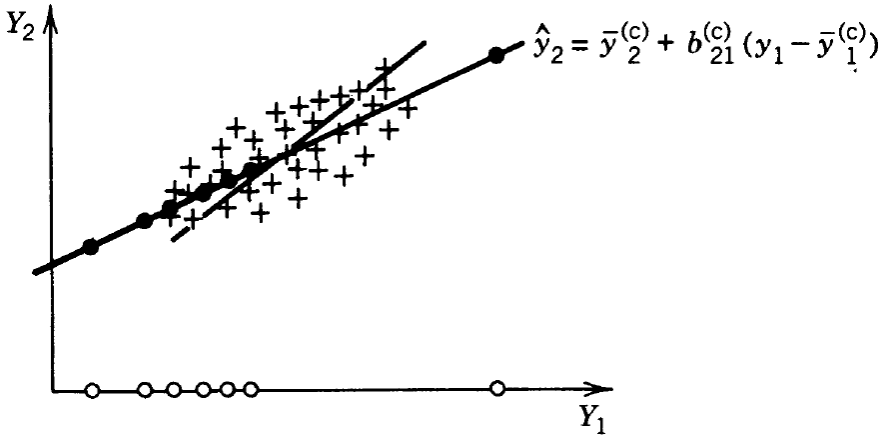


图 2: $K=2$ 时的回归填补

假设样本量为 n，Y_1 全都观测而 Y_2 有 r 各值观测，$n-r$ 个值缺失。在 r 个完全案例上计算回归模型，再用该模型对 $n-r$ 个缺失值做预测，用预测值填补缺失数据。二元正态下单变量缺失时地回归填补如图 (2) 所示。图中 $+$ 号表示 Y_1 和 Y_2 都观测到的案例，用这些案例来计算 $Y_2 \sim Y_1$ 的最小二乘回归 $\hat{y}_{i2} = \tilde{\beta}_{20\cdot1} + \tilde{\beta}_{21\cdot1} y_{i1}$。$Y_1$ 轴上的空心圈代表 Y_1 观测而 Y_2 缺失的案例。回归填补法用回归直线上的实心圈来代替 Y_1 轴上的空心圈。Y_2 观测而 Y_1 缺失的案例则会用 $Y_1 \sim Y_2$ 的回归直线上的点替代。

Buck's Method (1960) 将回归填补法拓展到了更普遍的一般缺失模式。首先从完全案例中估计样本均值和样本协方差，以此充分统计量计算每一个缺失变量对其他观测变量的最小二乘回归，用以预测值进行填补。当 K 很大时，这一计算量看似很大，但其实可以很轻松地利用扫描算法 [5] 实现。

Buck's Method 仍然低估了方差和协方差，虽然低估的程度比均值填补要小一些。具体来说 Buck's Method 填补后的 Y_j 的样本方差比 σ_{jj} 小了 $(n-1)^{-1}\sum_{i=1}^{n}\sigma_{jj\cdot obs,i}$，其中 $\sigma_{jj\cdot obs,i}$ 当 y_{ij} 缺失时是 Y_j 对观测到的变量回归后的残差方差，当 y_{ij} 观测时是 0。Y_j 和 Y_k 的样本标准差比 σ_{jk} 小了 $(n-1)^{-1}\sum_{i=1}^{n}\sigma_{jk\cdot obs,i}$，$\sigma_{jk\cdot obs,i}$ 在 y_{ij} 和 y_{ik} 都缺失的案例 i 中是 Y_j 和 Y_k 对观测到的变量回归后的残差的协方差，在其他案例中是 0。

如果我们在 Buck's Method 中按照如上分析修正了方差和协方差的偏差，我们得到的估计等价于V中的最大似然估计。所以 Buck's Method 在缺失数据理论的发展过程中被看作 MLE 估计的先驱。

总而言之。无论何种单一填补方法，都有一个不可避免的系统性缺陷——填补后数据的不确定度偏小。同 CC 分析一样，填补方法在实际中也应该谨慎使用。

V. 模型三：最大似然估计（MLE）

A. 单变量缺失的情况：Monotone missing pattern

缺失模式如图 (1a) 所示。以下仍沿用II D的记号，$(Y_1, Y_2) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$，$y_i 2$ 缺失当 $i = (r+1), \ldots, n$。在可忽略性条件下，对数似然函数为：

$$\begin{aligned} l_{ign}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | Y_{obs}) = &-\frac{1}{2} r \ln |\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{r}(\boldsymbol{y}_i - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu})' \\ &-\frac{1}{2}(n-r)\ln\sigma_{11} - \frac{1}{2}\sum_{i=r+1}^{n}\frac{(y_{i1}-\mu_1)^2}{\sigma_{11}} \end{aligned} \tag{15}$$

则通过最大化以上似然函数，求得 $\hat{\boldsymbol{\mu}}$ 和 $\hat{\boldsymbol{\Sigma}}$。这一似然函数看上去没有明显的解，但是其 Monotone 的缺失模式使问题得以分解。Anderson (1957)[6] 将 y_{i1} 和 y_{i2} 的联合分布

分解为 y_{i1} 的边际分布乘以给定 y_{i1} 时 y_{i2} 的条件分布：

$$f(y_{i1}, y_{i2}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = f(y_{i1}|\mu_1, \sigma_{11}) f(y_{i2}|y_{i1}, \beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1}) \tag{16}$$

由二元正态分布的性质，$f(y_{i1}|\mu_1, \sigma_{11})$ 是均值为 μ_1，方差为 σ_{11} 的正态分布，$f(y_{i2}|y_{i1}, \beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1})$ 是均值为 $\beta_{20\cdot1}, \beta_{21\cdot1} + y_{i1}$，方差为 $\sigma_{22\cdot1}$ 的正态分布。参数

$$\phi = (\mu_1, \sigma_{11}, \beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1})'$$

与原参数

$$\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})'$$

一一对应。具体来说，ϕ 和 θ 中的 μ_1 和 σ_{11} 相同，其他三个参数满足：

$$\begin{aligned} \beta_{21\cdot1} &= \sigma_{12}/\sigma_{11}, \\ \beta_{20\cdot1} &= \mu_2 - \beta_{21\cdot1}, \mu_1 \\ \sigma_{22\cdot1} &= \sigma_{22} - \sigma_{12}^2/\sigma_{11} \end{aligned} \tag{17}$$

或者相反地用 ϕ 中的参数表出 θ 中的参数：

$$\begin{aligned} \mu_2 &= \beta_{20\cdot1} + \beta_{21\cdot1}\mu_1, \\ \sigma_{12} &= \beta_{21\cdot1}\sigma_{11}, \\ \sigma_{22} &= \sigma_{22\cdot1} + \beta_{21\cdot1}^2\sigma_{11} \end{aligned} \tag{18}$$

Y_{obs} 可以作如下因子分解：

$$\begin{aligned} f(Y_{obs}|\theta) &= \prod_{i=1}^{r} f(y_{i1}, y_{i2}|\theta) \prod_{i=r+1}^{n} f(y_{i1}|\theta) \\ &= \left[\prod_{i=1}^{r} f(y_{i1}|\theta) f(y_{i2}|y_{i1}, \theta)\right] \left[\prod_{i=r+1}^{n} f(y_{i1}|\theta)\right] \\ &= \left[\prod_{i=1}^{n} f(y_{i1}|\mu_1, \sigma_{11})\right] \left[\prod_{i=1}^{r} f(y_{i2}|y_{i1}, \beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1})\right] \end{aligned} \tag{19}$$

在 θ 为自然参数空间（没有限制）时，(μ_1, σ_{11}) 和 $(\beta_{20\cdot1}, \beta_{21\cdot1}, \sigma_{22\cdot1})$ 满足 Distinctness 条件，所以求最大似然估计只需分别求这两个因子的最大值。

最大化第一个因子得到

$$\hat{\mu}_1 = n^{-1}\sum_{i=1}^{n} y_{i1},$$
$$\hat{\sigma}_{11} = n^{-1}\sum_{i=1}^{n}(y_{i1}-\hat{\mu}_1)^2 \tag{20}$$

由多元正态分布的性质，对第二个因子最大似然估计由线性回归给出：

$$\hat{\beta}_{21\cdot 1} = s_{12}/s_{11},$$
$$\hat{\beta}_{20\cdot 1} = \bar{y}_2 - \hat{\beta}_{21\cdot 1}\bar{y}_1,$$
$$\hat{\sigma}_{22\cdot 1} = s_{22\cdot 1} \tag{21}$$

其中 $y_j = r^{-1}\sum_{i=1}^{r} y_{ij}$，$s_{jk} = r^{-1}(y_{ij}-\bar{y}_j)(y_{ik}-\bar{y}_k)$，$s_{22\cdot 1} = s_{22} - s_{12}^2/s_{11}$。

利用式 (18)，可得到 θ 中剩余 3 个参数 $(\mu_2, \sigma_{12}, \sigma_{22})$ 的估计

$$\hat{\mu}_2 = \hat{\beta}_{20\cdot 1} + \hat{\beta}_{21\cdot 1}\hat{\mu}_1 = \bar{y}_2 + \hat{\beta}_{21\cdot 1}(\hat{\mu}_1 - \bar{y}_1) \tag{22}$$
$$\hat{\sigma}_{12} = \hat{\beta}_{21\cdot 1}\hat{\sigma}_{11} = s_{12}(\hat{\sigma}_{11}/s_{11}) \tag{23}$$
$$\hat{\sigma}_{22} = \hat{\sigma}_{22\cdot 1} + \hat{\beta}_{21\cdot 1}^2\hat{\sigma}_{11} = s_{22} + \hat{\beta}_{21\cdot 1}^2(\hat{\sigma}_{11} - s_{11}) \tag{24}$$

此外还能得到相关系数

$$\rho \equiv \sigma_{12}(\sigma_{11}\sigma_{22})^{-1/2} = \beta_{21\cdot 1}\sigma_{11}^{1/2}(\sigma_{22\cdot 1} + \beta_{21\cdot 1}^2\sigma_{11})^{-1/2} \tag{25}$$

的估计：

$$\hat{\rho} = [s_{12}(s_{11}s_{22})^{-1/2}](\hat{\sigma}_{11}/s_{11})^{1/2}(s_{22}/\hat{\sigma}_{22})^{1/2} \tag{26}$$

我们注意到 $\hat{\mu}_2$ 和 $\hat{\sigma}_{22}$ 的第一项，以及 $\hat{\sigma}_{12}$ 或 $\hat{\rho}$ 的第一个因子，等价于 r 个完全案例的最大似然估计。那么，其余的项或因子就代表基于其余的 $n-r$ 个值提供的信息，对 CC 估计值所作的修正。

特别需要注意的是，在二元正态分布下，**除了** $\hat{\sigma}_{22}$ 的估计，其他估计都等价于先做回归填补，再用完全数据估计参数。然而不同于均值填补法低估了 σ_{22}，MLE 给出了 σ_{22} 的相合估计。

B. 两个变量都有缺失的情况：General missing pattern

1. *EM 算法*

EM（Expectation Maximum）算法是一种求解含有缺失数据或隐变量的最大似然估计时通用的**迭代**算法。这里仅简述一下该算法。

每一次迭代分为两步：E step 和 M step。M step 就如同没用缺失数据，与普通的最大似然相同；E step 在当前的参数估计值下，求给定观测数据时“缺失值”的期望，然后用期望值来填补“缺失值”。这里的“缺失值”往往不是缺失值 Y_{mis} 本身，而是似然函数中线性依赖于 Y_{mis} 的函数的项（充分统计量）。

具体来说，如果 $\theta^{(t)}$ 为当前对 θ 的估计。E step 求出完全数据的期望似然：

$$Q(\theta|\theta^{(t)}) = \int l(\theta|y) f(Y_{mis}|Y_{obs}, \theta = \theta^{(t)}) \mathrm{d}Y_{mis} \tag{27}$$

M step 通过最大化完全数据的期望似然求出 $\theta^{(t+1)}$：

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geqslant Q(\theta|\theta^{(t)}), \quad \text{for all}\, \theta \tag{28}$$

2. *在本例中的具体实现 [7]*

本例中的缺失模式如图 (1b) 所示。此时的似然函数直接求最大似然估计是有困难的，不同于 Monotone 缺失，这里也不能进行式 (19) 的因子分解，所以要采用V B 1中介绍的 EM 算法。

完全数据的似然函数如式 (9) 所示。我们看到似然函数线性依赖的项为以下几个完全数据的**充分统计量**：

$$s_1 = \sum_{i=1}^{n} y_{i1}, \quad s_2 = \sum_{i=1}^{n} y_{i2}, \quad s_{11} = \sum_{i=1}^{n} y_{i1}^2, \quad s_{12} = \sum_{i=1}^{n} y_{i1}y_{i2}, \quad s_{22} = \sum_{i=1}^{n} y_{i2}^2 \tag{29}$$

（正态分布下的充分统计量：均值、方差、协方差）E step 给定 Y_{obs} 和 $\theta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 求式 (29) 的条件期望。

对于 y_{i1} 和 y_{i2} 都观测到，式 (29) 的期望就等于观测值；对于 y_{i1} 观测到而 y_{i2} 缺失的案例，y_{i1} 和 y_{i1}^2 就等于观测值，y_{i2}, y_{i2}^2 和 $y_{i1}y_{i2}$ 的期望由V A的结论可通过 y_{i2} 对

y_{i1} 的线性回归求出：

$$\mathrm{E}[y_{i2}|y_{i1},\boldsymbol{\mu},\boldsymbol{\Sigma}] = \beta_{20\cdot1} + \beta_{21\cdot1}y_{i1} \tag{30}$$

$$\mathrm{E}[y_{i2}^2|y_{i1},\boldsymbol{\mu},\boldsymbol{\Sigma}] = (\beta_{20\cdot1} + \beta_{21\cdot1}y_{i1})^2 + \sigma_{22\cdot1} \tag{31}$$

$$\mathrm{E}[y_{i1}y_{i2}|y_{i1},\boldsymbol{\mu},\boldsymbol{\Sigma}] = (\beta_{20\cdot1} + \beta_{21\cdot1}y_{i1})y_{i1} \tag{32}$$

$\beta_{20\cdot1}$，$\beta_{21\cdot1}$，$\sigma_{22\cdot1}$ 的定义见式 (17)，带入当前步的参数 $\theta^{(t)}$ 计算。y_{i2} 观测到而 y_{i1} 缺失的处理方法类似。把这三种情况加起来，我们就求得了式 (29) 的期望。

M step 就通过这些充分统计量求最大似然估计（正态分布下即为矩估计）：

$$\begin{gathered}\hat{\mu}_1 = s_1/n, \quad \hat{\mu}_2 = s_2/n, \\ \hat{\sigma}_1^2 = s_{11}/n - \hat{\mu}_1^2, \quad \hat{\sigma}_2^2 = s_{22}/n - \hat{\mu}_2^2, \quad \hat{\sigma}_{12}^2 = s_{12}/n - \hat{\mu}_1\hat{\mu}_2\end{gathered} \tag{33}$$

迭代执行 E step 和 M step 直到参数估计收敛。

关于参数初值的选取，可以用III或IV等简单方法得到的估计。

EM 算法被证明有收敛性保证，虽然收敛速度可能很慢。

VI. 随机模拟试验

A. 单变量缺失

本文用 R 软件进行随机模拟试验。我们先从二元正态分布

$$\begin{gathered}(Y_1, Y_2) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ \boldsymbol{\mu} = (0, 1), \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1.6 \\ 1.6 & 4 \end{pmatrix}\end{gathered} \tag{34}$$

中生成 $n = 100$ 个样本。这 100 个样本构成了完全数据。

我们人为地剔除一些数据来制造缺失，采用两个变量都有缺失的模式（即一般缺失）。我们分别在 MCAR 和 MAR 假设下剔除数据。记 $\Pr(M_{i1} = r, M_{i2} = s|y_{i1}, y_{i2}; \phi) = g_{rs}(y_{i1}, y_{i2}; \phi)$，我们假设 $g_{11} = 0$，在单变量 Y_2 缺失时，$g_{10} = 0$。在 MCAR 下，我们随机删除 $n/2$ 个 y_{i2}，即

$$g_{01} = \frac{1}{2}, \quad g_{00} = \frac{1}{2} \tag{35}$$

在 MAR 下，我们删除 y_{i1} 较大的 $n/2$ 个案例中的 y_{i2}，即

$$g_{01}(y_{i1}, y_{i2}; \phi) = \Pr(y_{i1}\text{为 } \{y_{ij}\} \text{ 中较大的 50 个}), \quad g_{00} = 1 - g_{01} \tag{36}$$

分别用正文中介绍的多种方法对参数

$$\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})'$$

进行估计。随机模拟 $B = 1000$ 次，求 1000 次估计的平均值，标准差，并与真值比对。

表 I: *MCAR* 单调缺失下各种方法对参数的估计，第一行用未删数据前的完全数据得到的估计，以供参考。括号内为 *1000* 次重复试验的标准差，加粗标出的为相合估计

Method	$\mu_1 = 0$	$\mu_2 = 1$	$\sigma_{11} = 1$	$\sigma_{12} = 1.6$	$\sigma_{22} = 4$	$\rho = 0.8$
完全数据	**-0.002(0.099)**	**1.00(0.19)**	**1.001(0.145)**	**1.604(0.262)**	**4.012(0.574)**	**0.799(0.037)**
完全案例	**0.003(0.140)**	**1.01(0.28)**	**0.987(0.201)**	**1.580(0.360)**	**3.970(0.776)**	**0.795(0.054)**
均值填补	**-0.002(0.099)**	**1.01(0.28)**	**1.001(0.145)**	0.782(0.178)	1.965(0.384)	0.555(0.066)
回归填补	**-0.002(0.099)**	**1.01(0.23)**	**1.001(0.145)**	**1.604(0.288)**	3.297(0.676)	0.883(0.033)
MLE	**-0.002(0.099)**	**1.01(0.23)**	**1.001(0.145)**	**1.604(0.288)**	**4.009(0.712)**	**0.800(0.048)**

MCAR 的试验结果如表I所示。作为参考，第一行为用完全数据（$n = 100$）得到的估计。在 MCAR 下，几乎所有的方法都得到了无偏估计，除了均值填补后的数据低估了 σ_{22}，σ_{12} 和 ρ。这一结论我们已在IV给出——完全案例分析的 $s_{22}^{(2)}$，$s_{12}^{(12)}$ 是相合估计，而均值填补后的 y_2 的样本方差估计为 $s_{22}^{(2)}(n^{(2)} - 1)/(n - 1)$，协方差估计为 $s_{12}^{(12)}(n^{(12)} - 1)/(n - 1)$，所以这里均值填补都以接近 1/2 的比例低估了不确定度 σ_{22} 和 σ_{12}，以约 $1/\sqrt{2}$ 的比例低估 ρ。

回归填补法虽然正确估计了其他参数，但是低估了 σ_{22}，从而使得 ρ 的估计偏大。设想这一低估不确定度的后果——如果用均值填补后的变量做多元回归等统计推断，很有可能高估其显著性。

完全案例分析在 MCAR 下给出了无偏估计，但是它的估计的方差都要大于 MLE——正如我们在III所分析的，完全案例分析丢失了一部分有用的信息，从而损失效率。按照公式14，对 μ_1 的估计 CC 方法的方差约为 MLE 的 2 倍，对 μ_2 的估计 CC 方法的方差约为 MLE 的 25/17，这与表I是完全吻合的。

表 II: *MAR* 单调缺失下各种方法对参数的估计，记号同表*I*

Method	$\mu_1 = 0$	$\mu_2 = 1$	$\sigma_{11} = 1$	$\sigma_{12} = 1.6$	$\sigma_{22} = 4$	$\rho = 0.8$
完全数据[a]	**-0.002(0.099)**	**0.99(0.19)**	**1.001(0.145)**	**1.604(0.262)**	**4.012(0.574)**	**0.799(0.037)**
完全案例	-0.794(0.118)	-0.273(0.251)	0.367(0.092)	0.586(0.176)	2.376(0.506)	0.622(0.090)
均值填补	**-0.002(0.099)**	-0.273(0.251)	**1.001(0.145)**	0.290(0.087)	1.176(0.250)	0.265(0.055)
回归填补	**-0.002(0.099)**	**0.99(0.33)**	**1.001(0.145)**	**1.600(0.361)**	3.339(1.006)	0.878(0.048)
MLE	**-0.002(0.099)**	**0.99(0.33)**	**1.001(0.145)**[b]	**1.600(0.361)**	**4.051(1.045)**	**0.794(0.067)**

[a] 完全数据与表I相同

[b] 这里的方差和协方差都做了自由度修正

MAR 的试验结果如表II所示。我们看到，原本有效的完全案例分析，在 MAR 下是完全错误的。而均值填补法对 Y_2 相关的参数也都给出了错误的估计。回归填补，除了低估了 σ_{22}，仍是有效的。MLE 始终给出相合的估计。

B. 两个变量都有缺失

如前所述，两变量都缺失的情况属于一般缺失（General missing pattern）。我们仍对比多种方法对参数的估计。

与VI A相同，我们先生成 $n = 100$ 个完全数据，再手动删除一些数据。对于 MCAR 情形，我们抽取 30 个 Y_{i1} 缺失而 Y_{i2} 观测，30 个 Y_{i1} 观测而 Y_{i2} 缺失，剩下 40 个不删减，即

$$\begin{aligned} g_{11} = 0, \quad g_{10} = \frac{30}{100}, \\ g_{01} = \frac{30}{100}, \quad g_{00} = \frac{40}{100} \end{aligned} \tag{37}$$

对于 MAR 情形，我们先抽取 50 个案例，从这个 50 案例中删除 y_{i1} 较大的 30 个案例中的 y_{i2}，从剩下的 50 个案例中删除 y_{i2} 较大的 30 个案例中的 y_{i1}，即：

$$\begin{aligned} g_{11} &= 0, \\ g_{10} &= \frac{50}{100} \times \Pr(y_{i2}\text{为第一批抽中的 50 个 } \{y_{ij}\} \text{ 中较大的 30 个}), \\ g_{01} &= \frac{50}{100} \times \Pr(y_{i1}\text{为第二批抽中的 50 个 } \{y_{ij}\} \text{ 中较小的 30 个}), \\ g_{00} &= 1 - (g_{10} + g_{01}) \end{aligned} \tag{38}$$

表 III: *MCAR 一般缺失下各种方法对参数的估计，记号同表I*

Method	$\mu_1 = 0$	$\mu_2 = 1$	$\sigma_{11} = 1$	$\sigma_{12} = 1.6$	$\sigma_{22} = 4$	$\rho = 0.8$
完全数据[a]	**-0.002(0.099)**	**1.00(0.19)**	**1.001(0.145)**	**1.604(0.262)**	**4.012(0.574)**	**0.799(0.037)**
完全案例	**-0.001(0.157)**	**1.00(0.31)**	**0.999(0.231)**	**1.591(0.416)**	**3.970(0.901)**	**0.795(0.063)**
均值填补	**-0.002(0.119)**	**1.01(0.23)**	0.700(0.123)	0.630(0.164)	2.781(0.467)	0.449(0.076)
Buck's	**-0.003(0.108)**	**1.00(0.21)**	0.899(0.161)	**1.604(0.290)**	3.576(0.629)	0.895(0.037)
EM 算法 [b]	**-0.003(0.108)**	**1.00(0.21)**	**0.994(0.163)**[c]	**1.590(0.284)**	**3.954(0.636)**	**0.802(0.053)**

[a] 完全数据与表I相同

[b] 参数初值选为 CC 估计值，迭代直到任意参数的变化小于 1e-7，平均迭代次数约为 25.3 次。

[c] 这里的方差和协方差没有做自由度修正，MLE 给出有偏但是相合的估计

随机模拟 $B = 1000$ 次。MCAR 试验的结果如表 (III) 所示。同VI A时一样，CC 法在 MCAR 下时可行的，虽然损失了一定的效率；填补法，无论是均值还是 Buck's

Method 都仍有低估不确定度的通病，虽然回归填补的低估要少一些。这里 EM 算法求的是最大似然估计，没有做不确定度修正，所以以 $(n-1)/n$ 倍低估了方差和协方差，是无偏但是相合的。

MAR 试验如表 (IV) 所示。在 MAR 一般缺失下，Buck's Method 对均值估计也不再相合。只有 EM 算法给出了相合的估计。

表 IV: *MAR 一般缺失下各种方法对参数的估计，记号同表 I*

Method	$\mu_1 = 0$	$\mu_2 = 1$	$\sigma_{11} = 1$	$\sigma_{12} = 1.6$	$\sigma_{22} = 4$	$\rho = 0.8$
完全数据[a]	**-0.002(0.099)**	**1.00(0.19)**	**1.001(0.145)**	**1.604(0.262)**	**4.012(0.574)**	**0.799(0.037)**
完全案例	-0.859(0.157)	-0.71(0.26)	0.458(0.120)	0.503(0.181)	1.829(0.467)	0.544(0.111)
均值填补	-0.219(0.116)	0.562(0.229)	0.699(0.120)	0.529(0.103)	2.781(0.475)	0.380(0.040)
Buck's	-0.164(0.123)	0.676(0.246)	0.759(0.139)	1.009(0.333)	3.030(0.550)	0.656(0.150)
EM 算法 [b]	**-0.001(0.112)**	**1.00(0.22)**	**0.992(0.166)**[c]	**1.585(0.276)**	**3.957(0.652)**	**0.800(0.050)**

[a] 完全数据与表I相同

[b] 参数初值选为 CC 估计值，迭代直到任意参数的变化小于 1e-7，平均迭代次数约为 51.5 次。

[c] 这里的方差和协方差没有做自由度修正，MLE 给出有偏但是相合的估计

VII. 结论：各模型的优缺点及适用范围

a. 完全案例分析：

适用条件：缺失值相比观测值非常少；MCAR；

优点：操作简单；计算速度快；不同变量的统计量具有可比性，因为它们基于相同的案例

缺点：MAR 下引入偏差；损失了一部分信息，估计的效率损失

b. 均值填补：

适用条件：MCAR

优点：操作简单，不损失其他非缺失变量的信息

缺点：低估与被填补变量相关的方差或者协方差；如果校正会使得协方差阵非正定

c. 条件均值填补（回归填补）：

适用条件：正态假设/变量之间线性依赖，大多数 MCAR 和少数 MAR 情况

优点：不损失其他非缺失变量的信息

缺点：低估了被填补变量的不确定度

d. MLE：

适用条件：已知分布（已知似然函数），满足可忽略性条件

优点：对所有参数给出相合估计，满足很多好的性质（渐进正态性，efficiency 等，见任一数理统计教材）；利用了数据中的所有信息

缺点：计算比较复杂；EM 算法需要迭代计算，可能收敛较慢，然而以今天的计算性能来看，这恐怕算不上问题。

---

[1] Molenberghs G, Kenward M. Missing data in clinical studies[M]. John Wiley & Sons, 2007.

[2] Little R J A. Regression with missing X's: a review[J]. Journal of the American Statistical Association, 1992, 87(420): 1227-1237.

[3] Casella G, Berger R L. Statistical inference[M]. Pacific Grove, CA: Duxbury, 2002.

[4] Little R J A, Rubin D B. Statistical analysis with missing data[M]. John Wiley & Sons, 2014.

[5] Weisberg S. Applied linear regression[M]. John Wiley & Sons, 2005.

[6] Anderson T W. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing[J]. Journal of the american Statistical Association, 1957, 52(278): 200-203.

[7] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the royal statistical society. Series B (methodological), 1977: 1-38.

附录 A: 说明

本组的邵智轩，曾上过“应用回归分析”课，这门课曾简要地介绍过缺失数据的概念及模型。本文更深入地探讨几个常用的处理缺失数据的数学模型，及其原理、适用范围、优缺点，使组员们对这一问题的处理有了更深刻的体会。

附录 B: 小组分工

a. 邵智轩：最大似然估计；随机模拟

b. 汪谷：完全案例分析；术语与框架

c. 石少宏：填补法；术语与框架

附录 C: EM 算法：R code

```
EM_algorithm <- function(sample.missing.both, eps = 1e-7) {
    n <- nrow(sample.missing.both)
    complete_cases <- complete.cases(sample.missing.both)
    y1_miss <- is.na(sample.missing.both[, 1])
    y2_miss <- is.na(sample.missing.both[, 2])
    mean.new <- colMeans(sample.missing.both[complete_cases, ])
    cov.new <- cov(sample.missing.both[complete_cases, ])[c(1, 2, 4)]
    mean.old <- rep(Inf, 2)
    cov.old <- rep(Inf, 3)
    #s的固定部分
    s_1.com <- sum(sample.missing.both[complete_cases, 1])
    s_2.com <- sum(sample.missing.both[complete_cases, 2])
    s_11.com <- sum(sample.missing.both[complete_cases, 1] ^ 2)
    s_22.com <- sum(sample.missing.both[complete_cases, 2] ^ 2)
    s_12.com <- sum(sample.missing.both[complete_cases, 1] *
                      sample.missing.both[complete_cases, 2])
    while (max(abs(c(mean.new - mean.old, cov.new - cov.old))) > eps) {
      mean.old <- mean.new
      cov.old <- cov.new
      #先计算y1观测而y2缺失的案例
      beta_21_1 <- cov.old[2] / cov.old[1]
      beta_20_1 <- mean.old[2] - beta_21_1 * mean.old[1]
      sigma_22_1 <- cov.old[3] - cov.old[2] ^ 2 / cov.old[1]
      s_1 <- s_1.com + sum(sample.missing.both[y2_miss, 1])
      s_11 <- s_11.com + sum(sample.missing.both[y2_miss, 1] ^ 2)
      hat_y2 <- beta_20_1 + beta_21_1 * sample.missing.both[y2_miss, 1]
      s_2 <- s_2.com + sum(hat_y2)
      s_22 <- s_22.com + sum(hat_y2 ^ 2 + sigma_22_1)
      s_12 <- s_12.com + sum(hat_y2 * sample.missing.both[y2_miss, 1])
      #再计算y2观测而y1缺失的案例
      beta_11_2 <- cov.old[2] / cov.old[3]
      beta_10_2 <- mean.old[1] - beta_11_2 * mean.old[2]
      sigma_11_2 <- cov.old[1] - cov.old[2] ^ 2 / cov.old[3]
      s_2 <- s_2 + sum(sample.missing.both[y1_miss, 2])
      s_22 <- s_22 + sum(sample.missing.both[y1_miss, 2] ^ 2)
      hat_y1 <- beta_10_2 + beta_11_2 * sample.missing.both[y1_miss, 2]
      s_1 <- s_1 + sum(hat_y1)
      s_11 <- s_11 + sum(hat_y1 ^ 2 + sigma_11_2)
      s_12 <- s_12 + sum(hat_y1 * sample.missing.both[y1_miss, 2])
      # 重新估计参数
      mean.new <- c(s_1, s_2) / n
      cov.new <- c(s_11 / n - mean.new[1] ^ 2,
                   s_12 / n - mean.new[1] * mean.new[2],
                   s_22 / n - mean.new[2] ^ 2)
    }
    c(mean.new, cov.new, cov.new[2] / sqrt(cov.new[1] * cov.new[3]))
  }
```