Data Imputation Notes

June 25, 2020

1 Introduction

Some notes regarding methods for data imputation.

2 Data Imputation

There are four well-used methods for data imputation. The first (which we will not discuss here) is to build a predictive model of the data and simply predict the missing values as though they were outputs - a neural network would be appropriate here. As we have little experience in this method, this will not be further discussed.

The three methods that are discussed are

1. Expectation Maximisation for missing values [1].
2. Multiple Imputation and Monte Carlo [2, 3, 4, 5].
3. Generalised Low Rank Models [6, 1].

Although these are, in some sense, three distinct methods, they are all reliant on a basic premise for the nature of the missing data, which ties together **latent variable modelling** and **data imputation**.

3 Probabilistic Framework

There are two general ways of thinking about imputation: maximum likelihood estimation of model parameters, or Bayesian treatment. These differ fundamentally in this area, not just at the usual Bayesian-Frequentist level.

The basic ethos of data imputation is **not** to simply estimate the missing values themselves. If we want to estimate some population statistic from our incomplete sample, then we would like to do two things: estimate this parameter in a way that would be consistent if we had more data, and model the uncertainty in our parameter estimate given that we are missing data. We cannot hope to have a "good" estimate of the parameter with missing data, but we can hope to be consistent and conservative in our estimation.

In the maximum likelihood setting, we try to estimate the parameters of a model whilst taking into account all possible values that the missing data might take. The ancillary statistics may then be estimated by the second derivatives in a standard way.

In the Bayesian setting, we try to take into account model uncertainty when trying to estimate some statistic. This is best understood in a regression setting. There are two sources of uncertainty here: the inherent noise in the generative process, and the uncertainty due to our missing data. Therefore our imputation mechanism needs to model both processes in order to be sufficiently conservative for our errors [3].

4 Maximum Likelihood and Gaussian Mixtures

In this section, we will review the maximum likelihood technique for missing data. Let the data be represented by X, in row format, consisting of M samples in $\mathbb{R}^N$, and hence $X \in \mathbb{R}^{M \times N}$[1]. Some of the elements of X are not observed. Let us denote the observed elements as X_o and the missing subset as X_m. Let us denote the parameter (or parameters) that we are inferring as θ, such that the likelihood is given by

$$L = P(X_o|\theta) = \sum_{X_m} P(X_o, X_m|\theta) \tag{1}$$

The basic idea is that we cannot make sense of the likelihood (or any estimator) without a full data set.

4.1 Example

As a simple example, consider tossing a coin N times, and then repeating this experiment M times. However, every experiment we run we somehow manage to loose a set of outcomes. As a particular example, consider the following with $N = 4$ and $M = 5$

$$X = \begin{pmatrix} \{H & H & ? & T\} \\ \{H & T & H & ?\} \\ \{H & H & ? & H\} \\ \{T & ? & T & T\} \\ \{? & T & H & T\} \end{pmatrix} \tag{2}$$

We cannot make sense of evaluating a likelihood here, and therefore the only thing we can do is evaluate the complete likelihood with the elements given by $?$ actually evaluated at points (in fact, all possible points). As we have one coin, this model is particularly easy: we can actually drop the matrix nature of the data and simply consider $\mathrm{Vec}(X)$, and we can reorder such that all of our missing data is placed at the end of the vector. Let the total number of heads **observed** be given by N_H (here $N_H = 8$) and similarly $N_T = 7$. Let us switch notation such that heads are $x = 1$ and tails are $x = 0$, and we will sum over the missing elements $x_m^1 \dots x_m^t$ for t missing elements (here $t = 5$ and $NM = N_H + N_T + t$).

If we denote n as the total number of heads in the unseen data, then the likelihood is given by

$$P(X_o|p) = \mathcal{L}(p|X_o) = \sum_{n=0}^{t} \begin{pmatrix} NM \\ N_H + n \end{pmatrix} \begin{pmatrix} t \\ n \end{pmatrix} p^{N_H+n}(1-p)^{NM-N_H-n} \tag{3}$$

[1]We usually use column format in order that our vectors are column vectors by default. This also means our Gram-matrix will be XX^T here.

in which the unseen data has been marginalised over. The summations can be represented in terms of special functions, and we note that this allows us to write the likelihood in the following way

$$P(X_o|p) = \begin{pmatrix} NM \\ N_H \end{pmatrix} p^{N_N}(1-p)^{NM-N_H} \ {}_2F_1\left(-t, N_H - NM, 1 + N_h; \frac{p}{1-p}\right) \tag{4}$$

where ${}_2F_1\left(a, b, c; z\right)$ is a hypergeometric function, and where $N_H \in \{0, 1, \ldots NM - t\}$. When $t \to 0$ we achieve the standard binomial[2].

Interpreting this likelihood, we can see that it consists of two important terms: a piece which depends only on the data that is seen, and is binomial in structure, and a correction from the hidden data. This correction means that the MLE for our parameter p is modified when there is missing data.

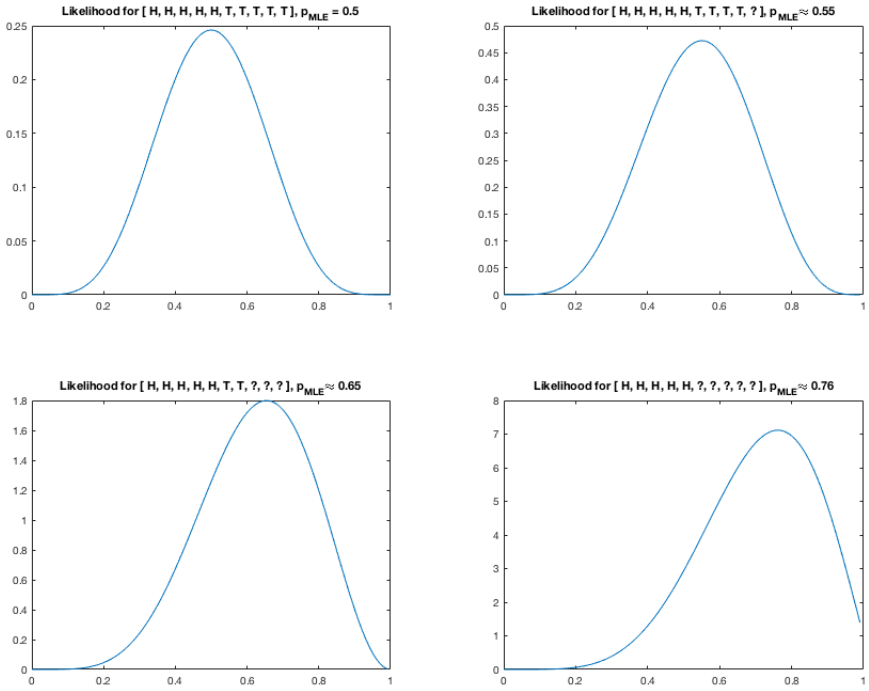

Figure 1: Plot of the likelihood function for various datasets, indicating the effect of the missing data imputation on both the likelihood, and the parameter estimation under MLE.

In Fig.1 we see four examples of the effect of imputation on the likelihood function. We can see that as the amount of unknown data increases, the MLE for the parameter p increases. For the case where all we have seen is all heads, the model suggests $p_{\text{MLE}} \approx 0.76$. As the fraction of missing data increases, we expect that the MLE will tend to the uninformed result of

[2]An interesting case is $t \to NM$ (and consequently $N_H = 0$), which corresponds to no data being seen, though no simplification of the likelihood has been found here for general parameters.

$p_{\text{uninformed}} = 0.5$, and certainly

$$\hat{p}_{\text{MLE}} = \left(\frac{N_H}{NM}\right) \underbrace{\frac{NM - t}{N_H}}_{p_{\text{Vis}}} + \left(1 - \frac{N_H}{NM}\right) p_{\text{uninformed}} \tag{5}$$

leads to a reasonable estimate of p_{MLE} in the large missing data limit, and where p_{Vis} is the value we would get if we simply ignored the missing data.

4.2 Expectation Maximisation for Missing Data

The exact summation over missing data in the binomial model is usually not possible, even in a likelihood setting (let alone a Bayesian treatment). Iterative procedures are therefore required, and expectation maximisation is the standard technique.

We will derive it in general based upon the above analysis. Let the likelihood of the observed data be given by

$$L = P(X_o|\theta) = \sum_{X_m} P(X_o, X_m|\theta) \tag{6}$$

If we denote the observed data by x_n and the missing data by y_m, and take the data log-likelihood of the above, then

$$\log L = \sum_n \log \left(\sum_{x_m} P(x_n, x_m|\theta) \right) \tag{7}$$

Note the summation over n is over the observed data (from the original product in the data likelihood), while the summation over x_m is over the values that the unobserved data can take. Fundamentally, a log of a sum is *hard* to control, and expectation maximisation is a way of re-ordering this.

Let us introduce a distribution over the hidden data, $P(x_m)$, which for now is arbitrary. Then

$$l \equiv \log L = \sum_n \log \left(\sum_{x_m} P(x_m) \frac{P(x_n, x_m|\theta)}{P(x_m)} \right) \tag{8}$$

Then Jensen's inequality tells us that for a concave function (such as the logarithm)

$$\mathbb{E}(f(x))_{P(x)} \geq f(\mathbb{E}(x)_{P(x)}) \tag{9}$$

and so our likelihood can be bounded as

$$l \geq \sum_n \sum_{x_m} P(x_m) \log \left(\frac{P(x_n, x_m|\theta)}{P(x_m)} \right) \tag{10}$$

To make the bound *tight*, we use the fact that Jensen's inequality becomes an equality if x is a constant, and therefore $P(x_m) \propto P(x_n, x_m|\theta)$. So, how do we choose how to make this proportionality? There are two natural choices: $P(x_m) = P(x_m, x_n|\theta)$ or $P(x_m) = P(x_m|x_n, \theta)$.

The first choice simply yields the original likelihood upon substitution, and so we use the second

$$
\begin{aligned}
l &= \sum_n \sum_{x_m} P(x_m|x_n,\theta) \log\left(\frac{P(x_n,x_m|\theta)}{P(x_m|x_n,\theta)}\right) && (11)\\
&= \sum_n \sum_{x_m} P(x_m|x_n,\theta) \log\left(P(x_n|x_m,\theta)\right) && (12)\\
&= \mathbb{E}_{P(x_m|x_n,\theta)}\left(\sum_n \log(P(x_n|x_m,\theta))\right) && (13)
\end{aligned}
$$

The final line indicates that we are taking the expectation value of the data log-likelihood, over the distribution of missing data. If we want the MLE, then we can perform the iterative procedure as follows:

- Start with estimates for the likelihood parameters, θ^{t-1}.
  - Calculate the expectation over $P(x_m|x_n,\theta^{t-1})$ yielding a function $Q(\theta,\theta^{t-1})$.
  - Update the parameters $\theta^t = \underset{\theta}{\operatorname{argmax}}\, Q(\theta,\theta^{t-1})$.

Obviously this is easier said than done. Forming the expected data likelihood (the interpretation of $Q(\theta,\theta^{t-1})$) can be difficult, and certainly finding the parameters which maximise this function can be troublesome. For the case of the multivariate normal distribution, because the marginals and conditionals are all calculable, we can carry out both steps in an analytic way.

4.3 Fitting Multivariate Normal Distributions with Missing Data

Following [7] Sec.11.6.1, we can derive the updates for our parameters by forming the function Q directly. First we form the expectation.

Let $P(\boldsymbol{x}|\theta) \sim \mathcal{N}(\boldsymbol{\mu},\Sigma)$, and then partition the variables into the observed elements $\{x_o\}$ and the missing elements $\{x_m\}$. Again, we envisage a data structure similar to that of Eq.2, but where our elements are in $\mathbb{R}$ rather than $\mathbb{Z}_2$. In general, we can calculate the marginal distribution for a subset of variables.

$$P(x_m|x_o,\boldsymbol{\mu},\Sigma) = \mathcal{N}(m,V) \tag{14}$$

5 EM-PCA and Generalised Low-Rank Models (GLRM)

Here, we discuss a direct method of applying expectation maximisation (EM) for data imputation. It was initially set up as an option for applying expectation maximisation to PCA in order to provide a proper density estimation for new data. A side effect of this was that PCA could be calculated via EM, which then allows it to deal with missing data [3].

Here we will work with low rank models instead. There is a choice of origin, which when taken to be $\mathbf{0}$ amounts to PCA, but in general is simply SVD. Denote the data matrix by X, with some

[3] The final algorithm for EM-PCA is actually simply alternating least squares.

missing values captured by Z. Let the set of elements for which $Z = 0$ be denoted by Ω. We seek to solve

$$\min_{U,L,Y} \sum_{(i,j)\in\Omega} (Y_{ij} - [UL^T]_{ij})^2 + \sum_{(i,j)\notin\Omega} (Y_{ij} - [UL^T]_{ij})^2 + \gamma \mathrm{Reg}(Y) \tag{15}$$

We also apply the constraint to this minimisation that

$$U^T U = I \tag{16}$$
$$L^T L = D \tag{17}$$
$$Y_{ij} = X_{ij} \quad (i,j) \in \Omega \tag{18}$$

where D is a diagonal matrix. We have also included a regularisation on Y which allows us to prevent values that are extreme from entering (by using, for example, an ℓ_1-norm). For now we will work with $\gamma = 0$.

Alternating methods are efficient ways of trying to solve this type of problem, though they often converge only to local extrema. The basic idea is to solve for one variable while holding rest constant. A natural solution is as follows, which follows the EM type updates

$$\text{Use SVD on the current } Y \text{ such that} \quad Y^T Y U = UD \text{ and } YY^T V = L \tag{19}$$
$$\text{For fixed } U \text{ and } L \quad Y_{ij} = [UL^T]_{ij} \ (i,j) \notin \Omega \tag{20}$$

We are using alternating variable updates, and then using the result to estimate the free elements of Y. Note that V are the orthonormal singular vectors, but we incorporate the singular values in such that $L^T = DV^T$. This algorithm sequentially minimises $\left\| Y - UL^T \right\|_2^2$, and imputes the missing values in X via Y.

Note that we have skipped over a significant idea here: rank. PCA (or SVD) is a way of representing data in a lower dimensional space, and we have not specified this here. Rank is incorporated into this algorithm by simply keeping a set number of singular values for the updates to U and L. However, the pattern of missing data Z has a rank, and the interaction between this and the chosen rank is inobvious.

5.0.1 Simulations, Results and Problems

Before we offer results, we first note that an important premise of PCA is that of Gaussian density. Following [6], the idea is as follows. We have some data $Y_{N\times P}$, which are N vectors $\boldsymbol{y}_n \in \mathbb{R}^p$. We argue that actually, each vector can be embedded in a K dimensional space via a linear transformation

$$\boldsymbol{y} = C\boldsymbol{x} + \boldsymbol{v} \tag{21}$$

where $\boldsymbol{x}$ is the projection of the data onto the lower dimensional space, $C \in \mathbb{R}^{N\times K}$ is the transformation, and $\boldsymbol{v} \in \mathbb{R}^K$ is additive Gaussian noise. We assume that $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, I)$ and $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, R)$. This implies that $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, CC^T + R)$. If we take the limit $R = \lim_{\epsilon\to 0} \epsilon I$, then we arrive at PCA.

References

[1] David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[2] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[3] Jared S Murray et al. Multiple imputation: A review of practical and theoretical findings. *Statistical Science*, 33(2):142–159, 2018.

[4] Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489, 1996.

[5] Yang C Yuan. Multiple imputation for missing data: Concepts and new development. In *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*, volume 267, 2000.

[6] Sam T Roweis. Em algorithms for pca and spca. In *Advances in neural information processing systems*, pages 626–632, 1998.

[7] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.