This document was created by students to fulfill a course requirement. Be aware of potential errors, and check with the original papers. The corresponding presentation file is at `http://www.slideshare.net/kaz_yos/multiple-imputation-joint-and-conditional-modeling-of-missing-data` and code example is at `http://rpubs.com/kaz_yos/mi-examples`.

1 Introduction

Missing data are an omnipresent problem that affects almost all real data set of substantial size. Complete case analysis (dropping any observations for which missing data exist) is often the default behavior of statistical packages, resulting in unconscious adoption of this option by many researchers. More sophisticated missing data handling methods are becoming more well-known and are being adopted in recent years. Such methods include inverse probability weighting, likelihood-based approaches, and multiple imputation.

Multiple imputation is becoming particularly popular, therefore, we will focus our discussion on this method. Roughly speaking there are two schools of thoughts in conducting multiple imputation. One is the joint distribution approach and the other is conditional distribution approach. Here we provide a brief overview of missing data methods, review in detail the two approaches to multiple imputation, and examine software packages implementing these methods.

2 Missing data mechanism

Consideration of the underlying mechanisms that gave rise to missing data is important in handling missing data appropriately [1]. Most importantly, assumptions have to be made as to whether the fact some variables are missing is related to the unobserved values themselves. The terms used to describe the assumptions are Missing Completely At Random (MCAR), Missing At Random (MAR), and Not Missing At Random (NMAR; sometimes also Missing Not At Random, MNAR) as coined by Rubin [2].

Let $\mathbf{Y}$ be the data matrix containing all data both observed and unobserved (missing) in reality. Following the biostatistics convention rows indexed by i are observations (often individuals) and the columns indexed by j are variables. Let $\mathbf{M}$ be the missing-data indicator matrix. That is, the entry M_{ij} is 1 if the corresponding entry in the $\mathbf{Y}$ matrix (Y_{ij}) is in fact missing in reality. It is 0 if the corresponding entry is observed. The conditional distribution of the missing indicator matrix $\mathbf{M}$ given the underlying data $\mathbf{Y}$, for example $f(\mathbf{M}|\mathbf{Y}, \phi)$ where ϕ characterizes the missing mechanism and is the unknown is used to formally define missing data mechanisms [1].

The MCAR missingness mechanism is defined as follows.

$$f(\mathbf{M}|\mathbf{Y},\phi) = f(\mathbf{M}|\phi) \ \ \forall \ \ \mathbf{Y},\phi$$

This expression implies that the missing data model that created the missing data does not actually depend on the true values of $\mathbf{Y}$ either observed or unobserved. Assuming MCAR is a rather strong assumption, however, it may be reasonable in some circumstances. This can occur by design, for example, conducting a detailed questionnaire among a random subset of the entire sample. If MCAR can be assumed, subsequent analyses are straightforward because observations that do not have missing variables (complete rows of $\mathbf{Y}$) are essentially a random sample of the entire observations (all rows of $\mathbf{Y}$). Notice the probability of having missingness is a constant related to ϕ for every observation (row) in the data set in this case.

A less restrictive mechanism, MAR is defined as follows. Let $\mathbf{Y}_{\text{obs}}$ be the observed components of $\mathbf{Y}$ and $\mathbf{Y}_{\text{mis}}$ be the missing components.

$$f(\mathbf{M}|\mathbf{Y},\phi) = f(\mathbf{M}|\mathbf{Y}_{\text{obs}},\phi) \ \ \forall \ \ \mathbf{Y}_{\text{mis}},\phi$$

This expression implies that the missing data model does depends on the true values of $\mathbf{Y}$, however, the dependence is only on the observed components. Assuming MAR is the essential part of most of the missing data handling methods. This mechanism means that within levels of observed variables, the probability of having missingness is a constant (i.e., MCAR within each subset) although across different subsets defined by observed variable values, the probability may differ. An example is a scale producing more missing weights when used on a soft surface [3].

The last mechanism, NMAR is defined as follows.

$$f(\mathbf{M}|\mathbf{Y},\phi) = f(\mathbf{M}|\mathbf{Y}_{\text{obs}},\mathbf{Y}_{\text{mis}},\phi) \ \ \forall \ \ \phi$$

This expression implies that the missing data model depends on the true values of $\mathbf{Y}$ both in the observed and missing parts. Notice this model is intractable in reality as we cannot condition on data that are not actually observed. In terms of the probability of having missingness, the probability varies even within levels of observed variables (as we need to define subsets using both observed and unobserved data). An example is a scale producing more missing weights for heavier objects (dependency on the missing weights themselves) [3].

3 Naive missing data methods and their problems

Complete case analysis is the default option for most statistical software packages and again as a result if is oftern unconsciously conducted by investigators [3].This is essentially sub-sampling of the full sample including observations

with missing variables. Only if MCAR is true, this sub-sample is a random sample of the full sample, which makes the sub-sample a random sample of the population, yielding unbiased estimates.

A minor improvement is available case analysis (also pairwise deletion), in which means are calculated from complete observations for each variables and pairwise co-variance are calculated using complete case analysis within each variable pair (complete cases can differ depending on variable pairs).

The indicator method extends the data set by a missingness indicator variable for each variable with missing data. In the original variable with missingness, missing data are replaced with 0. Then the analysis is proceeded using both the manipulated original variable and the corresponding indicator variable [3]. However, this has been demonstrated to result in severe bias even in the case of MCAR [4].

Single imputation is a technique in which missing data are replaced with plausible values. Examples include marginal mean imputation (or mode imputation for categorical variables), conditional mean imputation (regression imputation), and last observation carried forward (LOCF) for longitudinally observed variables [3]. All these approaches impute deterministically, thus, the imputed data have less variability than there should be. The marginal mean approach breaks the relationship between variables. The conditional mean approach is a slight improvement of the marginal mean approach as it leverages the observed data in the other variables. However, the imputed values are all conditional means and disregard variability of data beyond the systematic part of the regression model.

Stochastic regression imputation is a further improvement. Instead of imputing the conditional means, imputation is based on sampling from the conditional distribution given other observed variables. As seen in the summary table (adopted from [3]) regarding the assumptions and standard error estimates, this conditional distribution based approach is the best although the problem with the standard error underestimation persists. This method forms the foundation of multiple imputation.

	Mean	Regression parameters	Correlation	Standard error
Complete case	MCAR	MCAR	MCAR	Too large
Available case	MCAR	MCAR	MCAR	Complicated
Marginal mean imputation	MCAR			Too small
Conditional mean imputation	MAR	MAR		Too small
Conditional distribution imputation	MAR	MAR	MAR	Too small
LOCF				Too small
Indicator				Too small

4 Sophisticated methods other than multiple imputation

The best approach to missing data is actually prevention [5] although this is only possible if the study design and the data collection process is under control. If prevention is not possible, inverse probability weighting of the complete cases can be done. Inverse probability weighting of the complete cases is a possible method. Assuming MAR, the unobserved values in the incomplete cases can be represented by the observed values in the complete cases within levels of the observed variables. These complete cases are up-weighted by the inverse of the conditional probability of being observed. The analyses thereafter can be conducted as weighted analyses using these weighted complete cases that are representative of the entire cohort.

The direct likelihood and full information maximum likelihood (FIML) are methods that specify a model specific to the observed data[3], *i.e.*, direct modeling of parameters of intrests given data $f(Q|\mathbf{Y}_{\text{obs}})$ (left-handed side of the equation in the MI section). This allows skipping over missing data (no involvement of missing data in estimation process). Multiple imputation can be thought of as an extension of the likelihood-based approach. Drawing imputed values from the distribution estimated using the observed data is the added step.

5 Multiple imputation

As previously discussed the weakness of the single imputation methods is that the standard errors are too small as single imputation creates a complete data set, which eliminates true uncertainties related to having not observed some parts of the data set. Multiple imputation overcome this limitation by substituting the uncertainty of having missing data with uncertainty of having multiple imputed data sets [3].

Multiple imputation is a three step process. Mathematically, it can be described in the following expression, where Q is the quantity of interest we would like to estimate using the complete data set $\mathbf{Y}$.

$f(Q|\mathbf{Y}_{\text{obs}}) = \int f(Q|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}) f(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}) \mathrm{d}\mathbf{Y}_{\text{mis}}$

The left handed side is the posterior distribution of the quantity of interest given the observed data. The right handed side implies that we need a model for the quantity of interest given complete data and another model to relate the distribution of missing values to the observed data. The latter model is where MAR assumption is necessary because the missing values cannot be modeled in terms of only observed data in NMAR. The integration over the distribution of missing data given the observed data can be conducted using numerical simulation as follows[1].

First, imputed values are drawn from a distribution $f(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}})$ m times for each missing item, resulting in m

imputed complete data sets. Second, each imputed data set is analyzed as a complete data set without any missing data, corresponding to $f(Q|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$. Third, the results (both point estimate and variance) are pooled across imputed data sets using the following formula [6] , corresponding to the integration.

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m}\widehat{Q}_l$$
$$W = \frac{1}{m}\sum_{l=1}^{m}Var(\widehat{Q}_l)$$
$$B = \frac{1}{m-1}\sum_{i=1}^{n}(\widehat{Q}_l - \bar{Q})^2$$
$$T = W + B + \frac{1}{m}B$$

The pooled point estimate $\bar{Q}$ is just the mean of the m point estimates, whereas the pooled variance estimate consists of the mean within-imputation variance W and the variance across imputation data sets B. This comes from the variance decomposition formula $Var(Y) = E[Var(Y|X)] + Var(E[Y|X])$ (*i.e.*, $T = W + B$), with an added term $\frac{1}{m}B$ to account for the estimated nature of $\bar{Q}$. $\lambda = \frac{B+B/m}{T}$ can be interpreted as the proportion of the variation attributable to missing data[3]. Although there is an argument this formula is often conservative and in some special situations anti-conservative [7], it is still used as the standard method.

The two dominant approaches for implementing multiple imputation are the joint (multivariate normal; MVN) distribution approach and the conditional distribution approach [8]. Both approaches are reviewed here individually and then compared for their strengths and weaknesses.

5.1 Joint (multivariate normal) distribution multiple imputation

As the name suggests the joint distribution multiple imputation approach assumes a multivariate joint distribution of the data at hand, which is usually assumed to follow multivariate normal (MVN) distribution. Firstly, a MVN is assumed on the data, the parameters for this MVN are estimated, then imputed values are drawn from this estimated MVN of the data [8].

The actual imputation process proceed as follows [9]. A MVN is assumed on the entire data set $\mathbf{Y}$. This implies all marginal and conditional distributions are also assumed to be (multivariate) normal. The arguments for and against this assumption is discussed in the comparison part. Let $\mathbf{Y}_i$ be the i-th row ($i = 1, ..., n$) of the data matrix

$\mathbf{Y}_{n\times p}$. By the MVN assumption, $\mathbf{Y}_i \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *i.e.*, each observation follows an MVN with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The off diagonal elements of $\boldsymbol{\Sigma}$ dictates the pairwise correlation structure of columns (variables) in $\mathbf{Y}$. The likelihood function for the hypothetical complete data assuming independence between observations (rows) is:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{Y}) \propto \prod_{i=1}^{n} f(\mathbf{Y}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ where } f \text{ is MVN pdf}$$

Assuming MAR, the likelihood function conditional only on the observed data can be described as follows and should be parametrized by the same $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{Y}_{\text{obs}}) \propto \prod_{i=1}^{n} f(\mathbf{Y}_{i,\text{obs}}|\boldsymbol{\mu}_{i,\text{obs}}, \boldsymbol{\Sigma}_{i,\text{obs}})$$

$\mathbf{Y}_{i,\text{obs}}$ means the observed sub-vector of the full vector $\mathbf{Y}_i$ for the i-th observation. $\boldsymbol{\mu}_{i,\text{obs}}$ and $\boldsymbol{\Sigma}_{i,\text{obs}}$ are subset of the corresponding parameters necessary to describe the distribution from which the observed sub-vector $\mathbf{Y}_{i,\text{obs}}$ came from. Note depending on the specific i, $\boldsymbol{\mu}_{i,\text{obs}}$ and $\boldsymbol{\Sigma}_{i,\text{obs}}$ are different sub-vector and sub-matrix because each observation may have different elements of $\mathbf{Y}_i$ missing. Let $\mathbf{Y}_{i,\text{mis}}$ be the missing sub-vector for the i-th individual. Once $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ are obtained we can impute the sub-vector $\mathbf{Y}_{i,\text{mis}}$ by a random draw from the estimated MVN conditional on the observed sub-vector $\mathbf{Y}_{i,\text{obs}}$. This imputation process can be repeated m times for each observation to generate m imputed data sets, which can then be utilized for separate analyses and pooling. How to implement this computationally is another issue. The complicated nature of the observed data likelihood makes it difficult to maximize in classical methods. Roughly speaking there are two ways to implement this [9]. These are the Imputation-Posterior (IP) algorithm [10] and the Expectation Maximization (EM) algorithm [11].

5.1.1 Imputation-Posterior (IP) algorithm

The IP algorithm[9, 10] is a data augmentation method that enables drawing random simulations from the MVN parameter posterior. It is implemented as an iterative MCMC process with the following steps.

Imputation step: Draw imputed values $\widetilde{\mathbf{Y}}_{\text{mis}}$

$$\widetilde{\mathbf{Y}}_{\text{mis}}^{(t)} \sim f(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \widetilde{\boldsymbol{\mu}}^{(t)}, \widetilde{\boldsymbol{\Sigma}}^{(t)})$$

Posterior step: Draw updated parameters $\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}$

$$\widetilde{\boldsymbol{\mu}}^{(t+1)}, \widetilde{\boldsymbol{\Sigma}}^{(t+1)} \sim f(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{Y}_{\text{obs}}, \widetilde{\mathbf{Y}}_{\text{mis}}^{(t)})$$

Repeat until convergence

The I step start with some arbitrary initial values for $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\boldsymbol{\Sigma}}$ (EM estimates can be used). Then the missing data are drawn from the conditional distribution given the observed data and the estimated MVN. Using the resulting posterior distribution $f(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{Y}_{\text{obs}}, \widetilde{\mathbf{Y}}_{\text{mis}})$, the parameter estimates are updated. This is an MCMC process, thus theoretically, convergence to the stationary distributions of $\mathbf{Y}_{\text{mis}}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ occur after after infinite iterations. In practice, a sufficient initial burn-in period must be discarded from the chain to obtain the approximate stationary distribution, which must be assessed with empirically using diagnostic tools. Also the raw chain from an MCMC process violates independence because each step is dependent on the estimates from the previous step. Therefore, either thinning (keep only every r-th iteration) or use a separate MCMC chain for each one of m imputed data sets is necessary. In the implementation of `norm` package[12], only the last draw of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from the MCMC chain are used (last P step). $\mathbf{Y}_{\text{mis}}$ is then drawn (last I step) from the MVN determined by these last-step $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Thus, use of m separate MCMC chains are required to preserve uncertain arising from estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

5.1.2 Expectation Maximization (EM) algorithm

The EM algorithm[9, 11] is similar to the IP algorithm except that instead of stochastic draws from the entire posterior, the EM algorithm calculates posterior means deterministically– done explicitly in the expectation step. In the maximization (M) step corresponding to the P step, the posterior modes are used to update the estimated parameters $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\boldsymbol{\Sigma}}$.

Expectation step: Estimate

$$\widetilde{E}^{(t)}[\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \widetilde{\boldsymbol{\mu}}^{(t)}, \widetilde{\boldsymbol{\Sigma}}^{(t)}] \text{ Using } f(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \widetilde{\boldsymbol{\mu}}^{(t)}, \widetilde{\boldsymbol{\Sigma}}^{(t)})$$

Maximization step: Estimate

$$\widetilde{\boldsymbol{\mu}}^{(t+1)}, \widetilde{\boldsymbol{\Sigma}}^{(t+1)} \text{ Using } f(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{Y}_{\text{obs}}, \widetilde{E}^{(t)}[\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \widetilde{\boldsymbol{\mu}}^{(t)}, \widetilde{\boldsymbol{\Sigma}}^{(t)}])$$

Repeat until convergence

What we obtain after convergence is the point estimates of $E[\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \boldsymbol{\mu}, \boldsymbol{\Sigma}]$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$. Multiple imputed data sets can then be created by using point estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. This ignores the estimation uncertainty because only

point estimates are obtained and used. The `amelia` package[9, 13] preserves the estimation uncertainty by using m bootstrap re-samples and m EM processes to obtain m different point estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

5.2 Conditional distribution multiple imputation

Given the observed and imputed values in the data set, conditional MI aims to model the conditional distributions of partially observed variables (i.e variables with incomplete data) across observations. Assume $\mathbf{Y} = (Y_1, Y_2, ..., Y_k)$ is a vector of k random variables with a multivariate joint distribution fully specified by parameter θ. Further assume that each variable has n observations, and has missing-ness. The standard procedure for creating multiple imputations [14] is as follows:

(1) Calculate the posterior distribution of θ conditional on the observed variables

(2) Draw a value θ^* for the posterior

(3) Draw a value y^* from the conditional posterior distribution of the missing values given the observed values and $\theta = \theta^*$

For multivariate $\mathbf{Y}$ it is often difficult to get the multivariate distribution of θ. An option is to obtain posterior distributions of k separate θ by iteratively sampling from the conditional distributions of each variable. The conditional distribution are of the form

$$
\begin{aligned}
&P(Y_1|Y_{-1}, \theta_1) \\
&P(Y_1|Y_{-2}, \theta_2) \\
&\vdots \\
&P(Y_k|Y_{-k}, \theta_k)
\end{aligned}
$$

where the conditional distribution of the k-th variable is a function of θ_k and the other variables in the data set. Buuren, *et al* [14], provides more detailed insight for the t-th iteration of the method as follows:

$$\theta_1^{*(t)} \sim P(\theta_1|y_1^{\text{obs}}, y_2^{t-1}, ..., y_k^{(t-1)})$$

$$y_1^{(t)} \sim P(y_1^{\text{mis}}|y_1^{\text{obs}}, y_2^{t-1}, ..., y_k^{(t-1)}, \theta_1^{*(t)})$$

$$\vdots$$

$$\theta_k^{*(t)} \sim P(\theta_k|y_k^{\text{obs}}, y_1^t, y_2^t, ..., y_{k-1}^{(t)})$$

$$y_k^{(t)} \sim P(y_k^{\text{mis}}|y_k^{\text{obs}}, y_1^t, ..., y_{k-1}^{(t)}, \theta_k^{*(t)})$$

Here it is readily seen that no information on the missing values is used to sample values for any θ. This algorithm can be further explained as two nested loop [8]. The first loop iterates through the variables with missing-ness and imputes values for the missing entries. In order to determine the imputed values for a variable, first the variable specific theta is sampled from the posterior distribution conditional on the observed values within that variable strata, and the t-th iteration of the other imputed (or not) variables. Then the imputed values are sampled from the posterior distribution of the missing values; the conditioning is similar to that of the variable specific θ but also includes the θ. This first loop is nested inside of a second loop that repeats the process until there appears to be convergence.

5.3 Relative strengths and weaknesses

5.3.1 Violation of MVN

The joint MVN MI is based on a sound theoretical foundation assuming MVN-distributed data. However, in practice data can have varying degrees of departure from the MVN. As random samples from an MVN, the imputed values from the joint MVN MI follow normal distributions, meaning they are not bounded and are continuous. This can be awkward when the variables being imputed are bounded and/or discrete. For example, when "male" variable is coded 0,1, imputed values can be 0.3 or even 1.7. Rounding is one common approach, *e.g.*, 0.3 can be rounded to 0 (female) and 1.7 can be rounded to 1 (male). This will make the imputed values look "natural", however, there is an argument that this type of rounding may actually introduce bias[15]. One approach is to leave these in-feasible values as is because imputed values themselves are not of importance in multiple imputation. It is about estimating the parameters of interest with less bias due to missing data.

Another approach is to probabilistically assign feasible values[8, 16]. For example, out-of-range values are rounded to the nearest feasible values, and the imputed values within the possible range of values are subject to probabilistic draw, *i.e.*, 0.3 is replaced by the result of $Bernoulli(0.3)$ for the binary case. The joint MVN MI has been shown to be robust to departure from normality when imputing non-normal continuous variables[17].

It may be beneficial to transform non-normal continuous variables so that it is more amenable to the MVN, impute, and then transform back to the original form[3]. There are various extensions to the joint modeling approach to accommodate both continuous and categorical data[3]. The conditional MI approach is more flexible in this regard as specifying different conditional model (logistic, multinomial logistic, Poisson, etc) for each type of variables is possible. A simulation study found that the conditional approach was accurate in terms of imputed values and estimates of interest whenever the data included categorical variables [8].

5.3.2 Theoretical problem with conditional distribution MI

Although the conditional MI samples from the posterior distribution of the variable specific missing values, it does not use this information to sample the variable specific θ. This is a direct departure from the Gibbs sampler and the theoretical underpinnings have been largely critiqued [8]. Furthermore, when modeling k variables with missingness the variable specific parameters are likely to be dependent on each other, or over parameterized, and do not share a k dimensional joint distribution [14]. Therefore, unlike the Gibbs sampler, conditional MI is not always theoretically guaranteed to converge to the correct posterior distribution (or at all), and is possible to converge to different posterior distributions varying with the initial conditions. However, simulations appear to show conditional MI is robust against over parameterization, and the method produces reasonable imputations– justifying its use in practice. Continuing, conditional MI is relatively powerful and convenient. Combined with the added benefit of not having to specify a joint model, sampling from the variable specific distribution adds increased flexibility in model building [14]. In particular, transformations of variables can be incorporated without any added complexity.

5.3.3 Consideration of high dimensional data

Missing data are also problematic in the high-dimensional setting. Even if each variable has a small proportion of missing data, over many variables, many observations are likely to have at least some missing data. This makes handling of missing data more important. MI encounter difficulties with high-dimensional data where the number of variables are larger than the number of observations[18] ($p > n$). Estimation of distributional parameters are difficult in such settings for the joint MVN MI as the $p \times p$ variance-co-variance matrix is huge with numerous parameters. The conditional approach encounters difficulty of running regression model with a column-rank deficient design matrix.

To reduce the number of parameters for the $p \times p$ co-variance matrix in the joint MI, a hierarchical co-variance model has been examined[19]. In this method, a structure with parsimonious parameterization is introduced for the

co-variance matrix, and these relatively few parameters are estimated, instead of trying to estimate all individual parameters in the co-variance matrix. The regular conditional MI can only handle $p < n$ data, however, use of regularization technique incorporated into the conditional MI process has been examine in the high-dimensional data with some success[18]. Among these methods, the Bayesian LASSO regression performed the best. In one simulation study involving high-dimensional proteomics data with artificial missingness[20], the authors compared both the joint and the conditional approach. They approached the $p > n$ problem by randomly binning the co-variates into smaller subsets that are individually amenable to regular MI software. They found the joint approach performed somewhat better than the conditional approach. MI in the presence of high-dimensional data is still an active area of methodological research.

5.4 Software review

Here we summarize publicly available `R` packages for their distributional specification and implementation.

Package	Distribution	Algorithm	Note
`amelia`[16, 21]	Joint (MVN)	EM with bootstrap	Works with `Zelig`; Fast due to EM
`norm`[12]	Joint (MVN)	IP (MCMC)/EM	Most classical method
`cat`[22]	Joint (log-linear model)	MCMC	For categorical variables
`mix`[23]	Joint (general location model)	MCMC	For mixed categorical/continuous
`pan`[24]	? (linear mixed model)	MCMC?	For clustered data
`mice`[25, 26]	Conditional	Original FCS	Good user interface
`mi`[27]	Conditional	Bayesian regression	
`BaBooN`[28]	Conditional	Bayesian bootstrap	
`smcfcs`[29, 29]	Conditional	Modified FCS	Addresses theoretical problem with FCS
`mitools`[30]	-	-	Utility for Rubin's rule

6 Conclusion

In this paper, we reviewed the principles of missing data handling, focusing on multiple imputation and its popular algorithms. The two popular approaches are the joint approach and the conditional approach. The joint approach has more sound theoretical background as the joint distribution is directly estimated, whereas the conditional approach implies the joint distribution, which may or may not exist, by multiple conditional uni-variate distributions. The joint approach usually resort to the MVN, which may not closely represent the data. With the conditional approach, it is easy to incorporate different variable types as each variable is modeled one at a time. In simulation studies, they often perform similarly although some differences may exist in specific settings. Understanding the algorithms implemented in software and their assumption is important in making better choices.

References

[1] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. Hoboken, N.J: Wiley-Interscience, 2 edition ed., Sept. 2002.

[2] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, pp. 581–592, Dec. 1976.

[3] S. van Buuren, *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman and Hall/CRC, 1 edition ed., Mar. 2012.

[4] S. Greenland and W. D. Finkle, "A critical look at methods for handling missing covariates in epidemiologic regression analyses," *American Journal of Epidemiology*, vol. 142, pp. 1255–1264, Dec. 1995.

[5] R. J. Little, R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, J. D. Neaton, A. Rotnitzky, D. Scharfstein, W. J. Shih, J. P. Siegel, and H. Stern, "The Prevention and Treatment of Missing Data in Clinical Trials," *New England Journal of Medicine*, vol. 367, pp. 1355–1360, Oct. 2012.

[6] D. B. Rubin, "An Overview of Multiple Imputation," in *In Proceedings of the Survey Research Section, American Statistical Association*, pp. 79–84, 1988.

[7] J. M. Robins and N. Wang, "Inference for imputation estimators," *Biometrika*, vol. 87, pp. 113–124, Mar. 2000.

[8] J. Kropko, B. Goodrich, A. Gelman, and J. Hill, "Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches," *Political Analysis*, p. mpu007, Apr. 2014.

[9] G. King, J. Honaker, A. Joseph, and K. Scheve, "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review*, vol. 95, pp. 49–69, 2001.

[10] M. A. Tanner and W. H. Wong, "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 528–540, 1987.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[12] J. Schafer, "norm: Analysis of multivariate normal datasets with missing values," Feb. 2013.

[13] J. Honaker and G. King, "What to do About Missing Values in Time Series Cross-Section Data," *American Journal of Political Science*, vol. 54, no. 3, pp. 561–581, 2010.

[14] S. Van Buuren, J. P. L. Brand, K. Groothuis-Oudshoorn, and D. B. Rubin, "Fully conditional specification in multivariate imputation.," *Journal of Statistical Computation and Simulation*, vol. 76, no. 12, pp. 1049–1064, 2006.

[15] L. S. R. Horton N. J, "A Potential for Bias When Rounding in Multiple Imputation," *The American Statistician*, vol. 57, no. November, pp. 229–232, 2003.

[16] J. Honaker, G. King, and M. Blackwell, "Amelia II: A Program for Missing Data | Honaker | Journal of Statistical Software," *Journal of Statistical Software*, vol. 45, no. 7, 2011.

[17] H. Demirtas, S. A. Freels, and R. M. Yucel, "Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment," *Journal of Statistical Computation and Simulation*, vol. 78, pp. 69–84, Feb. 2008.

[18] Y. Zhao and Q. Long, "Multiple imputation in the presence of high-dimensional data," *Statistical Methods in Medical Research*, p. 0962280213511027, Nov. 2013.

[19] R. He and T. Belin, "Multiple imputation for high-dimensional mixed incomplete continuous and binary data," *Statistics in Medicine*, vol. 33, pp. 2251–2262, June 2014.

[20] X. Yin, D. Levy, C. Willinger, A. Adourian, and M. G. Larson, "Multiple imputation and analysis for high-dimensional incomplete proteomics data," *Statistics in Medicine*, Nov. 2015.

[21] J. Honaker, G. King, and M. Blackwell, "Amelia: Amelia II: A Program for Missing Data," Nov. 2014.

[22] J. Schafer, "cat: Analysis of categorical-variable datasets with missing values," Oct. 2012.

[23] J. Schafer, "mix: Estimation/Multiple Imputation for Mixed Categorical and Continuous Data," June 2015.

[24] J. Schafer, "pan: Multiple Imputation for Multivariate Panel or Clustered Data," Feb. 2015.

[25] S. v. Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011.

[26] S. v. Buuren, K. Groothuis-Oudshoorn, A. Robitzsch, G. Vink, L. Doove, and S. Jolani, "mice: Multivariate Imputation by Chained Equations," June 2014.

[27] A. Gelman, J. Hill, Y.-S. Su, M. Yajima, M. Pittau, B. Goodrich, Y. Si, and J. Kropko, "mi: Missing Data Imputation and Model Checking," Apr. 2015.

[28] F. Meinfelder and T. Schnapp, "BaBooN: Bayesian Bootstrap Predictive Mean Matching - Multiple and Single Imputation for Discrete Data," June 2015.

[29] J. Bartlett, "smcfcs: Multiple Imputation of Covariates by Substantive Model Compatible Fully Conditional Specification," Aug. 2015.

[30] T. Lumley, "mitools: Tools for multiple imputation of missing data," Sept. 2014.