Non-Parametric Gibbs Sampler with Kernel Based Conditional Distributions

Ryan M. Barnett and Clayton V. Deutsch

Many geostatistical techniques require samples to be drawn from a multivariate distribution. The distributions of geologic data are often described by a non-parametric distribution. One approach is to infer this distribution with kernel density estimation(KDE) on a multivariate grid. Unfortunately, the Curse of Dimensionality will make such methods computationally infeasible as the number of variables increases. A well established alternative for this problem is the Gibbs sampler, which effectively reproduces the joint distribution, while only requiring the conditional distributions. To apply the Gibbs sampler in a non-parametric setting, the conditional distributions must still be inferred. We propose a kernel integration method for determining these conditional distributions. Unlike the calculation and storage of kernel density estimates on a high dimensional grid, the iterative calculation and integration of kernels along conditional vectors is very manageable in massively multivariate settings. The algorithm is developed through a brief overview of its background theory. A discussion on implementation details motivates future research directions, with multiple data sets used for demonstration.

Introduction

The ability to sample from a the joint distribution of a multivariate geologic dataset is valuable in many areas of geostatistics. A non-parametric Gibbs Sampler algorithm is proposed for this task, which uses integrated kernels as a basis for the required conditional distributions. The essential background theory related to Gibbs sampling and kernel density estimation (KDE) will be provided, before applying them to the non-parametric Gibbs sampler algorithm. Details pertaining to the current implementation of this technique are outlined. As a great deal of further study is required for these details, future research is also discussed. A 2-D synthetic dataset is used for demonstration throughout these early sections due to its visual simplicity, with a more complex 6-D Ni laterite dataset used in the concluding case study. The parameters for the prototype standalone executable ɢᴍᴠ_sample are provided in the appendix. The remainder of this introduction will provide a brief discussion on the motivation for this technique, while introducing the 2-D synthetic data.

Motivation

While other potential applications for the proposed Gibbs sampler algorithm exist, including missing data replacement[1], this work was necessitated by recent advances in Gaussian mapping transforms [2,3]. These mapping transforms are interested in aquiring observations from the multivariate tails of a dataset`s underlying joint distribution, *beyond* the extents of the data itself. Consider the simulated drilling of a True synthetic model in Figure 1. As seen in this synthetic case, drill sample data will often fail to capture a great deal of information related to the underlying geologic deposit. If geostatistical simulation was performed using this data, it would be difficult to infer the multivariate tails for extrapolating the results. Capturing these extreme values in geostatistical models can be very important for certain petroleum and geometallurgical settings.

Gibbs Sampler

The Gibbs sampler is a method that allows for random variables of a multivariate distribution to be simulated without requiring the joint or even marginal densities. Although its foundations may be traced to work by Metropololis et al. [10], the Gibbs sampler came into wide spread use with the seminal paper by Geman and Geman [7]. Now applied in countless fields of study, it has already seen highly effective geostatistical applications with the spatial modeling of indicator variables [10,14].

Although the convergence proofs are not nearly as straightforward [4,7], the Gibbs sampler is extremely simple in its final algorithmic form. Following the theory given by Casella and George [4],

---

[1] Barnett, R., & Deutsch, C. (2012). Missing Data Replacement in a Complex Multivariate Context. *CCG Annual Report 14*, Paper112.

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 ith Gibbs sample.

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

After specifying the starting value as $Y_0^1 = y_0^1$ (which may be done randomly), the Gibbs sampler iteratively draws the remaining values from conditional probability distributions, that are formed based 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.

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

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. While the exact number of required Gibbs samples will potentially vary based on the complexity of the underlying random variables, it is expected based on statistical fundamentals [4] that less samples will be required for the Gibbs distributions to accurately represent simple statistics such the marginal means. As n becomes sufficiently large, however, the Gibbs distributions will converge to become representative of the True marginal and joint densities. The generalized multivariate representation of Equations 1 and 2 for k number of variables is given by Equations 3 and 4 respectively.

$$Y_0^1, Y_0^2, ..., Y_0^k$$
$$Y_1^1, Y_1^2, ..., Y_1^k$$
$$\vdots$$
$$Y_n^1, Y_n^2, ..., Y_n^k \tag{3}$$

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

It may be surpising upon first glance that that such a simple and seemingly naïve algorithm possesses these remarkable convergence properties. Readers are referred to the available references for a more thorough explanation of the background theory and convergence proofs [4,7,10], as only the essential basis is provided here.

The Gibbs sampler takes strong advantage of the Markov chain property[2], which is represented mathematically by Equation 5 [10]. A Markov chain is composed of a number of values for a random variable(s), which may vary with time t. As seen in Equation 5, the conditional distribution of a random variable at any given time is only dependent on its value at the immediately preceding time. The values occuring at times prior to this may be disregarded.

---

[2] The Markov chain property plays an important role in other areas of geostatistics, including colocated cokriging and the Markov model [6] where t is treated as spatial rather than temporal.

$$P(X_{t+1} \mid X_t, X_{t-1}, \dots, X_1, X_0) = P(X_{t+1} \mid X_t) \tag{5}$$

Literature treats X_t as the state of a random variable at time t, while all of the values that X can possibly assume are referred to as the state space. A state space may be categorical or continous, and of any dimension. Markov chains will explore this state space, with each step controled by transition probabilities [10]. It is not difficult to connect the Gibbs sampling algorithm to Markov chains, as the Gibbs sequence in Equation 3 represents the changing states of random variables, with their transition probabilities defined by the conditional distributions in Equation 4.

The Gibbs sampler was originally developed and is commonly applied with multivariate parametric models, [4,7] where it is difficult to attain the joint distribution through marginal integration. Parametric conditional distributions are combaritively easy to analytically define, motivating the Gibbs sampler approach. As gelogic data is non-parametric, however, its conditionals cannot be defined analytically. As stated earlier, this paper suggests that KDE can be used as a representative and computationally efficient means for inferring these distributions based on the data. Introductory kernel theory is provided in the next section, before applying it to attain the required conditional distributions.

Kernel Based Conditional Distributions
The kernel density estimation (KDE) calculation is given by Equation 6 [8], where K is any kernel satisfying the conditions $K(x) \geq 0$ and $\int K(x)dx = 1$ [8]. X_i denotes the ith data value, while x is the value or location of interest where the density must be estimated. n is the number of data used in its calculation, while h is the bandwidth (potentially subjective smoothing parameter that will be further outlined in the discussion).

$$f_{KDE}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{6}$$

Returning to the problem of inferring conditional distributions from potentially high dimensional data, consider discretizing a multivariate vector that represents the intercept of fixed conditional values with the random variable to be drawn. This vector is termed as L, which is k number of dimensions in accordance with the data, and discretized at m number of locations. Applying Equation 6, the KDE value at the discritized location L_j may be calculated based on the n number of data as seen in Equation 7. To attain the licit conditional distribution C, these integrated kernel values are then divided by their sum (Equation 8).

$$f_{KDE}(L_j) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{(L_j - X_i)}{h}\right) \tag{7}$$

$$C(L_j) = f_{KDE}(L_j) \bigg/ \sum_{j=1}^{m} f_{KDE}(L_j) \tag{8}$$

This concept is visually demonstrated in the schematic of Figure 2, where correlated Gaussian kernels are being integrated to form conditional distributions. In Figure 3, the outlined non-parametric Gibbs sampling algorithm is demonstrated within the same bivariate schematic setting. Moving away from schematics, a real step-by-step demonstration of the non-parametric Gibbs sampler is displayed in Figure 4. It is being executed on normal scores of the synthetic data presented in Figure 1.

Implementation
A great deal of research remains to be completed on details that fall out of the above described method. Implementation of the `gmv_sample` program (appendix) will be discussed, along with potential improvements that could be made in the future.

Normal Score Transform

Initial testing has found that better results are attained when this non-parametric Gibbs sampling algorithm is executed on unvariate Gaussian data. The normal score transform [1,6] is very commonly applied within geostatistics for modeling variables in Gaussian space. It is given by Equation 9, where quantiles are matched between the empirical data CDF, *F*, and the standard normal CDF, *G*. The sampled data from Figure 1 is normal scored in Figure 5.

$$y = G^{-1}(F(z)) \tag{9}$$

Gibbs sampling observations in normal score space may be back-transformed to the original data distribution. There are many potential reasons why the algorithm works better on normal score data, as opposed to the original skewed distributions. As outlined in the next sub-section, Gaussian kernels are currently used as a basis for the conditional distributions. It makes intuitive sense that Gaussian kernels would more accurately represent marginally Gaussian data, than irregular and skewed data. Further, while the Gibbs algorithm is proven to converge towards the underlying joint distribution, this non-parametric implementation may find it easier to reproduce marginally Gaussian variables due to the Central Limit Theorem [9]. This is all speculated reasoning that requires testing and validation.

Kernel Form
As the multivariate Gaussian distribution is generally considered the most widely applicable model in describing random variables, it has been applied as the initial kernel form in this algorithm. The multiGaussian model is given by Equation 10, where d is the number of random variables, Σ is the dxd dimension covariance matrix, and μ is the vector of means for each variable [9]. Inserting *f(x)* from Equation 10 into the conditional distribution kernel function of Equation 7, the multivariate Gaussian KDE equation is given by Equation 11. Additional kernel forms could be considered where alternative parametric models (e.g. lognormal) more accurately represent the distribution.

$$f(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \cdot \exp\left[\frac{-(x-\mu)'\Sigma^{-1}(x-\mu)}{2} \right] \tag{10}$$

$$f_{KDE}(L_j) = \frac{1}{nh(2\pi)^{d/2} |\Sigma|^{1/2}} \sum_{i=1}^{n} \exp\left[\frac{-(L_j - X_i)'\Sigma^{-1}(L_j - X_i)}{2h} \right] \tag{11}$$

Kernel Parameters
The potentially subjective kernel parameters in Equation 12 are of paramount importance with this algorithm. Recall that the bandwidth h is a smoothing parameter that is applied between and within all dimensions[3]. Optimal bandwidth is often considered to be a function of the variance and number of data in the underlying distribution. That is to say, the required bandwidth magnitude increases as the number of data decreases and the variance of the data increases [8]. Through methods such as bootstrapping and cross-validation, many different equations have been derived for determining the optimal bandwidth [12]. While disagreement on the best formula continues to exist in the univariate case, difficulty in determining the optimal bandwidth grows far greater in multivariate settings [8]. Scott advocates that density estimates may be plotted using various bandwidths [13], before choosing a final one based on the user's understanding of the distribution. A variety of optimal multivariate bandwidth equations were first applied with the non-parametric Gibbs sampler on data sets of varying sizes and dimensions. As no equation produced satisfactory results for all of the data sets, Scott's user knowledge suggestion was adopted. Bandwidth size is therefore left to the user as a manual input parameter in `gmv_sample`.

Another decision that is not yet fully understood, is whether the Gaussian kernels should match the covariance Σ of the normal score data. While it may make intuitive sense to do so, the multivariate arrangement of the data should naturally cause the Gibbs sampler to reproduce the underlying

---

[3] In the multivariate case, the bandwidth h may in fact be a matrix H of unique bandwidths between and within each dimension. This notation is avoided for simplicity since only a single smoothing parameters is currently applied.

covariance when using orthogonal kernels. Adding additional risk to the use of correlated kernels is that covariance is very sensitive to outlying values. Particularly with small data sets such as the one seen in Figure 5, these outliers may distort the covariance of the data away from that of the underlying True distribution. When this misrepresentative covariance is then imposed directly on the kernels, the problem may be exacerbated on the resultant Gibbs distribution. This remains to be thoroughly investigated. As with bandwidth, whether the kernels are orthogonal or correlated is left as a user defined input parameter for `gmv_sample`.

The dramatic effect that these kernel parameters will have on this non-parametric Gibbs sampler algorithm is demonstrated in Figure 6. 10,000 Gibbs sampler observations are presented using varying bandwidths with and without the use of the synthetic data covariance (Figure 5).

Extracting Observations from the Sampling Sequence

In its current implementation, the `gmv_sample` program selects a starting location based on a random number seed, with all subsequent samples aside from the first extracted as observations for the resultant Gibbs distribution. In other words, if N observations are required, n in Equation 3 is set as $N+1$, with the entire sequence extracted as observations aside from Y_0 (Equation 12).

$$Y_1^1, Y_1^2, \ldots, Y_1^k \ldots, Y_n^1, Y_n^2, \ldots, Y_{N+1}^k \tag{12}$$

Although the Gibbs sequence will likely converge with this simple extraction method, the distribution will be somewhat dependent on its starting location. To lessen this dependence and insure that the entire multivariate space is visited, a large variety of extraction techniques have been developed [4]. They may be summarized as using multiple starting locations (with associated independent Gibbs sequences) and/or sampling selectively from a Gibbs sequence that is far larger than the required N number of observations. These extraction methods will be tested in future study.

As mentioned, the required number of Gibbs samples will depend on properties that the user requires in the resultant distribution. As displayed in Figures 7 and 8 with the synthetic data, Gibbs distributions quickly converge toward the marginal means and standard deviations after only 100 observations, while the marginal and joint densities require closer to 5000 observations. Note that the summary statistics and histograms in Figure 7 should be compared with the normal score data in Figure 5. The number of Gibbs observations required for convergence on the various statistics was observed to change with dimension and complexity of the underlying data.

Case Study

An extremely complex and higher dimension Ni Laterite dataset will be used to demonstrate and challenge this non-parametric Gibbs sampler algorithm. Composed of 910 samples and six variables, the sampling will be performed on data that have been logratio[4] and normal score transformed. The histograms, summary statistics, scatterplots, and bivariate Gaussian KDE plots for this transformed Ni laterite data are presented in Figure 9. Gibbs sampling may be of interest for a variety of reasons on this transformed dataset[5], including the determination of multivariate tails, and the joint density of sparsely sampled multivariate space.

A distribution of 10,000 Gibbs observations was generated, with its statistics presented in Figure 10. Comparing the original data with the Gibbs distribution, good reproduction of the mean, as well as the marginal and joint densities is observed. Room for improvement remains in these statistics, however, which could potentially be tuned and improved with different kernel parameters, or a more robust sample extraction scheme for the Gibbs observations. Of greater concern, is the systematic inflation of standard deviation in all of the variables. This inflation of was also observed in the 2-D synthetic data, and

---

[4] The logratio transform [1] is not relevant to of this study, but reflects that this data was taken from a realistic geostatistical modeling framework that will honor compositional constraints. This transformed data was also used because of the exaggerated non-linear features that are present.

[5] e.g. If considering a Gaussian mapping [2,3] based simulation and requiring additional mapping points for the interpolation and extrapolation of back-transformed results

will be a focus of future research. Despite these concerns, the non-parametric Gibbs sampler has effectively reproduced this extremely complex multivariate data.

Conclusions

A non-parametric Gibbs sampler algorithm has been developed for multivariate sampling. It uses integrated data kernels as the basis for conditional distributions, which are required by the Gibbs sampler. Following development of the necessary theory, promising results were demonstrated using synthetic and real datasets. Future research is necessary into details of the algorithm, including the cause of variance inflation in the Gibbs distributions, kernel form, kernel bandwidth, kernel correlation, and alternative schemes for extraction of the Gibbs observations. Parameters for the `gmv_sample` program, which was used to produce the presented results, are outlined 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). MSNT Implementation Advances and Case Studies. *CCG Annual Report 14*, Paper 101.

3   Barnett, R., & Deutsch, C. (2012). Projection Pursuit Multivariate Transform. *CCG Annual Report 14*, Paper103.

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

5   Deutsch, C. (2011). Multivariate Standard Normal Transformation. *CCG Annual Report 13*, Paper 101.

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

7   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.

8   Hong, S. (2010). Multivariate Analysis of Diverse Data for Improved Geostatistical Reservoir Modeling. *PhD Thesis, University of Alberta, Canada*.

9   Johnson, R., & Wichern, D. (1988). Applied Multivariate Statistical Analysis. New Jersey: Prentice Hall.

10   Lyster, S. (2009). Simulation of Geologic Phenomena Using Multiple Point Statistics in Gibbs Sampler Algorithm. *PhD Thesis: University of Alberta, Canada*.

11   Metrolopolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (Journal of Chemical Physics). 1953. *Equations of State Calculations by Fast Computing Machines, vol.21*, pp.1087-1091.

12   Sain, S., Baggerly, K., & Scott, D. (1994). Cross-Validation of Multivariate Densities. *Journal of the American Statistical Association, vol.89*, pp.807-817.

13   Scott, D. (1992). *Multivariate Density Estimation: Thoery, Practice, and Visualization.* New York: John Wiley and Sons, Inc.

14   Srivastava, M. (1992). Iterative Methods for Spatial Simulation. *Stanford Center for Reservoir Forecasting*, No.5
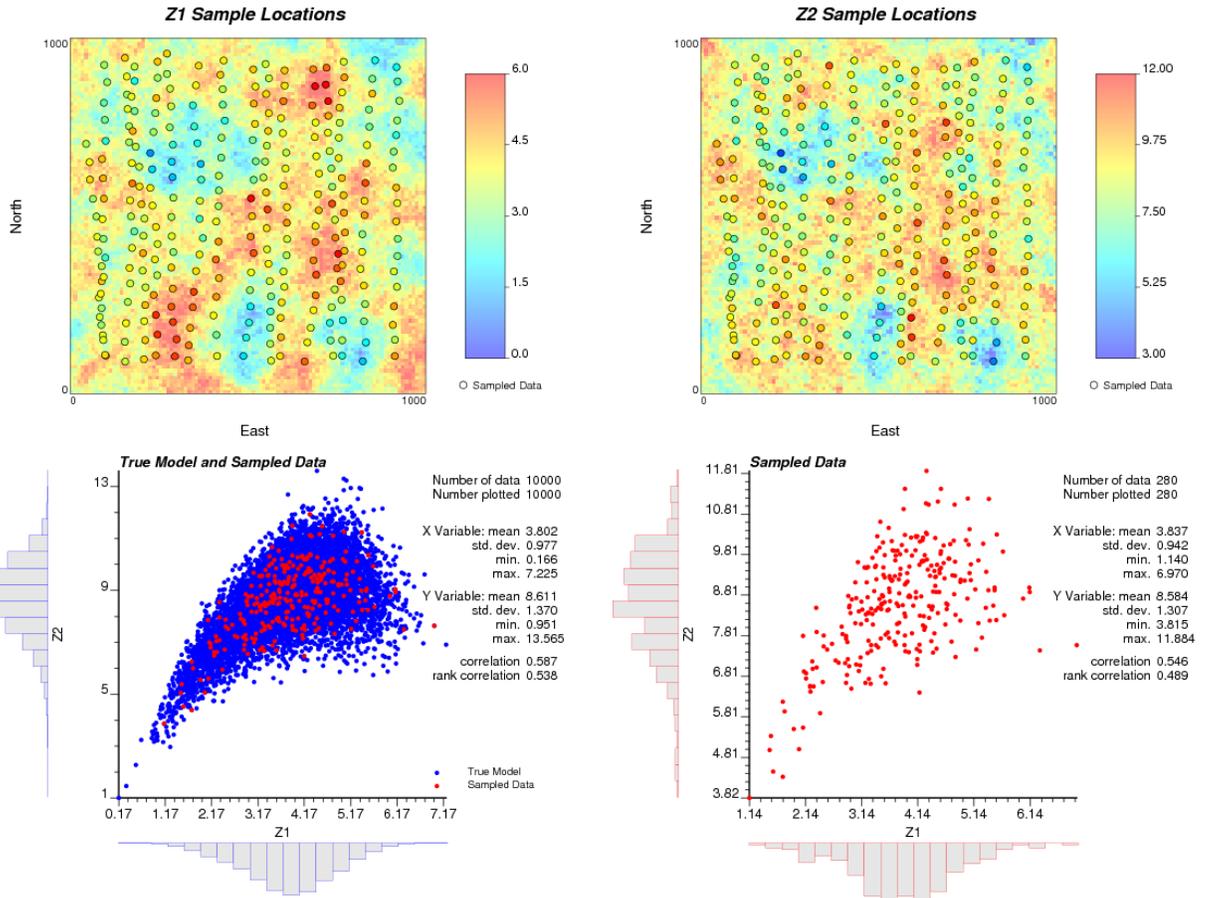
Figure 1: Overview of the synthetic model construction and resultant properties of Z1 and Z2. The top maps display the True model values (gridded) and data sampling locations (circles). The bottom left scatter plot displays the True model and sampled data with True model summary statistics/histograms, while the bottom right scatter plot displays summary statistics/histograms of the sampled data.
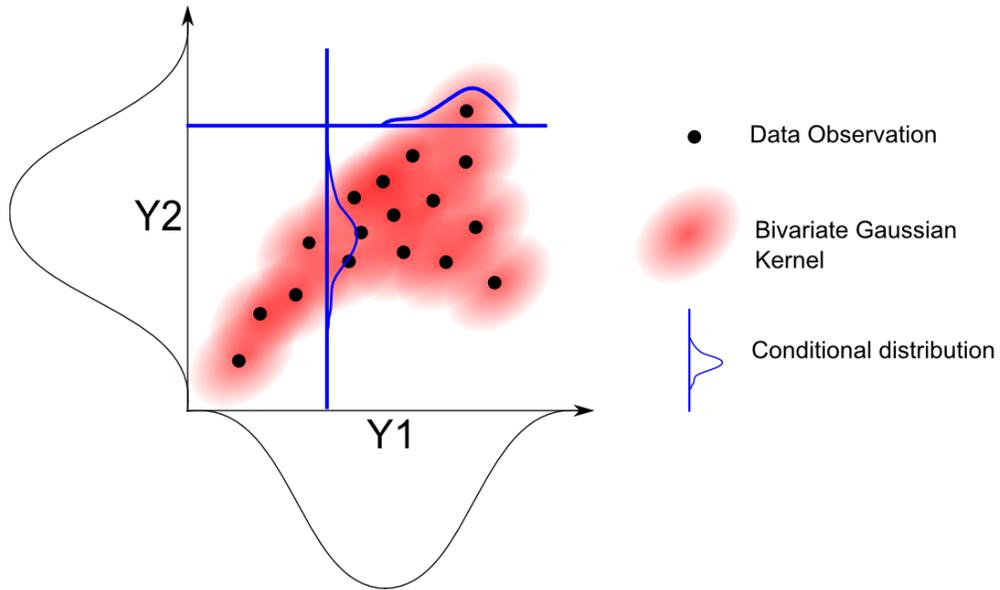
Figure 2: Schematic illustration of the data observations, their correlated bivariate Gaussian kernels, and the resultant conditional distributions. The opacity of the kernels corresponds with their approximate PDF value. Arbitrary locations are chosen for the displayed Y1 and Y2 conditional distributions, where the PDFs are calculated through integrating the data kernels at discretized locations along the line.
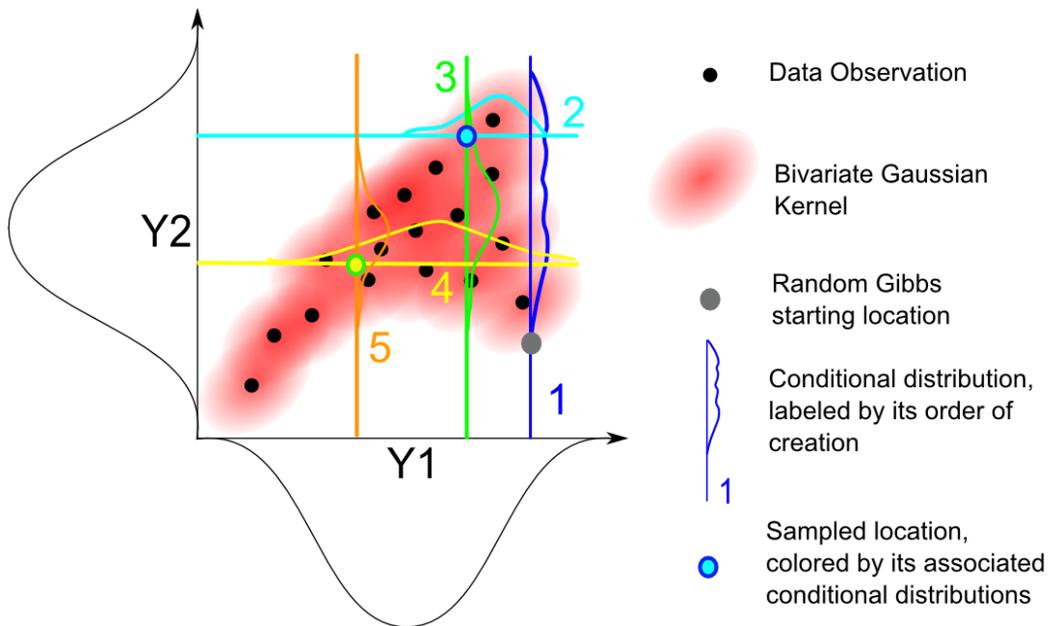
Figure 3: Schematic illustration of the non-parametric Gibbs sampling algorithm. The process begins at a random location (grey point), which is followed by the construction of the Y2 conditional distribution (blue). The randomly drawn Y2 value from the first conditional distribution is used to construct the Y1 conditional distribution (cyan). The drawn Y1 value from the second conditional distribution, as well as the drawn Y2 value from the first become the first Gibbs observation, from which the next iteration of the algorithm begins.
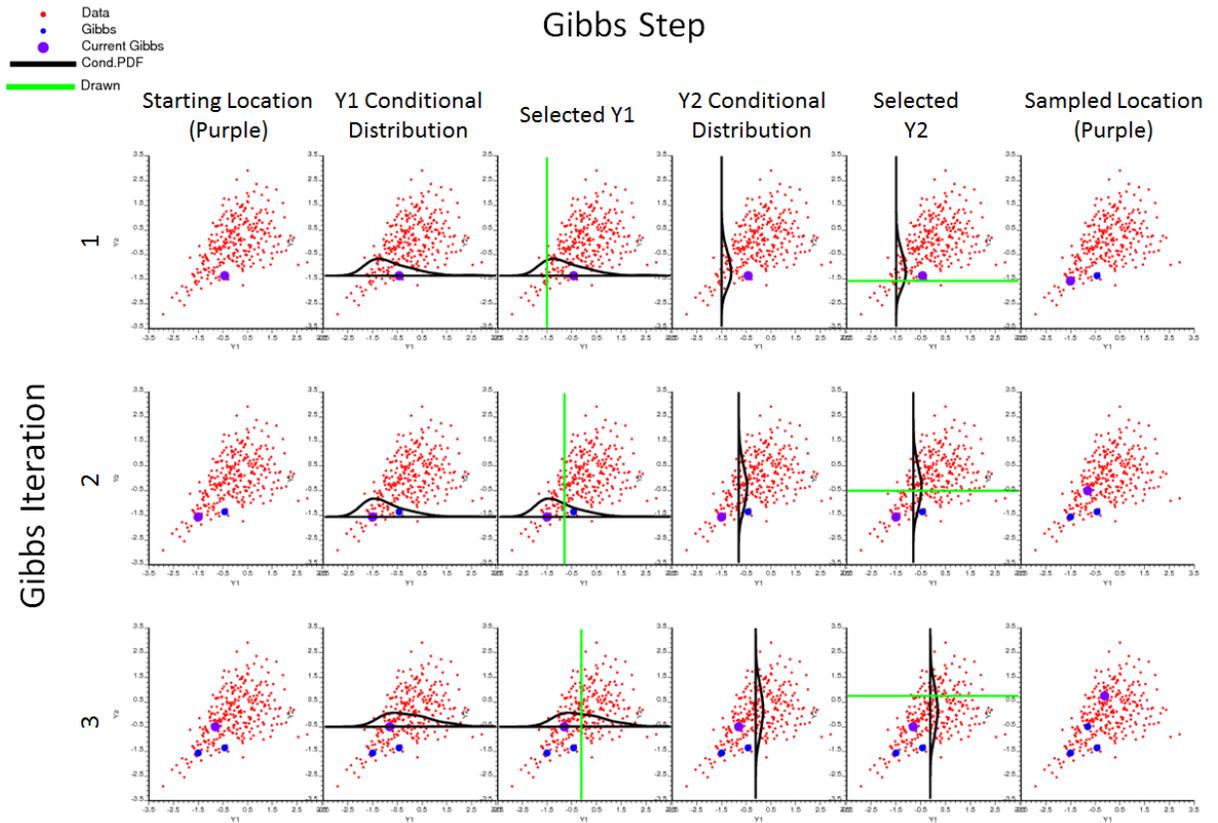
Figure 4: Breakdown of each step involved in non-parametric Gibbs sampling for the first three iterations.
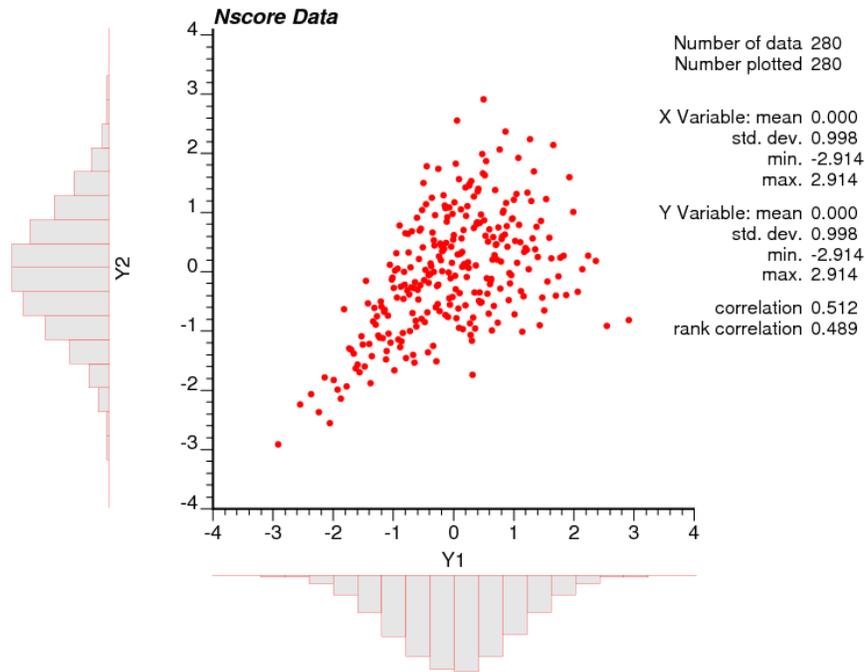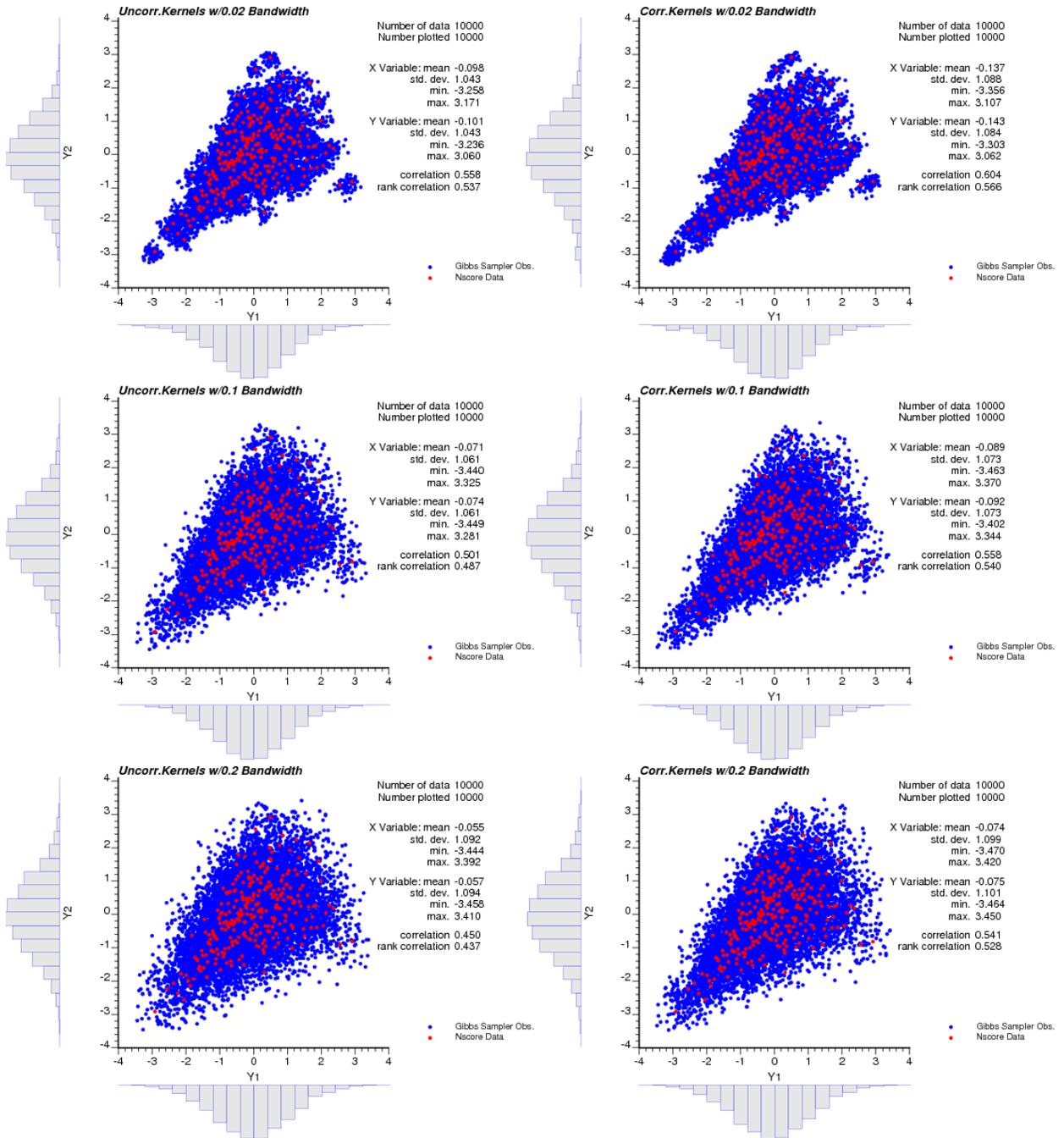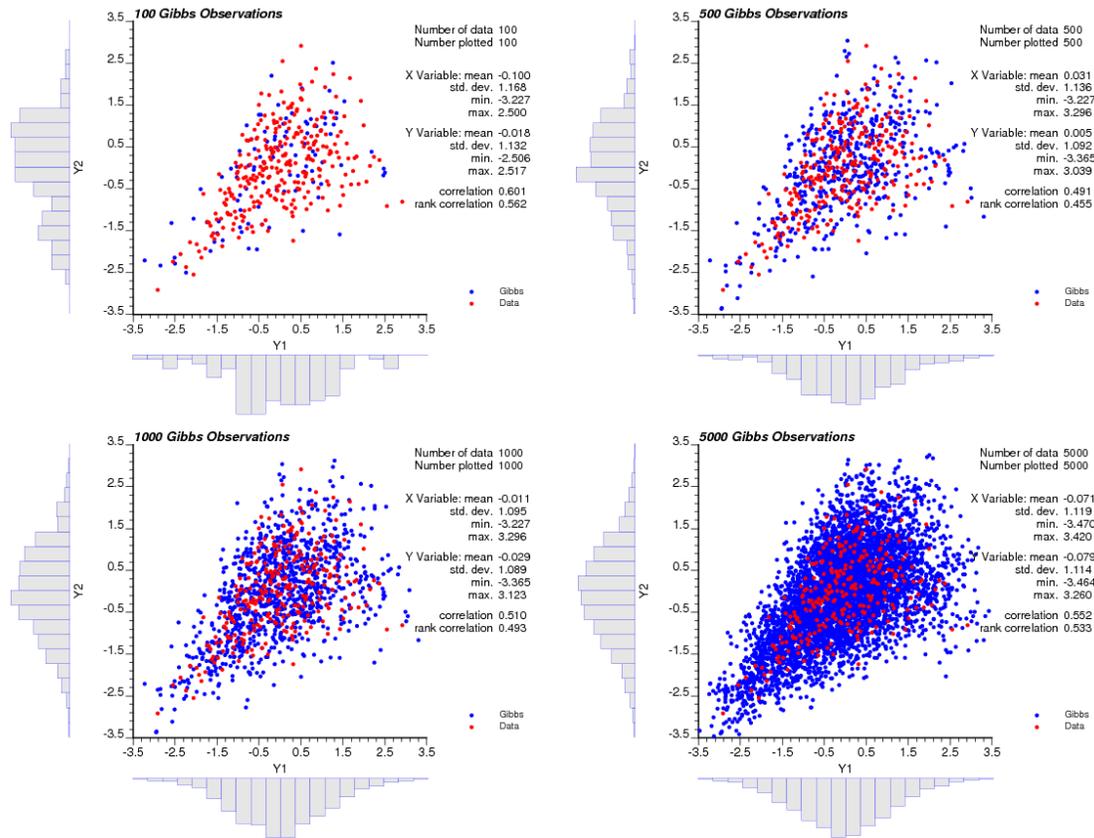

Figure 5: Scatterplot, histograms, and summary statistics of the normal scored data.
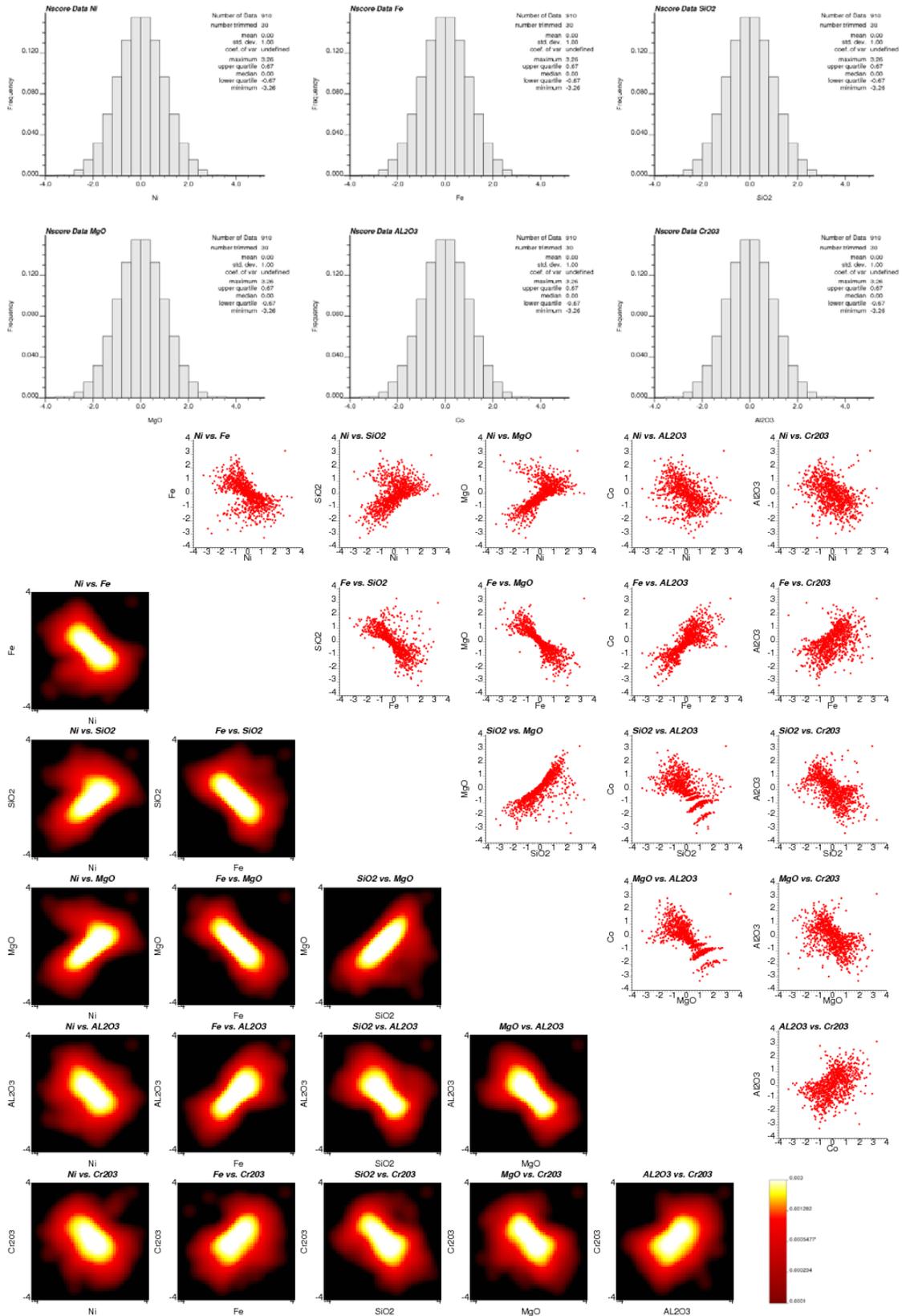
Figure 6: Scatterplots of the normal score data and Gibbs sampling observations using varying bandwidth with and without the normal score data correlation. Histograms and summary statistics are provided for the Gibbs distributions.

Figure 7: Scatterplots of the normal score data and Gibbs sampling distributions of varying observation quantities. Histograms and summary statistics are provided for the Gibbs distributions.
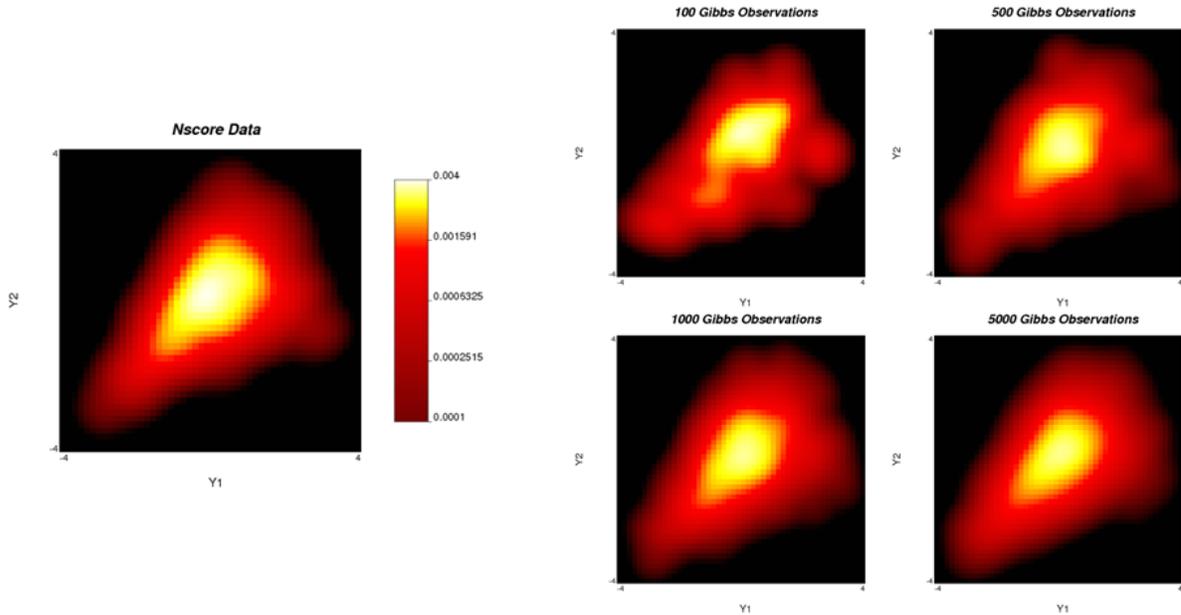
Figure 8: Gridded bivariate Gaussian KDE for the normal scored data and Gibbs sampling distributions of varying observation quantities.

Figure 9: Histograms (top), gridded bivariate Gaussian KDE (bottom covariance matrix triangle) and scatterplots (upper covariance matrix triangle) for the logratio and normal scored Ni Laterite data.

Figure 10: Histograms (top), gridded bivariate Gaussian KDE (bottom covariance matrix triangle) and scatterplots (upper covariance matrix triangle) for the Gibbs sampling distribution.
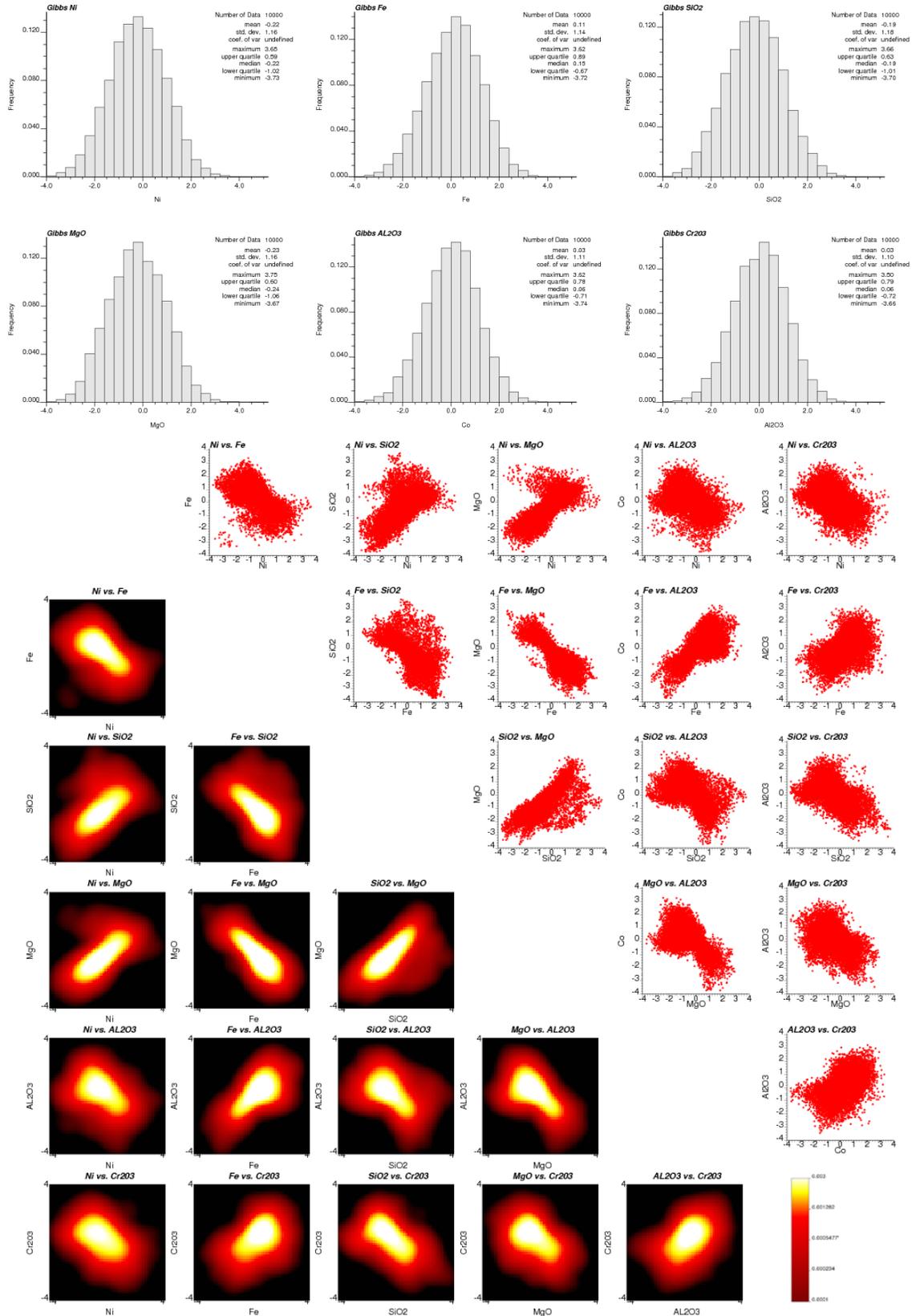
Appendix

The prototype standalone executable for non-parametric Gibbs sampling with kernel based conditional distributions is named `gmv_sample`. Its parameter file is displayed in Figure 11 and described below:

- **datafile:** file with the input data (recommended to be normal score transformed)
- **nvars,vcols(i), i=1,…,nvars:** number of variables and their column locations, for the distribution upon which Gibbs sampling will be performed
- **ngibb,nloc: ngibb** is the number of Gibbs observations that should be extracted, while **nloc** specifies the number of discretizations in each conditional distribution (both parameters effect execution time)
- **bandw:** kernel bandwidth that is applied across all dimensions
- **icor:** toggles whether correlation of the data should be applied to the kernels**(0=no,1=yes)**
- **ixv(1):** random number seed used to determine the starting location of the Gibbs sequence
- **tmin,tmax:** trimming limits that applied to all variables
- **outfl:** file containing the extracted Gibbs Observations

```
 1                    Parameters for GMV_SAMPLE
 2                    ***************************
 3
 4  START OF PARAMETERS:
 5  data.dat                 -  file with data
 6  3 2 3 4                  -  number of variables and columns
 7  200 100                  -  # of gibbs obs.,# cond. discretizations
 8  0.10                     -  kernel bandwidth
 9  0                        -  use nscore correlation?
10  69069                    -  random number seed
11  -998. 999.               -  trimming limits
12  gmv_sample.out           -  file for output gibbs observations
```

Figure 12: Parameter file for the `gmv_sample` program.