Missing Data Replacement in a Complex Multivariate Context

Ryan M. Barnett and Clayton V. Deutsch

While a challenge in the multiGaussian context[1], unequal sampling poses a much larger issue for the geostatistical analysis of complex multivariate data. Relative to reasonably multiGaussian data, complex multivariate data will more frequently motivate the use multivariate transforms to facilitate Gaussian based modeling. As these transforms may only use observations that are equally sampled, either data replacement or data elimination must be considered for the incomplete observations. Considering information loss and the potential for bias, data replacement rather than elimination is advocated by missing data methodologists. A companion paper[1] provided an overview of data replacement motivation, methodology, and common techniques, before demonstrating excellent results with a data replacement method for multiGaussian data. The following work will build on this previous study, proposing a potential method for data replacement in a complex multivariate context. A brief conceptual overview of this Gibbs Sampling Data Imputation (GSDI) method is provided, along with a synthetic case study for demonstration. While promising results are seen with GSDI, this exploratory work leaves room for a great deal of future research and improvement, as will be discussed throughout.

Introduction

A large variety of techniques are available for transforming complex multivariate data to be suitable for Gaussian based geostatistical modeling, including logratios [1,11], SCT [1,10], MSNT [4], and PPMT [5]. As these transforms may only be executed on equally sampled observations, incomplete observations must be eliminated from the modeling framework, or have their missing values replaced (imputed[2]). Obvious disadvantages exist for data elimination, as it may drastically reduce the information that is available for global statistics and local conditioning. As stated by methodologists in the field of missing data [8,13], this also makes strong assumptions of why data is missing in the first place, and could introduce a strong bias. A data replacement method that produces biased results, or does not reflect inherent uncertainty in the imputed values is no less dangerous.

With this in mind, the field of missing data theory [8,13] has focused on developing methods for unbiased data replacement that provide accurate uncertainty distributions for the imputed values. Following an overview of these data replacement methods by this study's companion paper[1] [2], Multiple Imputation was selected for geostatistical applications due to its natural suitability within spatial modeling frameworks. The basic notion of Multiple Imputation in a geostatistical modeling context, is to construct representative conditional distributions at every location (and for every variable) requiring replacement. Simulating based on these conditional distributions will then generate multiple realizations of the data, which may then be used for conditioning geostatistical simulation.

As the complexity of geologic data necessitates the use of differing modeling frameworks, so too does it motivate the use of varying data replacement techniques. If the data is not reasonably multivariate Gaussian (multiGaussian) following the normal score transform [1,7], the Bayesian Updating Data Imputation (BUDI) replacement method developed in the companion paper [3,12] will not be appropriately applied. The following study will introduce a data replacement technique for the complex multivariate setting.

When considering the imputation of missing data for a regionalized variable[3], its conditional distribution of potential replacement values may be infered based on (i) the values of spatially correlated samples of the same variable and (ii) the values of colocated and correlated secondary variables. Constructing conditional distributions from these two sources is well defined in the multiGaussian setting

[1] Barnett, R., & Deutsch, C. (2012). Data replacement in a multiGaussian context. *CCG Annual Report 14.* Paper 112.

[2] Rather than the word replace, missing data methodologists prefer the word impute. Replace and impute will be used interchangeably throughout this paper. From the Oxford English Dictionary: "Impute - to assign a value to something by inference based on the value of the products or processes to which it contributes".

[3] Working for now under the Markov model where colocated secondary data screens the effect of spatially correlated secondary data [7]

with methods such as cokriging [7] and Bayesian Updating [3,12]. Unfortunately, its construction is not nearly so straightforward in the complex multivariate setting. A potential method is presented for constructing and sampling a variable's conditional distribution based only on the colocated and correlated secondary variables. This leaves to future study, the very important task of integrating spatially related values of the same variable into the construction of these conditional distributions.

The conditional distributions will be determined using kernel density estimation (KDE), which are iteratively sampled within a Gibbs sampling style framework for multiple imputation. Following a brief overview of the essential Gibbs sampling theory, this closely related data replacement algorithm will be quickly developed. The accuracy of its imputed results will be demonstrated in a synthetic case study, along with the potential value it adds to subsequent geostatistical modeling. Parameters for the associated `gsdi` program are presented in the appendix.

Gibbs Sampling

The Gibbs sampler [6,9] is a widely applied method that allows for random variables of a multivariate distribution to be simulated without requiring the joint or marginal densities. Following the theory given by Casella and George [6], suppose one is interested in sampling from a bivariate distribution composed of Y^1 and Y^2 random variables. The Gibbs sampler iteratively draws random values for each variable, forming what is called the Gibbs sequence as seen in Equation 1. Here the subscript represents the i^{th} Gibbs sample.

$$Y_0^1, Y_0^2,\ Y_1^1, Y_1^2, \ldots, Y_n^1, Y_n^2 \tag{1}$$

After specifying the starting value as $Y_0^1 = y_0^1$, the Gibbs sampler iteratively draws the remaining values from conditional probability distributions, that are formed based only on the value of the previously sampled random value. In the bivariate case of Y^1 and Y^2, this conditional sampling is represented by Equation 2.

$$\begin{aligned} Y_i^1 &\sim f\left(y^1 \mid Y_{i-1}^2 = y_{i-1}^2\right) \\ Y_i^2 &\sim f\left(y^2 \mid Y_i^1 = y_i^1\right) \end{aligned} \tag{2}$$

Given a long enough sequence, the Gibbs sampled distributions of Y^1 and Y^2 will become statistically representative of the True Y^1 and Y^2 distributions [6]. The generalized multivariate representation of Equations 1 and 2 for k number of variables is given by Equations 3 and 4 respectively. It is truly remarkable that such a simple algorithm possesses these remarkable convergence properties. For additional Gibbs sampler background and convergence proofs, interested readers are referred to excellent sources including the Geman and Geman paper that led to its widespread use [6,9].

$$\begin{matrix} Y_0^1, Y_0^2, \ldots, Y_0^k \\ Y_1^1, Y_1^2, \ldots, Y_1^k \\ \vdots \\ Y_n^1, Y_n^2, \ldots, Y_n^k \end{matrix} \tag{3}$$

$$\begin{aligned} Y_i^1 &\sim f\left(y^1 \mid Y_{i-1}^k = y_{i-1}^k, Y_{i-1}^{k-1} = y_{i-1}^{k-1}, \ldots, Y_{i-1}^2 = y_{i-1}^2\right) \\ Y_i^2 &\sim f\left(y^2 \mid Y_i^1 = y_i^1, Y_{i-1}^k = y_{i-1}^k, \ldots, Y_{i-1}^3 = y_{i-1}^3\right) \\ &\vdots \\ Y_i^k &\sim f\left(y^k \mid Y_i^{k-1} = y_i^{k-1}, Y_i^{k-2} = y_i^{k-2}, \ldots, Y_i^1 = y_i^1\right) \end{aligned} \tag{4}$$

Gibbs Data Imputation

Returning to the task of missing data replacement, it is clearly attractive to impute values with an algorithm such as the Gibbs sampler, since it reproduces the underlying multivariate distribution. Particularly in cases where little information is available to condition the imputation of multiple variables,

it would be ideal for realizations to converge towards the stationary joint distribution of the data. Modifications to the original Gibbs sampling algorithm will be necessary, however, since observations with missing variables to be imputed (e.g. Gibbs sampled) have additional information that should not be left to random simulation. Applying the Gibbs sampler in its original form, where all variables are iteratively sampled according to Equation 3, would fail to reflect that certain variables at each location will have True values already sampled.

Consider the imputation of a single observation, where p number of variables have been sampled from the total k number of variables. Treating these sampled (fixed) variables as $Y^1,\ldots,Y^p$, Gibbs sampling represented by Equations 3 and 4 above, are modified to Gibbs Sampling Data Imputation (GSDI) in Equation 5 and 6 below. Here, each Gibbs sampling sequence is effectively restricted to the $p+1$ through k variables, since the remaining variables are fixed on every iteration to their True sampled values.

$$\begin{array}{c} y_0^1, y_0^2, \ldots, y_0^p, Y_0^{p+1}, Y_0^{p+2}, \ldots, Y_0^k \\ y_1^1, y_1^2, \ldots, y_1^p, Y_1^{p+1}, Y_1^{p+2}, \ldots, Y_1^k \\ \vdots \\ y_n^1, y_n^2, \ldots, y_n^p, Y_n^{p+1}, Y_n^{p+2}, \ldots, Y_n^k \end{array} \tag{5}$$

$$\begin{array}{c} Y^1 = y^1 \\ Y^2 = y^2 \\ \vdots \\ Y^p = y^p \\ Y_i^{p+1} \sim f\left(y^{p+1} \mid Y^p = y^p, Y^{p-1} = y^{p-1}, \ldots, Y_{i-1}^{p+2} = y_{i-1}^{p+2}\right) \\ Y_i^{p+2} \sim f\left(y^{p+2} \mid Y_i^{p+1} = y_i^{p+1}, Y^p = y^p, \ldots, Y_{i-1}^{p+3} = y_{i-1}^{p+3}\right) \\ \vdots \\ Y_i^k \sim f\left(y^k \mid Y_i^{k-1} = y_i^{k-1}, Y_i^{k-2} = y_i^{k-2}, \ldots, Y^1 = y^1\right) \end{array} \tag{6}$$

While this does not form an actual Gibbs sampling sequence, the imputed variables will now have their sampling restricted to a multivariate space that is likely to be far more representative of the True missing value (since the space is restricted based on their colocated and correlated secondary variables). In the most extreme case of this restricted multivariate space, when p is equal to k-1, only a single conditional distribution will be formed and repeatedly sampled. Everything else being equal, the resulting uncertainty distribution will be relatively narrow since this single conditional distribution is so well conditioned. Moving to the case where p is equal to k-2, the GSDI will sample from a restricted plane within the greater multivariate space. In the most extreme case of unrestricted space, p is equal to 1, meaning that the GSDI will be free to sample from the (k-1)-variate joint distribution. Although a high degree of uncertainty is likely to be seen in the resultant imputed realizations, no bias should result from the GSDI in these poorly conditioned cases so long as the provided multivariate distribution honors the standard first and second order stationarity assumptions [7].

Kernel Conditional Distributions

In order to execute GSDI on complex geologic data, a method will be required for inferring the non-parametric conditional distributions in Equation 6. A new but promising non-parametric Gibbs sampling algorithm `gmv_sample` [2] will be adapted for this purpose, where the conditional distributions are formed based on kernel density estimation (KDE). As KDE is only used for estimating discretizations of the iterating conditional distribution vectors, rather than a full multivariate grid, this algorithm allows for the efficient Gibbs sampling of massively multivariate data. Reader's are referred to the original paper for additional details of the algorithm [2], as only essential points that will impact the implementation and

results of this GSDS adaption are discussed. Following the original paper[2], important considerations for this data imputation application will include:

- It is recommended that the algorithm be executed on normal scores of the original data, due to a variety of reasons including standardized units and marginal Gaussian form. Gibbs samples may then be back-transformed to original space.
- Results will be very sensitive to user input parameters including kernel bandwidth and correlation (option for it to be orthogonal or based on the normal score data covariance matrix). Testing various ranges of the parameters and selecting the results that best match user knowledge of the data is advocated for now.
- A fairly rudimentary method is currently used for extracting observations from the Gibbs sequence. More robust starting locations and sampling extraction methods will be tested in the future which could lead to superior final convergence results.
- Overall, excellent reproduction of the marginal and joint densities were observed with the algorithm based on initial testing. The largest concern, however, was that a variance inflation (~10%) was observed in the Gibbs sampling results relative to the original data. The cause has not yet been determined.

As this Gibbs sampling algorithm will form the engine of GSDI, all of the shortcomings listed above are very likely to have a negative impact on data replacement results. Likewise, future improvements that are anticipated in the `gmv_sample` algorithm should have a positive effect in this application.

Case Study

The GSDI methodology will be demonstrated on a synthetic case study using the `gsdi` program (appendix). Exhaustive True synthetic models are first generated, creating five variables of varying spatial continuity, from which 283 homotopic observations are sampled (Figure 1). From these 283 samples, 30 observations of each variable are independently and randomly selected for removal (Figure 2). This results in a dataset of 172 complete observations, with 111 that are incomplete to varying degrees. The scatterplots of all five variables following this data removal are displayed for these sampled observations in Figure 3, before and after normal score transformation. Observe that despite marginal Gaussianity, this normal score data possesses strong non-linear features that would prevent the correct application of multiGaussian modeling (e.g. co-simulation [7]), or multiGaussian data replacement (e.g. BUDI [3]).

The `gsdi` program is applied next to the normal score data, forming 100 realizations of complete normal score data. Figure 4 displays scatterplots of the normal score data, with the 100 realizations of imputed values overlain. Note that the realizations forming 'stripes' across these scatterplots are the observations where only a single value must be imputed (the most common case). Since the remaining four variables are sampled and therefore fixed, only a single conditional distribution is constructed for sampling. In the remaining cases, where more than one value is missing from an observation, the Gibbs sequence is free to explore a multivariate space that is free in at least two dimensions, forming 'clouds' of imputed values. To provide a better indication of whether these Gibbs sequences honor the underlying multivariate distribution, scatterplots with 1000 imputed realizations are also displayed in Figure 4. Plots with this greater number of realizations suggests that the Gibbs sequences do honor the apparent concave hull of the underlying joint distribution.

To provide additional insight into the accuracy of the conditional distributions formed by the KDE method, arbitrary locations are chosen for displaying the conditional distributions, with their associated True values overlain. Recall that these True values were previously removed in the case of the variables/locations now requiring imputation, as seen in Figure 2. Figure 5 displays the conditional distribution where only one value is missing, while Figures 6 and 7 have two and three values missing respectively. Observe that the True values fall directly on the conditional distribution vector in the case of one missing value (Figure 5), since these are all known samples to the algorithm (excepting the one variable being imputed). It is also encouraging to see in this figure, that the relatively tight conditional distribution appears to nearly converge on the True removed value. This contrasts with Figures 6 and 7, where the location of the conditional distribution vectors do not always coincide with the True values since they are not entirely fixed based on sampled values. Likewise, their generated conditional

distributions are not always representative of the True removed value. Keep in mind, however, that unlike in Figure 5, these conditional distributions will change on every iteration, since the Gibbs sampling is exploring more than one dimension of the multivariate space.

Following back-transformation, the imputed realizations may be compared with the removed True values to judge whether the results are unbiased and accurate (Figure 8). While the mean of the realizations is relatively unbiased, it raises some concern that there is an overall inflation in variance. As this was a noted issue with the `gmv_sample` algorithm (anticipated to be resolved), however, it is attributed for now to the construction of these kernel based conditional distributions rather that the GSDI concept itself. Another observation from Figure 8, is the relatively weak correlation between the True removed data and the imputed realizations. One would hope to see greater correlation, indicating greater overall accuracy in the realizations. This is not necessarily surprising, however, as when comparing these results to their superior BUDI equivalent [3], one must consider that no information has yet been integrated for spatially correlated values of the same variable. Further, BUDI works with data that is far more behaved, and therefore is likely to represent the best potential results that could be hoped for in this complex multivariate setting.

Based on the above observations, the use of these data realizations in a geostatistical modeling framework are not necessarily expected to lead to measurable accuracy gains (as compared to geostatistical modeling with data elimination). Nevertheless, identical geostatistical modeling workflows will be executed with and without the use of data replacement. That is to say, one workflow will use the data replacement realizations attained above, while the other will eliminate the incomplete samples (necessary so that a multivariate transform may be applied). Comparing the resultant geostatistical models with and without data replacement to the True model from which the samples were originally drawn (Figure 1) will then provide an indication of value gained from the GSDI replacement.

Dealing first with details of the modeling workflow, the 100 data realizations are individually PPMT [5] transformed to form independent Gaussian variables. These 100 data realizations are used to condition an SGSIM [7] based simulation of 100 models, which are then back-transformed. An identical modeling workflow is then executed using a single dataset, where incomplete observations have been eliminated so that PPMT may be applied.

Next, to ascertain the value gained in terms of local accuracy, E-Type estimates are formed from 100 realizations of the two modeling workflows and compared with the True model. This comparison displayed uniformly better results for the workflow involving data replacement, which is summarized by Table 1 according to the MSE and Covariance improvement (as compared to the True model). Though this modeling improvement is not as substantial as the BUDI case study [2], they remain quite significant when considering the modest accuracy of the imputed data realizations according to Figure 8.

Table 1: Improvement in the MSE and covariance of E-Type estimates vs. the True model (using data replacement realizations rather than data elimination).

Variable	% Improvement	
	Mean Squared Error	Covariance
1	11.20	13.96
2	3.16	13.71
3	2.69	18.71
4	4.26	57.75
5	2.52	31.03

Conclusion

The replacement of missing data is important for multivariate geostatistical modeling in a complex setting. A companion paper selected Multiple Imputation as a suitable general method for data replacement in geostatistical frameworks. Multiple Imputation was then adapted to complex data in this study, using a modified version of the Gibbs Sampling algorithm. Using a synthetic case study for demonstration, reasonable accuracy was seen in the uncertainty distributions of the imputed realizations. In spite of these modest results, the imputed data was demonstrated to greatly improve geostatistical

modeling accuracy, as compared to a parallel workflow that used data elimination. As this initial algorithm is only based on colocated and correlated secondary data, a great deal of improvement is expected to result from the future integration of spatially correlated values of the same variable. Additionally, the non-parametric Gibbs sampling algorithm at the core of this data imputation method has several identified avenues for implementation improvement. Improvements to the algorithm will likely impact this data replacement application in a large and positive manner. Parameters for the `gsdi` program are provided in the appendix.

References

1 Barnett, R. (2011). *Guidebook on Multivariate Geostatistical Tools.* Edmonton, Alberta: Centre for Computational Geostatistics.

2 Barnett, R., & Deutsch, C. (2012). Gibbs Sampler for Non-Parametric Multivariate Sampling. *CCG Annual Report 14*, Paper 102.

3 Barnett, R., & Deutsch, C. (2012). Missing Data Replacement in a MultiGaussian Context. *CCG Annual Report 14*, Paper 112.

4 Barnett, R., & Deutsch, C. (2012). MSNT Advances and Case Studies. *CCG Annual Report 14*, Paper 101.

5 Barnett, R., Manchuk, J., & Deutsch, C. (2012). Projection Pursuit Multivariate Transform. *CCG Annual Report 14*, Paper 103.

6 Casella, G., & George, E. (1992). Explaining the Gibbs Sampler. *Journal of the American Statistician, vol.46*, pp.167-174.

7 Deutsch, C., & Journel, A. (1998). *GSLIB: A geostatistical software library and user's guide, second edition.* Oxford University Press.

8 Enders, C. (2010). *Applied Missing Data Analysis.* New York: Guilford Pess.

9 Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.6*, pp.721-741.

10 Leuangthong, O., & Deutsch, C. (2003). Stepwise conditional transformation for simulation of multiple variables. *Mathematical Geology, vol.35, no.2*, pp.155-173.

11 Pawlowsky-Glawh V, E. J. (2006). Compositional data and their analysis: an introduction. In A. M.-F.-G. Buccianti, *Compositional data analysis in the geosciences: from theory to practice.* (pp. vol.264, pp.1-10). London Geological Society Special Publication.

12 Ren, W. (2007). *Bayesian Updating for Geostatistical Analysis, PhD. Thesis.* Edmonton: University of Alberta.

13 Rubin, D., & Little, R. (2002). *Statistical analysis with missing data.* Hoboken, N.J.: Wiley.

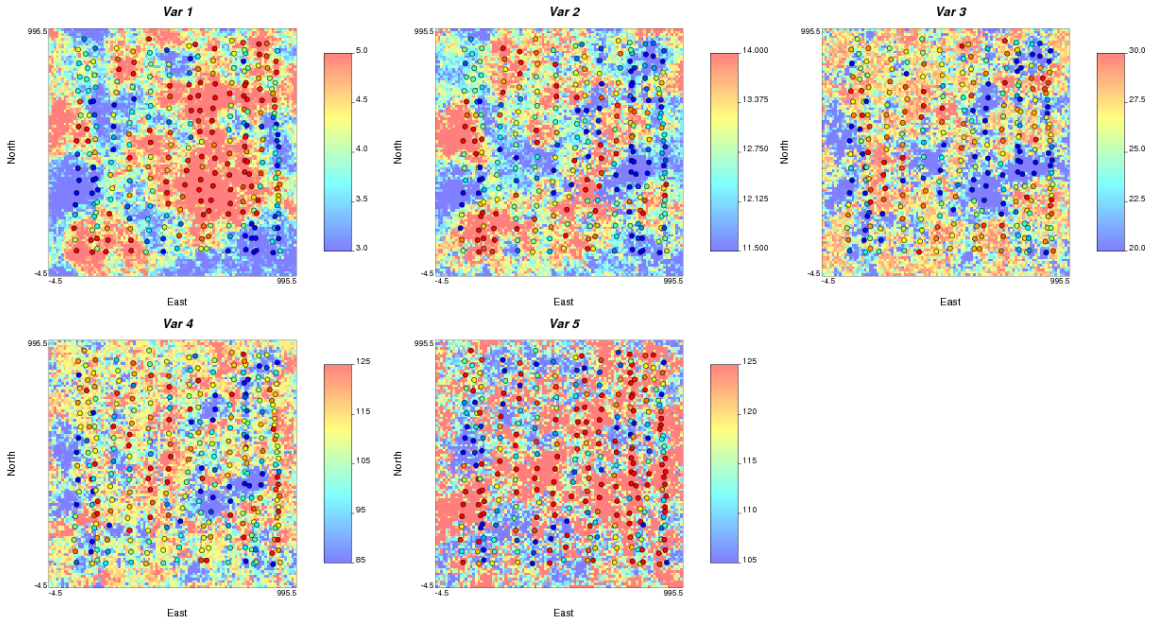


Figure 1: Synthetic true models, with circles indicating the locations of 283 homotopic samples.

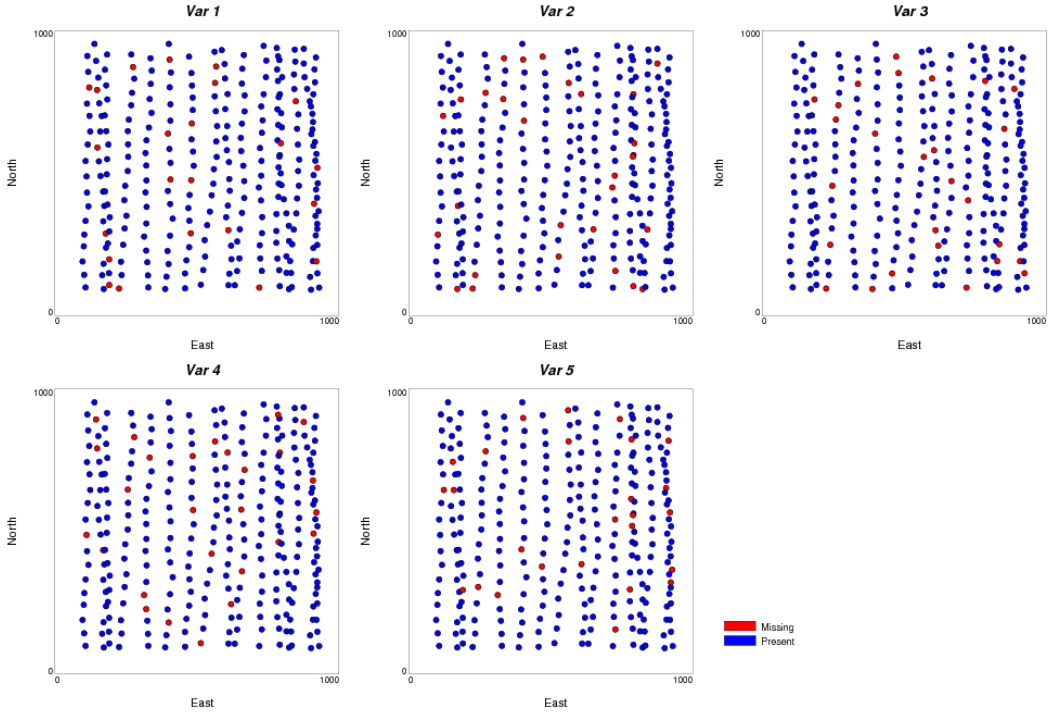


Figure 2: Locations of randomly removed observations for each variable.

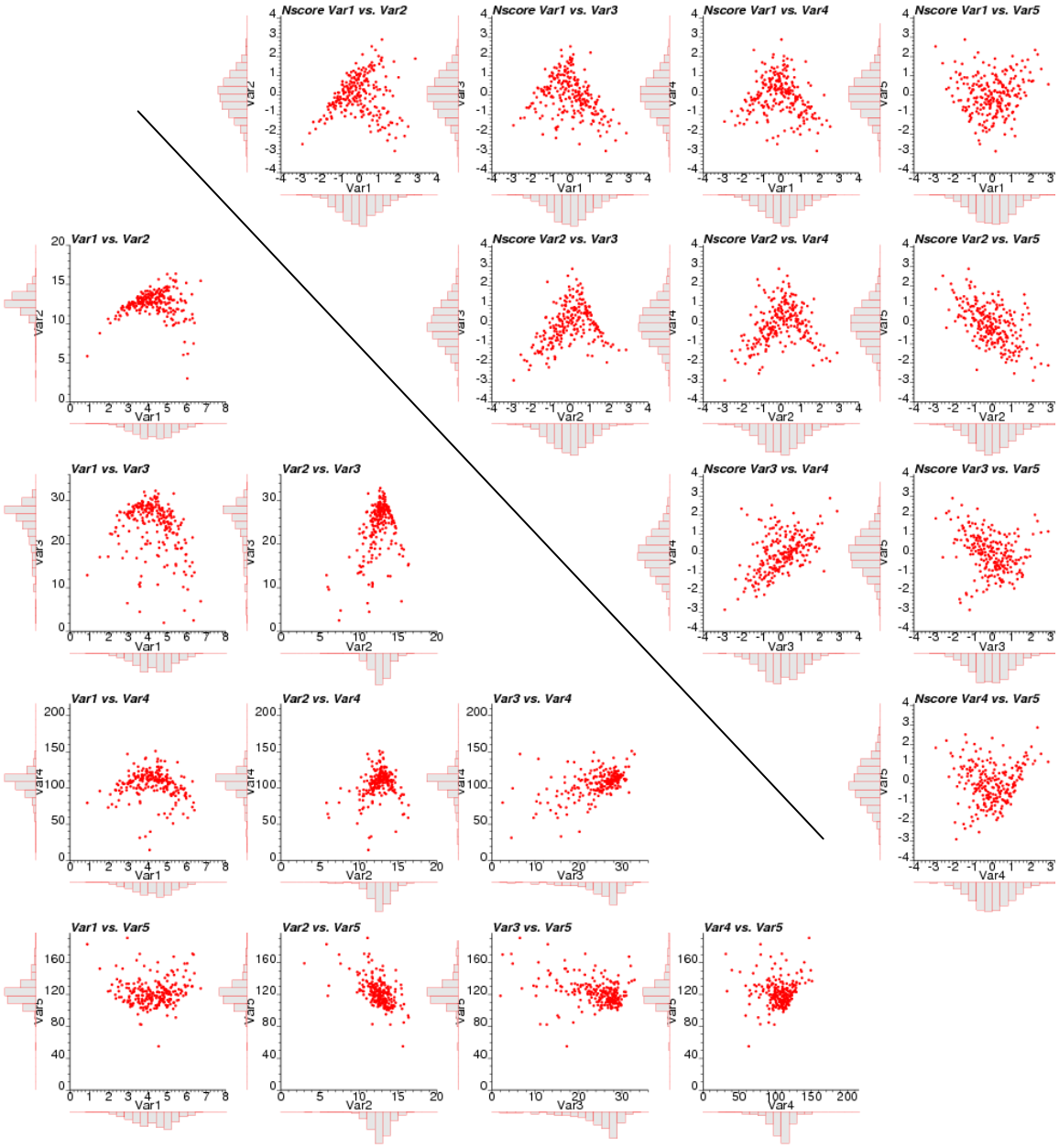


Figure 3: Scatterplots of the sampled observations before (bottom covariance triangle) and after (upper covariance triangle) normal score transformation.

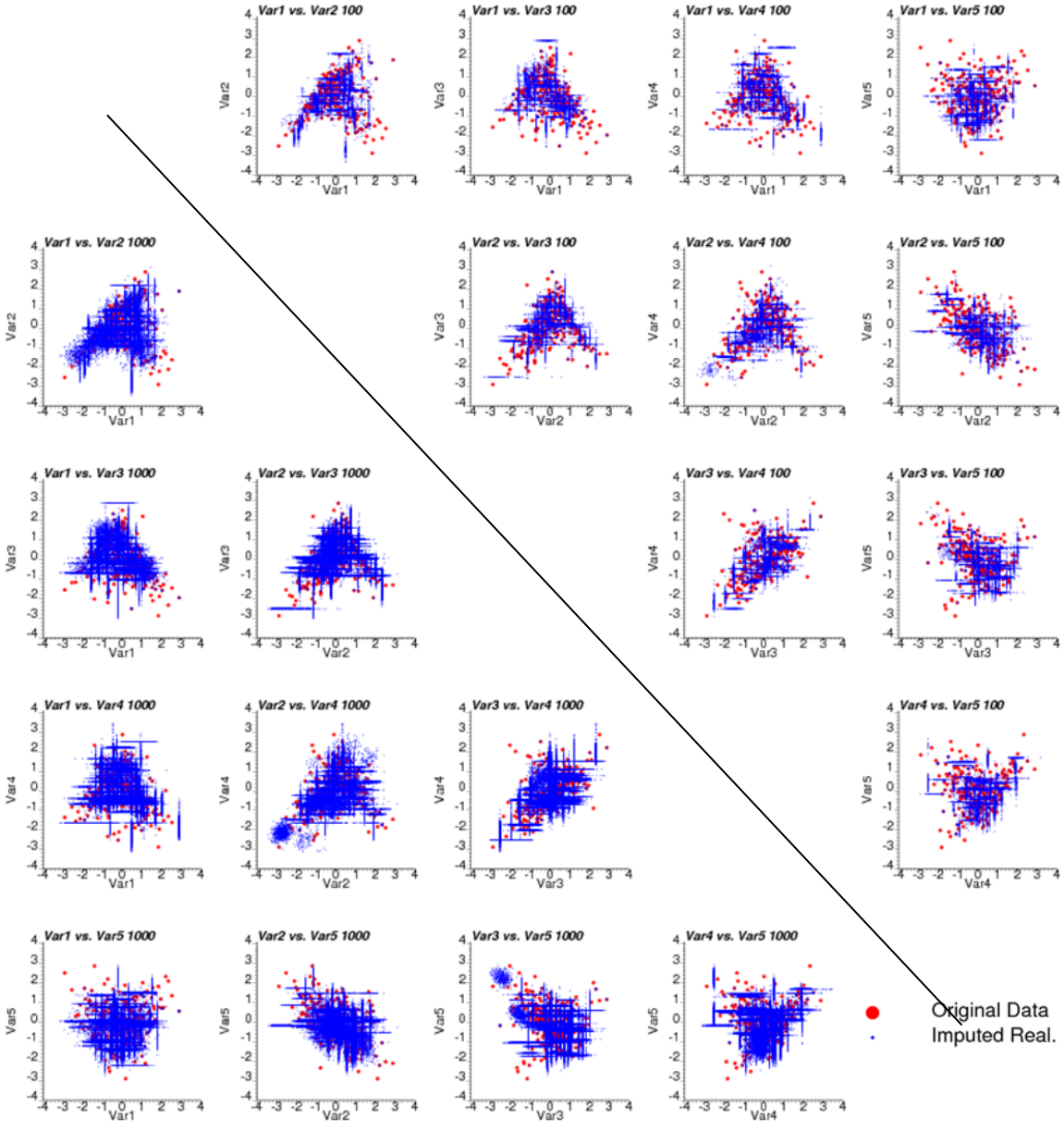


Figure 4: Scatterplots of the sampled observations and imputed observations for 1000 realizations (bottom covariance triangle) and 100 realizations (upper covariance triangle).

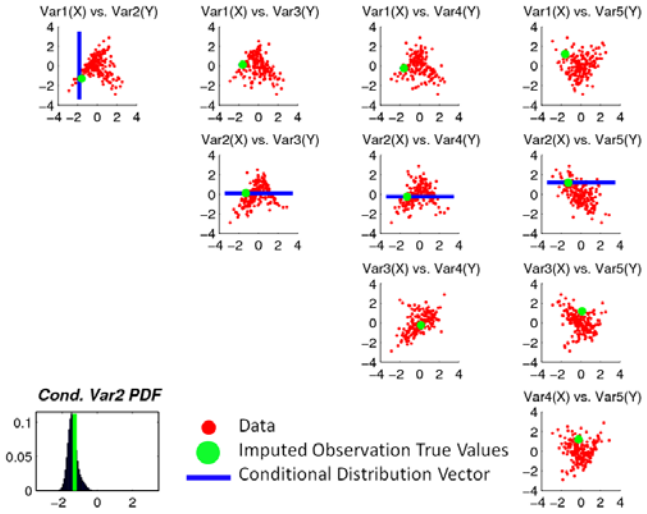


Figure 5: Conditional distribution (the only one that will exist) for an observation missing a single value.

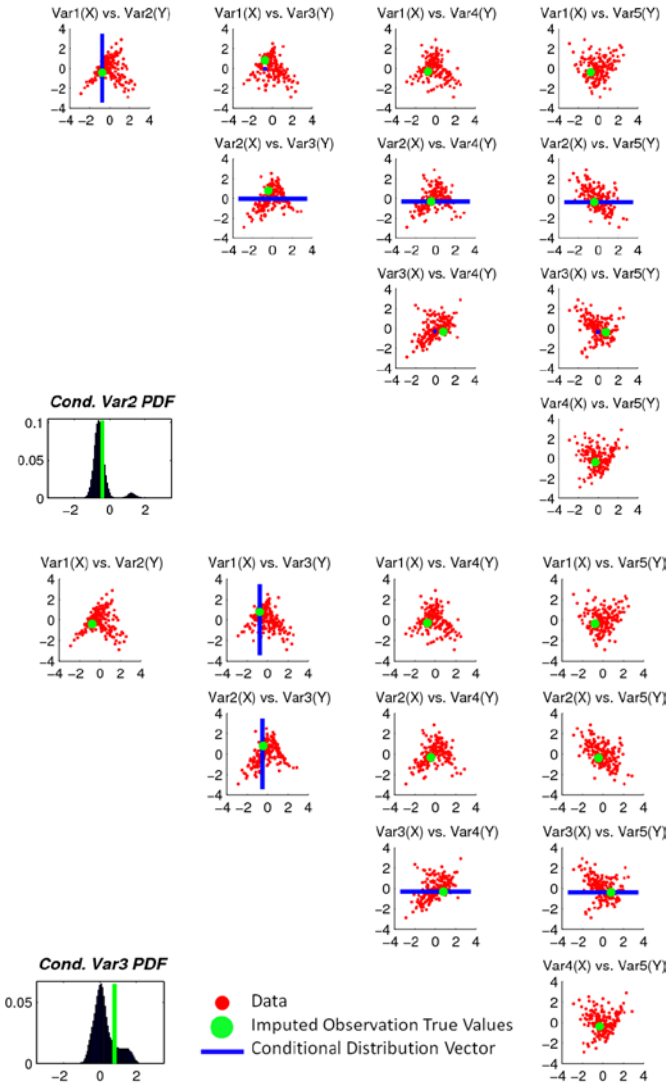


Figure 6: Conditional distributions for an observation missing two values.

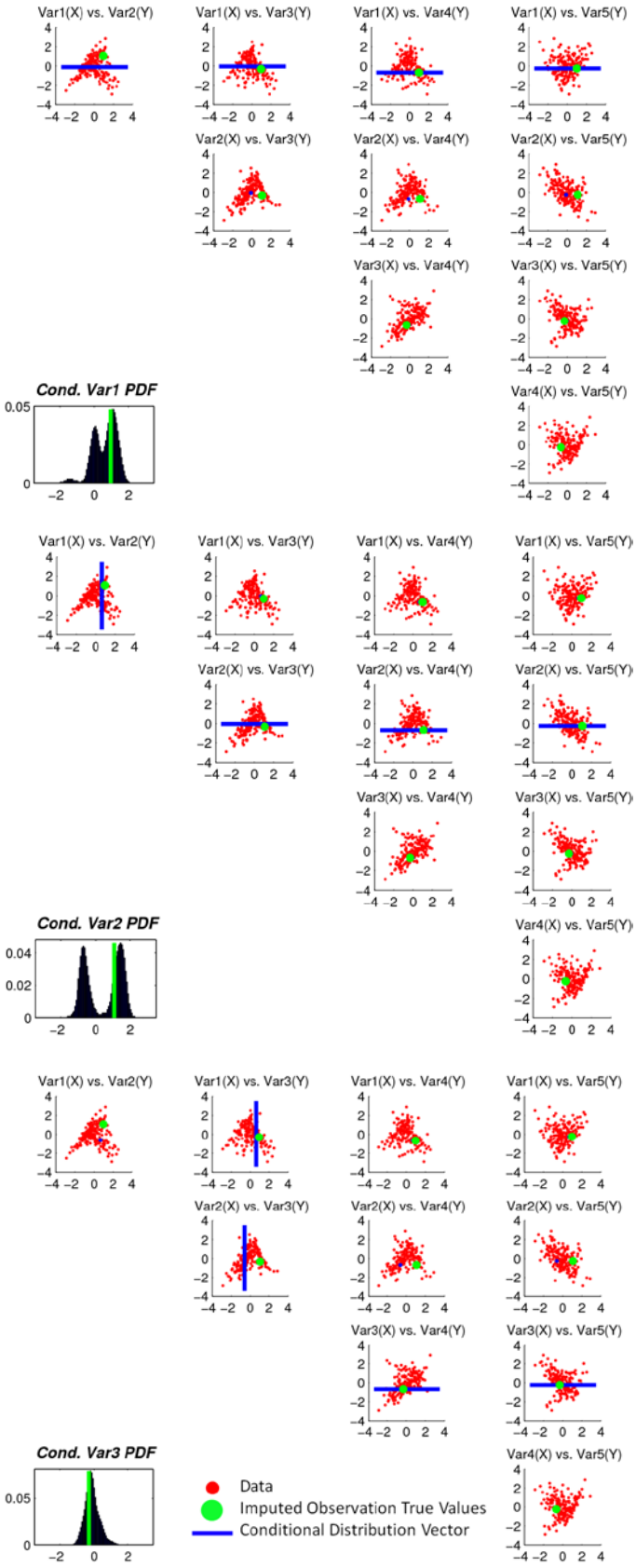
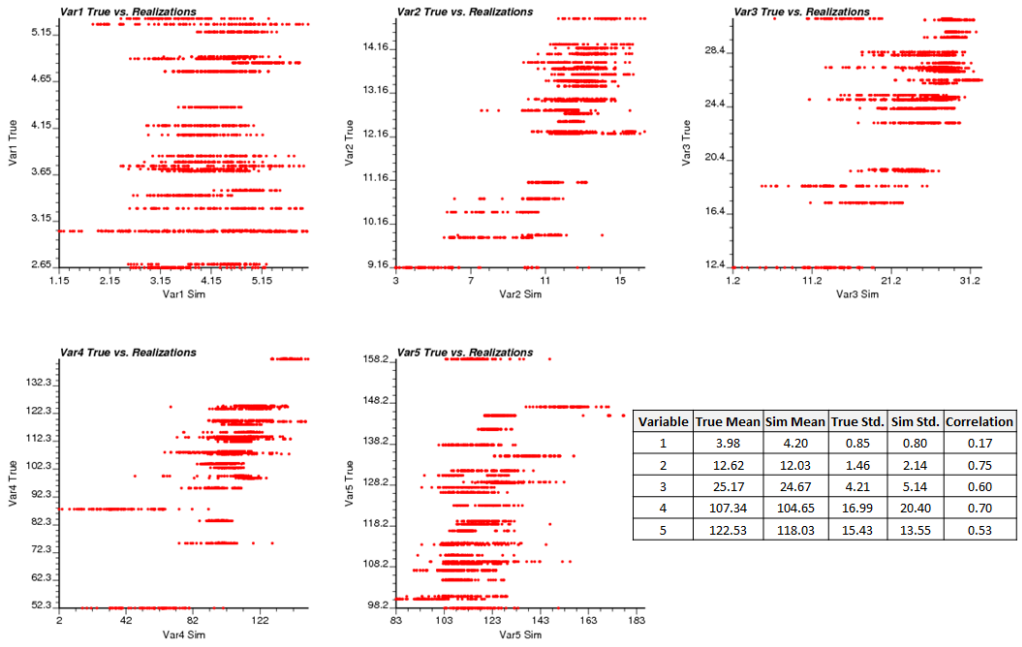


Figure 7: Conditional distributions for an observation missing three values.



Variable	True Mean	Sim Mean	True Std.	Sim Std.	Correlation
1	3.98	4.20	0.85	0.80	0.17
2	12.62	12.03	1.46	2.14	0.75
3	25.17	24.67	4.21	5.14	0.60
4	107.34	104.65	16.99	20.40	0.70
5	122.53	118.03	15.43	13.55	0.53

Figure 8: Comparison between removed True values (y-axis) and the associated 100 realizations of imputed values (x-axis). Summary comparison statistics are provided in the enclosed table.

Appendix: GSDI Program

Parameters for the Gibbs Sampling Data Imputation `gsdi` program are in Figure 9 and given below:

- **datafl:** input file containing the original normal score values. It is this file that will be made into realizations, where missing values are imputed
- **nvar:** number of variables requiring imputation
- **icol(i),i=1,…,nvar:** column locations within **datafl** for the variables to be imputed
- **tmin,tmax:** trimming limits that will determine which values are missing and requiring imputation
- **nreal:** number of realizations of the data that should be generated
- **nloc,bandw:** number of discretizations for the conditional distributions, and size of the kernel bandwidth as applied in all dimensions
- **seed:** random number seed
- **outname**: prefix name (may include a directory) for the output data realizations. This prefix will have the realization number and '.out' appended to form the final name of each file
- **indfl:** output file containing an indicator of whether a data value was imputed (1) or not (0)

```
                      Parameters for GSDI
                      *******************

START OF PARAMETERS:
../2-Nscore/nscore.out          - file with original nscore values
5                               - number of variables to simulate
3 4 5 6 7                       - columns of variables to simulate
-5   1.0e21                     - trimming limits
100                             - number of realizations
100 0.05                        - # cond. disc. and kernel bandwidth
69696                           - random number seed
./Data_Real/nscore_di           - output files prefix (will append real# and .out)
di_indicator.out                - output file for imputation indicator
```

Figure 9: Parameter file for the `gsdi` program.