A Generalized Linear Latent Variable Model to model Missing Non at Random Data

Guillaume Blanc

February 2021

1. Introduction

In this work, we propose a new method to model missing not at random (MNAR) data in a multivariate settings. Following **sportisse˙imputation˙2020**, we consider two settings: in the first, we model the joint density of the data together with the missing data "mask"; in the second, the mask is concatenated to the data: the density of the resulting augmented data is then modeled. In both cases, we propose a generalized linear latent variable model (GLLVM, [**skrondal˙generalized˙2004**]) to model the missingness mechanism. A GLLVM is an extension of factor analysis to accommodate responses within the exponential family, for instance binary responses; what's more, covariates information can also be incorporated. Similarly to factor analysis, it is assumed that a low-dimensional random variable, called the "latent" variable, explains the dependence within the data. By estimating the latent variables, one can then perform dimension reduction. The crucial assumption is that, conditional on the latent variable, the data are assumed independent. For instance, for setting 2, the data can be modeled as (conditional) independent Gaussian random variables, and the concatenated mask can be modeled as (conditional) Bernoulli independent random variables; by this conditional independence, the conditional density is simplified considerably.

An advantage of using a GLLVM, apart from the fact that it allows to model many types of responses, including the binary random variables that comprise the mask, is that it is a generative model (which may be advantageous for multiple imputation techniques?), and that its parameters are "almost" identifiable (up to a rotation, similarly to factor analysis). Should it be of interest, then, one could test for the significance of the loadings, in order to better understand the nature of the missingness mechanism in a particular application.

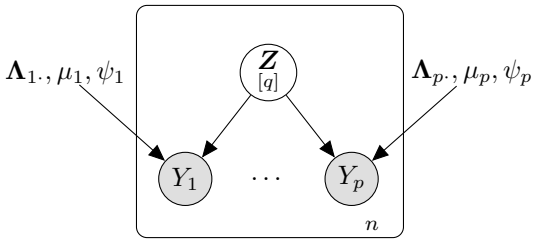

Figure 1: Generative model of the data: factor analysis.

2. Mathematical Setup

Following **little˙statistical˙2019**, we denote the complete data matrix by the matrix $Y = (y_{ij})$, and denote by the random matrix $M \in \{0,1\}^{n\times d} = (m_{ij})$ the missing-data pattern, where, for all $i = 1,\ldots,n$ and $j = 1,\ldots,d$, m_{ij} takes value 1 if y_{ij} is missing (NA), and 0 otherwise. The missingness mechanism is characterized by the conditional distribution of m_i given x_i, with density $f_{M|Y}(m_i|y_i,\phi)$ where ϕ denotes unknown parameters. In what follows, we will use a GLLVM to model this distribution. Let $F_{\boldsymbol{\theta}}$ be a model of the data parameterized by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

3. Estimation with Known Missingness Mechanism

We propose to model the multivariate data using exploratory factor analysis (EFA). Suppose that $Y_1, Y_2, \ldots, Y_p$ have a p-variate normal distribution with mean $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_p)$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{jk})$. We write $\mathbf{Y} = (\mathbf{Y}_{(0)}, \mathbf{Y}_{(1)})$, where $\mathbf{Y}$ represents a random sample of size n on $(Y_1,\ldots,Y_p)$, $\mathbf{Y}_{(0)}$ the set of observed values, and $\mathbf{Y}_{(1)}$ the missing data. For $i = 1,\ldots,n$, denote by $\mathbf{y}_{(0),i}$ the set of variables with values observed for unit i.

To estimate the parameter $\mathbf{\Lambda}, \boldsymbol{\mu}, \boldsymbol{\psi}$, we propose to impute the missing variables by their observed average and estimate the model by maximum likelihood. Unless the data are MCAR, this will in general result in a biased estimator. We propose Denote the estimator as a function of the data Y.

$$\widehat{\boldsymbol{\pi}}(Y) \tag{1}$$

When the missingness mechanism is known, we propose to estimate the model parameters

3.1 Modelling the mechanism

In this section, we largely follow the notations and exposition of **sportisse˙imputation˙2020**.

- Conditional on the data and an unobserved random variable Z of dimensions $q < p$, the

missing data indicators are independent.

- The distribution of a missing-data indicator M_{ij} is a function of Y_{ij} and Z only.

Given these assumptions, the missing-data mechanism can be written as:

$$p(M = \Omega|Y = y; \phi) = \prod_{i=1}^{n} \int \prod_{j=1}^{p} p(\Omega_{ij}|y_{ij}, z; \phi)h(z)dz, \quad \forall Y, \phi, \tag{2}$$

where the unobserved random variables Z with known density $h(\cdot)$ have been marginalized out. More specifically, for $i = 1, \ldots, n$, we assume the rows $Z_{i\cdot}$ of Z to be independent multivariate standard normal, $Z_{i\cdot} \sim MN(0, I_q)$, and, for all $i = 1, \ldots, n$ and $j = 1 \ldots, p$, that $M_{ij}|Y = y_{ij}, Z = z_i$ follows a Bernoulli distribution with mean $\mu_{ij} = g^{-1}(\eta_{ij}) := g^{-1}(\beta_j^0 + y_{ij}\beta_j + z_i^\top \Lambda_{j\cdot})$, for a vector $\beta^0 \in \mathbb{R}^p$ of intercepts, a vector $\beta \in \mathbb{R}^p$ of fixed effect coefficients and a matrix $\Lambda \in \mathbb{R}^{p\times q}$ of loadings, and where $g(\cdot)$ is the logit link function: $g(x) := \log(x)/(1-\log(x))$, whose inverse $g^{-1}(\cdot)$ is the sigmoid function $g^{-1}(x) = 1/(1+\exp(-x))$. We let ϕ collect the unknown model parameters β^0, β and Λ. This specification is a special case of a generalized linear latent variable model, and allows to model the dependence between the elements of the mask M through the latent variable Z. The joint model of the data Y and the mask M can be written, by the above independence assumptions, as

$$\begin{aligned} p(y, \Omega; \theta, \phi) &= p(y; \theta)p(\Omega|y; \phi) && (3) \\ &= \prod_{i=1}^{n} \int \prod_{j=1}^{p} p_j(y_{ij}; \theta_{ij})p(\Omega_{ij}|y_{ij}, z; \phi_j)h(z)dz, && (4) \end{aligned}$$

which can then be estimated with a Monte-Carlo EM algorithm (MC-EM, see [**sportisse˙imputation˙2020**]), where, for the M step, θ and ϕ can be separately estimated, and a dedicated algorithm is used to estimate ϕ. Here, Y can itself be modeled by probablistic PCA or factor analysis...

3.2 Adding the Mask

Here, again following **sportisse˙imputation˙2020**, we consider to augment the data Y by concatenating the mask, yielding the augmented data matrix $Y^A := [Y|M]$. We propose to model this matrix by a GLLVM, where, conditional on a latent variable Z, the elements of $Y_{i\cdot}^A$ are independent, for all $i = 1, \ldots, n$. Again, since the latent variable is not observed, it must be marginalized

out, yielding the following model for Y^A:

$$p(Y^A = y^A; \phi) = \prod_{i=1}^{n} \int \prod_{j=1}^{p} p_j(y_{ij}|z; \theta_{ij}) p(m_{ij}|z; \phi_j) h(z) dz, \quad \forall Y, \phi, \theta, \tag{5}$$

where θ_i denotes the parameters of the conditional density of $Y_i|Z$. Here again, we propose to estimate the parameters using an (MC)-EM algorithm, and where, in the M step, a dedicated algorithm is used to estimate the model parameters.

3.3 Summary and Open questions

We proposed to use a GLLVM to model the missingness mechanism, in two settings: in setting 1, the effect of the value of the variable on its probability of being missing is modeled via the linear predictor on the assumed conditional distribution of $M|Y$. Even so, $M|Y$ is not assumed independent: we propose to model the remaining dependence using a GLLVM. In a second setting, where the mask is added, we propose to model both the data and the mask using a single GLLVM: the dependence of m_{ij} and y_{ij} is thus captured by the latent variable. For both models, we propose to use an MC-EM algorithm to estimate the model parameters, where, in the M step, a dedicated algorithm is used to compute the parameters of the model (given the imputed values for the missing data).

Due to my limited knowledge in missing data mechanisms, here are open questions

1. I am not able to judge the potential of this methodology, and while I understand the differences between the two approaches, I am not able to judge which one shows the most promise for applications with missing data.

2. While I am condifent that both models can be estimated, I haven't yet implemented the algorithms (I anticipate that this requires some work, and would like to know if it is worth it). I have experience in estimating GLLVMs in large dimensions ($p \approx 10'000, n \approx 100'000$), but this works requires the estimation to be done within the MC-EM algorithm, probably with a Metropolis Hastings sampling procedure, which makes the estimation more complicated.

3. How could one judge the merits of the method? By which measure? Errors of prediction of the imputed missing values of Y, or MSE on the estimated θ?

4. Are there some typical datasets that one could analyze using such methods?

5. In setting 2, is there interest for confidence intervals on the loadings? One could devise a test where the null hypothesis is that all loadings are 0 (which, if the null hypothesis is rejected,

may potentially indicate that the data are not MCAR?)?

4. Estimation with Unknwon Missingness Mechanism