Multivariate Distribution Modeling of Geologic Variables

John G. Manchuk and Clayton V. Deutsch

Centre for Computational Geostatistics
Department of Civil & Environmental Engineering
University of Alberta

Multivariate Gaussianity is an assumption often utilized for multiple variables. It is assumed that all variables comprise a Gaussian random function after normal score transformation. Realistically, we are only know that the marginal distributions are Gaussian. Higher order distributions between variables at the same location or between variables distributed in space may not be Gaussian. The Gaussian model may be inadequate in some cases. This paper discusses a non-parametric technique to model multivariate distributions after normal score transformation. Marginal distributions are Gaussian. The sequence of polynomials known as Bernstein polynomials are used to model the multivariate relationship between a set of variables. The method is applied in an example involving full permeability tensors resulting in a nine-variable distribution.

Introduction

The relationship between a set of variables is often assumed multi-Gaussian after a normal score transformation. However, this is not always acceptable. Scatter plots of the normal scores often show non-linear features, constraint type features, or heteroscedasticity. These cannot be reproduced under a multi-Gaussian assumption. Techniques have been developed to produce multi-Gaussian relationships with transformations, for example the stepwise conditional transform (Leuangthong and Deutsch, 2000; Rosenblatt, 1952); however, this method requires a substantial amount of data, which grows exponentially with the number of variables. This problem is commonly referred to as the curse of dimensionality. For application to geological resource modeling, multivariate distribution modeling techniques must be insensitive to available data. This is especially true for certain reservoirs when there are typically very few samples due to the cost and risk in obtaining them.

Non-parametric techniques for smooth estimation of distribution functions have been available for some time, going back to the kernel technique (Rosenblatt, 1956). Their presence in the field of geostatistics has been limited to modeling univariate distributions, commonly with the empirical distribution, and change of support models. The empirical distribution is used in calculating the normal score transformation of a variable. A smooth estimate of the empirical distribution using Hermite polynomials has been implemented to transform distributions across different support volumes (Chiles and Delfiner, 1999). Non-parametric techniques are useful in other areas of geostatistics.

The sequence of polynomials known as Bernstein polynomials has several advantageous features that make them amenable to multivariate distribution modeling. In this paper, marginal distributions are Gaussian for all variables after normal score transformation, which is often the case in geostatistics. Bernstein polynomials are then used to model the multivariate empirical distribution. This paper first discusses the multi-Gaussian assumption in geostatistics. This is followed by the definition of Bernstein polynomials and the method of multivariate distribution modeling is described afterwards with bivariate examples. Finally, an application to model the multivariate distribution of permeability tensor elements is covered.

Background

Several geostatistical modeling techniques that involve multiple variables are based on the assumption of multivariate Gaussianity. After normal score transformation, the variables are assumed to follow a Gaussian random function (GRF). This is an important assumption since Gaussian distributions are fully defined by a mean and covariance function, the two requirements for Kriging (Chiles and Delfiner, 1999).

The spatial distribution of a GRF can be determined knowing these two components. It is common to see other assumptions on the mean, typically tied to a particular form of Kriging. Simple Kriging assumes the mean is known and is more often assumed to be zero with Gaussian data. Ordinary kriging assumes the mean is unknown, but constant. This is a special case of Universal kriging, which assumes the mean is a polynomial expansion.

There are several geostatistical modeling methods that involve the multi-Gaussian assumption; three will be touched on here: cokriging, collocated cokriging and Bayesian updating. Cokriging involves relating a set of variables together with a linear model of coregionalization. Essentially this is a set of covariance functions for each variable (direct covariance) and between the variables (cross covariance). Different variables are combined into a single covariance matrix calculated from the direct and cross covariance functions. Cokriging is typically implemented to estimate one primary variable; however, there is a variant to estimate two or more variables simultaneously. Estimates for each variable are also associated with an estimation variance and these parameters describe Gaussian distributions.

Collocated cokriging is a special case of cokriging where only two variables are considered. One is treated as secondary and is co-located with each location to be estimated. Rather than using numerous conditioning data of each variable type as in cokriging, this variant uses only the co-located secondary. Due to this simplification, this model is known as a Markov-type model (Chiles and Delfiner, 1999). No cross covariance function is needed; rather we assume it is equivalent to the primary variable's covariance function at lag h scaled by the covariance between the variables at lag h=0.

Bayesian updating is slightly different than the variety of cokriging estimators discussed above. Several primary and secondary variables are combined into prior and likelihood distributions that are used in Bayes' theorem to calculated updated distributions for the primary variables (Deutsch and Zanon, 2007). Secondary are assumed discrete variables. Each primary variable is kriged independently providing the prior distributions. Correlations between secondary to secondary and primary to secondary are used to calculate the likelihood. The outcome is a set of estimates and estimation variances that describe non-standard normal distributions for each primary variable.

All three multivariate methods discussed result in an estimate and estimation variance for each variable. Problems arise when considering simulation, which is typically performed assuming variables define a GRF. Results will reproduce this function; however, after normal score transformation only the marginal distributions are truly Gaussian. Simulation with collocated cokriging involves only one distribution from which a random value is drawn. Simulation of the other multivariate methods can be done by drawing a set of correlated values from a standard normal multivariate distribution, which is described by the covariance matrix between those variables (Johnson and Wichern, 2002). Values must then be conditioned by the marginal distributions from kriging (Deutsch, Ren and Leuangthong, 2005).

What if the structure beyond marginal distributions was not Gaussian? The actual distribution between variables would not be reproduced by the multivariate geostatistical methods discussed above. In fact, there are some variable combinations that result in very peculiar relationships after normal score transformation. Multivariate modeling will be demonstrated later using elements of a permeability tensor and the relationships between some of the variables will be looked at here to show how non-Gaussian they are. Full permeability tensors calculated from unstructured grid blocks are used (Equation 1).

$$K = \begin{bmatrix} K_{xx} & K_{xy} & K_{xz} \\ K_{yx} & K_{yy} & K_{yz} \\ K_{zx} & K_{zy} & K_{zz} \end{bmatrix} \tag{1}$$

For visualization purposes, bivariate relationships between pairs of tensor entries will be looked at. Relationships that are observed here are specific to the scenario from which tensors were derived. Different conditions on the grid cells, geology, porosity and scalar permeability will lead to different relationships. Also, using other flow simulators and conditions on flow will affect tensor solutions.

Three different forms of bivariate relationships were determined to exist between tensor entries; diagonal to diagonal (Figure 1A), diagonal to off-diagonal (Figure 1B), and off-diagonal to off-diagonal (Figure 1C). Each type of relationship shows similar shape and form of marginal distribution, so there is no need to show all relationships. Trivariate relationships are too cumbersome to visualize effectively here.

Reproduction of these relationships would be difficult if not impossible assuming a multivariate Gaussian distribution. A model that can capture the relationships, yet have Gaussian marginal distributions will provide better statistical reproduction. This paper proposes the use of Bernstein polynomials to model bivariate and higher dimensional relationships such as those exhibited by permeability tensors

Bernstein Polynomials and Cumulative Distributions

Bernstein polynomials have been applied to modeling univariate distribution functions (Kakizawa, 2003; Babu, Canty, and Chaubey, 2002; Petrone, 1999) as well as multivariate distributions (Sancetta and Satchel, 2004; Sancetta, 2007; Kolev, Anjos, and Mendes, 2006). There are several characteristics of these polynomials that make them attractive for modeling cumulative distribution functions (cdf). The general form of Bernstein polynomials will be provided first in their univariate and multivariate forms.

Any continuous function can be approximated by a sequence of Bernstein polynomials; however, the functions must be defined in the space [0,1]. Various transformations exist for functions that do not meet this condition. An approximation to a function, $f(x)$, by Bernstein polynomials, $B_n(f : x)$, of degree n with x in [0,1] is given by Equation 2. B_n also interpolates the endpoints of the interval being approximated, that is $B_n(f : 0)=f(0)$ and $B_n(f : 1)=f(1)$.

$$\begin{aligned} B_n(f : x) &= \sum_{a=0}^{n} f\left(\tfrac{a}{n}\right)\binom{n}{a}x^a(1-x)^{n-a} \\ &= \sum_{a=0}^{n} f\left(\tfrac{a}{n}\right)B_{a,n}(x) \\ \binom{n}{a} &= \frac{n!}{a!(n-a)!} \end{aligned} \tag{2}$$

The function f must be known at $x=a/n$, $a=0,\ldots,n$, and it is approximated by the combination of Bernstein basis functions, $B_{a,n}(x)$. Extending Equation 2 to multiple dimensions amounts to a product of Bernstein basis functions (Equation 3) that are defined on the hypercube $[0,1]^d$. A set of Bernstein polynomials define each axis of a multidimensional function $f(x_1,x_2,\ldots,x_d)$. All functions do not have to be known for the same set, a/n, $a=1,\ldots,n$. This function, with some constraints to be discussed, will be used to capture the bivariate structure of Figure 1 as well the multivariate structure of permeability tensor elements.

$$B_n(f : x_1,\ldots,x_d) = \sum_{a_1=0}^{n_1}\cdots\sum_{a_d=0}^{n_d} f\left(\tfrac{a_1}{n_1},\ldots,\tfrac{a_d}{n_d}\right)\prod_{k=1}^{d} B_{a_k,n_k}(x_k) \tag{3}$$

In reference to modeling cdf's for geostatistics, there are two properties of Bernstein polynomials that are amenable:

1. If $f(x)$ is positive on [0,1], $B_n(f : x)$ is also positive: B_n is a monotone operator. If $f(x)$ is bounded by $[y,Y]$, then so is $B_n(f : x)$. For cumulative distributions, $y = 0$ and $Y = 1$ giving

$$0 \le f(x) \le 1 \Rightarrow 0 \le B_n(f : x) \le 1, x \in [0,1] \tag{4}$$

2. If $f(x)$ is monotonically increasing in [0,1], so is $B_n(f : x)$. for a proof see (Phillips, 2003).

$$f'(x) \ge 0 \Rightarrow 0 \le B_n'(f : x) \ge 0, x \in [0,1] \tag{5}$$

For a cumulative distribution, $B_n(f : x)$ will be approximating the probability, $P(X \le x)$. These probabilities are typically approximated in geostatistics with the empirical distribution (Equation 6). Data may either be given equal weights (λ's) or by other means such as declustering. Normal score transformation is applied to make this empirical distribution Gaussian. The problem now lies in modeling structure beyond the marginal distributions when considering more than one variable. This structure can also be approximated with a multivariate empirical distribution; however, before getting into the details on the use of Bernstein polynomials, some comments on empirical distributions and histograms must be made.

$$F(x) = \sum_{i=1}^{N} \lambda_i \delta\left(x_i \le x\right) \tag{6}$$

Modeling the empirical distribution can be done with values from Equation 6 directly; however, this poses a problem for Bernstein polynomials when extending to higher dimensions. Acquiring $f(a_1/n_1,\ldots,a_d/n_d)$ in Equation 3 would require resampling the empirical distribution on regular intervals in each dimension. Consider the nine variables to be modelled in permeability tensors and resampling with only 5 values per variable. Each sample would be defined by a vector, $\boldsymbol{x}$, with 9 elements and a probability. The empirical distribution would require storage for every $\boldsymbol{x}$ amounting to 5^9=1,953,125. This would rapidly become unacceptable with increased variables and number of resample vectors. However, if the density is considered rather than the cumulative probability, one will find that in almost all cases the set is sparse, containing many zero density points. This result is shown for the cross-plot of Figure 1A (Figure 2) using 20 resample points along each axis. All 400 points are assigned values larger than zero for the empirical distribution whereas only 98 of them have a density greater than zero. Note that the maximum number of populated bins possible will always be equal to the size of the data set.

To accommodate this issue, modeling of the empirical distribution can be accomplished by integrating the Bernstein approximation of the density. Densities having a value of zero can be omitted from Equation 3. Integration of Equation 3 can be done by integrating each polynomial component separately, since each only depends on one of $x_1,\ldots,x_d$. Calculating the cumulative distribution for a particular x is done by integrating Equation 2 from 0 to x (Equation 7), which can be evaluated recursively by parts. Equation 8 is the multivariate form of Equation 7, which is the same as Equation 3 with $B_{a,n}(x)$ replaced with $\beta_{a,n}(x)$.

$$\begin{aligned} B_n(f:x) &= \sum_{a=1}^{n} f\left(\tfrac{a}{n}\right) n \binom{n-1}{a-1} \int_0^x t^{a-1}(1-t)^{n-a+1}\,dt \\ &= \sum_{a=1}^{n} f\left(\tfrac{a}{n}\right) \beta_{a,n}(x) \end{aligned} \tag{7}$$

$$B_n(f:x_1,\ldots,x_d) = \sum_{a_1=0}^{n_1} \cdots \sum_{a_d=0}^{n_d} f\left(\tfrac{a_1}{n_1},\ldots,\tfrac{a_d}{n_d}\right) \prod_{k=1}^{d} \beta_{a_k,n_k}(x_k) \tag{8}$$

Methodology

Recall that Bernstein polynomials are defined for all continuous functions on [0,1] and that the multivariate distribution model to be generated has Gaussian marginals, which are defined for [-∞,∞]. Transformation to [0,1] is accomplished by evaluating the standard normal cdf. In effect we are now dealing with modeling a copula function. Marginal distributions are uniform and defined in [0,1]. To ensure the marginals are uniform, two requirements of Equation 8 must be met. The first (Equation 9) is guaranteed if $f(x_1,\ldots,x_{k-1},0,x_{k+1},\ldots,x_d)$=0, which it should for a Gaussian marginal distribution or equivalently its probabilities over [-∞,∞]. The second requirement (Equation 10) ensures the densities sum to a uniform distribution along each dimension (Sancetta and Satchell, 2004).

$$B_n(f:x_1,\ldots,x_{k-1},0,x_{k+1},\ldots,x_d) = 0 \tag{9}$$

$$B_n(f:1,\ldots,1,x_k,1,\ldots,1) = x_k \tag{10}$$

A series of operations are involved in preparing data to model the multivariate structure, most of them are common to geostatistical modeling workflows.

1. Data preparation (cleaning, outlier, missing sample and duplicate point detection, etc…)
2. Declustering to accommodate irregularities in sample spacing
3. Normal score transformation using declustering weights if available

Probabilities or quantiles for distribution modeling can either be retrieved during normal score transformation or by evaluating the standard normal cdf with normal scores from 3 (Figure 3). Remaining

components of the modeling process involve calculating densities from the quantile space for Equation 8, and ensuring those values meet requirements of Equation 9 and 10. When dealing with normal scores without weights, or those that share the same set of declustering weights, which is often the case, quantiles will be distributed such that Equation 10 is honoured. Equation 9 is honoured by construction.

Resampling of the distribution, as mentioned previously, will be done with a multivariate histogram. It must capitalize on the sparse nature of the problem, especially when data sets and number of variables become large. Moreover, the Bernstein approximation does not require zero-valued densities. A bivariate histogram (Figure 4) was calculated for the quantiles in Figure 3. Note that the marginals are uniform cdf's.

Having calculated densities in quantile space provides all the information for modeling the multivariate cdf with Bernstein polynomials. Cumulative distributions are used for several purposes in geostatistics, two being the evaluation of probabilities from events and drawing events given probabilities. The latter is carried out in simulation such as sequential Gaussian simulation. A quantile is randomly drawn from a uniform distribution and the resulting event is calculated as the inverse of the cdf model. For multivariate simulation, events must be drawn such that any underlying correlation structure is reproduced. This is possible from the Bernstein polynomial model described above. Operations such as these will be discussed in the following section.

Operations

In reference to geostatistical modeling applications, several operations are carried out with cumulative distributions: evaluating probabilities, inversion, and extracting conditional distributions. These operations will be described in reference to Equation 8 for a vector of quantiles $\boldsymbol{x}$ that are associated with a vector of Gaussian random variables $\boldsymbol{y}$. Evaluating probabilities is straightforward and amounts to evaluating Equation 8 with $\boldsymbol{x}$. Inversion can be accomplished by line search techniques. For a single variable and given a probability p, we must find x such that Equation 7 evaluates to p. For multiple variables, conditional distributions are required.

Given a multivariate distribution $F(\boldsymbol{x})$, the conditional distribution in terms of x_j is given by Equation 11. Extracting the conditional distribution from Equation 8 is done by differentiating the integral component that involves x_j (Equation 12). Note that $\boldsymbol{x}_{(j)}$ indicates that element j is not in $\boldsymbol{x}$.

$$F\left(\mathbf{x}_{(j)} \mid x_j\right) = \frac{d}{dx_j} F\left(\mathbf{x}\right) \tag{11}$$

$$B_n(f : \mathbf{x}_{(j)} \mid x_j) = \sum_{a_1=0}^{n_1} \cdots \sum_{a_d=0}^{n_d} f\left(\tfrac{a_1}{n_1}, \ldots, \tfrac{a_d}{n_d}\right) \frac{1}{a_j} B_{a_{j-1}, n_j - a_j + 1}(x_j) \prod_{k=1, k\neq j}^{d} \beta_{a_k, n_k}(x_k) \tag{12}$$

Inversion of Equation 8 for d variables involves inverting the first variable's marginal distribution for a probability p_1. Since the marginals are all uniform on [0,1], the probability is equal to the first value, $x_1=p_1$. Equation 12 with j=1 is the conditional distribution from which x_2 can be determined from a second probability p_2. For variable 3, Equation 12 is of the form $B_n(f : \boldsymbol{x}_{(1,2)} \mid x_1, x_2)$. This process continues for all variables to give $\boldsymbol{x}$. Since $\boldsymbol{x}$ is a set of quantiles, $\boldsymbol{y}$ can be readily calculated knowing the actual marginal distributions.

Application: Permeability Tensors

A distribution model will be generated for nine variables that describe a set of full permeability tensors. Full tensors result because non-linear boundary conditions were applied in flow simulation. Permeability tensors were calculated for a set of 3-dimensional voronoi cells using a finite difference single phase flow solver. There were a total of 99 cells and 50 realizations resulting in 4950 tensors to use in generating a distribution model. 20 bins per variable were used for a total of 20^9 bins. Only 4902 actually contained data for calculating the densities.

Visualization will only be done on the bivariate level. Since it is somewhat complicated to decipher the quality of the distribution model from bivariate cumulative plots, sets of conditional distributions will be extracted for the three cases observed in Figure 1. 20 conditional distributions were calculated for each case: K_{zz} conditional to K_{xx}, K_{xy} conditional to K_{xx} and K_{xz} conditional to K_{xy}. These were joined with contours in Figure 5. Relationships that are apparent in Figure 1 are reproduced by the contour plots in Figure 5. There is some break down towards the extremes of the conditioning variables (those on the x axes). The bimodal structure in the $K_{xx} - K_{xy}$ and $K_{xy} - K_{xz}$ relationships is recovered by the model.

Some features of the input data were smoothed out by using Bernstein polynomials, which tend to have a slow convergence to the true underlying function. More accurate fit results can be obtained by using more bins in calculating the densities. However, processing time to evaluate cumulative probabilities from the model increases with number of bins (up to a maximum when each data point falls only in 1 bin). Some comments on improving model quality will be made below.

Conclusions and Future Work

The multivariate Gaussian assumption can be unacceptable for certain geostatistical applications. The normal score transformation only ensures univariate Gaussianity. By using non-parametric methods, more accurate relationships exhibited by multivariate data sets can be captured in bivariate and higher dimensions. The method described here using Bernstein polynomials can be applied to a large number of variables and can capture relationships with relatively few data. The number of samples used to model the nine-variable distribution for permeability tensors was low for that number of dimensions. Not having access to the true underlying distribution did not permit statistical testing for model quality; however, a simple visual study shows substantial information gain over the multi-Gaussian assumption.

The current methodology for modeling multivariate distributions can be improved. For example, the only knowledge that is applied to the model is Gaussian univariate marginals and the sample data. This can be enhanced by modeling multivariate marginals, starting with bivariate or dimension d=2, in a more accurate manner and using this to constrain the d=3 distributions and so on. Consider modeling the distributions shown in Figure 5 just as bivariate models with 100 bins per dimension rather than 20. Results are enhanced (Figure 6). Due to the current indexing methods 100 bins in all nine dimensions is not feasible. Indexing methods impede the use of too large a number of bins in all dimensions since 1-dimensional indexing is used. Use of vector indices will accommodate the use of more bins and provide a slight improvement in processing time since 1-dimensional indices will not have to be resolved to all other dimensions. There would be an increase in memory usage of $n(d\text{-}1)$ with d the number of dimensions and n the number of populated bins. It should also be noted that in order to evaluate a single cumulative probability, the number of populated bins must be cycled over completely. Methods to avoid this operation may be required especially for Monte Carlo simulation, when thousands of samples may be drawn from the distribution model.

References

Babu, J., Canty, A. J., Chaubey, Y. P., 2002, Applications of Bernstein polynomials for smooth estimation of a distribution and density function: Journal of Statistical Planning and Inference, No. 105, p. 377–392

Chiles, J. P., and Delfiner, P., 1999, Geostatistics: modeling spatial uncertainty: Wiley, 720 p.

Deutsch, C. V., Ren W., and Leuangthong O., 2005, Joint uncertainty assessment with a combined Bayesian updating/LU/P-field approach: Seventh Annual Report of the Centre for Computational Geostatistics, No. 203, p. 1–6

Deutsch, C. V., and Zanon S. D., 2007, Direct prediction of reservoir performance with Bayesian updating: JCPT, Vol. 46, No. 2

Johnson, R. A., and Wichern, D. W., 2002, Applied multivariate statistical analysis: Prentice Hall, 767 p.

Kakizawa, Y., 2004, Bernstein polynomial probability density estimation: Journal of Nonparametric Statistics, Vol. 16, No. 5, p. 709–729

Kolev, N., Anjos, U., and Mendes, B. V., 2006, Copulas: A review and recent developments: Stochastic Models, Vol. 22, No. 4, p. 617 – 660

Leuangthong,O., and Deutsch, C. V., 2000, Stepwise conditional transform for simplified cosimulation of reservoir properties: Second Annual Report of the Centre for Computational Geostatistics, No. 6, p. 1–14

Petrone, S., 1999, Bayesian density estimation using Bernstein polynomials: The Canadian Journal of Statistics, Vol. 27, No. 1, p. 105–126

Phillips, G. M., 2003, Interpolation and approximation by polynomials: Springer-Verlag, 328 p.

Rosenblatt, M., 1956, Remarks on some non-parametric estimates of density functions, Annals of Mathematical Statistics, Vol. 27, p. 832–837

Rosenblatt, M., 1952, Remarks on a multivariate transformation: Annals of Mathematical Statistics, Vol. 23, p. 470–472

Sancetta, A., 2007, Nonparametric estimation of distributions with given marginals via Bernstein–Kantorovich polynomials: L1 and pointwise convergence theory: Journal of Multivariate Analysis, No. 98, p. 1376 – 1390

Sancetta, A., and Satchell, S., 2004, The Bernstein copula and its applications to modeling and approximations of multivariate distributions: Economic Theory, No. 20, p. 535–562

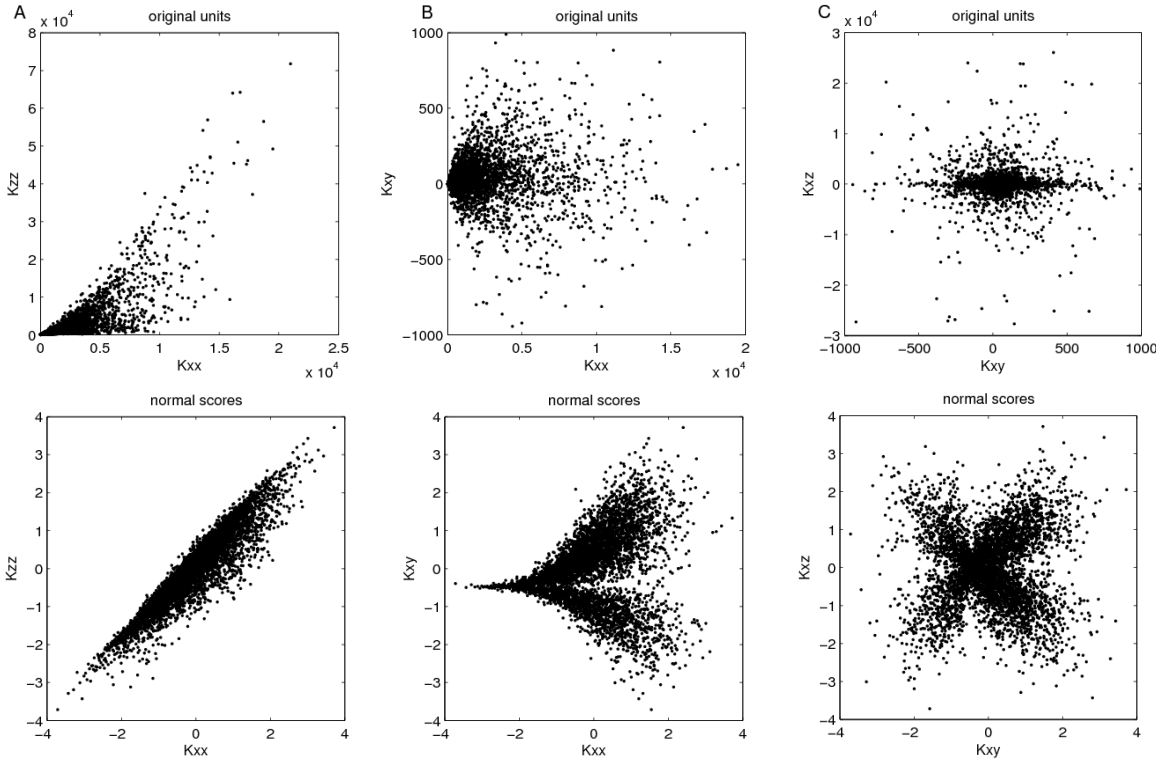

Figure 1: Normal score transform resulting in non-Gaussian bivariate distributions.

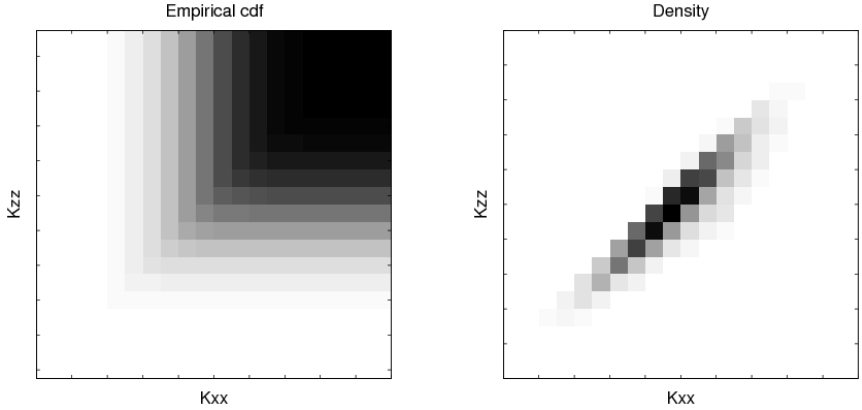

Figure 2: Bivariate empirical distribution and density function for Kxx with Kzz.

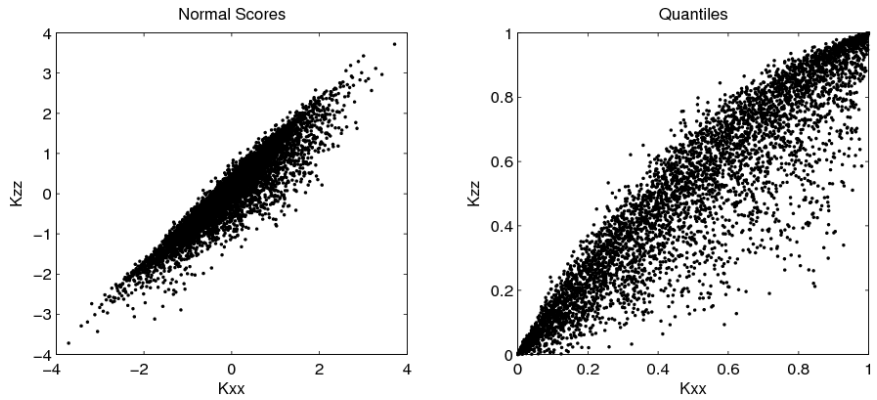

Figure 3: Transform from normal scores to quantiles for Kxx and Kzz.

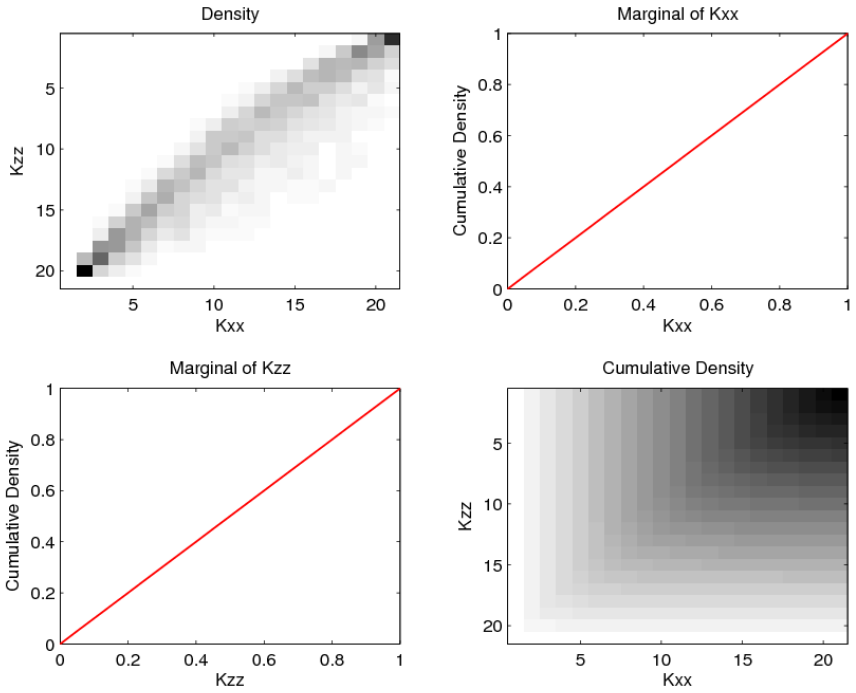

Figure 4: Quantile density, marginals and cumulative density for Kxx and Kzz.

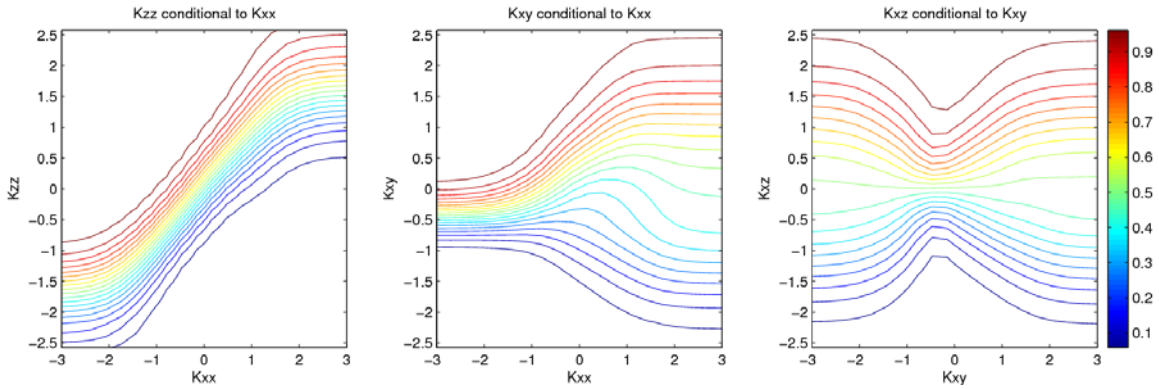

Figure 5: Cumulative conditional distributions using 20 bins per dimension.

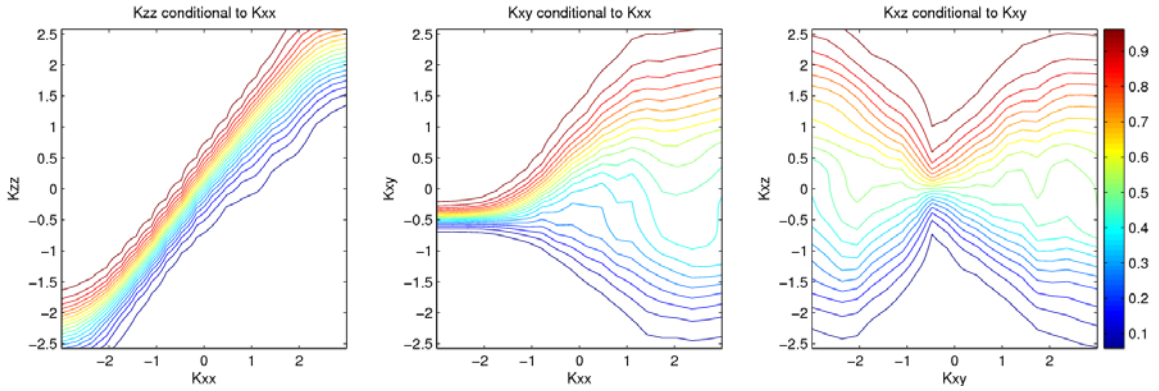

Figure 6: Cumulative conditional distributions using 100 bins per dimension