An Analysis on Heart Disease Data

Pengkun Gu

Abstract

In this article, I mainly studied the relation between the occurrence of heart disease and some other variables by Bayesian regression algorithm, using an open dataset.

Raw Data Processing

The original data is downloaded from *UCI Machine Learning Repository*, since the data form of the unprocessed version is confusing, I finally choose to use the processed version. There are four blocks of data, respectively collected in $\{Cleveland\}$, $\{Hungarian\}$, $\{VA\}$ and $\{Switzerland\}$.

In each block there are 14 attributes, respectively representing (1) **[AGE]** for *Age*, (2) **[SEX]** for *sex*, (3) **[CHESTP]** for *chest pain type*, (4) **[RESTBPS]** for *resting blood pressure*, (5) **[CHOL]** for *serum cholesterol level*, (6) **[FBS]** for *fasting blood sugar*, (7) **[RESTECG]** for *resting electrocardiographic results*, (8) **[ACHHR]** for *maximal heart rate achieved*, (9) **[EXANG]** for *exercise induced angina*, (10) **[OLDPEAK]** for *ST depression induced by exercise relative to rest*, (11) **[SLOPE]** for *slope of the peak exercise ST segment*, (12) **[COLOR]** for *number of major vessels colored by fluoroscopy*, (13) **[THAL]** for *Thallium stress test result*, (14) **[DIAG]** for *diagnosis of heart disease*.

It's notable that there are missing values included in the data.

The [DIAG] column, ranging from 0 to 4, represents the degree of severity of symptom. Since in $\{Hungarian\}$ there are only reports "0"s and "1"s in [DIAG] column, so for unity and simplicity, in this article I only distinguish the presence or absence of symptom, that is, identify all of $\{1, 2, 3, 4\}$ with 1 in [DIAG] column.

For another 13 columns of data, the translation from description to digits is shown in the following table in detail:

Columns	Translation
[AGE]	30 to 0, 60 to 1, linear
[SEX]	"Female" to 0, "Male" to 1
[CHESTP1]	"Asymptomatic" to 0, "Typical Angina"(1) to 1
[CHESTP2]	"Asymptomatic" to 0, "Atypical Angina"(2) to 1
[CHESTP3]	"Asymptomatic" to 0, "Non-anginal Pain"(3) to 1
[RESTBPS]	90 to 0, 180 to 1, linear
[CHOL]	150 to 0, 300 to 1, linear
[FBS]	"<120" to 0, ">120" to 1
[RESTECG1]	"Normal"(0) to 0, "ST-T Wave Abnormality" to 1
[RESTECG2]	"Normal"(0) to 0, "Left Ventricular Hypertrophy" to 1
[ACHHR]	100 to 0, 200 to 1, linear
[EXANG]	"No" to 0, "Yes" to 1
[OLDPEAK]	-1 to 0, 1 to 1, linear
[SLOPE+]	"Flat"(2) to 0, "Upsloping"(1) to 1

Columns	Translation
[SLOPE-]	"Flat"(2) to 0, "Downsloping"(3) to 1
[COLOR]	0 to 0, 3 to 1, linear
[THAL1]	"Normal"(3) to 0, "Fixed"(6) to 1
[THAL2]	"Normal"(3) to 0, "Invertible"(7) to 1
[DIAG]	"No"(0) to 0, "Yes"(1,2,3,4) to 1

We regard the [DIAG] term as the label, denoted by $\mathbf{y}$, and the other 18 columns of data as factors to be examined, denoted by $\mathbf{z}$. The whole data (including the missed) is considered as an array of dimension 3, that is, $\mathbf{z}_{h,t,p}$, were $h \in \{1:4\}$ represents the hospitals, $t \in \{1:18\}$ represents the features, and $p \in \{1:P_h\}$ represents the people included in the investigation, where explicitly $\{P_1, P_2, P_3, P_4\} = \{303, 294, 123, 200\}$.

Methodology

The first main algorithm is a meta-version of Bayesian logistic random effect regression model, which is endowed with L^1-prior in purpose of sparsifying the inferred parameters.

The traditional logistic regression for binary data is derived from the assumption that $P(y_{h,p} = 1|\mathbf{z}_{h,p}) = 1/(1 + e^{-(\mathbf{w}^\top \mathbf{z}_{h,p} + v)})$, which, is further derived from some normal assumptions. If we consider the random effects and Bayesian settings, a Bayesian logistic random effect regression model can be set up as follows:

$$\begin{aligned} y_{h,p}|\mu_{h,p} &\overset{i.}{\sim} Bernoulli(\mu_{h,p}) \\ \mu_{h,p} &= \frac{1}{1 + e^{-((\mathbf{w}^\top, v)\cdot(\mathbf{z}_{h,p}^\top, 1)^\top + r_h)}} \end{aligned}$$

Here r_h represents the random effect of different hospitals, $\mathbf{w}$ is the regression coefficient with length 13 corresponding to 13 items, while v refers to the constant term.

$\mathbf{w}$ is endowed with LASSO prior with prior λ, that is, $p(\mathbf{w}) \propto exp(-\lambda \sum |w_t|)$. Noticing that the two-sided exponential distribution can be regarded as a Gaussian with its variance under a exponential distribution, it can be implemented as a hierarchical structure:

$$\begin{aligned} \mathbf{w} &\sim LASSO(\lambda) \\ &\Updownarrow \\ \tau_t^2 &\overset{i.}{\sim} Exp(\frac{\lambda^2}{2}) \\ w_t|\tau_t^2 &\sim N(0, \tau_t^2) \end{aligned}$$

And it is not harmful setting λ as another random parameter with a broad prior distribution in order to avoid the discordant cross-validation procedure for tuning λ.

We do not assume that we have any knowledge on the distribution of data $\mathbf{z}$, so we here take non-informative prior $\mathbf{z} \sim 1$. So does $v \sim 1$. Without any prior knowledge on those 4 different hospitals, we regard r_h as exchangeable parameters, that is, they are *i.i.d.* conditioning on some common parameter σ. And σ can be assumed with a broad prior distribution.

The total prior distributions is demonstrated as follows:

$$
\begin{aligned}
\mathbf{w} &\sim LASSO(\lambda) \\
\lambda &\sim 1_{\mathbb{R}^+} \\
v &\sim 1 \\
z_{h,t,p} &\overset{i.}{\sim} 1 \\
r_h|\sigma &\overset{i.}{\sim} N(0,\sigma) \\
\sigma &\sim \Gamma^{-1}(1,1)
\end{aligned}
$$

But I found a regretful fact that some data provider missed several columns of data, leading to difficulties in defining the regression parameters. To be explicit, in $\{Hungarian\}$, [SLOPE], [COLOR], [THAL] is mostly missing; in $\{Switzerland\}$, [FBS], [CHOL], [COLOR] is mostly missing and [THAL] is severely missing; in $\{VA\}$, [COLOR], [THAL] is mostly missing. Besides of that, there are also lot of occasional missing values.

The missing values have to be considered. The whole data $\mathbf{z}$ can be viewed as related to a missing variable $\mathbf{m} = \{m_{h,t,p}\}$ where $m_{h,t,p} = 1$ means $z_{h,t,p}$ can be observed and $m_{h,t,p} = 0$ means $z_{h,t,p}$ is missed. According to $\mathbf{m}$, $\mathbf{z}$ can be divided into 2 parts: $\mathbf{z}_{obs}$ and $\mathbf{z}_{mis}$. In this case, without any negative evidence found or negative hypothesis proposed, it's reasonable to assume that the missing mechanism is so-called *missing completely at random* (MCAR), that is, $\mathbf{m}$ is independent to both observed and missing data. Or at least *missing at random* (MAR), that is, $\mathbf{m}$ is independent to the missing data. Here we postpone the controversy to later sections and here we only need to agree on that directly deleting the defective data would not lead to inferential bias.

And in that later section, I'll try to use EM algorithm to take advantage of the defective data and carry out regression with missing values. The main methodology there, is to "refill the missing data using current parameters" and "do regression and choose new parameter based on the filled data" alternatively until it seems to converge.

Regression on a Single Dataset.

Since the data in $\{Cleveland\}$ have few missing values and the data is adequate for an elementary logistic regression, we first ignore all other 3 dataset and delete all defective data in $\{Cleveland\}$. In this simplified case the random effect can be ignored and the problem turns to traditional Bayesian logistic regression. Here's the model:

$$
\begin{aligned}
y_p|\mu_p &\overset{i.}{\sim} Bernoulli(\mu_{h,p}) \\
\mu_p &= \frac{1}{1+e^{-(\mathbf{w}^\top, v)\cdot(\mathbf{z}_p^\top, 1)^\top}} \\
\mathbf{w} &\sim LASSO(\lambda) \\
\lambda &\sim 1_{\mathbb{R}^+} \\
v &\sim 1 \\
z_{t,p} &\overset{i.}{\sim} 1
\end{aligned}
$$

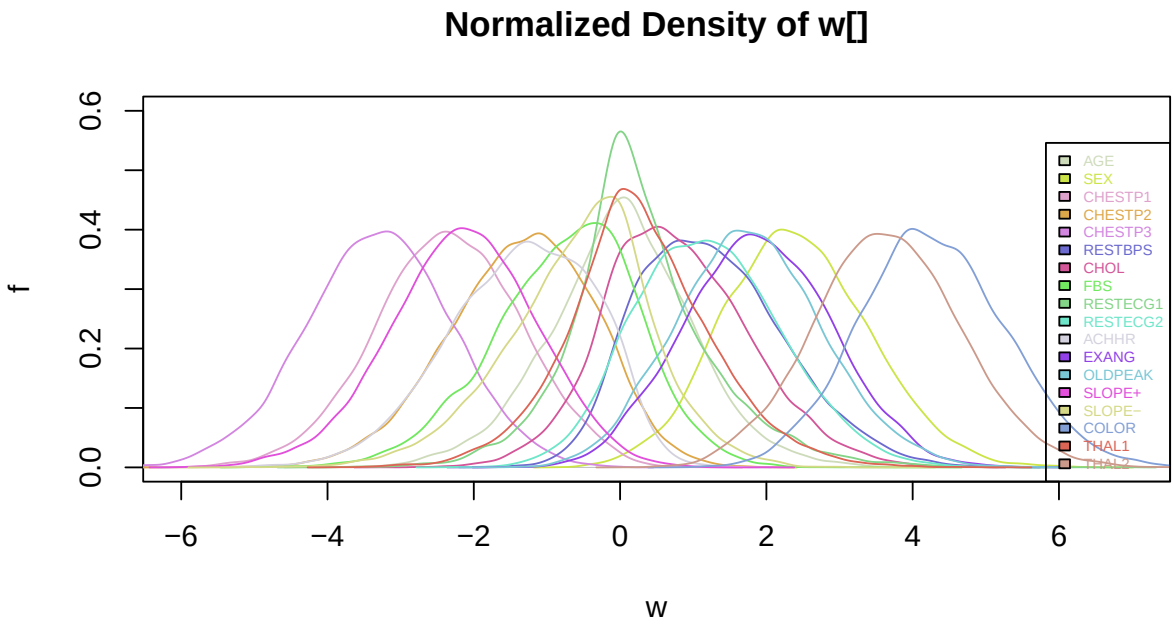
The posterior distribution of such parameters given data $\mathbf{y}$ and $\mathbf{z}$ is

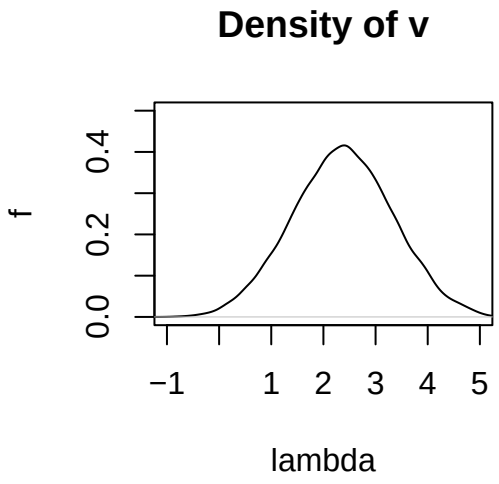
$$
\begin{aligned}
p(\mathbf{w}|\mathbf{y},\mathbf{z}) &= \int_{\mathbb{R}^2} p(\mathbf{w}, v, \lambda|\mathbf{y},\mathbf{z})\ dv\ d\lambda \\
p(\mathbf{w}, v, \lambda|\mathbf{y},\mathbf{z}) &\propto p(\mathbf{y}|\mathbf{w}, v, \lambda, \mathbf{z}) \cdot p(\mathbf{w}, v, \lambda|\mathbf{z}) \\
&= p(\mathbf{w}|\lambda) \cdot p(\mathbf{y}|\mathbf{w}, v, \mathbf{z}) \cdot 1_{\lambda>0} \\
&= p(\mathbf{w}|\lambda) \cdot \prod_p (\mu_p \cdot 1_{y_p=1} + (1-\mu_p)\cdot 1_{y_p=0}) \cdot 1_{\lambda>0}
\end{aligned}
$$

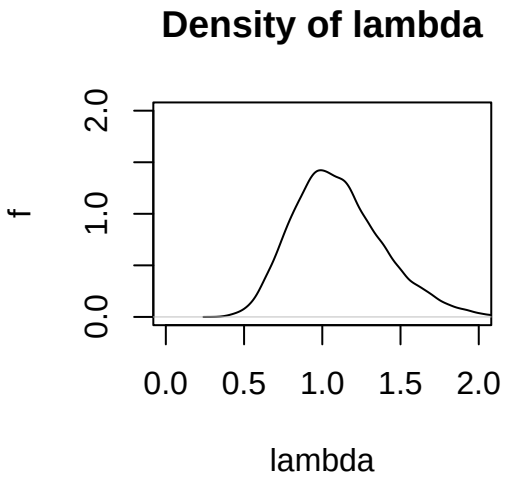
I used Gibbs sampling tool *WinBUGS* and R package `R2WinBUGS` here to sample from the posterior distribution. Here shows the result:

node	mean	sd	MC error	2.5%	median	97.5%
λ	1.11	0.3	0.009895	0.6229	1.077	1.782
v	2.374	0.9637	0.04263	0.4856	2.378	4.257
[AGE]	0.007958	0.5224	0.0156	-1.062	0.01224	1.052
[SEX]	1.117	0.4629	0.01289	0.2441	1.101	2.06
[CHESTP1]	-1.483	0.63	0.01164	-2.743	-1.48	-0.268
[CHESTP2]	-0.6667	0.4945	0.007617	-1.695	-0.6417	0.2133
[CHESTP3]	-1.544	0.4668	0.008192	-2.488	-1.538	-0.642
[RESTBPS]	1.052	0.8366	0.02171	-0.3452	0.9819	2.86
[CHOL]	0.3766	0.4813	0.0112	-0.5105	0.3433	1.395
[FBS]	-0.347	0.4643	0.006087	-1.329	-0.3102	0.492
[RESTECG1]	0.2743	0.9751	0.009851	-1.556	0.1667	2.493
[RESTECG2]	0.4101	0.3376	0.004728	-0.2	0.396	1.103
[ACHHR]	-1.158	0.86	0.02717	-2.959	-1.103	0.272
[EXANG]	0.7659	0.4044	0.006962	-0.0003	0.761	1.563
[OLDPEAK]	0.741	0.4169	0.01199	-0.0413	0.7314	1.593
[SLOPE+]	-0.9263	0.4307	0.008951	-1.805	-0.9194	-0.1106
[SLOPE-]	-0.332	0.5954	0.007551	-1.613	-0.2733	0.7661
[COLOR]	3.236	0.7559	0.01323	1.805	3.21	4.756
[THAL1]	0.1537	0.5573	0.006899	-0.9402	0.1274	1.322
[THAL2]	1.433	0.3919	0.0061	0.6697	1.425	2.217

Here I plot the normalized density of coefficients $\{w_1, ..., w_{18}\}$, hyperparameter v and λ:





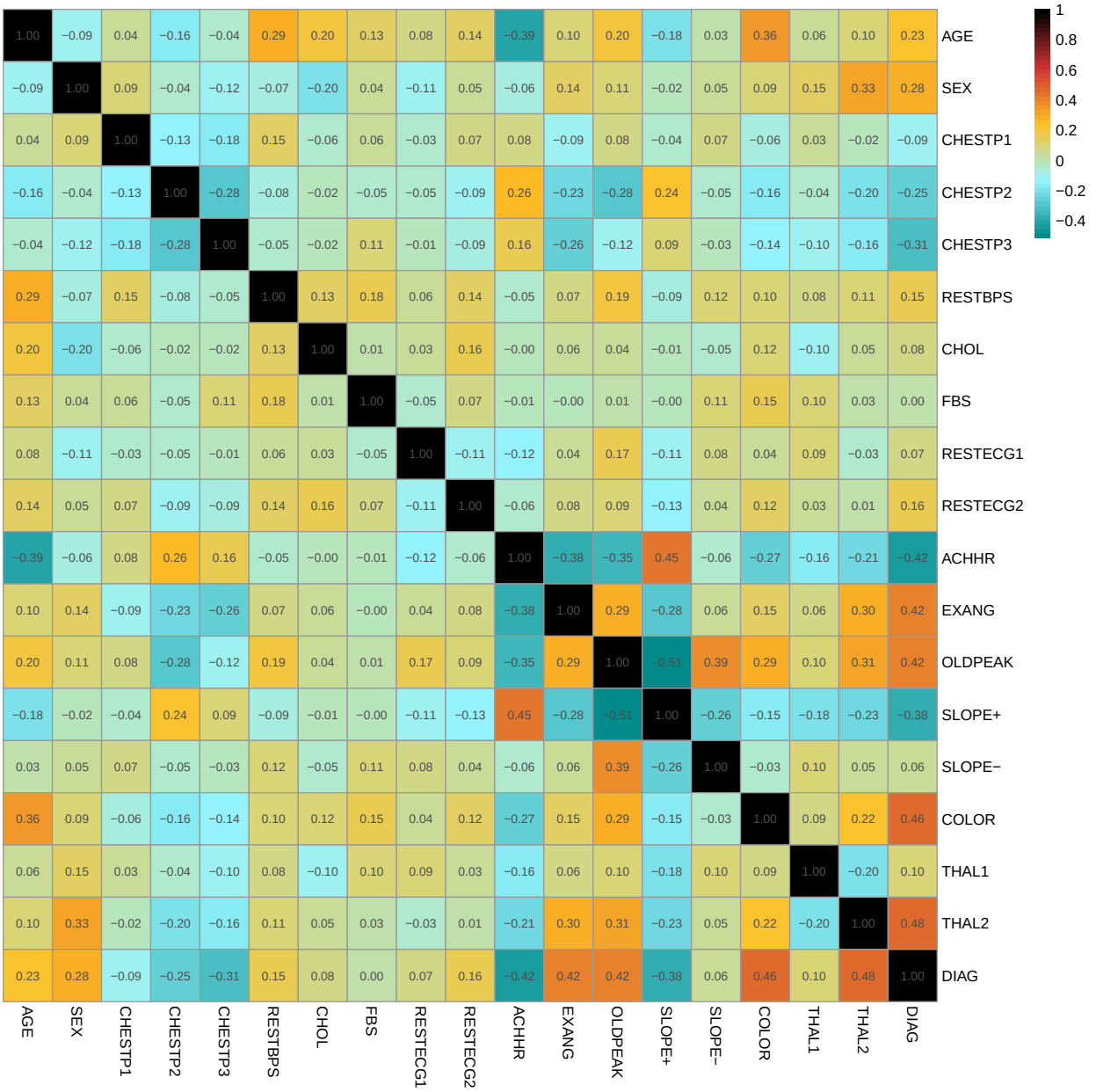


As a primary conclusion, the 95% confidence region for [COLOR], [THAL2], [CHESTP3] is far from 0, which means with very high confidence [COLOR] and [THAL] have significant correlation with the diagnosis of heart disease. Besides these, there are also some features such as [SEX], [EXANG], [OLDPEAK], [SLOPE+], [CHESTP1], whose 95% confidence region cannot be definitely separated from 0, it's still evidently that they have strong correlation with the diagnosis of heart disease. Some other features, such as [CHESTP2], [ACHHR], [RESTBPS], [RESTECG2] still have a relatively large confidence being nonzero. The features [AGE], [RESTECG1], [THAL1], [SLOPE-] have definitely no evident relation with heart disease, providing only this dataset. And the other features, such as [CHOL], [FBS], maybe are relative to heart disease, but not evidently enough.

If we are willing to propose an automatic procedure, to calling attention to patients highly possible with heart disease, and if data of all checking items are attainable (which is unlikely true), [COLOR], [THAL2], [ACHHR], [CHESTP2], [CHESTP3], [EXANG], [SEX], [OLDPEAK] can be considered. The most 3 significant ones, [COLOR], [THAL2] and [CHESTP3], can only assured in certain hospital, though some of the auxiliary items can be obtained without going to hospital. It's not feasible for an at-home heart disease early-warning algorithm based on this dataset.

Thought evidently correlative, it is far from enough to talk about exact illness from some complex body measurements, such as [FBS], [ACHHR], [SLOPE] and so on. In this analysis we only used a linear model under a certain translation of the original data, but the interpretation of such measurements can be really sophisticated. For more exact prediction or inference, maybe a more complex model should be implemented, such as *quadratic discriminant analysis*. But a more complex model lead to higher chance of overfitting, I'm still skeptical about whether that will perform better on this dataset with such a limited amount of data.

But there are some strange things, let us have a look at how the variable are empirically correlated, for comparing with the posterior inference:



It seems that the posterior distribution of $\mathbf{w}$ is largely congruent with the correlation of [DIAG] and different items. But it is interesting that the inference seems far from the common sense in our life. For example, coefficients of 3 [CHESYP_] items are very likely with negative coefficient, which suggests that compared with "asymptomatic" result in [CHESTP_], the 3 "symptomatic" result all tends to suggest that there are not heart disease, or just conversely, "asymptomatic" result strongly tends to suggest there is heart disease. And the same peculiarity happens on [ACHHR], which shows that the coefficient of that item is likely to be negative, leading to a strange result that high achieved heart rate suggests lower chance of heart disease, but we know that human with heart disease usually have a too-high heart rate. And about "fluoroscopy" item, I found some materials claiming that some fluoroscopy methods like CT can be used to detect "very bright" regions in their arteries, corresponding to calcified lesions. So it seems reasonable that higher number of colored vessel are related to higher chance of heart disease, which is shown in the posterior inference.

But it's hard to say whether thare is anything wrong in the original data, which is indeed downloaded from *UCI Machine Learning Repository*. And the same peculiarity also happens in *hungarian*, *switzerland* and *VA*. It is possible that original interpretation or raw data processing provided by *UCI Machine Learning*

Repository have something wrong or improper.

Dealing with Missing Data

Since in *Hungarian*, *VA* and *Switzerland*, most of key item, that is, [COLOR] and [THAL] are missing, and that the dataset have peculiarity in result, it is not too valuable in processing a more complicated meta-analysis. But here I still choose to state the methodology of doing meta-regression with missing data.

Here we fix λ for simplicity. Usually missing data can be regarded as parameters whose posterior distribution is to be simulated, that is, to find $p(\mathbf{z}_{mis}, [\mathbf{w}, v, \mathbf{r}]|\mathbf{z}_{obs}, \mathbf{y})$. But when dimension of $\mathbf{z}_{mis}$ is too high relative to dimension of other parameters, the usual MCMC methods become inefficient. A common solution for missing data is EM algorithm and data augmentation, though through which we can only approximately calculate the *maximum a posterior* (MAP) solution of $\mathbf{w}, v, \mathbf{r}$.

A common EM algorithm requires both $p([\mathbf{w}, v, \mathbf{r}]|\mathbf{z}_{obs}, \mathbf{z}_{mis}, \mathbf{y})$ and $p(\mathbf{z}_{mis}|[\mathbf{w}, v, \mathbf{r}], \mathbf{z}_{obs}, \mathbf{y})$ easy to compute. In the k-th iteration, first we draw samples $\widehat{\mathbf{z}_{mis}}_k$ from distribution $p(\mathbf{z}_{mis}|\widehat{[\mathbf{w}, v, \mathbf{r}]}_k, \mathbf{z}_{obs}, \mathbf{y})$, then do maximize the posterior density from the augmented data, obtaining $\widehat{[\mathbf{w}, v, \mathbf{r}]}_{k+1}$ from $p([\mathbf{w}, v, \mathbf{r}]|\mathbf{z}_{obs}, \widehat{\mathbf{z}_{mis}}_k, \mathbf{y})$.

The way of drawing $\widehat{\mathbf{z}_{mis}}_k$ is usually taking expectation, which arises from the fact that

$$\begin{aligned}
(\mathbf{w}, v, \mathbf{r})^*_{map} &= \underset{[\mathbf{w},v,\mathbf{r}]}{argmax}\ p((\mathbf{w}, v, \mathbf{r})|\mathbf{z}_{obs}, \mathbf{y}) \\
&= \underset{[\mathbf{w},v,\mathbf{r}]}{argmax}\ p(\mathbf{w}, v, \mathbf{r}) \cdot p(\mathbf{y}|(\mathbf{w}, v, \mathbf{r}), \mathbf{z}_{obs}) \\
&= \underset{[\mathbf{w},v,\mathbf{r}]}{argmax}\ \mathbb{E}_{\mathbf{z}_{mis}|(\mathbf{w},v,\mathbf{r}),\mathbf{z}_{obs},\mathbf{y}}\left[p(\mathbf{w}, v, \mathbf{r}) \cdot p(\mathbf{y}|(\mathbf{w}, v, \mathbf{r}), \mathbf{z}_{obs}, \mathbf{z}_{mis})\right]
\end{aligned}$$

The intuition is to iteratively optimize $[\mathbf{w}, v, \mathbf{r}]$ which occurs 2 times in the expression above, and an iterative algorithm can converge to the optimal solution. There are also some variative algorithms derived from fundamental EM, for example, stochastic EM, which directly sample $\mathbf{z}_{mis}$ from $p(\mathbf{z}_{mis}|[\mathbf{w}, v, \mathbf{r}], \mathbf{z}_{obs}, \mathbf{y})$ instead of taking expectation, which is more convenient sometimes.

In this case the MAR (or MCAR) assumption ensured that the missing variable $\mathbf{m}$ is independent with the missing value: $p(\mathbf{m}|\mathbf{z}_{obs}, \mathbf{z}_{mis}, \mathbf{y}, \phi) = p(\mathbf{m}|\mathbf{z}_{obs}, \mathbf{y}, \phi)$, which makes it feasible to talk about $p(\mathbf{z}_{mis}|[\mathbf{w}, v, \mathbf{r}], \mathbf{z}_{obs}, \mathbf{y})$. And this distribution can be seen as marginalized from $p(\mathbf{z}|[\mathbf{w}, v, \mathbf{r}], \mathbf{y})$. Based on this, some approximate assumption can be adopted, for example, $p(\mathbf{z}_{h,p}|[\mathbf{w}, v, \mathbf{r}], y_{h,p})$ are independently multivariate normal, witch is decided by 2 mean vectors $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1$ and 2 covariance matrix $\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1$ with dimension 18, which is related to parameters $[\mathbf{w}, v, \mathbf{r}]$.

For example, if a wide normal prior is endowed for $\mathbf{z}$,

$$\begin{aligned}
p(\mathbf{z}_{h,p}|[\mathbf{w}, v, \mathbf{r}], y_{h,p}) &\propto p(y_{h,p}|[\mathbf{w}, v, \mathbf{r}], \mathbf{z}_{h,p})\ p(\mathbf{z}_{h,p}|[\mathbf{w}, v, \mathbf{r}]) \\
&= p(y_{h,p}|[\mathbf{w}, v, \mathbf{r}], \mathbf{z}_{h,p})\ p(\mathbf{z}_{h,p}) \\
&\propto \left[\frac{1_{y_p=1}}{1 + e^{-[(\mathbf{w}^\top, v)\cdot(\mathbf{z}_{h,p}^\top, 1)^\top + r_h]}} + (1_{y_p=0} - \frac{1_{y_p=0}}{1 + e^{-[(\mathbf{w}^\top, v)\cdot(\mathbf{z}_{h,p}^\top, 1)^\top + r_h]}})\right] \cdot e^{-\epsilon\cdot\|\mathbf{z}_{h,p}\|_2^2}
\end{aligned}$$

Then use Laplace approximation to decide $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$ or $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1$. for different value of $y_{h,p}$.

The expectation procedure for partly-given data can be done by marginalizing the given part. Obviously this approximation is too coarse to implement, but some finer discrete model can be figured out, and is omitted here.

Then in this case, in each iteration, the "Expectation" part requires about linear time related to the number of rows with missing values and the "Maximization" procedure require at most an optimization mission under 20 dimensions. The time cost is acceptable.

References

[1]. UCI Machine Learning Repository: Heart disease Data Set

[2]. {Gelman.A, Carlin.J, Stern.H, Dynson.D, Vehtari.A, Rubin.D}, 2013, *Bayesian Data Analysis*, 3rd ed, CRC Press.

[3]. {Pelberg.R}, 2015, *Cardiac CT Angiography Manual*, 2nd ed, Springer-Verlag.