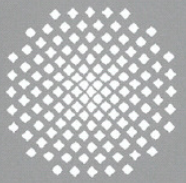
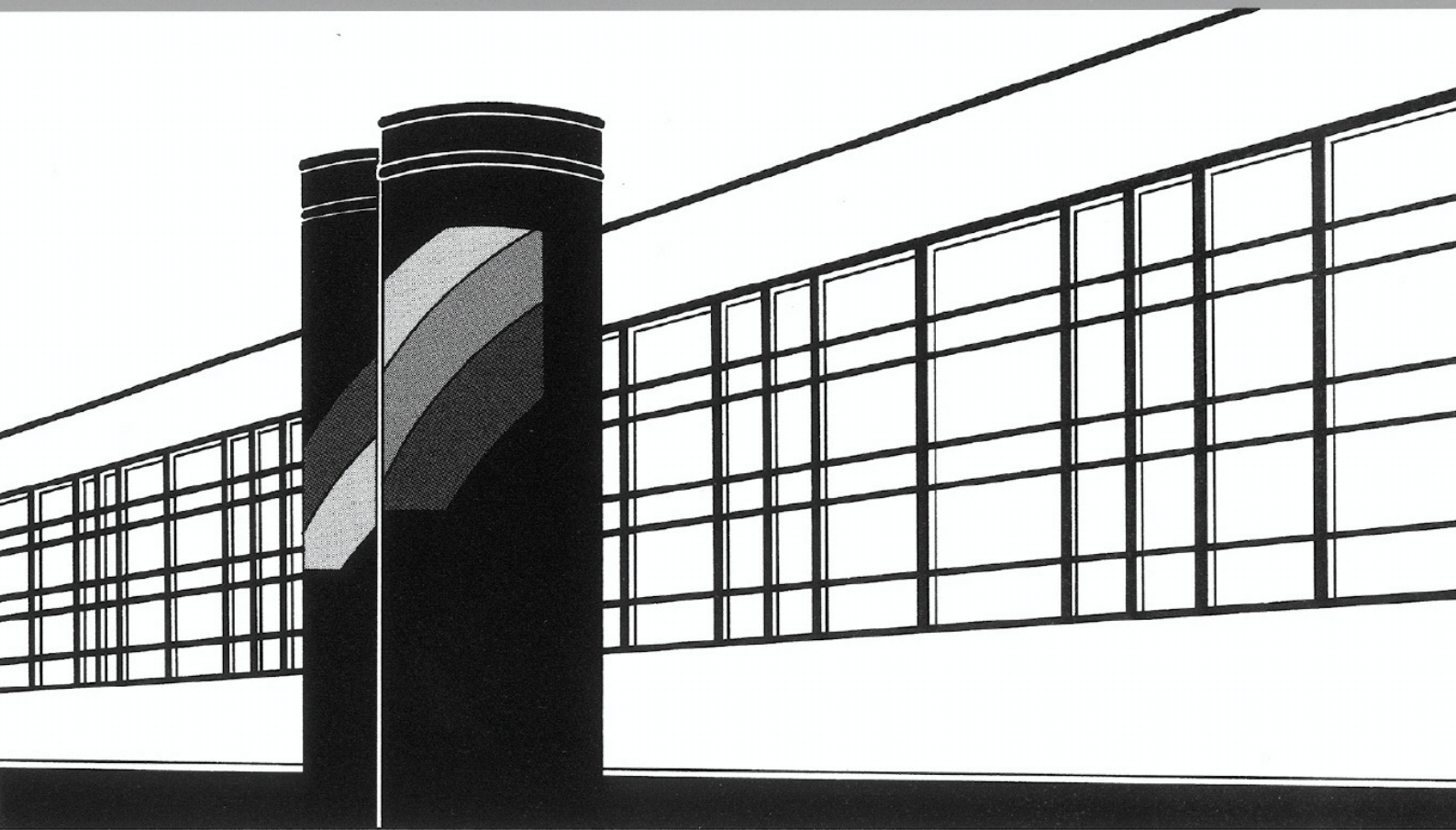


Universität Stuttgart



Institut für Wasser- und Umweltsystemmodellierung

Mitteilungen



Heft 225 Jhan Ignacio Rodríguez Fernández

High Order Interactions among
environmental variables: Diagnostics
and initial steps towards modeling

High Order Interactions among environmental variables: Diagnostics and initial steps towards modeling

Von der Fakultät Bau- und Umweltingenieurwissenschaften und dem
Stuttgart Research Centre for Simulation Technology
der Universität Stuttgart zur Erlangung der Würde eines
Doktor-Ingenieurs (Dr.-Ing.) genehmigte Abhandlung

Vorgelegt von
Jhan Ignacio Rodríguez Fernández
aus Caracas - Venezuela

Hauptberichter:	Prof. Dr. rer.nat. Dr.-Ing. András Bárdossy
Mitberichter:	Prof. Lelys Bravo de Guenni, Ph.D.

Tag der mündlichen Prüfung: 15. Oktober 2013

Institut für Wasser- und Umweltsystemmodellierung
der Universität Stuttgart
2013

Heft 225 High Order Interactions among
environmental variables:
Diagnostics and initial steps
towards modeling

von
Dr.-Ing.
Jhan Ignacio Rodríguez
Fernández

D93 High Order Interactions among environmental variables: Diagnostics and initial steps towards modeling

Gedruckt mit Unterstützung des Deutschen Akademischen Austauschdienstes

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://www.d-nb.de> abrufbar

Rodríguez Fernández, Jhan Ignacio:
High Order Interactions among environmental variables: Diagnostics and initial steps towards modeling von Jhan Ignacio Rodríguez Fernández. Institut für Wasser- und Umweltsystemmodellierung, Universität Stuttgart. - Stuttgart: Institut für Wasser- und Umweltsystemmodellierung, 2013

(Mitteilungen Institut für Wasser- und Umweltsystemmodellierung, Universität Stuttgart: H. 225)

Zugl.: Stuttgart, Univ., Diss., 2013

ISBN 978-3-942036-29-0

NE: Institut für Wasser- und Umweltsystemmodellierung <Stuttgart>: Mitteilungen

Gegen Vervielfältigung und Übersetzung bestehen keine Einwände, es wird lediglich um Quellenangabe gebeten.

Herausgegeben 2013 vom Eigenverlag des Instituts für Wasser- und Umweltsystemmodellierung

Druck: Document Center S. Kästl, Ostfildern

Acknowledgements

Many people should be thanked for their support to this work, I do not even pretend to be exhaustive in these acknowledgements. I would like to express my gratitude to the following people and institutions: to professors András Bárdossy and Lelys Bravo de Guenni for their guidance and recommendations, I have certainly learned much from them; to my colleagues of the IWS, who helped me feel at home at the Institute; to the DAAD for its extensive support during this project; to Dr. Hartmann of the international ENWAT program. On a more personal side, I thank my dear wife, Yosandra Sandoval for all her support and constant encouragement along these years in this new home we have had. This attainment is certainly hers, too. I also thank greatly my mother, Maritza Fernández, for instilling in me, many years ago (sometimes even now), the idea of the importance of learning and perseverance.

Contents

List of Figures	VII
List of Tables	XI
Abstract	XIII
Zusammenfassung	XV
1. Introduction	1
1.1. Relevance of the topic	1
I. Theory and Methodology	5
2. Preliminaries	7
2.1. Brief summary of the concept of statistical dependence	7
2.2. Dependence in the context of Spatial Statistics	9
2.2.1. Generalities	9
2.2.2. The Covariance Function	9
2.2.3. Dealing with Anisotropy	12
2.2.3.1. Geometric Anisotropy	12
2.2.3.2. Deformation Approach	12
2.2.4. Some Remarks	14
2.3. Difficulties in Extending Covariance	15
2.3.1. Conceptual	15
2.3.2. The "curse" of dimensionality	16
2.3.2.1. The number of parameters problem	16
2.3.2.2. The Number of data features problem	16
2.4. Dependence Quantification in 2-D	17
2.4.1. Product moment correlation coefficient	17
2.4.2. Spearman's and Kendall's coefficients	18
2.4.2.1. Spearman's correlation coefficient	18
2.4.2.2. Kendall's correlation coefficient	20
2.4.3. Copulas	21
2.4.3.1. Sklar's Theorem	23
2.5. The Edgeworth-Sargant distribution	25

3. The Proposed Approach: General	29
3.1. Application-relevant interaction manifestations	29
3.1.1. Interaction manifestations versus dependence structure	30
3.2. Joint cumulants as interdependence parameters	31
3.2.1. Definition and preliminaries	31
3.2.2. "Lancaster interactions" and joint cumulants	34
3.2.3. Relation of the Additive Interaction Measure with Joint Cumulants . .	36
3.2.3.1. Small digression: Alternative definition of joint cumulants .	38
3.3. Joint cumulants as parameters and relation to "interaction manifestations" . .	38
3.3.1. Preliminary: The Edgeworth Expansion and the Saddlepoint Approx- imation	39
3.3.2. Connection of dependence structure with interaction manifestations .	41
3.3.2.1. Connection of dependence structure with "joint" quantiles .	41
3.3.2.2. Connection of dependence structure with entropy	42
3.3.2.3. Connection of dependence structure with the distribution of the components sum	43
3.3.3. Putting the pieces together	46
4. The Proposed Approach: Spatial Statistics	49
4.1. Archetypal Dependence Structure	49
4.1.1. Moment generating function	51
4.1.2. Some advantages of the archetypal dependence structure	53
4.2. Useful representation of the archetypal model	54
4.2.1. Relation between R^2 and the archetypal c.g.f.	55
4.2.2. A convenient model for R^2	57
4.2.2.1. The data for estimating f_{R^2}	58
4.3. Parameter estimation	59
4.3.1. First step: Mean and Covariance determination	60
4.3.2. Second step: manifestations of higher order	61
4.4. More flexibility: Transformations on marginals	63
4.4.1. Quantile-Quantile Transformations	64
4.4.1.1. One-dimensional marginals of $\mathbf{X} \in \mathbb{R}^J$	64
4.4.1.2. The Transformation	64
4.4.2. Polynomial Transformations	66
4.4.2.1. Polynomial transformations	67
4.4.2.2. Dependence structure of the transformed vector	68
4.4.2.3. Fitting parameters "orthogonally"	69
4.4.3. Combinations of both types of transformations	70
4.5. Dealing with censored/truncated data	70
4.6. Simulation	71
4.6.1. Gibbs Sampler / Sequential simulation	71
4.6.2. Generating Variable method	72
4.6.2.1. Deviance from Normality	73

II. Examples and Illustrations	75
5. Two Random fields	77
5.1. Random Fields Set 1	77
5.1.1. Empirical Variograms	77
5.1.2. Marginal Distributions	78
5.1.3. Two, Three and Four dimensional Marginals	80
5.1.4. Interaction Manifestation: Sums of components	83
5.1.5. Interaction manifestations: A statistic built on the marginal joint probability distributions	84
5.2. Random Fields Set 2	90
6. Inference in a quasi-real setting	101
6.1. The simulated fields	101
6.2. Analysis	102
6.3. Estimating the parameters of the field	107
6.4. Inference for the whole field	112
6.4.1. (Partial) Inferential results	118
7. Summary and outlook	121
A. Joint cumulants derivation	123
B. Joint cumulants of transformed vectors	125
C. Outline of Estimation Procedure at section 6.3	129
D. Parameters fitted	131
Bibliography	133

List of Figures

2.2.1.Example site configuration. Circles represent sites with observed data, the “x” represent a site with not observation where an estimation is required. . .	10
2.2.2.Schematic process of deformation from Geographic to Dispersion spaces . . .	14
2.4.1.Two Data sets simulated from two different distributions having the same product moment correlation coefficient.	19
2.4.2.Plot of ranks of example data sets	19
2.4.3.Scaled ranks of 5000 values from a t-distribution (left) with 3 degrees of freedom and dispersion matrix $\Gamma = \begin{pmatrix} 1 & .472 \\ .472 & 1 \end{pmatrix}$, and a Gaussian distribution with mean $\mu = (0, 0)$ and covariance matrix Γ	22
5.1.1.Perfect Gaussian Random field (left). Sequentially simulated random field (right)	79
5.1.2.Field with 4-th (left) and 6-th (right) order non-zero joint cumulants	79
5.1.3.Field with 8-th (left) and 10-th (right) order non-zero joint cumulants	80
5.1.4.Empirical variograms from simulated fields. More convenient variant corresponds to scaling fields 4-D, 6-D, 8-D and 10-D. In this way all sequentially simulated fields possess the same empirical variogram.	81
5.1.5.Quantile-Quantile plots of values of simulated fields: all data (left), a randomly selected sample of size n=200 (right).	82
5.1.6.2-dimensional approximate representation of the six distances involved in the 4-dimensional marginal analysis	84
5.1.7.Histograms for Multivariate Shapiro’s test for Normality applied to the 2-dimensional marginal distributions. The test rightly rejects most of the time the hypothesis of normality.	85
5.1.8.Histograms for Multivariate Shapiro’s test for Normality applied to the 3-dimensional marginal distributions. The test rightly rejects most of the time the hypothesis of normality.	86
5.1.9.Histograms for Multivariate Shapiro’s test for Normality applied to the 4-dimensional marginal distributions. The test rightly rejects most of the time the hypothesis of normality.	87
5.1.10Comparison of sums of components for 2-dimensional marginals. Increase is appreciable for sums of components of non-Gaussian fields at the uppermost quantile.	88
5.1.11Comparison of sums of components for 3-dimensional marginals. Increase is appreciable for sums of components of non-Gaussian fields at the uppermost quantile.	88

5.1.12	Comparison of sums of components for 4-dimensional marginals. Increase is appreciable for sums of components of non-Gaussian fields at the uppermost quantile.	89
5.1.13	Comparison of congregation measure for marginals of different dimensions and data from the different fields. From left to right results of the comparison procedure are given for marginals of dimension 2, 3 and 4. Congregation seems to be always greater for the non-Gaussian fields.	90
5.2.1.	Perfect Gaussian Random field (left). Sequentially simulated random field (right)	91
5.2.2.	Field with 4-th (left) and 6-th (right) order non-zero joint cumulants	91
5.2.3.	Field with 8-th (left) and 10-th (right) order non-zero joint cumulants	92
5.2.4.	Empirical variograms from simulated fields. More convenient variant corresponds to scaling all fields: 2-D, 4-D, 6-D, 8-D and 10-D. In this way all sequentially simulated fields possess the same empirical variogram.	93
5.2.5.	Quantile-Quantile plots of values of simulated fields: all data (left), a randomly selected sample of size $n=200$ (right).	94
5.2.6.	Histograms for Multivariate Shapiro's test for Normality applied to the 2-dimensional marginal distributions. The test rightly rejects most of the time the hypothesis of normality, except for field 10D, which exhibits a similar rejection pattern as the 2D and the true Gaussian fields.	95
5.2.7.	Histograms for Multivariate Shapiro's test for Normality applied to the 3-dimensional marginal distributions. The test rightly rejects most of the time the hypothesis of normality. Random field 10D is roughly Gaussian, however.	96
5.2.8.	Histograms for Multivariate Shapiro's test for Normality applied to the 4-dimensional marginal distributions. The test rightly rejects most of the time the hypothesis of normality. Random field 10D is roughly Gaussian, however.	97
5.2.9.	Comparison of sums of components for 2-dimensional marginals. Increase is appreciable for sums of components of non-Gaussian fields at the uppermost quantile.	98
5.2.10	Comparison of sums of components for 4-dimensional marginals. Increase is appreciable for sums of components of non-Gaussian fields already at the 75% quantile.	98
5.2.11	Comparison of sums of components for 3-dimensional marginals. Increase is appreciable for sums of components of non-Gaussian fields at the uppermost quantile.	99
5.2.12	Comparison of congregation measure for marginals of different dimensions and data from the different fields. From left to right results of the comparison procedure are given for marginals of dimension 2, 3 and 4. Congregation seems to be always greater for the non-Gaussian fields.	99
6.2.1.	Simulation Scheme and kernel smoothing approximation to the density of $\sqrt{R_*^2}$	103
6.2.2.	Original Gaussian and Scaled sample Fields, after QQ transformation	105

6.2.3. From left to right and downwards: Box-plots of the sum of positive values and of number of components above 1.04, 1.28 and 2.5 on each field's realization. Regarding the number of components (locations) above the given thresholds, divergence between the Gaussian and non-Gaussian fields become more and more apparent as one moves towards the uppermost part of the marginal distribution.	106
6.2.4. Empirical copulas of data from locations 1,3 and 5 from the random fields simulated.	107
6.2.5. Entropy congregation measure applied to data from randomly selected triplets of locations. Box-plots are organized in terms of the size of the catheti of the right-angled triangles constituting the triplets. Fifty triplets were selected per distance category and their entropies computed. <i>Blue boxes</i> represent the results for the Gaussian field data. Selected quantile threshold was 99.5%. . . .	108
6.2.6. Ratio of the entropy measure computed for data from Gaussian to the entropy computed from Non-Gaussian data, at thresholds 90%, 99%, 99.5% and 99.9%. Locations for the triplet are 1,3 and 5. The increase in considerable above the 99% quantile.	109
6.3.1. Locations with "gaging stations" are marked with an x . Negative values were regarded as zero.	110
6.3.2. Data at five of the sites from which the field's characteristics are to be estimated.	111
6.3.3. St-EM produced chains for the mean, Nugget and range parameters.	113
6.3.4. Probability density (left) and distribution (right) functions of the squared generating variable, as estimated from the 30 Stations' dataset. Estimated distribution for R^2 appears in red as compared in its upper quantiles with the squared generating variables of a multivariate normal distribution (black) and that of a mult. Student r.v. with 15 degrees of freedom (blue).	114
6.3.5. Density (left) and probability distribution (right) of the squared deviance from normality random variable R_*^2	114
6.4.1. Left: Probability density function of the squared generating variable estimated for dimension $J = 90000$. Right: squared generating variables for a Gaussian (black), a Student-t with 15 degrees of freedom (blue), and the fitted generating variable (red). These variables are adapted to dimension $J = 90000$ of the random field.	117
6.4.2. Estimated density (left) and probability distribution (right) of the squared scaling variable for dimension $J = 90000$	117
6.4.3. From left to right and downwards: Box-plots of the sum of positive values and of number of components above 1.04, 1.28 and 2.5 on each field's realization. Regarding the number of components (locations) above the given thresholds, divergence between the Gaussian and non-Gaussian fields become more and more apparent as one moves towards the uppermost part of the marginal distribution. This figure corresponds to figure (6.2.3) of the original fields.	119

6.4.4. Entropy congregation measure applied to data from the same selected triplets of locations as in figure 6.2.5. Box-plots are organized in terms of the size of the catheti of the right-angled triangles constituting the triplets. <i>Blue boxes</i> represent the results for the Gaussian field data. Selected quantile threshold was 99.5%.	120
---	-----

List of Tables

2.4.1. Two example data sets for the sake of Kendall's τ illustration	20
2.4.2. The same as 2.4.1 but after rank ordering according to "x"	20
5.1.1. Random fields c.g.f coefficients configurations	78
5.1.2. Tests for equality in marginal distributions of the different fields analyzed. Proportion of times, out of 1000, in which p-value of the test was smaller than 0.05.	82
5.1.3. Interesting distances used for the multivariate marginals analysis	83
5.2.1. Random fields c.g.f coefficients configurations	90
5.2.2. Tests for equality in marginal distributions of the different fields analyzed. Proportion of times, out of 1000, in which p-value of the test was smaller than 0.05.	92
6.1.1. Coefficients producing R_*^2 for the fields example. Coefficients were fitted by the method of moments.	102
6.3.1. First nine m_k and c_k coefficients fitted on the basis of the 30 sites data. Original c_k interdependence coefficients of the c.g.f are also presented for compar- ison. These coefficients are clearly underestimated for $k \geq 2$, which calls for a second estimation step addressing specifically interdependence manifestations.115	
6.4.1. Estimated parameters for the squared generating variable, R^2 , of the 90000- dimensional non-Gaussian field. Estimated weights $\hat{\pi}_1$ through $\hat{\pi}_{86}$ are 0.00. .	116
D.0.1 Estimated weights for Gamma mixture representing squared generating vari- able of 30-D model.	132

Abstract

In the field of geostatistics and spatial statistics, variogram based models have proved a very flexible and useful tool. However, such spatial models take into account only interdependencies between pairs of variables, mostly in the form of covariances. In the present work, we point out to the necessity to extend the interdependence models beyond covariance modeling; we summarize some of the difficulties arising when attempting such extensions; and propose an approach to address these difficulties.

The necessity for extending covariance models, apart from the common sense notion that there can be more structure in a data-set than that expressed in terms of pairwise relations, has been suggested recently in the hydrological literature (Bárdossy and Pegram (2009, 2012)). For example, two multivariate data-sets/models with identical correlation matrices can exhibit systematically different congregation patterns, as expressed by entropy based measures applied to multivariate ($d \geq 3$) marginals.

An initial difficulty in trying to consider interdependence measures which go beyond pairwise measures, is to conceptualize what, say, a three-wise correlation coefficient might mean, or how is it to be interpreted. We suggest that joint cumulants are legitimate extensions of the covariance coefficients, since both represent the integral of a well known interaction measure (the Lancaster Interaction Measure); the covariance being the special case for $d = 2$. Then, from a more practical point of view, we suggest to address the issue of higher order interdependence via subject-matter relevant manifestations of such interdependence. Three example manifestations are provided, and their connection with multivariate joint cumulants is exhibited, namely: the distribution of the sum, the joint survival function, and the differential entropy of subsets S of the random vector representing the random field under study, where $\|S\| > 2$. The importance of the first of these for rainfall modeling is illustrated.

An important difficulty in trying to consider extensions to covariance models is the high dimensionality incurred. This high dimensionality is palliated by the use of low dimensional variogram models in traditional spatial statistics. By considering a cumulant generating function (c.g.f.) as a dependence structure, and introducing an archetypal c.g.f., we show that much of this low-dimensional approach can be kept, while allowing the explicit consideration of higher order interdependence. The issue of parameter estimation is dealt with, and three examples illustrate the consequences of manipulating joint cumulants on diverse interaction manifestations.

Finally, it is indicated how we can use this archetypal dependence structure (i.e., c.g.f.) together with marginal transformations, both monotonic and non-monotonic, in order to give more flexibility to the method, while retaining its low-dimensional desirable properties.

Zusammenfassung

Im Bereich der Geostatistik und der räumlichen Statistik haben sich variogrammbasierte Modelle in der Praxis als nützlich und flexibel erwiesen um räumliche Zusammenhänge zu beschreiben. Allerdings beziehen diese Modelle nur Wechselwirkungen zwischen Paaren von Variablen, vor allem in Form von Kovarianzen, mit ein. In dieser Dissertation weisen wir auf die Notwendigkeit hin, räumliche Modelle jenseits der Kovarianz Modellierung zu erweitern. Wir fassen einige Schwierigkeiten zusammen, die bei solch einer Erweiterung entstehen. Letztendlich präsentieren wir einen Ansatz, mit dem man diesen Schwierigkeiten behandeln kann.

Dass mehrdimensionale Datensätze eine Zusammenhangstruktur aufweisen können, die nicht ausschließlich mit Kovarianzen erfasst werden kann, mag selbstverständlich sein. Die Notwendigkeit, Kovarianz-basierte Modelle zu erweitern, ist beispielsweise neulich in der hydrologischen Fachliteratur aufgetaucht (siehe Bárdossy and Pegram (2009, 2012)). In dieser Arbeit wurde gezeigt, dass zwei multivariate Datensätze bzw. Modelle, identische Kovarianzmatrizen aufweisen, aber trotzdem sehr unterschiedliche Eigenschaften oder Cluster-Bildungen aufweisen können. Cluster-Bildungen werden dabei anhand eines Entropiemaßes, das dreidimensionale Randverteilungen umfasst, quantifiziert.

Eine der ersten Hürden für die Bildung eines Maßes, das die gegenseitige Abhängigkeit von mehr als zwei Variablen betrachtet, ist das Problem der Konzeptualisierung. Was soll eine Korrelation zwischen drei Variablen bedeuten? Wie soll man so etwas interpretieren? Wir schlagen in dieser Arbeit vor, dass multivariate Kumulanten eine legitime Erweiterung des Korrelationskoeffizienten bereitstellen, denn sie repräsentieren das Integral eines bekannten Wechselwirkungsmaßes, des "Lancaster Interaction Measure". Die Kovarianz ist dabei der Sonderfall für die Dimension $d = 2$.

Aus einer praktischeren Sicht betrachten wir die anwendungsspezifischen Wechselwirkungseigenschaften von Daten als Schlüsselkonzept zur Quantifizierung von Interaktionen. Die Kumulanten werden dabei daraufhin angepasst, die beschriebenen Wechselwirkungseigenschaften richtig widerzuspiegeln. Zur Veranschaulichung werden drei Wechselwirkungseigenschaften aufgeführt und deren Verbindung mit multivariate Kumulanten aufgezeigt. Die drei Wechselwirkungseigenschaften umfassen Teilmengen S von Komponenten eines stochastischen Vektors: die Verteilung der Summe, die multivariate Verteilung, und die Differentialentropie mehrerer Komponenten. Wobei $\|S\| > 2$. Die Relevanz der beschriebenen Wechselwirkungseigenschaften wird exemplarisch mit Niederschlagsmodellen veranschaulicht.

Eine weitere Hürde für die Erweiterung von kovarianzbasierten Modellen für Räumliche Statistik ist die Anzahl an Parametern, die angepasst werden müssen. Bei der Kovarianz Anpassung wird die Anzahl an Parametern üblicherweise mit Hilfe von niedrig-dimensionalen Variogrammodellen stark reduziert und kontrolliert. Wir stellen eine Abhängigkeitsstruktur in Form einer kumulantenerzeugenden Funktion (K.e.F) vor, die eine Anpassung von

Kovarianzen durch das Variogrammodell zulässt, aber auch das Anpassen von Kumulanten höherer Ordnungen (also, höher als 2) erlaubt. Solche Kumulanten höherer Ordnung können angepasst werden, um beobachtete, relevante Wechselwirkungseigenschaften besser zu reproduzieren. Möglichkeiten der Parameteranpassung für die vorgeschlagene Abhängigkeitsstruktur werden in dieser Dissertation ebenfalls behandelt.

Zur Veranschaulichung werden drei Beispielfelder betrachtet. Dabei experimentieren wir mit verschiedenen Konstellationen von Kumulanten höherer Ordnung und werten die Konsequenzen für bestimmte Wechselwirkungseigenschaften aus.

Abschließend wird gezeigt, wie die vorgeschlagene Abhängigkeitsstruktur (K.e.F.) in Verbindung mit monotonen oder nicht-monotonen Komponenten-Transformationen verwendet werden kann. Die vorgeschlagene Methode zur Generierung von hochdimensionalen Abhängigkeiten gewinnt damit an Flexibilität, während sie ihre niedrig-dimensionalen Vorteile beibehält.

1. Introduction

1.1. Relevance of the topic

From a purely methodological viewpoint, in statistics generally and in spatial statistics, in particular, very little work has been done to explicitly take into account inter-dependencies of more than two variables simultaneously. By taking it into account we mean diagnosing its existence, identifying its manifestations within specific subjects (e.g. in meteorology), quantifying its intensity and modeling it.

To our knowledge, only in the field of computational neuroscience there have been a consistent effort to address the issue of simultaneous interactions: Modern theories of the brain intend to use ensemble or groups of neurons as building blocks for representation and processing of information, rather than individual neurons. It is thus of interest to determine the cardinality or size of such groups, and the nature and dynamics of the interaction among its members Grün and Rotter (2010). The idea, according to Grün and Rotter (2010), dates back to an influential 1949 theory of behavior Hebb (2002). The fact that the area is still one of intense research suggests its complexity and potential. Though inspiring for the work below, the methods employed in parallel spike train (such is the field-name in Neuroscience) are not entirely adequate for spatial statistics.

In a context more related to spatial statistics, namely in that of rainfall modeling, Bárdossy and Pegram (2009) and Bárdossy and Pegram (2012) have raised the question of the need to consider dependence among more than two variables (i.e. Rainfall amounts at two locations) simultaneously. Bárdossy and Pegram (2009) propose a novel weather generation model. Their model reproduces well many important characteristics of rainfall adequately, such as mean daily precipitation, wet-dry spells, rank correlation among rainfall values, etc. However, interdependence is systematically underestimated, according to the entropy measure they employ for validation.

A similar situation was found in Bárdossy and Pegram (2012, 2011), where the authors deal with the rainfall output of three RCMs having a spatial resolution of $25\text{Km} \times 25\text{Km}$, and daily temporal resolution. They have also gauging-stations' daily rainfall data at their disposal, which they aggregate block-wise into a resolution of $25\text{Km} \times 25\text{Km}$. The RCMs outputs are corrected in such a way that the (marginal) probability distribution at each block, and the correlations among every two blocks of the area under analysis are exactly matched to those of the gauging stations-based block aggregations. By analyzing the sum of every four blocks of the bias-corrected variable, and comparing these sums with those found from gauge-based interpolation, it is found that the sums behaves in a different way regarding extreme values. That is, matching the correlation (pair-wise) characteristics of data does not automatically mean that we are having four-wise characteristics of data right, rainfall data is essentially higher dimensional than that. Bárdossy and Pegram (2012) employ the same entropy-based congregation measure as in their former research, Bárdossy and Pegram

(2009), and find also in this case that the congregation of every three sites interpolated values is significantly higher than the reconstructed values, even though the correlation among sites is exactly the same.

For the sake of clarity, we mention some instances of what we mean by d -wise characteristics: The entropy of d -dimensional marginals distributions, the behavior (distribution) of the sum of values at d sites, the simultaneous trespassing by d components of a vector over a threshold value, etc. The interest in one characteristic or the other will depend on the specific application motivating the analysis. We call in this research such characteristics *interaction manifestations*. Not considering such interaction manifestations may lead to considerable underestimation of subject-matter relevant statistics: consider in the context of rainfall modeling, for example, statistic T = "99% quantile of the sum of the positive components of a vector".

Although we deal in the present research with the typical problem of field estimation and interpolation, accounting for interactions that go beyond covariance is likely to be useful in various research areas where statistics plays an important role, such as Dowsaling, weather generator models, time series analysis and empirical finance, among others.

All the above suggests considering models that can reproduce relevant interaction manifestations. Some issues come immediately to mind. For example, given a random vector $\mathbf{X} \in \mathbb{R}^J$, how can one go from, say, the entropy of a 4-dimensional marginal to a model that reproduces such an entropy value? Even more important for Spatial Statistics: how to obtain a model that stays manageable as dimension increases (potentially letting $J \rightarrow +\infty$), and that can be extended consistently, so as to allow for interpolation into non-gauged sites? How can one build models that preserve the well-established techniques for first and second order statistics (e.g. mean and co-variance)?

In this research, we propose joint cumulants as building blocks for models that can address the questions posed above. In the context of spatial statistics, we propose a basic model, defined in terms of a cumulant generating function (c.g.f.), adequate for tackling the need of low dimensionality and consistent extension. This model is seen to be a natural extension to the Gaussian model, which currently dominates spatial and spatio-temporal statistics (cf. Cressie and Wikle (2011)). Parameter estimation and random simulation are addressed.

Two simulation based examples illustrate the possible implications for real environmental variable modeling. Gaussian and non-gaussian fields are simulated that look very similar regarding their one and two dimensional marginal distributions, but which exhibit very different interaction manifestations, as defined for each example. All the additional non-gaussianity is induced by manipulating joint cumulants of order greater than 4.

Extension possibilities, in order to give more flexibility to the model are considered. These are given in the form of monotonic and non-monotonic transformations applied to each marginal component.

Finally, the need to take into account subject-matter relevant interaction manifestations *explicitly* in estimation is acknowledged and illustrated. A course of action is given in the form of a two-step procedure, whereby the whole model is estimated at a first step (i.e. via Maximum Likelihood estimation). At a second step, lower order statistics (e.g. mean and co-variance) are held fixed, while parameters connected with higher order statistics are optimized, so as to make the interaction manifestations expected from the model as similar as possible as those observed in data. The structure of the suggested c.g.f. defining our basic

model is such that statistics of increasing order can be fitted “orthogonally”, that is, without altering the statistics of lower order (e.g. mean and co-variance).

We conclude this work with an outline of desirable future research.

Part I.

Theory and Methodology

2. Preliminaries

2.1. Brief summary of the concept of statistical dependence

The study of statistical dependence was begun by Francis Dalton (1822-1911) in the context of two related problems Pearson (2011): The influence of parents height on adult children's height, which lead to the concept of "regression", and the association among anatomical measures of the same individual, such as foot length, head length, stature, etc. This second problem led to the concept of correlation, since the different anatomical measurements had different scales; Galton had to standardize them (by subtracting the observed median of the respective measure to each observation and dividing the result by one half of the inter-quartile distance) and then apply his formerly found measure of "regression". In this way, even both interpretations assigned in practice to correlation, i.e. that of "partial causation" and that of association measure (not implying any sort of causation) were considered by Galton. Elementary though the correlation concept may appear today, it meant a revolutionary invention for the experimental science of Galton's time. We quote the view of Karl Pearson, mathematician and philosopher, usually called the founder of modern statistics Pearson (2011):

Up to 1889 men of science had thought only in terms of causation, in future they were to admit another working category, that of correlation, and thus open to quantitative analysis wide fields of medical, psychological and sociological research. [...] Galton, turning over two different problems in his mind, reached the conception of correlation: *A* is not the sole cause of *B*, but it contributes to the production of *B*; there may be other, many or few, causes at work, some of which we do not know and may never know. Are we then to exclude from mathematical analysis all such cases of incomplete causation? Galton's answer was: "No, we must endeavour to find a quantitative measure of this degree of partial causation". This measure of partial causation was the germ of the broad category, that of correlation, which was to replace not only in the minds of many of us the old category of causation, but deeply to influence our outlook on the universe. [...] The idea Galton placed before himself was to represent by a single numerical quantity the degree of relationship, or of partial causality, between the different variables of our ever-changing universe.

Thus, it was possible analyze rationally objects belonging to more complex systems than were conceivable before, knowing that we could not identify all causes determining their development (which for practical purposes are infinite) but that at least it was possible to quantify the partial influence from postulated important causes or "factors". The correlation coefficient proposed by Galton was further developed by K. Pearson into the so-called product moment correlation coefficient, which is widely used today. The paradigmatic step had

been made by Galton, though, acknowledging the importance and clarification potential of a “soft” or “probabilistic” or “average” type of causality and of association.

The concept was general enough to allow its application to the most diverse disciplines, from psychology to economics. Since the concept was already discovered, new measures expressing the concept, but adapted to specific applications or created to solve any specific data type difficulty, appeared in the course of time. Both the Spearman’s ρ and the Kendall’s τ coefficients were developed by Spearman and Kendall, respectively, with a view to their application in Psychology (Spearman (1904), Kendall (1938)), whereas Gini’s γ was developed in the context of economics. Other measures of dependence include Blomqvist Beta Blomqvist (1950), Goodman and Kruskal’s τ and γ (for categorical data), etc.

As usual in mathematics, concepts are first developed in connection with a specific application and then they are generalized and their theory systematized. In a theoretical paper, Rényi (1959) formulated seven “rather natural postulates” to be fulfilled by a reasonable measure of dependence. The postulates of Rényi are rather restrictive and he mentions only one measure of dependence fulfilling his postulates (the “maximal correlation coefficient”, due to Gebelein (1941)). Schweizer and Wolff (1981) build on Rényi’s work and come up with copula-based measures of dependence which fulfill “reasonable modifications” of Rényi’s seven postulates.

As a final step in this brief summary of theoretical works on stochastic dependence, we mention the paper due to Schmid et al. (2010). In this paper, the authors gather a set of desirable properties for dependence measures proposed in the literature and introduce extensions to well-known measures of bi-variate dependence, such as Spearman’s ρ , Kendall’s τ , Blomqvist’s Beta, Gini’s γ . They also survey other types of measures, such as measures based on information/entropy and on distances between distributions. For each explained dependence measure, the authors check which of the listed properties are fulfilled; it is noteworthy that none of the measures fulfill all properties listed, but different measures fulfill different sub-sets of them. The unifying concept of all these measures is that they are expressed in terms of the copula of the random vector’s distribution under study. Estimation is performed non-parametrically via estimation of the empirical copula. The paper of Schmid et al. (2010) was inspiring for the present work: they generalize many well known measures of dependence using the copula as unifying element and pay attention to theoretically attractive properties such a measure should have, at the same time. Moreover, these measures consider more than just pair-wise dependence. Concerning spatial statistics, however, the non-parametric nature of these dependence measure’s estimation makes them unsuitable for model-building, which in turn is a complication if we are to impute the value of the random variable in an ungauged site (interpolation), or if we want to have a forecast of the random field conditioned on additional variables (say, circulation patterns), since in this case we have no parameters to “connect” the response (say, rainfall) with the conditioning variables.

The approach or guiding principles employed in this work can be summarized as follows: We try to provide a theoretical, logically appealing basis for our dependence quantification and modeling method, but we are also concerned with the specific case of spatial data analysis, as outlined in the next sections. Application-relevance is for us primary, theory compliance is complementary.

2.2. Dependence in the context of Spatial Statistics

2.2.1. Generalities

In the statistical practice, “we typically start with a subject-matter question. Data are or become available to address this question” Cox (2006). Sometimes, such data considered relevant for the research question(s) possess labels indicating their location in space. If statistics can be described as a methodology for “utilizing data to make inferences about un-measured quantities” Kitanidis (1997), then the geographic principle that locations in space that are closer tend to look or react more similar must be taken into account. This effect of the location at which an event or measurement occurs, its identification, quantification and modeling is the object of Spatial Statistics.

A generic Spatial model can be defined as follows Cressie (1991): Let $s \in \mathbb{R}^g$ be a generic location in the G -dimensional Euclidean space and suppose the potential datum $\mathbf{Z}(s)$ at spatial location s is a random quantity. Let s vary over an index set $\mathbf{D} \subset \mathbb{R}^g$, then a multivariate random field is generated

$$\{\mathbf{Z}(s) : s \in \mathbf{D}\} \quad (2.2.1)$$

For example, if $\mathbf{D} = \{s_1, s_2, s_3, \dots, s_N\}$ is a fixed finite set, we obtain the random field

$$(\mathbf{Z}(s_1), \dots, \mathbf{Z}(s_N)) \quad (2.2.2)$$

in case set $\mathbf{D} \subset \mathbb{R}^g$ is countably infinite and fixed, the generated field can be written as

$$(\mathbf{Z}(s_1), \mathbf{Z}(s_2), \dots) \quad (2.2.3)$$

The above two instances are usually called “lattice data models”. If set $\mathbf{D} \subset \mathbb{R}^g$ is fixed and non-countable, then the model is a “Geostatistical model”, the name coming from the original context in which this type of model was first developed. If the index set $\mathbf{D} \subset \mathbb{R}^g$ is not provided in advance, and the model for the process of interest can be decomposed into two steps: 1. a location s is generated on \mathbb{R}^g ; 2. given location s , a random quantity is generated $\mathbf{Z}(s)$. Then the model is a “Point pattern” model; important instances of this model are the “Poisson point process” and the “Cox Process” Cressie and Wikle (2011), in which only the location generation mechanisms are random, and $\mathbf{Z}(s) = 1, \forall s$. Those models do not exhaust the Spatial Statistics types of models. The reader is referred to Cressie and Wikle (2011) for more details. We focus in this work on the Lattice and the Geostatistics models.

The key issue regarding these models is that the field is only partially observed, i.e. observed at some locations or sites only. The values at sites where no observations are available must be estimated on the basis of the observed values and the dependence structure of the multivariate field. This dependence structure, which takes into account the position of the sites, must also be inferred from the observed data.

2.2.2. The Covariance Function

There follows an example that intends to introduce the topic and possible issues in the types of Spatial Statistics models dealt with in this dissertation. It is simple enough not to distract the flow of reading excessively, but is likely to provide the necessary background for the

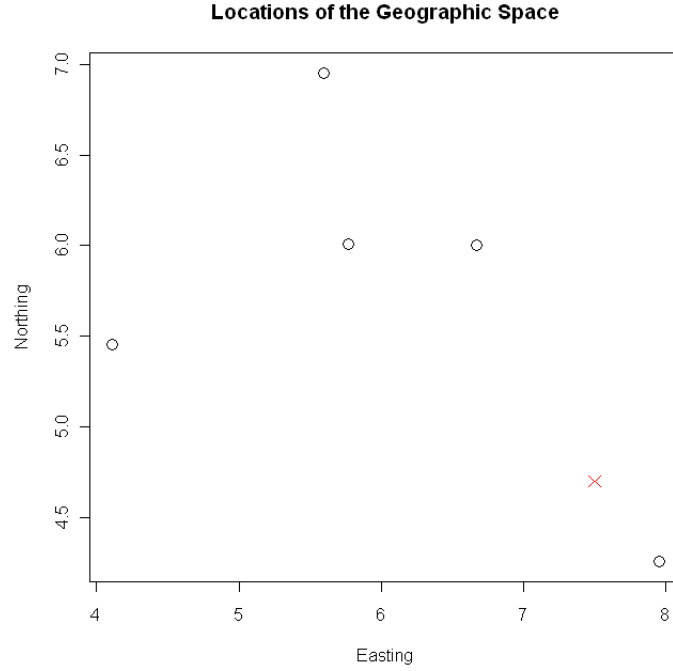


Figure 2.2.1.: Example site configuration. Circles represent sites with observed data, the “x” represent a site with not observation where an estimation is required.

rest of the work. The kind of model dealt with in this research falls in the sub-class of those for which a valid covariance function can be found, probably after describing the expected value of $\mathbf{Z}(s)$ by means of a deterministic function of external variables and location s (such as geographical variables: elevation, coordinates of site location, etc.). This kind of model is called a “second order stationary model” in the literature; the reader is referred to the specialized literature for more details (Cressie and Wikle (2011); Cressie (1991); Diggle and Ribeiro (2007)).

We see in figure 2.2.1 five sites on the plane at which observations are available, represented by circles, and one site at which no observation is available but an estimation is required, represented by an “x”. This is an example of a lattice model with a location set $\mathbf{D} \subset \mathbb{R}^g$, $g = 2$ of 6 elements, where the value of the random quantity $\mathbf{Z}(s)$ is not observed at one location. A possible solution to this problem is to set up a multidimensional probability model such as, for example, the multivariate Gaussian model:

$$(X_1, \dots, X_6)' \sim N_6(\mathbf{0}, \Gamma_{6 \times 6}) \quad (2.2.4)$$

where $X_j = \mathbf{Z}(s_j)$, $j = 1, \dots, 6$, $\mathbf{0}$ is a vector of zeros, and Γ is a 6×6 covariance matrix. This would imply that the marginal distribution of the random quantity at each location is Gaussian with zero mean. Moreover, it is a well known result that the conditional distribution of $X_6 \mid X_j, j \neq 6$, is normally distributed with mean

$$\mu_6^* = \Gamma_{12} \Gamma_{22}^{-1} (x_1, \dots, x_5)'$$

and variance

$$\sigma_6^{2*} = \Gamma_{66} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}$$

where each x_j represent the observed value of $X_j = \mathbf{Z}(\mathbf{s}_j)$, and

$$\begin{aligned}\Gamma_{12} &= (\Gamma_{62}, \dots, \Gamma_{65}) \\ \Gamma_{22} &= \begin{pmatrix} \Gamma_{22} & \dots & \Gamma_{25} \\ \vdots & \ddots & \vdots \\ \Gamma_{52} & \dots & \Gamma_{55} \end{pmatrix} \\ \Gamma_{21} &= (\Gamma_{26}, \dots, \Gamma_{56})'\end{aligned}$$

In this manner, it is possible to obtain an estimate for $X_6 = \mathbf{Z}(\mathbf{s}_6)$ and even a measure of the quality of the estimation, namely σ_6^{2*} . If the association between values of the random quantity at two locations, $\mathbf{Z}(\mathbf{s}_i)$ and $\mathbf{Z}(\mathbf{s}_j)$, can be assumed to be reasonably determined by the distance between \mathbf{s}_i and \mathbf{s}_j , then the geographic principle demanding that closer things look more similar can be represented by a correlation matrix $\rho_{6 \times 6}$, with entries $\rho_{ij} = \Gamma_{ij} / \sqrt{\Gamma_{ii}\Gamma_{jj}}$ that are described by some decreasing function h of the (euclidean) distance between sites $\mathbf{s}_i = (s_{i1}, s_{i2})$ and $\mathbf{s}_j = (s_{j1}, s_{j2})$:

$$\rho_{ij} = \text{corr}(\mathbf{Z}(\mathbf{s}_i), \mathbf{Z}(\mathbf{s}_j)) = h(\text{Dist}(\mathbf{s}_i, \mathbf{s}_j)) \quad (2.2.5)$$

A field whose correlations can be expressed in this way is said to be isotropic. If all variances Γ_{jj} , $j = 1, \dots, 6$, are considered equal on the basis of some analysis specific subject-matter considerations, then an adequate and equivalent means of representing the geographic principle mentioned above is the *isotropic covariance function*, C :

$$\Gamma_{ij} = \text{cov}(\mathbf{Z}(\mathbf{s}_i), \mathbf{Z}(\mathbf{s}_j)) = C(\text{Dist}(\mathbf{s}_i, \mathbf{s}_j)) \quad (2.2.6)$$

and it is clear that

$$h(\text{Dist}(\mathbf{s}_i, \mathbf{s}_j)) = C(\text{Dist}(\mathbf{s}_i, \mathbf{s}_j)) / C(0) \quad (2.2.7)$$

Since usually matrix Γ of the example is not known in advance, it must be estimated from observed data (which data comprises the locations of the observed values). The representation of covariance at (2.2.6) is desirable, since: 1. If only one observation is available at each of the sites providing the data, then it is impossible to estimate the covariance matrix of the model without postulating assumptions on the nature of the spatial dependence. Dependence as a function of distance seems a harmless assumption. 2. If an adequate function C is found which depends on a reduced number of parameters (such as two or three, as below), then the number of parameters to estimate in order to have an estimation of $\mathbf{Z}(\mathbf{s}_6)$ is small, providing a parsimonious model; 3. If the dependence structure, i.e. the covariance matrix, can be represented as a function of distance, then in principle the value of the random quantity $\mathbf{Z}(\mathbf{s})$ can be estimated at any point on the plane \mathbf{s} . Thus a whole field on the plane can be obtained, as in the Geostatistics models.

However, this covariance function C cannot be an arbitrary function, since the resulting covariance matrix Γ must be positive definite, namely

$$\mathbf{v}\Gamma\mathbf{v}' > 0, \quad \forall \mathbf{v} \in \mathbb{R}^6$$

A number of functions satisfying this requirement have been developed, two of which are:

Powered-exponential: given by equation

$$C(d) = \sigma_0^2 \cdot I(d = 0) + \sigma_1^2 \exp\left(- (d/\theta_1)^{\theta_2}\right) \quad (2.2.8)$$

where $I(*)$ stands for the indicator function.

Matern's: given by equation

$$C(d) = \sigma_0^2 \cdot I(d = 0) + \sigma_1^2 \left[2^{\theta_2-1} \Gamma(\theta_2) \right]^{-1} [d/\theta_1]^{\theta_2} K_{\theta_2}(d/\theta_1) \quad (2.2.9)$$

where $\Gamma(*)$ stands for the Gamma function and $K_{\theta_2}(d/\theta_1)$ for the modified Bessel function of the second kind of order θ_2 (see, for example Abramowitz (1972)).

Parameters $(\theta_1, \theta_2, \sigma_0^2, \sigma_1^2)$ are the covariance function parameters, and must usually be estimated on the basis of observed data by means of a computerized optimization procedure. If a model such as (2.2.4) is employed, maximum likelihood estimation is a possibility.

2.2.3. Dealing with Anisotropy

Sometimes it is unreasonable to assume that the association between the values at two sites is merely a function of distance. For example, if the circles at figure 2.2.1 represent air pollution gauges and wind flow is known to be predominantly in a specific direction for the times estimation at point “x” is supposed to be made, then the isotropic assumption is not reasonable. If this wind direction is, for example, from North-West to South-East, then the four locations at the upper part of the figure are somehow drawn “closer” to the estimation or prediction location, by the effect of the wind, than the site located at the lowest part of the plane. The random field is said to be anisotropic in such a case. Sometimes it is possible to transform the original “geographic plane” into a “dispersion plane”, where isotropy holds, compute the estimation at the ungauged location, and then transform back into the geographic space. This is the idea underlying the two methods mentioned subsequently.

2.2.3.1. Geometric Anisotropy

Define $\mathbf{d}_{ij} = \mathbf{s}_i - \mathbf{s}_j$ to be the lag vector obtained by (component-wise) subtracting two location vectors. This lag has not only magnitude $Dist(\mathbf{s}_i, \mathbf{s}_j)$ as above, but also a direction. If there exists an invertible matrix \mathbf{A}_{gsg} , such that process $\mathbf{Z}(\mathbf{A}\mathbf{s})$ is isotropic, then the process is said to be *geometrically anisotropic*, and its covariance function can be written as

$$C^*(\mathbf{d}_{ij}) := C(Dist(\mathbf{A}\mathbf{s}_i, \mathbf{A}\mathbf{s}_j)) \quad (2.2.10)$$

by using any of the available isotropic covariance functions.

2.2.3.2. Deformation Approach

Another approach, which we shall name the *Sampson-Guttorp* or *deformation approach* (see Sampson and Guttorp (1992) and Schmidt and O'Hagan (2003)), considers more general transformations. It is assumed in this case that a series of observations is available at each

gauging site, and thus it is possible to compute sample covariances for the values at these sites, $\hat{\Gamma}_{ij}$, $i, j = 1, \dots, J$. The covariance matrix thus obtained can be interpreted as a similarity matrix, and this idea is implemented by constructing a *distance matrix* given by $D_{ij} = \sqrt{\hat{\Gamma}_{ii} + \hat{\Gamma}_{jj} - 2\hat{\Gamma}_{ij}}$. Note that in the presence of anisotropy, the distances at D_{ij} do not correspond to a monotonic function of the euclidean geographical distances $Dist(s_i, s_j)$, as would be the case in the isotropic case. This new distance matrix is inputted into Kruskal and Shepard's non-metric multidimensional scaling procedure (see Kruskal (1964)), thus obtaining a new 2-dimensional representation of the original J gauged locations, that we can call dispersion locations, in accord with Sampson and Guttorg's exposition. The idea is that, by using multidimensional scaling on the *distance matrix*, the original geographic locations have been mapped from the *Geographic Space* into a *Dispersion Space*. On this Dispersion Space, the covariance between values at any two sites is approximately a function of the distance between them¹, so that an isotropic covariance function such as in (2.2.8) or (2.2.9) can be fitted on the basis of available data. It is relatively simple now to work in the dispersion space, by using an isotropic model: at any new location on the Dispersion Space, the value of the random quantity can be estimated as above. However, it is at the Geographic Space that the estimation is required, and thus one must be able to identify the coordinates transformation T implicitly performed by multidimensional scaling. Namely, we have proceeded as represented from left to right at figure 2.2.2, implicitly defining a Dispersion Space $\mathbf{D}^* = T(\mathbf{D})$, where \mathbf{D} is our original geographic locations set. Transformation T has been implicitly defined by finding a distance matrix on the basis of observed covariances and then producing a 2-dimensional representation of locations possessing (approximately) such distances among them. Sampson and Guttorg use multidimensional thin-plate Splines to approximate the mapping

$$T : \mathbb{R}^2 \rightarrow \mathbb{R}^2 \quad (2.2.11)$$

on the basis of the two sets of coordinates available. That is, we have two sets of locations (s_1, \dots, s_J) and $(T(s_1), \dots, T(s_J))$, the second one obtained from the multidimensional scaling procedure. With the Splines, we find a parsimonious representation \hat{T} of T , such that $\hat{T}(s_i) \approx T(s_i)$. The theory of Splines implies that extrapolation to points outside the observed ones, (s_1, \dots, s_J) , is reasonable.

Summarizing, assume a series of observations of an interesting random quantity is available at each of J sites, (s_1, \dots, s_J) . For a new ungauged site such as "x" in figure 2.2.1, an estimation of the random quantity can be obtained as follows:

1. Identify the new location at which an estimation is required, s_{new} , on the Geographical Space.
2. Use observed data (possibly after some pre-processing to remove temporal effects, etc.) to compute sample covariances, $\hat{\Gamma}_{ij}$, for $i, j = 1, \dots, J$.
3. Build distance matrix $D_{ij} = \sqrt{\hat{\Gamma}_{ii} + \hat{\Gamma}_{jj} - 2\hat{\Gamma}_{ij}}$, from the computed covariances.

¹Given a $J \times J$ distance matrix M , Multidimensional Scaling returns a set of J vectors (v_1, \dots, v_J) on the n -dimensional space, $n \leq J$, such that the distances matrix $M_{ij}^* = Dist(v_i, v_j)$ is as close as possible to M . In general, however, equality $M_{ij} = M_{ij}^*$ for all $i, j = 1, \dots, J$ can only be obtained when $n \geq J - 1$.

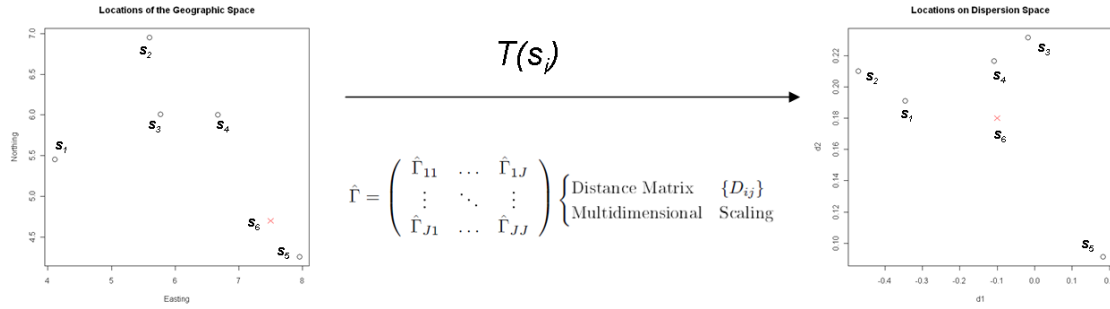


Figure 2.2.2.: Schematic process of deformation from Geographic to Dispersion spaces

4. Obtain J new locations by performing multidimensional scaling on matrix $\{D_{ij}\}$. This results in a new set of locations on the Dispersion Space, $(T(s_1), \dots, T(s_J))$.
5. Find the estimate \hat{T} of T , such that $\hat{T}(s_i) \approx T(s_i)$, for $i = 1, \dots, J$, using Splines.
6. Find the corresponding location of s_{new} on the Dispersion Space by using \hat{T} , namely set $s_{\text{new}}^* = \hat{T}(s_{\text{new}})$.
7. Compute the estimation of the random field at the new dispersion location, s_{new}^* , by using an isotropic model in the Dispersion Space. The estimation obtained is our estimation for $\mathbf{Z}(s_{\text{new}})$.

2.2.4. Some Remarks

Second order stationary models must not be necessarily Gaussian, as in (2.2.4), although this is a widely used model, even after suitable transformation of observed data. Copula based models are also available, whereby the distribution of the ranks is addressed independently of the marginal distributions in a multivariate model such as (2.2.4). This type of model has been used recently in Spatial Statistics (e.g. Bárdossy and Li (2008); Bárdossy and Pegram (2009)) with very good results.

It was seen that the covariance function (or the variogram, in the more general “intrinsically stationary” model, not addressed here) is a powerful tool for dealing with a potentially infinite random field in a parametrically low dimensional way. Actually, computing covariances among the random field’s components becomes unfeasible, both computationally and in terms of the data required, without the aid of such a tool. For example, a set up with $J = 100$ sites would require to estimate $\frac{J(J+1)}{2} = 5050$ components of the covariance matrix, whereas in the isotropic case, using one of the covariance functions of the above section, the number of parameters to estimate is just 4.

Covariance quantification is a central topic in identifying dependence structure in Spatial Statistics, even though it addresses interdependence between two variable at a time only. The issue of the possible high dimensionality must be kept in mind if we intend to extend the concept of interdependence to more than two variables simultaneously.

2.3. Difficulties in Extending Covariance

2.3.1. Conceptual

The possibilities for finding a measure for interdependence among a set of variables are infinite. Even for the case of two variable, many measures are available, as stated in section 2.1. It seems reasonable to begin with the most common measures of covariance and correlation. Additionally, these measures are key measures for Spatial Statistics, which is the focus of the present research.

Having chosen to extend the concept of covariance in order to have a more detailed picture of dependence, the natural question is: how could the covariance coefficient be extended? The covariance coefficient is defined by

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y) \quad (2.3.1)$$

Asking themselves this very question, Staude et al. (2010) note that a straightforward extension of (2.3.1) could be

$$\text{cov}(X, Y, Z) = E(XYZ) - E(X)E(Y)E(Z) \quad (2.3.2)$$

If two random variables, X and Y are independent, then it is well known that $\text{cov}(X, Y) = 0$. If three random variables are mutually independent, it can be seen that the covariance coefficient (2.3.2) would be zero. However, if X is independent of both Y and Z , but $\text{cov}(Y, Z) \neq 0$, then it can be seen that $\text{cov}(X, Y, Z) \neq 0$. Thus, our provisional 3-wise "covariance" coefficient is not zero even though there is no set of three variables interacting. If, given a random vector (X_1, \dots, X_J) , we seek to find the size of the smallest set of non-independent components, a coefficient fulfilling $\text{cov}(X_{i_1}, \dots, X_{i_k}) = 0$ whenever a subset of the random variables in the set $\{X_{i_1}, \dots, X_{i_k}\}$ (for $1 \leq i_k \leq J$) is independent of another can be stated to be a desirable measure. With it, a kind of interdependence index = "size of greatest subset of interacting components", for example, can be used to rank the random vectors in term of dependence.

Assuming for the moment that we have found such a coefficient, the next question is: How is it supposed to be interpreted? The interpretation of the standard covariance is relatively straightforward, and its scaled version, the correlation coefficient

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}} \quad (2.3.3)$$

allows us to form a quick idea of the type of association between X and Y , by means of its sign and its proximity to -1 or 1, even though one must always remember that ρ is not a good measure for every type of dependence. If data is available that can be considered as realizations of X and Y , ρ is of course more revealing when an x-y dispersion plot shows data falling roughly on a line on the x-y plane. A simple x-y dispersion plot of the data under study helps find out whether ρ is reliable or not as a dependence indicator.

But to interpret a value of, say, $\text{cov}(X_1, X_2, X_3, X_4) = -1$, requires much imagination. Is this a "high" or a "low" value, indicating "intense" or "weak" association? Is it possible to find a scaled, more useful, analog to ρ ? If our coefficient can be interpreted as a measure of

linear dependence, just like the standard covariance, best suited for the case in which data values fall roughly on a 4-dimensional hyper-plane $P \subset \mathbb{R}^5$: how can this be visualized? It is clear that there are conceptual difficulties, without saying anything yet about applicability, when trying to extend even the simple concept of correlation to more than two dimensions.

2.3.2. The "curse" of dimensionality

2.3.2.1. The number of parameters problem

Interpreting a coefficient such as $cov(X_1, X_2, X_3, X_4)$ has its difficulties. Another difficulty is: How many coefficients of this type can be calculated for a random vector of J components? Only coefficients of the form $cov(X_i, X_j)$ and $cov(X_i, X_j, X_k, X_l)$, with $1 \leq i, j, k, l \leq J$, are selected for exploration. The number of covariance coefficients for a vector of J components is of course $\frac{J(J+1)}{2}$. Assuming symmetry on the 4-wise covariance coefficient (the order of the indexes does not alter the value of the measure), the number of four-wise covariance coefficients can be found to be $\frac{J^3(J+1)}{2}$. These numbers, for J set at 4, 6 and 8, are: 10, 21 and 36 for the case of coefficients of the form $cov(X_i, X_j)$; and 160, 756 and 2304 for coefficients of the form $cov(X_i, X_j, X_k, X_l)$. Thus, the number of coefficients to estimate increases rapidly with the dimension of the field, J .

The above is an instance of the so called "curse of dimensionality" (e.g. kot (2006)), in the form of huge data requirements if all these coefficients are to be estimated without any further assumption. An assumption such as isotropy and the subsequent introduction of valid covariance functions was found to be essential for recovering spatial dependence, as quantified by covariances among the components of the random field. An analogous idea is required for quantification of higher order dependence parameters, or at least a judicious exploitation of the covariance function in combination with other assumptions (a suggestion in this last direction, can be found in this work).

2.3.2.2. The Number of data features problem

Another instance of the "curse of dimensionality", is the number of potentially interesting characteristics of the random field representing the process of interest. In the one dimensional case, these characteristics are relatively few, and they are easy to conceptualize or "paraphrase": We have location, dispersion, skewness, and kurtosis as characteristics of the probability distribution representing the output of some process, which convey usually much of the information required in practice. They can be easily visualized with the aid of a plot of the frequency curve of the distribution, and so "paraphrased": around what point are the realizations of the random variable are mostly concentrated (location), how much information gives us the location measure (dispersion), the degree of symmetry of the distribution (skewness), to what extent one can expect occasional values observed "far away" (i.e. in terms of a Normal distribution with the same dispersion) from the location value (kurtosis). Additionally, these measures are readily expressed in terms of the first four moments of the distribution; one could practically say that the measures indicating to what extent those interesting characteristics belong to the probability distribution *are* the moments of the distributions.

As dimension of the random variable increases, the number of potentially interesting characteristics increases. The richness of higher dimensional models can be seen, even staying within the mind-set of characteristics of one-dimensional probability distributions, by considering the full conditional distributions of the form

$$\Pr(X_j \leq x_j \mid X_1 = x_1, \dots, X_{j-1} = x_{j-1}, X_{j+1} = x_{j+1}, \dots, X_J = x_J) \quad (2.3.4)$$

There are as many as J distributions of this form, with their specific characteristics, for *every* observed vector $(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_J)$. Conditioning variables can of course be lumped into classes, reducing the amount of possible conditional distributions to analyze. But there can be anyway a great number of 1-dimensional conditional distributions, each of which one may wish to analyze in terms of its first four moments, for example. Additionally, in practice, some measure quantifying association is necessary among at least two variables, and this cannot be addressed conveniently by moments of the 1-dimensional conditional distributions.

For the two dimensional case, the characteristics of the association can be captured to some extent with the aid of dependence coefficients such as Pearson's product moment correlation, Spearman's rank correlation ρ , Kendall's τ , etc. But the richness of the possible types of dependence can be hardly captured by a single coefficient: several distributions can possess the same dependence coefficients. We focus on the three coefficients just mentioned in order to make this point clear, since they are often used in practice, and because they will be used later on in this work.

2.4. Dependence Quantification in 2-D

In this section, we present some issues related to dependence quantification in two dimensions. The approach is as applied as possible, addressing measures employed in practice. We shall see that, unlike the one-dimensional case, in two dimensions single coefficients are but an incomplete measure of the possible types of dependence. Copulas are introduced at the end as excellent pictures of 2-dimensional dependence.

2.4.1. Product moment correlation coefficient

This is the most used correlation coefficient, its definition is given by (2.3.3), and it is assumed that the reader is acquainted with it. In figure 2.4.1, we have two data-sets of size $N = 5000$ simulated from distributions having the same correlation coefficients, namely 0.472.

Data values $(x_{i,1}, x_{i,2})$, $i = 1, \dots, 5000$, on the right panel shown at figure 2.4.1 is a random sample from a 2-dimensional Gaussian distribution with mean $\mu = (0, 0)'$ and covariance matrix

$$M = \begin{pmatrix} 1 & 0.472 \\ 0.472 & 1 \end{pmatrix}$$

Data values $(y_{i,1}, y_{i,2})$, $i = 1, \dots, 5000$, shown on the left panel were simulated from a probability distribution constructed as follows:

1. Sample (ψ_1, ψ_2) from a 2-dimensional Gaussian distribution with covariance coefficient 0.546, and unit variance on each marginal distribution.
2. Set $Y_j = \psi_j + 0.01\psi_j^2 + 0.5\psi_j^3$, for $j = 1, 2$. A technique for computing covariances and other characteristics-related coefficients of transformations of this form is provided in section 4.4.2.

The type of association is of course different. Accordingly, data values with different characteristics are to be expected from these two distributions. For example, the “x” shape of the distribution on the left panel of figure 2.4.1 makes the observations of very high values of variable Y_2 given values of Y_1 relatively close to the median of its sample marginal distribution (which is 0.01); this phenomenon is not observable at the data on the right panel. On this right panel, higher values of X_2 are associated with higher values of X_1 . If these distributions are supposed to model jointly the behavior of two environmental variables, and X_2 can be somehow considered as a response variable, then the expected highest or lowest values of this response variable are to be expected at different regions of the distribution of X_1 , depending on whether the assumed distribution is of the type in (a) or in (b). Summarizing, the correlation coefficient does not tell everything about the kind of dependence between two variables.

2.4.2. Spearman's and Kendall's coefficients

2.4.2.1. Spearman's correlation coefficient

Computation of the sample product moment coefficient from the 5000 samples shown at figure 2.4.1 results in 0.472 and in 0.471 for the data sets plotted with an (a) and a (b) label, respectively. This similarity is not surprising, since the theoretical common correlation coefficient is 0.472. We now consider the ranks of the data, namely values $u_{i,1} = \text{rank}(x_{i,1} | \mathbf{x}_1)$ and $u_{i,2} = \text{rank}(x_{i,2} | \mathbf{x}_2)$, and $v_{i,1} = \text{rank}(y_{i,1} | \mathbf{y}_1)$ and $v_{i,2} = \text{rank}(y_{i,2} | \mathbf{y}_2)$, for $i = 1, \dots, 5000$. Vectors in the rank function argument indicate the that ranks are taken on each marginal sample separately. Since there are no ties in the simulated data sets, all ranks computed are unique. The resulting data is shown in figure 2.4.2. The plots look more similar to one another than in figure 2.4.1, though they are not entirely equal.

In connection with figure 2.4.2, the sample correlations of data sets $(v_{i,1}, v_{i,2})$ to the left and $(u_{i,1}, u_{i,2})$ to the right, $i = 1, \dots, 5000$, are $\rho_v = 0.539$ and $\rho_u = 0.452$, respectively. A Bootstrap-based 95% confidence interval for the value of $\rho_v - \rho_u$, created by using 2000 Bootstrap samples, was computed to be $(q_{2.5\%}, q_{97.5\%}) = (0.057, 0.117)$. So, the difference between both rank correlation coefficients, also known as “Spearman's correlation coefficients”, is significant. It is clear then that the information provided by the ranks of data is different from the one provided by the absolute values of data, which one to use is a consideration connected with the research question. However, also when working with ranks, it is possible to obtain equal correlation values for distributions whose ranks behave differently, or which distributions are different regarding important dependence-related characteristics. This aspect will be shown shortly but before that, Kendall's τ coefficient of correlation is introduced, following Long (2006).

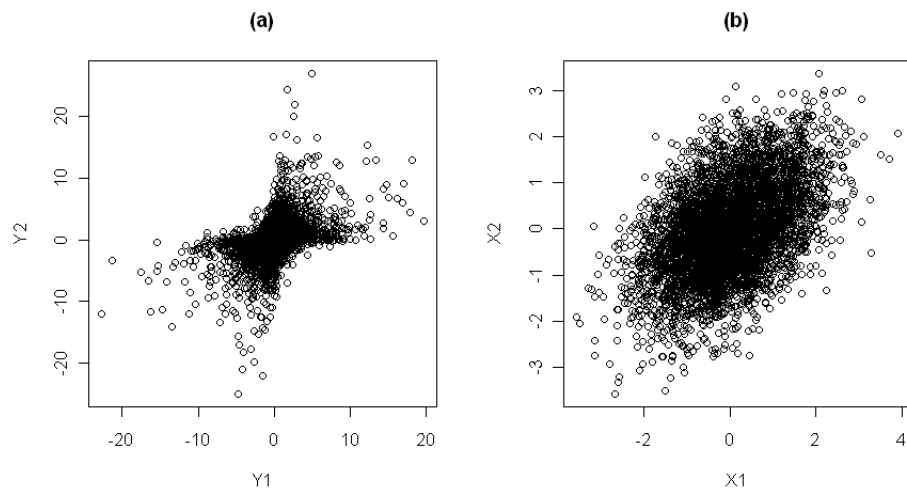


Figure 2.4.1.: Two Data sets simulated from two different distributions having the same product moment correlation coefficient.

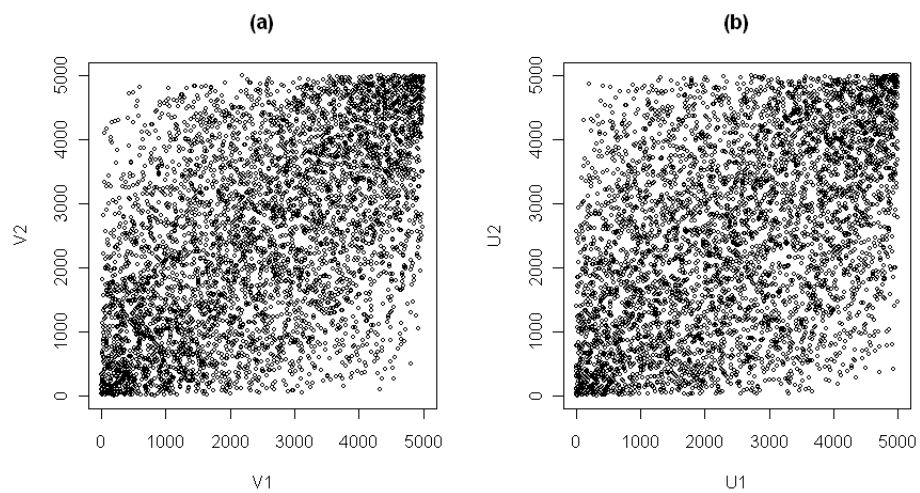


Figure 2.4.2.: Plot of ranks of example data sets

2.4.2.2. Kendall's correlation coefficient

Kendall's τ is "a measure of agreement or disagreement between two sets of rankings" Long (2006). A concrete example can help to work out the definition of this coefficient. Assume you are given two data sets, corresponding to realizations of two random variables X and Y :

Y:	14	5	8	11	7
X:	19	41	12	26	17

Table 2.4.1.: Two example data sets for the sake of Kendall's τ illustration

By rank ordering according to "x", table 2.4.1 is converted into table 2.4.2.

y:	8	7	14	11	5
x:	12	17	19	26	41

Table 2.4.2.: The same as 2.4.1 but after rank ordering according to "x"

Data from the "x" row is ordered increasingly, the question is: to what extent has this ordering also rank ordered the other data set? Or, equivalently: to what extent the ranks of the "x" row coincide with the ranks of the "y" row?. If it has induced a perfect increasing (decreasing) rank order, then there is evidence of high positive (negative) association. As the ordering of this "y" row becomes more erratic, the degree of association is considered smaller. In order to have a quantitative answer to the question just posed, the *dominance score* function is introduced now, which is defined for every two real values v_i and v_j , as:

$$ds(v_i, v_j) = \text{sign}(v_j - v_i) \quad (2.4.1)$$

It is convened that $ds(v_i, v_j) = 0$, if $v_i = v_j$. Applying this function to every pair of data from Y results in values:

$$\begin{aligned} ds(8, 7) &= -1 \\ ds(8, 14) &= 1 \\ &\vdots \\ ds(11, 5) &= -1 \end{aligned}$$

The whole list of values, excluding results of the form $ds(y_i, y_i)$, is

$$d_y = (-1, 1, 1, -1, 1, 1, -1, -1, -1) \quad (2.4.2)$$

Since all values of variable X are rank ordered increasingly, the result for this variable is

$$d_x = (1, 1, 1, 1, 1, 1, 1, 1, 1) \quad (2.4.3)$$

As when computing a correlation matrix, the number of pair-wise comparisons included in vectors d_y and d_x is $\frac{n(n-1)}{2} = \frac{5 \times 4}{2} = 10$. A predominant number of positive (negative) values

in vector d_y at (2.4.2) indicate a higher intensity of positive (negative) association between X and Y . In other words, the value $\tau^* = \frac{1}{10} \sum_{i=1}^{10} d_y(i)$ can be taken as an answer to the question of to what extent the ranks of the two data sets coincide: if it is very close to 1, it means that they coincide for most of the data values X (positive association); a value close to -1 indicates that the ranks of Y decrease for increasing ranks of X (negative association); and a value near 0 indicate that there is little association between the pair sample.

However, since some values from X can be tied, introducing zeros into d_x . This information must be taken into account. Hence, it is better to define the τ correlation coefficient as $\tau = \frac{1}{10} \sum_{i=1}^{10} (d_y(i) d_x(i))$, for our example.

In general, it is defined as

$$\tau = \frac{2}{n(n-1)} \sum_{i=1}^{\frac{n(n-1)}{2}} (d_x(i) d_y(i)) \quad (2.4.4)$$

Since Kendall's τ coefficient is computed using information of the ranks of data only, it is invariant to transformations on data that preserve rank ordering. That is, it is said to be invariant with respect to increasing monotonic transformations on data.

2.4.3. Copulas

The fact that rank-based correlation coefficients convey only a partial picture of 2-dimensional dependence, can be illustrated as follows. A sample of size 5000 from a Gaussian random vector (X_1, X_2) with the same characteristics as in section 2.4.1, and a sample from a random vector (Y_1, Y_2) having a t-distribution with three degrees of freedom and the same correlation were obtained by simulation. Their ranks were taken as in section 2.4.2.1. Additionally, each value was divided by 5001 in order to ensure it would fall in interval $(0, 1) \subset \mathbb{R}$. The plots are presented at figure 2.4.3. Even though it is well-known that these two distributions possess different strengths of dependence far away from the mean (see below), this fact is not captured neither by Spearman's nor by Kendall's coefficient. To see that this is the case, we resort again to a Monte Carlo technique:

1. A sample of size 5000 was simulated from each distribution, Gaussian and Student.
2. The ranking and scaling was performed, as explained in the paragraph above.
3. The Sample values of Kendall's and Spearman's coefficients were computed for each model and subtracted, obtaining $\tau_v - \tau_u$ and $\rho_v - \rho_u$.
4. After iterating steps 1 through 3 a total of 2000 times, an approximate 90% confidence interval was created on the basis of the 2000 computed coefficients subtractions.

The approximate confidence intervals thus obtained for the differences are: $I_\tau(5\%, 90\%) = (-0.022, 0.021)$, for the Kendall coefficients; and $I_{rho}(5\%, 90\%) = (-0.012, 0.046)$, for the Spearman coefficient. That is, the coefficients are practically the same for the two data sets. This time we used 90% instead of 95%, since the point is to show the incapacity of these coefficients to grasp the difference in association between data of these two distributions.

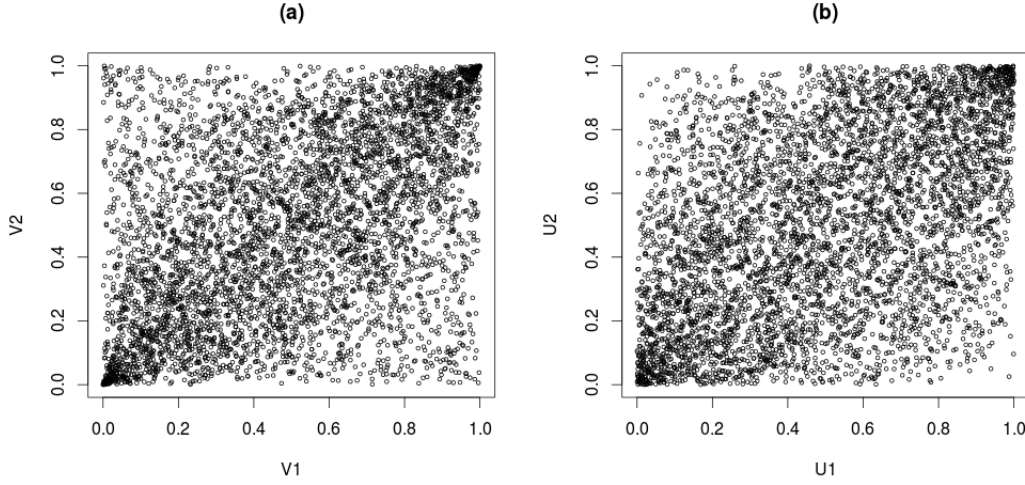


Figure 2.4.3.: Scaled ranks of 5000 values from a t-distribution (left) with 3 degrees of freedom and dispersion matrix $\Gamma = \begin{pmatrix} 1 & .472 \\ .472 & 1 \end{pmatrix}$, and a Gaussian distribution with mean $\mu = (0, 0)$ and covariance matrix Γ .

Then, the fact that the 90% confidence interval does contain zero, provides a more convincing argument.

The difference in association occurs most importantly at the upper and lower joint tails of the distribution. To see this, select a value close to 100%, say 0.95. Compute values $\hat{q}_{y_j,95\%}$ such that the proportion of observations of Y_j , $j = 1, 2$, smaller than or equal to $\hat{q}_{y_j,95\%}$ is 0.95, and do the same for the observations from X_j , $j = 1, 2$, obtaining $\hat{q}_{x_1,95\%}$ and $\hat{q}_{x_2,95\%}$. The proportion of values $(y_{i,1}, y_{i,2})$ from (Y_1, Y_2) *simultaneously* trespassing $\hat{q}_{y_1,95\%}$ and $\hat{q}_{y_2,95\%}$ is greater than the proportion of values $(x_{i,1}, x_{i,2})$ from (X_1, X_2) simultaneously trespassing $\hat{q}_{x_1,95\%}$ and $\hat{q}_{x_2,95\%}$. In mathematical notation, and labeling $N = 5000$

$$\frac{|\{(x_{i,1}, x_{i,2}) : x_{i,1} > \hat{q}_{x_1,95\%}, :x_{i,2} > \hat{q}_{x_2,95\%}\}|}{N} < \frac{|\{(y_{i,1}, y_{i,2}) : y_{i,1} > \hat{q}_{y_1,95\%}, :y_{i,2} > \hat{q}_{y_2,95\%}\}|}{N}$$

That is, even in their respective scales, the probability of simultaneous trespassing is higher for the t-distribution. For the example, the quantiles values are $(\hat{q}_{x_1,95\%}, \hat{q}_{x_2,95\%}) = (1.678, 1.610)$ and $(\hat{q}_{y_1,95\%}, \hat{q}_{y_2,95\%}) = (2.454, 2.348)$. The respective proportions are

$$\begin{aligned} \frac{|\{(x_{i,1}, x_{i,2}) : x_{i,1} > \hat{q}_{x_1,95\%}, :x_{i,2} > \hat{q}_{x_2,95\%}\}|}{N} &:= \hat{p}_x = 0.012 \\ \frac{|\{(y_{i,1}, y_{i,2}) : y_{i,1} > \hat{q}_{y_1,95\%}, :y_{i,2} > \hat{q}_{y_2,95\%}\}|}{N} &:= \hat{p}_y = 0.0186 \end{aligned}$$

A Monte Carlo confidence interval for the difference based on 2000 simulations is found to be, for data from the given distributions, $I_{\hat{p}_y - \hat{p}_x}(5\%, 95\%) = (0.0032, 0.009)$. This difference in the upper tail (or lower tail) dependence, can be crucial, depending on the specific subject matter questions posed (compare Embrechts et al. (2002)).

Figure 2.4.3 of standardized ranks allows to visualize quickly the difference in the tail dependence association. And the technique of using the quantile of each marginal distribution, as in the example above, allows to make comparisons in the very scales of the variables involved: the question is whether each variable is “big” or “small” with respect to its *own* particular distributions. This idea has proved very useful in multivariate statistics applications, as implemented via the use of copula methodology, which we summarize subsequently. We can cite, just for illustrating the applicability of the concept: Jaworski (2010), Singh and (ed.), Cherubini et al. (2004).

2.4.3.1. Sklar’s Theorem

A bi-variate probability distribution $F_{XY}(x, y) = \Pr(X \leq x, Y \leq y)$ describing random variables X and Y , having marginal probability distribution functions F_X and F_Y , can always be written in terms of a particular type of probability distribution function, $C(*, *): [0, 1] \times [0, 1] \rightarrow [0, 1]$, as

$$F_{XY}(x, y) = C(F_X(x), F_Y(y)) \quad (2.4.5)$$

This result is known as Sklar’s Theorem (see, for the topics in this section Nelsen (1999)). Function $C(*, *)$ must fulfill, in order to be a well defined probability distribution function:

1. $C(u, 1) = C(1, u) = u$, for every $u \in [0, 1]$.
2. $C(u, 0) = C(0, u) = 0$, for every $u \in [0, 1]$.
3. $C(u_1, u_2) + C(v_1, v_2) - C(u_1, v_2) - C(v_1, u_2) \geq 0$, for $1 \geq u_1 \geq v_1 \geq 0$ and $1 \geq u_2 \geq v_2 \geq 0$.

Additionally, if random variables X and Y are continuous, then representation (2.4.5) is unique. Any function having domain $[0, 1] \times [0, 1]$ and fulfilling conditions 1 through 3 above is called *Copula*. There are many well-known parametric models for Copulas used in practice: Clayton Copula, Frank Copula, Gumbel Copula, Gaussian Copula, etc., the parameters of which can then be considered as quantitative association coefficients. The number of these parameters is not inflated by the complexity of the marginals, which can be considered separately. The reader should see Nelsen (1999) or the applied literature on Copulas mentioned above.

Copulas can be studied on their own, merely as probability distribution functions of random values $(U, V) \in [0, 1] \times [0, 1]$. But these values are almost always in practice interpreted to be (approximations to) random variables resulting from other random variables, $X \sim F_X$ and $Y \sim F_Y$, in the form

$$\begin{aligned} U &= F_X(X) \\ V &= F_Y(Y) \end{aligned} \quad (2.4.6)$$

Thus, given a data set considered to be representable by a random vector $(X, Y) \sim F_{XY}$, it is possible to model this data by: First, estimate marginal distributions \hat{F}_X and \hat{F}_Y of X and Y , respectively. Second, estimate a suitable Copula $\hat{C}(*, *)$, such that $\hat{C}(\hat{F}_X(*), \hat{F}_Y(*))$ is a good approximation to $F_{XY}(*, *)$. These method is called the *Inference Function for Margins*

(IFM) method (see Joe (1997, 2005)). The strength of using Copulas for modeling multivariate random variables becomes then clear. Instead of attempting to fit a joint distribution that fits all characteristics of data, the task is divided into two simpler sub-tasks. So, in principle, more flexibility in the fitted distribution is attained.

The estimates for $F_X(X)$ and $F_Y(Y)$ can be either parametric or non-parametric, dealing with each marginal independently. In the parametric case, subject-matter adequate probability distributions are fitted using standard parametric methods, such as maximum likelihood or the method of moments. As non-parametric estimates, the empirical probability distribution function can be used. For a generic marginal X of which a realized sample x_1, \dots, x_N is available, this function would be:

$$\hat{F}_X(a) := \frac{|\{x_i : x_i \leq a\}|}{N+1} \quad (2.4.7)$$

for any given $a \in \mathbb{R}$.

The Copula density,

$$c(u, v) = \frac{\partial^2}{\partial v \partial u} C(u, v) \quad (2.4.8)$$

is mostly used for estimation via the IFM method. Inference on the Copula can thus begin with a data-set of values on $[0, 1] \times [0, 1]$. As usual in statistic, regions with more concentration of points indicate an underlying probability density function with higher values. Whereas regions with fewer or no points indicate a low value of the underlying density function. The reader can then have an idea of the copula density by examining the data illustrated at figure 2.4.3.

To sum up, copulas provide a flexible means of modeling 2-dimensional data. By fitting marginals first and the Copula second, the parameters of the copula do not have to manage the potential complexity of marginals; its parameters focus on dependence. Copulas can also be defined analogously in higher dimensions, $J > 2$. Then, each of the J marginals is fitted separately. The domain or support of the Copula is then $[0, 1] \times \dots \times [0, 1] := [0, 1]^J$. Apart from parametric models, such as those mentioned in the literature provided in this section, it is also possible to fit non-parametric copulas to rank-scaled data. Standard, unmodified kernel smoothing (e.g. Scott (1992)) are inconvenient, since these methods work best when there is no restriction on the support of the distribution. For more details see Charpentier et al. (2006). To circumvent this problem, modifications to the original kernel smoothing technique must be performed Marron and Ruppert (1994). Other approaches include wavelets-based estimation Genest et al. (2009), and using a special kind of kernel smoothing Chen (1999). However, all those non-parametric techniques suffer from the curse of dimensionality. They become useless as dimension increases. For the case of $J > 10$ (typical in spatial statistics) they are almost useless.

More in connection with Spatial Statistics, the problem arises of what kinds of Copulas can be extended in dimension as much as necessary. The Gaussian and the t-Copula allow this, with the aid of a covariance function, such as 2.2.8, for example. A recent model, that allows much flexibility in the form of the dependence modeled, is the V-transformed Copula (Bárdossy and Pegram (2009); Bárdossy and Li (2008)), which has been used in Geostatistical applications. Vine Copulas (Joe (1996); Bedford and Cooke (2001); Aas et al. (2009)),

whereby high-dimensional Copulas are built pair-wise, can also be extended indefinitely. They have already been applied to Spatial interpolation in Gräler and Pebesma (2011). The approach presented in this work, attempts to focus on aspects of dependence that are directly relevant for applications, at the same time providing acceptable theory at its foundation.

2.5. The Edgeworth-Sargant distribution

This section intends to provide a bridge between the topics just discussed and the approach proposed in this work for dealing with interdependence of more than two variables simultaneously.

We saw for the 1-dimensional case, that some coefficients or parameters were directly connected with distribution characteristics of interest. Namely the parameters were the moments, and the characteristics were the location, dispersion, skewness and kurtosis of the distribution. This simple approach has consistently paid results in practice, since the origins of mathematical statistics. It was the reason for the introduction of systems of distributions that could match the sample characteristics of data, such as Pearson's (see, e.g. Kendall and Stuart (1969)) and Johnson's Johnson (1949); Slifker and Shapiro (1980) systems.

The "Edgeworth-Sargant" (E-S) distribution, used in econometrics (Mauleon and Perote (2000); Sargan (1976)), is so defined that moments are explicitly expressed in terms of the parameters of the distribution, to be estimated by maximum likelihood, for example. Moments of all orders exist. The number of moments to use is application-dependent and can, in principle, be found on the basis of an information criterion, such as the AIC criterion Akaike (1974). The distribution provides great flexibility. Thus, a great spectrum of distribution characteristics can be modeled.

The density of the E-S distribution is given by

$$f_{ES}(x) = \phi(x) \left\{ 1 + \sum_{r=1}^R \delta_r H_r(x) \right\} \quad (2.5.1)$$

where parameters $\delta_r, r = 1, \dots, R$, are the parameters to estimate; $\phi(*)$ stands for the standard Normal probability density function, and $H_r(x)$ is the r -th order Hermite Polynomial (see Kendall and Stuart (1969)), obtainable by means of identity

$$(-1)^r H_r(x) \phi(x) = \frac{d^r}{dx^r} \phi(x) \quad (2.5.2)$$

It can be seen with little work that the moments of the distribution whose density is (2.5.1), are given by

$$\begin{aligned} E(X) &= \delta_1 \\ E(X^2) &= 1 + 2\delta_2 \\ E(X^3) &= 6\delta_3 + 3\delta_1 \\ E(X^4) &= 24\delta_4 + 12\delta_2 + 3 \end{aligned} \quad (2.5.3)$$

Moments of order k are obtained similarly as functions of $\delta_1, \dots, \delta_k$. We can consider as many moments (i.e. distribution characteristics) as data complexity demands, avoiding

over-fitting with the aid of the AIC criterion, for example. We see from (2.5.3) that higher order moments (i.e. data characteristics) can be fitted *without* altering lower order ones, “orthogonally”, by fitting only the parameter corresponding to the highest index. This convenient aspect is imitated by our approach in the multidimensional case, with a specific implementation adequate for Spatial Statistics.

As Mauleon and Perote (2000) report, the E-S can model conveniently heavy tailed data, and is flexible in terms of asymmetry. It can, under minor modification, also represent truncated data.

A closely related distribution was introduced by Gallant and Nychka (1987). We call it the N-G distribution for brevity. It is considered a variant of the E-S distribution by Mauleon and Perote (2000), circumventing the problem of careful selection of $\delta_1, \dots, \delta_R$ in order to ensure non-negativity of (2.5.1). The density in the one-dimensional can be written as

$$f_{NG}(x) = \phi(x) \left\{ w_0 + \left(\sum_{r=1}^R w_r x^r \right)^2 \right\} = \phi(x) \left\{ w_0 + \sum_{r=1}^R \sum_{s=1}^R w_r w_s x^{r+s} \right\} \quad (2.5.4)$$

This function is clearly non-negative everywhere, provided that $w_0 \geq 0$. In order to ensure integration to unity, we can impose the following parameters constraints:

$$0 < w_0 = 1 - \sum_{r=1}^R \sum_{s=1}^R w_r w_s \mu_{r+s}^* \quad (2.5.5)$$

where μ_{r+s}^* is the moment of order $r+s$ of a standard Normal distribution. Moments of this distribution are

$$\begin{aligned} E(X^k) &= \int_{-\infty}^{\infty} x^k f_{NG}(x) dx = w_0 \int_{-\infty}^{\infty} x^k \phi(x) dx \\ &\quad + \sum_{r=1}^R \sum_{s=1}^R \left\{ w_r w_s \int_{-\infty}^{\infty} x^{r+s+k} \phi(x) dx \right\} \\ &= w_0 \mu_k^* + \sum_{r=1}^R \sum_{s=1}^R w_r w_s \mu_{r+s+k}^* \quad (2.5.6) \end{aligned}$$

Again, moments of all orders exist and can be written explicitly in terms of w_0, \dots, w_R . Maximum likelihood estimation and inference can be then performed on distribution parameters w_0, \dots, w_R . The aspect to highlight for this distribution, as for the “basic” S-E distribution, is its flexibility in reproducing important characteristics of data, and the “extendable” capacity of increasing the order R in w_0, \dots, w_R , as data complexity requires it.

The flexibility of E-S type of distributions has lead to research for its generalization to more than one dimensions. We summarize this research briefly.

The paper Gallant and Nychka (1987) deals with the multivariate case directly. A simple version of the two-dimensional case can be written as

$$f_{NG}(x_1, x_2) = \frac{1}{2\pi} \exp(-x_1^2 - x_2^2) \left\{ \sum_{r_1=0}^{R_1} \sum_{r_2=0}^{R_2} \alpha_{r_1 r_2} x_1^{r_1} x_2^{r_2} \right\}^2$$

Then the dependence parameters $\alpha_{r_1 r_2}$, for $1 \leq r_1 \leq R_1$ and $1 \leq r_2 \leq R_2$, are estimated on the basis of data. A restriction must be set on α_{00} such that the density integrate to unity. The joint moments can be readily found in terms of the moments of a 1-dimensional standard Normal distribution, as:

$$E(X_1^{k_1} X_2^{k_2}) = \sum_{r_1=0}^{R_1} \sum_{r_1^*=0}^{R_1} \sum_{r_2=0}^{R_2} \sum_{r_2^*=0}^{R_2} \left\{ \alpha_{r_1 r_2} \alpha_{r_1^* r_2^*} \left(\mu_{r_1+r_1^*+k_1}^* \right) \left(\mu_{r_2+r_2^*+k_2}^* \right) \right\}$$

The computational demand of the density increases rapidly with dimension. Additionally it is necessary to constrain the $\alpha_{r_1 \dots r_J}$ values to ensure integration to unity. However, this approach might be more accurately explored in future research.

A direct generalization of (2.5.1) has been proposed by Perote (2004), and its application tested in financial returns data. Unfortunately, as in the case of (2.5.1), the resulting “probability density” can take on negative values. This by no means demerits the usefulness of the distribution, but increases the computational and analytical effort required for its application. Of course, this effort increases dramatically with dimension of the random vector modeled, which is inconvenient for Spatial Statistics modeling.

In order to circumvent the problem of eventual negativity met by Perote (2004), and to frame this distribution in a more general family, Del Brio et al. (2009) deal with densities they call “Multivariate Gram-Charlier densities”. The flexibility of the approach, the possibility of adding parameters as data complexity requires, makes of this a very attractive approach. Marginal distributions of different shapes and tail thickness can be, in particular, conveniently represented. We consider the full model as tow high-parametric for Spatial applications, though we made no serious attempt to find a convenient re-parametrization. A simplified density model for $\mathbf{X} = (X_1, \dots, X_J)$ is of the form (equation (10) in Del Brio et al. (2009)):

$$f_{GC}(\mathbf{X}) = \frac{1}{J+1} G_{\Gamma}(\mathbf{X}) + \frac{1}{J+1} \left\{ \prod_{j=1}^J g_j(x_j) \right\} \times \left\{ \sum_{j=1}^J \frac{1}{1 + s_{jt}^2/6} \left[1 + \frac{s_{jt}}{6} (x_j^3 - 3x_j) \right]^2 \right\} \quad (2.5.7)$$

where $G_{\Gamma}(\mathbf{X})$ stands for the probability density function of a multivariate Normal distribution with covariance matrix $\Gamma_{J \times J}$, $g_j(*)$ represent each marginal density of $G_{\Gamma}(\mathbf{X})$, and (s_{1t}, \dots, s_{Jt}) are J parameters to fit. These parameters are intended to model the various features of data, as expressed in terms of moments, and conditional moments, which can be readily derived for this distribution. We have used the notation of Del Brio et al. (2009) here, including the possible dependence of (s_{1t}, \dots, s_{Jt}) on time t . Thus, important features, relevant for the evolution in time of financial variables of interest can be modeled.

These generalizations of the E-S distribution into multivariate distributions provided part of the inspiration to the approach presented in this work, and can be considered as paths to be explored in the future. However, we have attempted to come up with an implementation more closely related to a type of measure or coefficient of multivariate interdependence given below, namely the joint cumulants of a random vector. This measure we consider theoretically appealing and practicably connectable with important data features, as explained subsequently.

3. The Proposed Approach: General

In chapter 2 it has been seen that the richness of multivariate interdependence demands that we focus on subject-matter relevant features of data, and come up with flexible models, parametrically extendible and shrinkable as the complexity of these (interdependence) features demands.

But in addition to this ad-hoc approach, we want also to introduce some foundational basis for the concept of multivariate interdependence, to provide some minimal requirements for a measure of interdependence, and to show how a measure fulfilling such minimal requirements can be connected with subject-matter relevant characteristics. These are the topics of the present chapter.

In the present chapter we refer to Spatial Statistics for the sake of clarity and illustration of ideas, but in principle any other area of applied Statistics where we need to consider interdependence among random variables might be adapted for illustration.

3.1. Application-relevant interaction manifestations

The following aspects are inherent to spatial statistics, and must be taken into account:

1. High-dimensionality, which implies a need for really parsimonious models. Just think of a full model with 10 variables; it would have, assuming symmetry (the order of the random variables does not alter the value of the dependence measure), some $O(10^{10})$ interdependence parameters!
2. The need to interpolate or extrapolate the variable of interest to ungauged sites.

We consider the behavior of the following parameters as *manifestations* of multivariate interdependence, relevant for spatial statistical analysis. They are related to a random vector $\mathbf{X} \in \mathbb{R}^J$ representing the phenomenon under study (e.g. rainfall, temperature, mineral quantity in geostatistical applications, etc., at several locations).

Joint probabilities such as

$$\Pr(X_{j_1} \leq q_{j_1, \alpha}, \dots, X_{j_k} \leq q_{j_k, \alpha})$$

where $\{j_1, \dots, j_k\} \subseteq \{1, \dots, J\}$ is a set of indices, and vector $(q_{1, \alpha}, \dots, q_{J, \alpha})$ is such that $\Pr(X_1 \leq q_{1, \alpha}) = \alpha, \dots, \Pr(X_J \leq q_{J, \alpha}) = \alpha$, for a given α . For example, $\alpha \in \{0.5, 0.75, 0.9, 0.99\}$ might be of interest. The modeling of these probabilities is relevant, for example, in the context of flood forecasting. Our model should reproduce, as well as possible, these probabilities and they should be considered upon estimation of the model's parameters, whatever this model and these parameters may be.

The value of the differential entropy of \mathbf{X} or of subsets of its components. This represents an omnibus measure of dependence, which relies on the very generally applicable concept of entropy. It is a powerful concept, helpful for verification, but usually not helpful for model-building or elucidation of the processes causing the interdependence among variables. For a random vector with probability density function f , the entropy is given by $H(\mathbf{X}) = \int f(\mathbf{X}) \cdot \log(f(\mathbf{X})) d\mathbf{X}$. If we take a (low-dimensional) sub-set $\mathbf{V} = (X_{j_1}, \dots, X_{j_k})$ of components of \mathbf{X} and estimate their joint entropy $\hat{H}(\mathbf{V})$ by using, for example, kernel smoothing Joe (1989a), we want our fitted model to have similar values of differential entropy for the marginal distribution of $\mathbf{V} = (X_{j_1}, \dots, X_{j_k})$ as the estimated $\hat{H}(\mathbf{V})$. We would like to be able to “fit” $\hat{H}(\mathbf{V})$ as we estimate our model’s parameters.

The distribution of sums of components of $\mathbf{X} \in \mathbb{R}^J$, as expressed in conveniently selected parameters of the resulting random variable $S_{\mathbf{X}} = \sum_{j=1}^J X_j$, such as its moments or cumulants (see below). As seen in the previous chapter, these parameters convey important information about the 1-dimensional distribution of the sum. We consider this a most important set of parameters in the context of rainfall modeling, in that they can help better understand the relationship between the rainfall field and the expected discharge at the outlet of the basin covered by such rainfall field. Since we are mostly concerned with rainfall modeling in this research, we consider these parameters as of primary importance, the other two above, though also considered, are considered complementary. As seen shortly, the cumulants of $S_{\mathbf{X}} = \sum_{j=1}^J X_j$ can be found straightforwardly in terms of the joint cumulants of random vector $\mathbf{X} \in \mathbb{R}^J$.

None of the above three manifestations-related parameters of interdependence qualify as suitable dependence measures, according to the postulates of Rényi or of Schmid, mentioned at the introduction, even though Joe (1989b) proposed a set of entropy-based measures having theoretically appealing properties. These parameters are also not very helpful, on their own, for model building; they are not nice “building blocks” for a model nor suggest any manageable dependence structure (in terms of parameter estimation, interpretation, and with a view to interpolation). Model building is desirable since it implies, if the model is validated, a better understanding of the phenomenon studied. That is why we shall call the values listed from 1 through 3, just “interaction *manifestation* parameters”. We shall look somewhere else for a dependence structure and “*dependence* structure parameters”, which may provide a manageable structure and in terms of which the manifestation-related parameters can be expressed.

3.1.1. Interaction manifestations versus dependence structure

This section intends to serve as a guide to the topics next dealt with in section 3.2, and its subsections.

We introduce in section 3.2 our suggested “*dependence* structure parameters” for a random vector $\mathbf{X} \in \mathbb{R}^J$.

We shall see that the interaction manifestation parameters of section 3.1 can be represented in terms of quantities or parameters related to \mathbf{X} ’s distribution which:

1. Are logically appealing as dependence parameters, in that: a.) They are capable of an interpretation in terms of a very elementary set of axioms desirable for a dependence measure; b.) they have sometimes, and depending on the application, a physical interpretation in terms of subject-matter relevant parameters.
2. Comprise both mean and covariance/correlation as particular cases, thus providing potential for the extension of techniques based on correlation to groups of more than two random variables.
3. Work as better building blocks for a model, as compared to interaction manifestations listed at section 3.1, since it is easier to establish a functional relation between these quantities and the distribution of \mathbf{X} and thus can be used as parameters to estimate.

We suggest that such quantities are the joint cumulants of \mathbf{X} , which are introduced in section 3.2.1.

We show subsequently that they are a suitable basis for addressing the issues inherent to spatial statistics mentioned above, and that they provide a link between model building (and estimation) and the interaction manifestation parameters mentioned on the last section. We state that they are logically appealing measures of dependence, in that, departing from simple first principles, we can consider them as summary measures of interaction among the components of a random vector, as seen in section 3.2.2.

For these reasons, we shall call the joint cumulants “dependence parameters”, and the cumulant generating function “dependence structure”.

The knowledge of the cumulant generating function of a random vector implies an approximation to its probability density function or distribution function, via the Edgeworth Expansion or the Saddle-point Approximation (section 3.3.1). Hence this dependence structure seems to be flexible enough to tackle different subject-matter relevant *interaction manifestations*. The connection between joint cumulants and the interaction manifestations listed in section 3.1, is dealt with at section 3.3.2.

The new idea in this work is to use the cumulants as parameters to fit that can reproduce as good as possible the manifestation parameters at hand, in a “method of moments” fashion. Thus the dependence structure may be simple (and with relatively few parameters), as long as the manifestations parameters are modeled properly.

3.2. Joint cumulants as interdependence parameters

3.2.1. Definition and preliminaries

Moments and cumulants are constants summarizing important information about a probability distribution and sometimes, even determining it completely Kendall and Stuart (1969). In this section we deal with random variables having a probability density function. The development is also valid for discrete distributions, under simple modifications. The reader is referred to Kendall and Stuart (1969); Muirhead (1982); Billingsley (1986) for more details on moments and cumulants.

In this section, we intend to make cumulants a little more known to the reader, by explaining its connection with the more common concept of moments. Some detail is given to the uni-

variate case, since the analysis will be useful in section 4.1.1, when dealing with moments of a multivariate model that is suggested as convenient for Spatial Statistics.

In the one dimensional case, let a random variable $X \in \mathbb{R}$ with probability density function $f_X(X)$ have a moment generating function

$$M_X(t) = \int_{-\infty}^{+\infty} e^{xt} f_X(x) dx$$

This moment generating function is here assumed to exist, at least for t in a sufficiently small interval around zero, $t \in [-\epsilon, \epsilon]$.

The k -th moment of random variable X with probability density function f_X is given by

$$\mu_k := E(X^k) = \int_{-\infty}^{+\infty} x^k f_X(x) dx \quad (3.2.1)$$

These quantities can determine, to some extent, the distribution of X .

Due to the exponential representation $\exp(x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots$, and the linearity property of the integral,

$$\begin{aligned} M_X(t) &= \int f_X(x) dx + t \int x f_X(x) dx + \frac{t^2}{2!} \int x^2 f_X(x) dx + \frac{t^3}{3!} \int x^3 f_X(x) dx + \dots \\ &= 1 + t\mu_1 + \frac{t^2}{2!}\mu_2 + \frac{t^3}{3!}\mu_3 + \dots \end{aligned} \quad (3.2.2)$$

and so, moments are also defined to be derivatives of function $M_X(t)$ evaluated at zero, that is,

$$E(X^k) = \mu_k := \frac{d^k}{dt^k} M_X(t) \big|_{t=0} \quad (3.2.3)$$

The cumulant generating function of X is given by

$$K_X(t) = \log(M_X(t))$$

For t in a sufficiently small interval around zero, $t \in [-\epsilon, \epsilon]$, one has that $M_X(t) \in (0, 2)$, and then it is possible to use the series representation for the logarithm (see eq. 3.3.7)

$$\begin{aligned} K_X(t) &= \log(M_X(t)) = \log\left(1 + t\mu_1 + \frac{t^2}{2!}\mu_2 + \frac{t^3}{3!}\mu_3 + \dots\right) = \\ &= \left(t\mu_1 + \frac{t^2}{2!}\mu_2 + \frac{t^3}{3!}\mu_3 + \dots\right) - \frac{\left(t\mu_1 + \frac{t^2}{2!}\mu_2 + \frac{t^3}{3!}\mu_3 + \dots\right)^2}{2} + \\ &\quad - \frac{\left(t\mu_1 + \frac{t^2}{2!}\mu_2 + \frac{t^3}{3!}\mu_3 + \dots\right)^3}{3} + \frac{\left(t\mu_1 + \frac{t^2}{2!}\mu_2 + \frac{t^3}{3!}\mu_3 + \dots\right)^4}{4} - \dots \end{aligned} \quad (3.2.4)$$

Finally, after re-arranging (3.2.4) in terms of powers of t , the series can be written as

$$K_X(t) = t\kappa_1 + \frac{t^2}{2!}\kappa_2 + \frac{t^3}{3!}\kappa_3 + \frac{t^4}{4!}\kappa_4 + \dots \quad (3.2.5)$$

and the coefficients κ_r of this series are called “cumulants”. These can be obtained by differentiating $K_X(t)$ and evaluating the result at $t = 0$,

$$\kappa_r = \left. \frac{d^r K_X(t)}{dt^r} \right|_{t=0} \quad (3.2.6)$$

Since

$$\log \left(1 + t\mu_1 + \frac{t^2}{2!}\mu_2 + \frac{t^3}{3!}\mu_3 + \dots \right) = t\kappa_1 + \frac{t^2}{2!}\kappa_2 + \frac{t^3}{3!}\kappa_3 + \dots \quad (3.2.7)$$

it is possible to differentiate on both sides of (3.2.7) and evaluate derivatives at zero, in order to obtain moments in terms of cumulants, and vice versa. The first four cumulants are given by

$$\begin{aligned} \kappa_1 &= \mu_1 \\ \kappa_2 &= \mu_2 - \mu_1^2 \\ \kappa_3 &= \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3 \\ \kappa_4 &= \mu_4 - 4\mu_3\mu_1 - 3\mu_2^2 + 12\mu_2\mu_1^2 - 6\mu_1^4 \end{aligned} \quad (3.2.8)$$

A computationally convenient algorithm for finding cumulants in terms of moments, and vice versa, is given by Smith (1995).

The cumulant generating function and cumulant coefficients can be, depending on the statistical problem at hand, more convenient tools of analysis than the moment generating function and the moments. In this work we shall see actually part of their usefulness.

An important characteristic of cumulants, is that they are location invariant (except for κ_1), and they are not distorted by affine transformations. Namely, if random variable X has cumulants $\kappa_1, \kappa_2, \kappa_3, \kappa_4, \dots$, then $Y = m + aX$ has cumulants κ_r^* of the form

$$\begin{aligned} \kappa_1^* &= m + a\kappa_1 \\ \kappa_2^* &= a^2\kappa_2 \\ &\vdots \\ \kappa_r^* &= a^r\kappa_r \end{aligned}$$

Quite similarly in the *multivariate case*, with $\mathbf{X} \in \mathbb{R}^J$, and departing from the moment generating function

$$M_{\mathbf{X}}(\mathbf{t}) = \int \dots \int e^{\mathbf{t}' \mathbf{x}} f_{\mathbf{X}}(\mathbf{x}) dx_1 \dots dx_J$$

we obtain the cumulant generating function

$$K_{\mathbf{X}}(\mathbf{t}) = \log(M_{\mathbf{X}}(\mathbf{t}))$$

The joint moments of \mathbf{X} ,

$$\mu_{r_1, \dots, r_J} = E(X^{r_1} \dots X^{r_J})$$

are the coefficients of expansion,

$$M_{\mathbf{X}}(\mathbf{t}) = \sum_{r_1=0}^{\infty} \dots \sum_{r_J=0}^{\infty} \frac{\mu_{r_1, \dots, r_J} t_1^{r_1} \dots t_J^{r_J}}{r_1! \dots r_J!} \quad (3.2.9)$$

Similarly, we find the joint cumulants to be the coefficients of the expansion

$$K_{\mathbf{X}}(\mathbf{t}) = \sum_{r_1=0}^{\infty} \dots \sum_{r_J=0}^{\infty} \frac{\kappa_{r_1, \dots, r_J} \cdot t_1^{r_1} \dots t_J^{r_J}}{r_1! \dots r_J!} \quad (3.2.10)$$

Where $\kappa_{0, \dots, 0} = 0$. In this case, too, the joint cumulants κ_{r_1, \dots, r_J} can be computed by derivation of $K_{\mathbf{X}}(\mathbf{t})$ and evaluation at $\mathbf{t} = \mathbf{0}$, that is

$$\frac{\partial^{r_1 + \dots + r_J}}{\partial t_{r_J} \dots \partial t_{r_1}} K_{\mathbf{X}}(\mathbf{t}) \big|_{\mathbf{t}=\mathbf{0}} = \kappa_{r_1, \dots, r_J}$$

The relationship between joint cumulants and moments can also be found, as in the univariate case, by noticing that

$$\log \left(\sum_{r_1=0}^{\infty} \dots \sum_{r_J=0}^{\infty} \frac{\mu_{r_1, \dots, r_J} \cdot t_1^{r_1} \dots t_J^{r_J}}{r_1! \dots r_J!} \right) = \sum_{r_1=0}^{\infty} \dots \sum_{r_J=0}^{\infty} \frac{\kappa_{r_1, \dots, r_J} \cdot t_1^{r_1} \dots t_J^{r_J}}{r_1! \dots r_J!}$$

and differentiating on both sides, followed by evaluation at vector zero, $\mathbf{t} = \mathbf{0}$, we can find moments in terms of cumulants and vice versa. This is the basis for an alternative definition for joint cumulants. Since it requires the definition of some notation which, anyway, will be developed for section 3.2.2, we postpone this alternative definition until section 3.2.3.1.

The most well-known types of joint cumulants are: the one having only two indexes, corresponding to indexes j_1 and j_2 say, set to 1 and all others set to zero, which is the covariance, $Cov(X_{j_1}, X_{j_2})$; and the one having only one index distinct from zero and set to 1 or 2, which corresponds respectively to the mean or the variance of a specific component of vector $\mathbf{X} \in \mathbb{R}^J$.

For example, if $J = 4$, then $\kappa_{0,1,1,0} = Cov(X_2, X_3)$, $\kappa_{1,0,0,1} = Cov(X_1, X_4)$, $\kappa_{0,2,0,0} = Var(X_2)$, $\kappa_{0,0,0,1} = E(X_4)$, and so on.

An alternative (and equivalent) definition of the joint cumulants based on the moments, together with properties of the joint cumulants can be found in Brillinger (1974).

Joint cumulants and cumulant generating functions have found application within statistics in several forms: In time series Analysis Brillinger (1974) and signal analysis Mendel (1991), where the important joint cumulant property of vanishing in case of a set of independent variables is exploited; in the asymptotic analysis of the covariance matrix distribution Muirhead (1982); in distribution approximation, by means of the Edgeworth Expansion and the Saddlepoint approximation Barndorff-Nielsen and Cox (1990); and in other, more specific applications.

3.2.2. "Lancaster interactions" and joint cumulants

We deal now with a function, called "additive interaction measure" or "Lancaster interaction measure", introduced by Lancaster (1969) and later modified by Streitberg (1990).

An additive interaction measure $\Delta F(\mathbf{X})$ is a signed¹ measure determined by a given distribution $F(\mathbf{X})$ on \mathbb{R}^J . Its defining characteristic is that it is equal to zero for all $\mathbf{X} \in \mathbb{R}^J$, if $F(\mathbf{X})$ can be written as the non-trivial product two or more of its (multivariate) marginal

¹This just means that the measure is allowed to be negative.

distributions (Streitberg (1990)). For example, if $J = 4$ and F can be written as $F_{124}F_3$, being F_{124} and F_3 the marginal distributions of (X_1, X_2, X_4) and X_3 , respectively, then $\Delta F \equiv 0$.

An alternative explanation is that $\Delta F \equiv 0$ if one subset of \mathbf{X} 's components is independent of another subset of components.

The measure ΔF is called "additive" because it is a measure written as a linear combination of products of (univariate and multivariate) marginal distributions of \mathbf{X} , as we shall see shortly. It was studied by Lancaster and further modified by Streitberg (1990) and Streitberg (1999) to address a number of issues in the analysis of interactions in high-dimensional contingency tables, most notably: The need to be able to analyze interactions in a sub-set $(X_{j_1}, \dots, X_{j_k})$ of variables of \mathbf{X} without having to impose any conditions on the joint distribution of all \mathbf{X} .

We make an effort to explain here briefly the ideas, but the reader is referred to Streitberg's papers for details. We introduce first some preliminary notation.

To our random vector $\mathbf{X} \in \mathbb{R}^J$ under study, corresponds a set of indexes $C = \{1, \dots, J\}$. This set of indexes can be partitioned into $|\pi|$ non-overlapping subsets, $C = C_1 \cup \dots \cup C_{|\pi|}$. The set of non-overlapping sets of which the union is C , i.e. $\pi = \{C_1, \dots, C_{|\pi|}\}$, is called a "partition" of C . A set of J elements has a total of B_J possible partitions², where $B_0 = B_1 = 1$ and any subsequent $B_{k>1}$ can be found Rota (1964) by the recurrence relation $B_{k+1} = \sum_{r=0}^k \binom{k}{r} B_r$.

For example, for $C = \{1, 2, 3, 4\}$ there are 15 partitions, three of which are: $\pi_1 = \{\{1\}, \{2\}, \{3, 4\}\}$, $\pi_2 = \{\{1, 4\}, \{2, 3\}\}$, $\pi_3 = \{\{1, 2, 3, 4\}\}$.

It is convenient to use shorthand notation, the one we use is illustrated as follows for the partitions above: $\pi_1 = 1 \mid 2 \mid 34$, $\pi_2 = 14 \mid 23$ and $\pi_3 = 1234$. Additionally, we illustrate the meaning of $|\pi|$ by noting that $|\pi_1| = 3$, $|\pi_2| = 2$ and $|\pi_3| = 1$.

Now, concerning probability distribution F on \mathbb{R}^J , we define F_π to be the factorization of F implied by partition π . For instance,

$$F_{\pi_1}(\mathbf{X}) = F_1(X_1) F_2(X_2) F_{34}(X_3, X_4) \quad (3.2.11)$$

where the factors correspond to the respective marginal distributions. It will be convenient to define partition operator J_π , to be applied to F for a given partition π , by

$$J_\pi F \rightarrow F_\pi \quad (3.2.12)$$

where F_π is as in the example at equation (3.2.11).

As a final piece of nomenclature, a distribution is called "decomposable" if there exists $\pi \neq \{C\}$, such that $F = F_\pi$. That is, if it can be written as a product of at least two of its (multidimensional) marginal distributions. For example, if $F(\mathbf{X}) = F_{\pi_1}(\mathbf{X})$ at (3.2.11), then F is said to be decomposable.

We are ready for the uniqueness result that provides a definition for our additive interaction measure Streitberg (1990, 1999):

Theorem 1. *Let F be a probability distribution function on \mathbb{R}^J and let ΔF be a function fulfilling the following conditions:*

²The number B_J is often called Bell's number.

1. ΔF is a linear combination of all factorizations of F implied by the partitions of $C = \{1, \dots, J\}$, that is, $\Delta F = \sum_{\pi} a_{\pi} F_{\pi}$, for some real numbers a_{π} .
2. For partition $\pi^* = \{C\}$, also called "unity partition", the corresponding coefficient is one: $a_{\pi^*} = 1$.
3. (Interaction property) If F is decomposable, then $\Delta F(\mathbf{X}) = 0$, for all $\mathbf{X} \in \mathbb{R}^J$.

Then ΔF is uniquely given by:

$$\Delta F = \sum_{\pi} \left\{ \left((-1)^{|\pi|-1} (|\pi| - 1)! \right) F_{\pi} \right\} \quad (3.2.13)$$

Or equivalently: each coefficient a_{π} , corresponding to partition π , is uniquely defined by $a_{\pi} = (-1)^{|\pi|-1} (|\pi| - 1)!$.

Condition 1 above states the additive nature of ΔF , note that this definition makes ΔF existent for every distribution, since it just utilizes its marginal distributions and adds them in a weighted manner. Condition 2 makes trivial forms of non-uniqueness impossible, as would be the case in obtaining a "different" interaction measure ΔF^* satisfying 1 and 3, by multiplication by a constant, $\Delta F^* = c \cdot \Delta F$. Condition 3 is a most reasonable requirement one might demand from an interaction measure: it should vanish whenever the components of \mathbf{X} are completely or group-wise independent.

Paraphrasing: for every probability distribution function F , we have identified the *only* function ΔF , built as a linear combination of products of (multivariate) marginal distributions of F , such that $\Delta F(\mathbf{X}) := 0$, whenever F is decomposable. "Decomposable" can be read: "one subset of \mathbf{X} 's components is independent of another subset". Such a function is given by equation (3.2.13).

Since the interaction measure is defined in terms of a given distribution F , it is convenient sometimes to define the interactions operator:

$$\Delta = \sum_{\pi} \left\{ \left((-1)^{|\pi|-1} (|\pi| - 1)! \right) J_{\pi} \right\} \quad (3.2.14)$$

which, upon application to the distribution in question, returns the additive interaction measure.

3.2.3. Relation of the Additive Interaction Measure with Joint Cumulants

In this section we show the relationship between Lancaster interaction measure and joint cumulants. Building on the Lancaster interaction measure, we explain why joint cumulants can be interpreted as legitimate interdependence measures.

We begin by indicating two alternative notations for the cumulants. The most explicit notation is the one that includes the random variables as arguments; this is the notation used e.g. by Brillinger (1974). Thus, the joint cumulants of vector $(X_{j_1}, \dots, X_{j_k})$ are expressed by $cum(X_{j_1}, \dots, X_{j_k})$. The second notation type is the "index" notation used e.g. by McCullagh (1987), where the indexes of the vector's components, of which the joint cumulants are computed, appear as superscripts: κ^{j_1, \dots, j_k} .

Both of those notation variants have the advantage of not having to indicate explicitly the size of the vector, of which the components subset is taken. A few examples with $\mathbf{X} \in \mathbb{R}^4$ should suffice to make the relation among the three notation systems clear:

$$\begin{aligned}\kappa_{1,1,0,0} &= cum(X_1, X_2) = \kappa^{1,2} \\ \kappa_{0,0,2,0} &= cum(X_3, X_3) = \kappa^{3,3} \\ k_{1,2,0,1} &= cum(X_1, X_2, X_2, X_4) = \kappa^{1,2,2,4} \\ k_{1,1,1,0} &= cum(X_1, X_2, X_3) = \kappa^{1,2,3}\end{aligned}$$

In the following, either of the two new notations will be used.

If we concentrate for new on the case $\mathbf{X} \in \mathbb{R}^2$, then Lehmann (1966) reports that:

$$Cov(X_1, X_2) = cum(X_1, X_2) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [F_{12}(x_1, x_2) - F_1(x_1)F_2(x_2)] dx_1 dx_2 \quad (3.2.15)$$

under the condition that $E(|X_1^{k_1} X_2^{k_2}|) < +\infty$, for $k_j = 0, 1$.

This equation is often called "Hoeffding's formula" since it was discovered by Hoeffding (1940). Of course, the above equation can be written in terms of Lancaster interaction measure (3.2.13), as

$$cum(X_1, X_2) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \Delta F(x_1, x_2) dx_1 dx_2 \quad (3.2.16)$$

It turns out that this equation can be extended to higher dimensions.

Let $\mathbf{X} \in \mathbb{R}^J$ be a random vector. According to Block and Fang (1988), we have that (after suitable identification of ΔF on page 1808):

$$cum(\mathbf{X}) = (-1)^J \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \Delta F(\mathbf{X}) d\mathbf{X} \quad (3.2.17)$$

under the condition that $E(|X_j^J|) < +\infty$, for $j = 1, \dots, J$.

Thus, joint cumulants are equal (up to a known constant) to the integral of Lancaster Interaction measure; they are "summary" or "integral" measures of additive interaction. It goes without much explanation that the joint cumulants of a random vector \mathbf{X} vanish whenever a subset of the vector is independent of another, since then the integrating function is identically zero. This property is well-known and oftentimes the reason why joint cumulants are used. The only contribution here is that they are seen as the integral of the Lancaster interaction measure.

Now, as seen in Theorem 1 it is the only additive measure, built very elementarily with the marginal distributions of the random vector, which fulfills the interaction property (condition 3).

We have provided a theoretical basis for declaring joint cumulants "dependence parameters", and the cumulant generating function a "dependence structure".

3.2.3.1. Small digression: Alternative definition of joint cumulants

With the aid of the interactions operator (3.2.14) one can present, with little additional work, an alternative definition for joint cumulants. This definition is the one given, for example, at Brillinger (1974). It has the advantage of expressing joint cumulants in terms of joint moments. Sample estimates of the latter are readily obtainable.

This definition is also useful for section 4.4.2, when dealing with transformations of random vectors.

Let $\mathbf{X} \in \mathbb{R}^J$ be a random vector. For a set $(X_{j_1}, \dots, X_{j_d})$ of \mathbf{X} 's components, where some sub-indexes j_r may be repeated, consider joint moments

$$E(X_{j_1} \dots X_{j_d})$$

and a partition operator J_π^* , analogous to (3.2.12), related to each partition π of (j_1, \dots, j_d) . This operator converts $E(X_{j_1} \dots X_{j_d})$ into the product of the factors determined by partition π .

For example, for $d = 4$, (j_1, j_2, j_3, j_4) and $\pi = 1 \mid 23 \mid 4$, one has partition components $v_1 = \{1\}$, $v_2 = \{2, 3\}$ and $v_3 = \{4\}$. Upon application of J_π^* , we have,

$$J_\pi^* E(X_{j_1} \dots X_{j_d}) = E(X_{j_1}) E(X_{j_2} X_{j_3}) E(X_{j_4})$$

In the general case

$$J_\pi^* E(X_{j_1} \dots X_{j_d}) = \prod_{v \in \pi} E\left(\prod_{j_r \in v} X_{j_r}\right)$$

The alternative definition of joint cumulants can now be given.

For random variables $(X_{j_1}, \dots, X_{j_d})$, their joint cumulant of order d is given by,

$$\text{cum}(X_{j_1}, \dots, X_{j_d}) := \sum_{\pi} \left\{ \left((-1)^{|\pi|-1} (|\pi| - 1)! \right) J_\pi^* \right\} E(X_{j_1} \dots X_{j_d}) \quad (3.2.18)$$

Two examples are:

$$\text{cum}(X_1, X_2) = E(X_1 X_2) - E(X_1) E(X_2)$$

and

$$\begin{aligned} \text{cum}(X_1, X_2, X_3) = & E(X_1 X_2 X_3) - E(X_1 X_2) E(X_3) - E(X_1 X_3) E(X_2) \\ & - E(X_2 X_3) E(X_1) + 2E(X_1) E(X_2) E(X_3) \end{aligned}$$

3.3. Joint cumulants as parameters and relation to "interaction manifestations"

In this section we exhibit the relation between the cumulant generating function (our dependence structure) and the interaction manifestation parameters introduced in subsection 3.1. As before, none of the results is new, the difference is how we interpret and employ objects and results.

3.3.1. Preliminary: The Edgeworth Expansion and the Saddlepoint Approximation

The following two approximations are relevant for this work: The Edgeworth Approximation and the Saddlepoint approximation. These approximate the probability density of \mathbf{X} in terms of its joint cumulants and c.g.f, respectively.

The *Edgeworth Expansion* is a series expansion of the probability density and of the probability distribution in terms of the joint cumulants (performing as coefficients) and of the multivariate normal distribution (performing as basis function). Details for all topics of this subsection can be found in Barndorff-Nielsen and Cox (1990); we present here just the expansion, in the context of a distribution having a probability density.

We employ in this section the shorthand notation for summations used in Barndorff-Nielsen and Cox (1990), in order to avoid an overflow of symbols in these pages. Arrays are represented by symbols with superscripts and under-scripts. For example a matrix is represented by $a^{i,j}$ or by b_{ij} . A three dimensional array would be $c^{i,j,k}$ or d_{ijk} , and so on. The product of these symbols indicates summation along all dimensions for which the index is repeated. For example the term $\frac{1}{6\sqrt{n}}\kappa^{j_1,j_2,j_3}h_{j_1j_2j_3}$, to be used below, should be interpreted as

$$\frac{1}{6\sqrt{n}}\kappa^{j_1,j_2,j_3}h_{j_1j_2j_3} = \frac{1}{6\sqrt{n}} \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \sum_{j_3=1}^{J_3} \kappa^{j_1,j_2,j_3}h_{j_1j_2j_3} \quad (3.3.1)$$

We can now introduce the Edgeworth Expansion. Let $\mathbf{Z} \in \mathbb{R}^J$ be a random vector with probability density function f . Assume also, without loss of generality, that \mathbf{Z} has mean a vector of zeros, a $J \times J$ covariance matrix $\kappa^{i,j} = \Gamma$, and joint cumulants $\{\kappa^{j_1,j_2,j_3}\}, \{\kappa^{j_1,j_2,j_3,j_4}\}, \dots$. If we have a random sample of n i.i.d. random vectors with the same distribution as \mathbf{Z} , namely $\mathbf{Z}_1, \dots, \mathbf{Z}_n$, then we can form the average random vector $\mathbf{X} = \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i$. This latter random vector has a density function $f_{\mathbf{X}}$ which can be formally³ written as the following series expansion, in terms of the summation shorthand notation:

$$f_{\mathbf{X}}(\mathbf{x}) = \phi_{\Gamma}(\mathbf{x}) \left\{ 1 + \frac{1}{6\sqrt{n}}\kappa^{j_1,j_2,j_3}h_{j_1j_2j_3}(\mathbf{x};\Gamma) + \frac{1}{24n}\kappa^{j_1,j_2,j_3,j_4}h_{j_1j_2j_3j_4}(\mathbf{x};\Gamma) + \frac{1}{72n}\kappa^{j_1,j_2,j_3}\kappa^{j_4,j_5,j_6}h_{j_1j_2j_3j_4j_5j_6}(\mathbf{x};\Gamma) \right\} + O\left(n^{-\frac{3}{2}}\right) \quad (3.3.2)$$

Where ϕ_{Γ} is the multivariate Normal density function with zero mean and covariance matrix Γ , and $h_{j_1\dots j_k}(\mathbf{x};\Gamma)$ represents the evaluation at \mathbf{x} of the k -order Hermite polynomial determined by the identity

$$\phi_{\Gamma}(\mathbf{x}) h_{j_1\dots j_k}(\mathbf{x};\Gamma) = (-1)^k \frac{\partial^k \phi_{\Gamma}(\mathbf{x})}{\partial x_{j_1} \dots \partial x_{j_k}} \quad (3.3.3)$$

The reader has surely noticed that we have considered only the case of an average $\mathbf{X} = \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i$ of random vectors. However, if the distribution of \mathbf{Z} is unimodal and not wildly skewed or leptokurtic, then the Edgeworth Approximation given in 3.3.2 is often a good

³That is, without caring at the moment for issues of convergence, non-negativity, or the conditions \mathbf{Z} must fulfill in order for this to be a valid expansion.

approximation in practice even with $n = 1$, as we shall use it. The reason is that a random variable does not have to be the result of averaging n variables in order to have cumulants as such an average variable. This is the case of the chi-squared distribution with n degrees of freedom, for example, which can be interpreted as the sum of n standard Normal variables after raising each to the second power.

The usefulness of retaining the dependence on n is that we are reminded of when the Edgeworth Expansion is useful in practice: When the cumulants of \mathbf{X} , of which the density must be approximated, do not explode as their order increases, i.e. they behave as if \mathbf{X} were approximately an average.

The Edgeworth expansion is more accurate near the expected value of the distribution, but degenerates as one moves towards the tails of the distribution.

The *Saddlepoint Approximation*, also called “tilted” Edgeworth Approximation, is a more accurate approximation to the density of \mathbf{X} at the tails, which we can apply if we know its cumulant generating function $K_{\mathbf{X}}(\mathbf{t})$. In the context of considering \mathbf{X} as the mean of n copies of \mathbf{Z} , the relation between the cumulant generating functions is $K_{\mathbf{X}}(\mathbf{t}) = nK_{\mathbf{Z}}\left(\frac{\mathbf{t}}{\sqrt{n}}\right)$. As mentioned above, we shall be using this approximations as if we were dealing with a variable being the average of $n = 1$ random variables. Thus we remove in the following the dependence on such an underlying n and work directly with $K_{\mathbf{X}}(\mathbf{t})$.

We begin by a mathematical trick: we try to find the Edgeworth Expansion not of $f_{\mathbf{X}}(\mathbf{x})$, but of a related family of density functions, defined in terms of an auxiliary vector $\lambda \in \mathbb{R}^J$,

$$f_{\mathbf{X}}(\mathbf{x}; \lambda) = \exp(\mathbf{x}^T \cdot \lambda - K_{\mathbf{X}}(\lambda)) f_{\mathbf{X}}(\mathbf{x}) \quad (3.3.4)$$

The idea is, for *each* $\mathbf{x} \in \mathbb{R}^J$ to choose the most advantageous value $\hat{\lambda}$ of $\lambda \in \mathbb{R}^J$ in order to make the Edgeworth approximation $\hat{f}_{\mathbf{X}}(\mathbf{x}; \lambda)$ to $f_{\mathbf{X}}(\mathbf{x}; \lambda)$ as accurate as possible. Of course, this will provide automatically an approximation

$$\hat{f}_{\mathbf{X}}(\mathbf{x}) = \exp(K_{\mathbf{X}}(\hat{\lambda}) - \mathbf{x}^T \cdot \hat{\lambda}) \hat{f}_{\mathbf{X}}(\mathbf{x}; \hat{\lambda})$$

which is in fact what we want.

The optimum value $\hat{\lambda}$ can be proved to be the one fulfilling $\mathbf{x} = \nabla K_{\mathbf{X}}(\hat{\lambda})$, for the particular $\mathbf{x} \in \mathbb{R}^J$ in question, because then density $f_{\mathbf{X}}(\mathbf{x}; \hat{\lambda})$ corresponds to a random vector having its mean at \mathbf{x} , where the Edgeworth Approximation is most accurate. Now, under suitable regularity conditions, the leading term of the Edgeworth expansion of $f_{\mathbf{X}}(\mathbf{x}; \hat{\lambda})$ is a multivariate Normal density with covariance matrix with entries

$$(\hat{\Sigma}_{i,j}) = \frac{\partial^2 K_{\mathbf{X}}(\lambda)}{\partial \lambda_i \partial \lambda_j} \Big|_{\lambda=\hat{\lambda}}$$

evaluated at its mean; that is,

$$f_{\mathbf{X}}(\mathbf{x}; \hat{\lambda}) \approx \frac{e^0}{(2\pi)^{J/2} \det(\Sigma)^{1/2}}$$

Thus, the looked for approximation is given by

$$f_{\mathbf{X}}(\mathbf{x}) = \exp(K_{\mathbf{X}}(\hat{\lambda}) - \mathbf{x}^T \cdot \hat{\lambda}) f_{\mathbf{X}}(\mathbf{x}; \hat{\lambda}) \approx \frac{\exp(K_{\mathbf{X}}(\hat{\lambda}) - \mathbf{x}^T \cdot \hat{\lambda})}{(2\pi)^{J/2} \det(\hat{\Sigma})^{1/2}} \quad (3.3.5)$$

The error of this approximation is of order $O(n^{-1})$ for all $\mathbf{x} \in \mathbb{R}^J$, if the joint cumulants of random vector \mathbf{X} behave like an average of n iid random vectors. Suitable normalization can bring this order to $O(n^{-2})$.

In spite of the apparent disadvantage of having to re-compute the density estimation for each \mathbf{x} , the computational cost becomes considerably smaller than that of the Edgeworth Approximation as dimension increases, since the number of multivariate Hermite polynomials at 3.3.2 to evaluate increases exponentially with the dimension of \mathbf{x} .

3.3.2. Connection of dependence structure with interaction manifestations

We saw at section 3.3.1 that joint cumulants, by themselves or arranged in the form of a cumulant generating function, can be "inverted" in order to find approximately the probability density to which they correspond, via either the Edgeworth expansion or the Saddlepoint approximation. We shall see now explicitly the connection of joint cumulants with the three interaction manifestation parameters listed at section 3.1.

3.3.2.1. Connection of dependence structure with "joint" quantiles

In order to find probabilities of the form $\Pr(\mathbf{X} \geq \mathbf{x}) = 1 - F_{\mathbf{X}}(\mathbf{x})$, one should in principle integrate expression 3.3.5.

In the univariate case, it is a well-established practice Huzurbazar (1999) to employ instead an accurate approximation to that integral, which is due to Lugannani and Rice (1980). Namely, in the univariate case, we have:

$$F_X(x_0) \approx \int_{-\infty}^{x_0} \frac{\exp\left(K_X(\hat{\lambda}(x)) - x\hat{\lambda}(x)\right)}{(2\pi)^{1/2} \left(\frac{d^2 K_X(\lambda)}{d\lambda^2} \Big|_{\lambda=\hat{\lambda}(x)}\right)^{1/2}} dx \approx \Phi(r) + \phi(r) \left\{ \frac{1}{r} - \frac{1}{q} \right\} \quad (3.3.6)$$

Where $\hat{\tau}$ is such that $K'_X(\hat{\tau}) = x_0$, and:

$$\begin{aligned} r &= \text{sign}(\hat{\tau}) \{2[\hat{\tau}x_0 - K_X(\hat{\tau})]\}^{\frac{1}{2}} \\ q &= \hat{\tau} \left\{ \frac{d^2 K_X(\lambda)}{d\lambda^2} \Big|_{\lambda=\hat{\tau}} \right\}^{\frac{1}{2}} \end{aligned}$$

Thus, we must not perform the numerical integration at all.

For the multivariate case, Kolassa and Li (2010) have provided a generalization of the Lugannani-Rice formula, which produces an approximation to probability $\Pr(\mathbf{X} \geq \mathbf{x})$ of order $O(n^{-1})$, for $\mathbf{X} \in \mathbb{R}^J$. This formula is extremely complicated and writing it here will most likely obscure rather than clarify anything. Only the probability distribution function of a multivariate Normal distribution with covariance matrix given by

$$\Gamma_{ij} = \frac{\partial^2}{\partial t_i \partial t_j} K_{\mathbf{X}}(\mathbf{t}) \Big|_{\mathbf{t}=\mathbf{0}}$$

must be computed. For this task there are accurate methods available for up to 20 dimensions Genz (1993).

Since we intend to deal with vectors of dimension 3 or 4, corresponding to multidimensional marginals of the random field modeled, we consider more convenient to use numerical integration of (3.3.5). For higher dimensions it would be better to use the result of Kolassa and Li (2010) in order to avoid difficult and inaccurate integrations.

3.3.2.2. Connection of dependence structure with entropy

It is possible to approximate the entropy of a distribution via the Edgeworth Expansion presented above, by using the technique presented at Hulle (2005). All the mathematical ammunition we need is the classical expansion of the (natural) logarithm when $0 < x < 2$, the orthogonality property of the multivariate Hermite Polynomials defined by 3.3.3, and a property of the entropy of a Gaussian random vector.

For the natural logarithm, it is well known that,

$$\log(x) = (x-1) - \frac{(x-1)^2}{2} + \frac{(x-1)^3}{3} - \frac{(x-1)^4}{4} + \dots \quad (3.3.7)$$

Concerning the Hermite polynomials, regardless of covariance matrix Γ ,

$$\int \dots \int h_{i_1 \dots i_k}(\mathbf{x}; \Gamma) h_{j_1 \dots j_l}(\mathbf{x}; \Gamma) \phi_\Gamma(\mathbf{x}) d\mathbf{x} = 0, \text{ if } k \neq l \quad (3.3.8)$$

Note also the particular case $h_0(\mathbf{x}; \Gamma) = 1$ for one of the Hermite Polynomials above.

Finally, for any random vector \mathbf{X} with mean zero, covariance matrix Γ and probability density function $f_{\mathbf{X}}$, we have

$$\int f_{\mathbf{X}}(\mathbf{x}) \log(f_{\mathbf{X}}(\mathbf{x})) d\mathbf{x} = H(f_{\mathbf{X}}) = H(\phi_\Gamma) - \int f_{\mathbf{X}}(\mathbf{x}) \log\left(\frac{f_{\mathbf{X}}(\mathbf{x})}{\phi_\Gamma(\mathbf{x})}\right) d\mathbf{x} \quad (3.3.9)$$

where ϕ_Γ is the multivariate Normal density with mean zero and covariance matrix Γ .

Using the shorthand notation of 3.3.1, define $Z(\mathbf{x}) := \frac{1}{3!} \kappa^{j_1, j_2, j_3} h_{j_1 j_2 j_3}(\mathbf{x}; \Gamma)$. Now we can easily follow argument of Hulle (2005), which utilizes only the first correction term in 3.3.2:

$$\begin{aligned} \int f_{\mathbf{X}}(\mathbf{x}) \log(f_{\mathbf{X}}(\mathbf{x})) d\mathbf{x} &= H(\phi_\Gamma) - \int f_{\mathbf{X}}(\mathbf{x}) \log\left(\frac{f_{\mathbf{X}}(\mathbf{x})}{\phi_\Gamma(\mathbf{x})}\right) d\mathbf{x} \\ &\approx H(\phi_\Gamma) - \int \phi_\Gamma(\mathbf{x}) (1 + Z(\mathbf{x})) \log(1 + Z(\mathbf{x})) d\mathbf{x} \\ &\approx H(\phi_\Gamma) - \int \phi_\Gamma(\mathbf{x}) \left(Z(\mathbf{x}) + \frac{1}{2} Z(\mathbf{x})^2 \right) d\mathbf{x} = H(\phi_\Gamma) - \frac{1}{12} \left\{ \sum_{j=1}^J (\kappa^{j,j,j})^2 \right. \\ &\quad \left. + 3 \sum_{i,j=1, i \neq j}^J (\kappa^{i,i,j})^2 + \frac{1}{6} \sum_{i,j,k=1, i < j < k}^J (\kappa^{i,j,k})^2 \right\} \quad (3.3.10) \end{aligned}$$

The value of $H(\phi_\Gamma)$ can be found in closed form, $H(\phi_\Gamma) = \frac{1}{2} \log(\det(\Gamma)) + \frac{J}{2} \log(2\pi) + \frac{J}{2}$. The approximation (3.3.10) is accurate to order $O(n^{-2})$. At Hulle (2005), the properties of this approximation are studied.

It might be necessary sometimes to apply a simple transformation, such as $c\mathbf{X} := \mathbf{Y}$, with $c < 1$ a small constant, in order to ensure the validity of the logarithm expansion. Then the resulting random vector \mathbf{Y} has joint cumulants $\tilde{\kappa}$ given by

$$\begin{aligned}\tilde{\kappa}^{j_1, j_2} &= c^2 \kappa^{j_1, j_2} \\ \tilde{\kappa}^{j_1, j_2, j_3} &= c^3 \kappa^{j_1, j_2, j_3} \\ \tilde{\kappa}^{j_1, j_2, j_3, j_4} &= c^4 \kappa^{j_1, j_2, j_3, j_4} \\ &\vdots\end{aligned}\tag{3.3.11}$$

with which we can work.

3.3.2.3. Connection of dependence structure with the distribution of the components sum

We address now a connection that is very relevant for rainfall modeling and its impact quantification. It is also the most straightforward connection between the dependence structure of a distribution, as given by its cumulants or c.g.f., and the interaction manifestation parameters described in section 3.1.

Given a random vector $\mathbf{X} \in \mathbb{R}^J$ representing, say, rainfall at a given time at several locations on a basin, we are interested in the distribution of variable $S_{\mathbf{X}} = \sum_{j=1}^J X_j$. The characteristics of this new random variable can be, to a great extent, be identified on the basis of its moments or cumulants.

Now, two of the properties of joint cumulants are Brillinger (1974) symmetry and multilinearity. Symmetry means that $\kappa^{j_1, \dots, j_k} = \kappa^{P(j_1, \dots, j_k)}$ for any permutation $P(j_1, \dots, j_k)$ of the indexes (j_1, \dots, j_k) . Concerning multilinearity, write joint cumulants more explicitly as $\text{cum}(X_{j_1}, \dots, X_{j_k}) := \kappa^{j_1, \dots, j_k}$.

Then, for any random variable $Z \in \mathbb{R}$,

$$\text{cum}(Z + X_{j_1}, \dots, X_{j_k}) = \text{cum}(Z, \dots, X_{j_k}) + \text{cum}(X_{j_1}, \dots, X_{j_k})$$

With the aid of these two properties, it can be shown that

$$\kappa_r(S_{\mathbf{X}}) = \sum_{j_1=1}^J \left[\sum_{j_2=1}^J \dots \left[\sum_{j_r=1}^J \kappa^{j_1, \dots, j_r} \right] \right] \tag{3.3.12}$$

where $\kappa_r(S_{\mathbf{X}})$ denotes the r -th cumulant of random variable $S_{\mathbf{X}} = \sum_{j=1}^J X_j$, and κ^{j_1, \dots, j_r} denote the joint cumulants of the random field under analysis, $\mathbf{X} \in \mathbb{R}^J$.

3.3.2.3.1. Cumulant generating function of sums Additionally, given a random vector $\mathbf{X} = (X_1, \dots, X_J)$, one can study the joint distribution of aggregated variables of the form:

$$\begin{aligned}\xi_1 &= \sum_{j_1 \in I_1} X_{j_1} \\ \xi_2 &= \sum_{j_2 \in I_2} X_{j_2} \\ &\vdots \\ \xi_l &= \sum_{j_l \in I_l} X_{j_l}\end{aligned}\tag{3.3.13}$$

$$\begin{aligned}&\vdots \\ \xi_l &= \sum_{j_l \in I_l} X_{j_l}\end{aligned}\tag{3.3.14}$$

where I_k , for $k = 1, \dots, l$ represent non-overlapping index sets such that

$$I_1 \cup \dots \cup I_l = \{1, \dots, J\}$$

The cumulant generating function of the l -dimensional vector so obtained is given by

$$\begin{aligned}K_\xi(\mathbf{t}) &= \log \left(E \left(\exp \left(\mathbf{t} \cdot \boldsymbol{\xi}' \right) \right) \right) = \\ &\quad \log \left(E \left(\exp \left(t_1 \xi_1 + \dots + t_l \xi_l \right) \right) \right) = \\ &\quad \log \left(E \left(\exp \left(t_1 \sum_{I_1} X_{j_1} + \dots + t_l \sum_{I_l} X_{j_l} \right) \right) \right) = \\ &\quad \log \left(E \left(\exp \left(g_1(\mathbf{t}) X_1 + \dots + g_J(\mathbf{t}) X_J \right) \right) \right) = \\ &\quad \log \left(E \left(\exp \left(g(\mathbf{t}) \cdot \mathbf{X}' \right) \right) \right) = K_{\mathbf{X}}(g(\mathbf{t}))\end{aligned}\tag{3.3.15}$$

Function $g : \mathbb{R}^l \rightarrow \mathbb{R}^J$ is a vector function defined by

$$\begin{aligned}g(\mathbf{t}) &= (g_1(\mathbf{t}), \dots, g_J(\mathbf{t})) \\ g_j(\mathbf{t}) &= \mathbf{t} \cdot (\mathbf{1}(j \in I_1), \dots, \mathbf{1}(j \in I_l))'\end{aligned}\tag{3.3.16}$$

where

$$\mathbf{1}(j \in I_k) = \begin{cases} 1, & j \in I_k \\ 0, & j \notin I_k \end{cases}$$

Summarizing, it is possible to find the cumulant generating function of random vector $\boldsymbol{\xi} \in \mathbb{R}^l$ in terms of that of the original vector $\mathbf{X} \in \mathbb{R}^J$. Then, if we know the c.g.f. of the original random vector \mathbf{X} , the cumulants, the cumulant generating function (and hence the approximate density, via Saddlepoint approximation) of $\boldsymbol{\xi} \in \mathbb{R}^l$ can be found. In this way it is possible to deal with interaction manifestations of these aggregate variables, as well. An example should help clarified these statements.

Example 2. Runoffs to a dam

For example, let $\mathbf{X} = (X_1, \dots, X_4)$ represent runoffs to a dam, during a specific 6-hour time-span. At Mathai and Moschopoulos (1991), it is suggested that this process can be modeled

by a random vector having dependence structure

$$K_{\mathbf{X}}(s_1, \dots, s_4) = \left(\mathbf{h} + \left(\frac{\gamma_0}{\beta_0} \right) \mathbf{b} \right) \cdot \mathbf{s}' - \alpha_0 \log(1 - \mathbf{b} \cdot \mathbf{s}') - \sum_{j=1}^4 \alpha_j \log(1 - \beta_j s_j) \quad (3.3.17)$$

for parameters $\alpha_0, \beta_0 > 0$, $\gamma_0 \geq 0$, $\mathbf{h} = (\gamma_1, \dots, \gamma_4) \geq \mathbf{0}$, $\mathbf{b} = (\beta_1, \dots, \beta_4) \geq \mathbf{0}$, and $(\alpha_1, \dots, \alpha_4) \geq \mathbf{0}$. The explanation and physical interpretations of these parameters can be found at Mathai and Moschopoulos (1991).

We are interested in the c.g.f. of $\mathbf{Y} = (Y_1, Y_2)$, where

$$\begin{aligned} Y_1 &= X_1 + X_2 \\ Y_2 &= X_3 + X_4 \end{aligned} \quad (3.3.18)$$

Then, we have

$$\begin{aligned} K_{\mathbf{Y}}(t_1, t_2) &= \log(E(\exp(t_1 Y_1 + t_2 Y_2))) = \\ &= \log(E(\exp(t_1(X_1 + X_2) + t_2(X_3 + X_4)))) = \\ &= \log(E(\exp(t_1 X_1 + t_1 X_2 + t_2 X_3 + t_2 X_4))) = \\ &= K_{\mathbf{X}}(g(t_1, t_2)) \end{aligned}$$

where

$$g(t_1, t_2) = (t_1, t_1, t_2, t_2)$$

Then, the dependence structure of \mathbf{Y} is given by

$$\begin{aligned} K_{\mathbf{Y}}(t_1, t_2) &= \left(\mathbf{h} + \left(\frac{\gamma_0}{\beta_0} \right) \mathbf{b} \right) \cdot (t_1, t_1, t_2, t_2)' \\ &= \alpha_0 \log(1 - \mathbf{b} \cdot (t_1, t_1, t_2, t_2)') - \sum_{j=1}^2 \alpha_j \log(1 - \beta_j t_1) \\ &\quad - \sum_{j=3}^4 \alpha_j \log(1 - \beta_j t_2) \end{aligned} \quad (3.3.19)$$

Joint cumulants (our interdependence) parameters can be found by differentiation of (3.3.19) and evaluation at zero. For example, covariance is given by

$$\begin{aligned} \text{cum}(Y_1, Y_2) &= \frac{\partial^2}{\partial t_2 \partial t_1} K_{\mathbf{Y}}(t_1, t_2) \big|_{(t_1, t_2) = (0, 0)} = \\ &= \alpha_0 (\beta_1 + \beta_2) (\beta_3 + \beta_4) \end{aligned} \quad (3.3.20)$$

Higher order cumulants can be found similarly.

If the original vector \mathbf{X} has now dimension 9, for example, and at (3.3.18) we are also interested in a third variable, say

$$Y_3 = X_5 + \dots + X_9$$

then along similar lines as above, and now considering $\mathbf{b} = (\beta_1, \dots, \beta_9)$, etc., we would have

$$\begin{aligned}
K_{\mathbf{Y}}(t_1, t_2, t_3) = & \left(\mathbf{h} + \left(\frac{\gamma_0}{\beta_0} \right) \mathbf{b} \right) \cdot (t_1, t_1, t_2, t_2, t_3, \dots, t_3)' \\
& - \alpha_0 \log \left(1 - \mathbf{b} \cdot (t_1, t_1, t_2, t_2, t_3, \dots, t_3)' \right) - \sum_{j=1}^2 \alpha_j \log (1 - \beta_j t_1) \\
& - \sum_{j=3}^4 \alpha_j \log (1 - \beta_j t_2) - \sum_{j=5}^9 \alpha_j \log (1 - \beta_j t_3) \quad (3.3.21)
\end{aligned}$$

and would have 3rd order interdependence parameters of the form:

$$\begin{aligned}
cum(Y_i, Y_j, Y_k) = & \frac{\partial^3}{\partial t_i \dots \partial t_k} K_{\mathbf{Y}}(t_1, t_2, t_3) |_{\mathbf{t}=\mathbf{0}} \\
= & 2\alpha_0 (\beta_1 + \beta_2) (\beta_3 + \beta_4) (\beta_5 + \dots + \beta_9)
\end{aligned}$$

The case of higher dimensions and more summary variables can be found analogously. In the example above, the distribution of the “parent” random vector \mathbf{X} uses many parameters, which is inconvenient for Spatial Statistics. Below we see a parametrically lower dimensional model that we consider suitable for Spatial Statistics.

3.3.3. Putting the pieces together

We have seen in this section that joint cumulants can be rightfully defined as “summary” or “integral” measures of multivariate interaction. For this reason, we considered joint cumulants and cumulant generating functions as dependence parameters and dependence structures, respectively. We have seen how to express the three (subject-matter specific) interaction manifestations presented in subsection 3.1 in terms of these dependence parameters and structure.

The idea of our approach is to fit the dependence parameters and/or the dependence structure in such a way that the observed interaction manifestations are faithfully reproduced. This amounts to a method of moment estimation procedure, or “method of cumulants”, should we say.

The interaction manifestations to be reproduced may refer to those of low-dimensional marginals, such as the 4 or 5 dimensional marginal distributions. For example, the entropy of the four-dimensional marginals of the distribution. Actually, only up to such dimensions can we estimate anything from the sample with precision.

But the model will be “glued” together by the dependence structure (our c.g.f.), which can span hundreds or thousands of dimensions. It is usual in Geostatistics to deal with components of the field two at a time, by using some kind of covariance function, such as (2.2.8) or (2.2.9), trying to model properly covariance. In this way a model of hundreds or thousands of dimensions is “glued” by its 2-dimensional marginals. In this work, we attempt to work with 4 or 5 components at a time, trying to model properly both covariance and other interaction manifestations (say, entropy) appearing in the 4 or 5 dimensional marginals, for example.

The dependence structure provided by the cumulants or the cumulant generating function can be controlled to be low-dimensional, since it is a parametric function. Thus we counteract the worst enemy of interaction quantification: the "curse" of dimensionality. The issue of interpolation to ungauged sites mentioned before, to which the 4 or 5 dimensional manifestations must also be carried over as faithfully as possible, can also be tackled in this way (see chapter 4).

In spite of the issues mentioned in section 2.4, in two dimensions dependence parameters are often considered practically the same thing as the interaction manifestations parameters: both are summarized by any of the several correlation coefficients available, such as Kendall's τ , Spearman's ρ , the product moment correlation coefficient, etc., and so they can be readily integrated into modeling.

We mentioned the use of Copulas as an alternative to represent and analyze bi-variate interactions. A Copula conveys a tremendous amount of information about dependence among two variables, in that its density informs at what ranges dependence is higher, lower, etc. Unfortunately, non-parametrically estimated Copulas are useful only in relatively low dimensions. In higher dimensions Copulas themselves require some parametric specification to model the observed dependence, these parameters being usually either the correlation parameters just mentioned, or some version of the same concept.

We suggested in this work, that data features which can be interpreted as dependence features ("interaction manifestations") also suffer from a version of the curse of dimensionality: they increase exponentially. The reader may wish to add his/her favorite interaction manifestation to the list given in subsection 3.1. But "this" or "that" high dimensional interaction manifestation usually does not provide any guide for model building, let alone for low-dimensional model building.

For this reason, we propose to look for a flexible dependence structure, in terms of the cumulants as building blocks, and try to see how well or bad the manifestations are recovered. In this, joint cumulants are employed as the correlation coefficients above, of which they are a generalization, but we don't expect to compute them directly on the basis of the data; that would require too many data. We rather attempt to match the interaction manifestations: joint cumulants are estimated, so that these manifestations are better reproduced.

4. The Proposed Approach: Spatial Statistics

In this chapter, the ideas put forward in the previous chapters are further elaborated with a view to their application in spatial statistics. As our model, a low dimensional, “archetypal” cumulant generating function (dependence structure) is provided. This dependence structure resembles that of the Gaussian distribution usually used in Geo-statistics, but allows for non-zero interdependence parameters of order greater than two.

By using the Saddlepoint Approximation method introduced in section 3.3.1, this c.g.f. can be (approximately) inverted. In this sense, we are actually proposing a density model for \mathbf{X} . Alternatively, we shall see that the c.g.f. can be identified to be that of an elliptically contoured random vector, and accordingly a useful density model corresponding to this c.g.f. will be presented in this chapter.

A method for giving more flexibility to the archetypal dependence structure is presented and illustrated. It consists in applying polynomial transformations to the one-dimensional marginal random variables of the random vector having a simple, low-dimensional dependence structure. Joint cumulants of the transformed random vector can then be found in terms of the polynomial transformation coefficients and the original dependence structure parameters.

This chapter is mostly theoretical. The reader is referred to chapter 5, where two examples are presented that illustrate the differences in interaction manifestations that can occur between distributions indistinguishable from their one or two-dimensional marginal distributions, but having different interaction parameters of order 4 or 6.

4.1. Archetypal Dependence Structure

We introduce in this section a series expansion that represents a cumulant generating function (dependence structure) for a random vector $\mathbf{X} \in \mathbb{R}^J$. Examples of the characteristics of data obtainable from this model can be seen in chapter 5.

We begin with a definition due to Cambanis et al. (1981): A random vector $\mathbf{X} \in \mathbb{R}^J$ is said to have an elliptically contoured distribution if there exists a vector $\mu \in \mathbb{R}^J$ and an $J \times J$ non-negative definite (covariance) matrix Γ such that the characteristic function Ψ of $\mathbf{X} - \mu$ can be written as

$$\Psi_{\mathbf{X}-\mu}(\mathbf{t}) = \Upsilon(\mathbf{t}^T \Gamma \mathbf{t}) \quad (4.1.1)$$

for some function $\Upsilon : \mathbb{R} \rightarrow \mathbb{R}$.

We suggest that elliptical distributions provide a departing point for implementing the ideas presented in this dissertation. We shall be assuming the existence of sufficiently many joint cumulants (or product moments) so as to provide a practically useful approximation to the

processes modeled. Then it is more convenient, for our purposes, to conceptualize elliptical distributions in terms of their moment generating function: We say that random vector $\mathbf{X} \in \mathbb{R}^J$ is “elliptically distributed” if and only if its moment generating function can be written as

$$M_{\mathbf{X}-\mu}(\mathbf{t}) = \Upsilon(\mathbf{t}^T \Gamma \mathbf{t}) \quad (4.1.2)$$

for some function $\Upsilon : \mathbb{R} \rightarrow \mathbb{R}$, and some $\mu \in \mathbb{R}^J$.

The most famous distribution of the elliptical family is the multivariate Normal distribution, for which $\Upsilon(y) = \exp(\frac{1}{2}y)$. For ease of notation and argument we assume in the following, without loss of generality, that $\mu = \mathbf{0}$.

We shall be considering moment generating functions reminiscent of that of the multivariate Normal distribution. The corresponding cumulant generating function will be an archetypical dependence structure for spatial statistics applications, for reasons that will be explained shortly.

Consider a moment generating function of the form

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\delta\left(\frac{1}{2}\mathbf{t}^T \Gamma \mathbf{t}\right)\right) \quad (4.1.3)$$

for some function $\delta : \mathbb{R} \rightarrow \mathbb{R}$. Then the cumulant generating function of \mathbf{X} is given by

$$K_{\mathbf{X}}(\mathbf{t}) = \delta\left(\frac{1}{2}\mathbf{t}^T \Gamma \mathbf{t}\right) \quad (4.1.4)$$

This function $\delta(y)$ can be formally expanded in its Taylor Series around zero,

$$\begin{aligned} \delta(y) &= c_0 + \frac{c_1}{1!}y + \frac{c_2}{2!}y^2 + \frac{c_3}{3!}y^3 + \frac{c_4}{4!}y^4 + \dots \\ &= c_0 + \frac{c_1}{1!}\left(\frac{1}{2}\mathbf{t}^T \Gamma \mathbf{t}\right) + \frac{c_2}{2!}\left(\frac{1}{2}\mathbf{t}^T \Gamma \mathbf{t}\right)^2 + \frac{c_3}{3!}\left(\frac{1}{2}\mathbf{t}^T \Gamma \mathbf{t}\right)^3 + \dots \\ &= \delta\left(\left(\frac{1}{2}\mathbf{t}^T \Gamma \mathbf{t}\right)\right) \end{aligned} \quad (4.1.5)$$

where $c_r = \frac{d^r}{dy^r} \delta(y) |_{y=0}$.

A little thought shows that the assumption $\mu = \mathbf{0}$ implies that $c_0 = 0$. Thus, by virtue of 4.1.4 and 4.1.5 combined, we have that the c.g.f can be written as

$$K_{\mathbf{X}}(\mathbf{t}) = c_1 \frac{1}{2}\mathbf{t}^T \Gamma \mathbf{t} + \frac{1}{2!}c_2 \left[\frac{1}{2}\mathbf{t}^T \Gamma \mathbf{t}\right]^2 + \frac{1}{3!}c_3 \left[\frac{1}{2}\mathbf{t}^T \Gamma \mathbf{t}\right]^3 + \dots \quad (4.1.6)$$

Steyn (1993) introduces cumulant generating functions of the form 4.1.6. The idea of Steyn (1993) is to come up with distributions similar to those in the elliptical family, but with added flexibility, in that they can have different kurtosis for the different marginal distributions. The apparently capricious inclusion of $\frac{1}{2}$ in the argument of δ allows us to use directly Steyn’s illustrative result, but it could also be “absorbed” by the coefficients c_r .

As far as we are concerned, we shall use an expansion of the c.g.f up to some practical order R , which shall constitute the *dependence structure* later to be connected with the interaction manifestations, presented in the previous chapters. The parameters c_r are extremely

important in our approach, since they provide the means for quantifying the (summary) interactions of order greater than two, with a minimum of additional parameters.

At the same time, the form of $K_{\mathbf{X}}(\mathbf{t})$, built on expression $(\mathbf{t}^T \Gamma \mathbf{t})$, allows us to use available spatial statistics techniques for estimation of correlation, i.e. based on covariance functions. Then these correlations can be “enhanced” or complemented with higher order interdependence, via non-zero values for coefficients $c_{r>1}$.

The joint cumulants of a random vector having a c.g.f as in 4.1.6 are readily found by differentiating $K_{\mathbf{X}}(\mathbf{t})$ with respect to the indexes of the joint cumulant, and evaluating the result at $\mathbf{t} = \mathbf{0}$. This is entirely analogous to finding moments with the aid of the moment generating function.

All joint cumulants of odd order, κ^{j_1, \dots, j_k} (k odd), are zero for our dependence model. Some of the non-zero joint cumulants are:

$$\begin{aligned} \kappa^{j_1, j_2} &= c_1 \Gamma_{j_1 j_2} \\ \kappa^{j_1, j_2, j_3, j_4} &= c_2 \{ \Gamma_{j_1 j_2} \Gamma_{j_3 j_4} + \Gamma_{j_1 j_3} \Gamma_{j_2 j_4} + \Gamma_{j_1 j_4} \Gamma_{j_2 j_3} \} \\ \kappa^{j_1, j_2, j_3, j_4, j_5, j_6} &= c_3 \{ \Gamma_{j_1 j_2} \Gamma_{j_3 j_4} \Gamma_{j_5 j_6} + \dots + \Gamma_{j_1 j_6} \Gamma_{j_2 j_4} \Gamma_{j_5 j_3} \} \end{aligned} \quad (4.1.7)$$

and so on (see the appendix). In this manner, interaction among sets of four or six variables can be conveniently summarized.

Joint cumulants of order r , κ^{j_1, \dots, j_r} , are similar. They are the product of c_r times the summation of the product of all covariances involved. It will be convenient to introduce “covariance interdependence factor” $\varrho(j_1, \dots, j_k)$ defined as the sum of the products of the covariance coefficients at (4.1.7). Specifically,

$$\begin{aligned} \varrho(j_1, j_2) &= \Gamma_{j_1 j_2} \\ \varrho(j_1, \dots, j_4) &= \Gamma_{j_1 j_2} \Gamma_{j_3 j_4} + \Gamma_{j_1 j_3} \Gamma_{j_2 j_4} + \Gamma_{j_1 j_4} \Gamma_{j_2 j_3} \\ \varrho(j_1, \dots, j_6) &= \Gamma_{j_1 j_2} \Gamma_{j_3 j_4} \Gamma_{j_5 j_6} + \Gamma_{j_1 j_3} \Gamma_{j_2 j_4} \Gamma_{j_5 j_6} + \dots + \Gamma_{j_1 j_6} \Gamma_{j_2 j_4} \Gamma_{j_5 j_3} \end{aligned}$$

and so on. This is a “potential” interdependence factor, since its effect on higher order interdependence parameters (i.e. joint cumulants of order greater than 2), is only present if its corresponding coefficient $c_{k/2}$ is non-zero. So every joint cumulant at (4.1.7) can be written as

$$\kappa^{j_1, \dots, j_k} = c_{\frac{k}{2}} \times \varrho(j_1, \dots, j_k) \quad (4.1.8)$$

Our interdependence parameter of order $k > 2$ can then be conceptually split into two components: On the one hand, a “covariance interdependence component”, $\varrho(j_1, \dots, j_k)$, that can be estimated via covariance function fitting, as usual in Geo-statistics. On the other hand, an interdependence “enhancing” parameter $c_{k/2}$, whose departure from zero determines the departure from zero of the k -th order joint cumulant.

4.1.1. Moment generating function

The moment generating function of the archetypal dependence structure will be now introduced. The form of the dependence structure, makes this function readily obtainable.

For random vector $\mathbf{X} \in \mathbb{R}^J$, our dependence structure is given by (4.1.6). By setting shorthand notation

$$y := \frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t}$$

the dependence structure can be written

$$K_{\mathbf{X}}(\mathbf{t}) = \frac{c_1}{1!} y + \frac{c_2}{2!} y^2 + \frac{c_3}{3!} y^3 + \dots \quad (4.1.9)$$

On the other hand, the definition of our dependence structure, given originally by (4.1.3) implies that we can write, using the same shorthand notation as above,

$$\begin{aligned} \exp(K_{\mathbf{X}}(\mathbf{t})) &:= M_{\mathbf{X}}(\mathbf{t}) = \\ \exp(\delta(y)) &= 1 + \frac{m_1}{1!} y + \frac{m_2}{2!} y^2 + \frac{m_3}{3!} y^3 + \dots \end{aligned} \quad (4.1.10)$$

for some coefficients m_1, m_2, m_3, \dots , at least for y in a neighborhood of zero (that is, for \mathbf{t} in a sufficiently small neighborhood of $\mathbf{0}$). Summarizing, we have that

$$\log\left(1 + \frac{m_1}{1!} y + \frac{m_2}{2!} y^2 + \frac{m_3}{3!} y^3 + \dots\right) = \frac{c_1}{1!} y + \frac{c_2}{2!} y^2 + \frac{c_3}{3!} y^3 + \dots \quad (4.1.11)$$

and then we can obtain, as in the case of the one-dimensional cumulants in terms of the one-dimensional moments, coefficients m_1, m_2, m_3, \dots in terms of c_1, c_2, c_3, \dots (see section 3.2.1). According to (3.2.8), we have,

$$\begin{aligned} c_1 &= m_1 \\ c_2 &= m_2 - m_1^2 \\ c_3 &= m_3 - 3m_2m_1 + 2m_1^3 \\ c_4 &= m_4 - 4m_3m_1 - 3m_2^2 + 12m_2m_1^2 - 6m_1^4 \end{aligned} \quad (4.1.12)$$

which after some algebraic manipulation, returns,

$$\begin{aligned} m_1 &= c_1 \\ m_2 &= c_2 + c_1^2 \\ m_3 &= c_3 + 3c_2c_1 + c_1^3 \\ m_4 &= c_4 + 4c_3c_1 + 3c_2^2 + 6c_2c_1^2 + c_1^4 \end{aligned} \quad (4.1.13)$$

So, we have shown, that the moment generating function at (4.1.3) can be written as

$$M_{\mathbf{X}}(\mathbf{t}) = 1 + \frac{m_1}{1!} \left(\frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t}\right) + \frac{m_2}{2!} \left(\frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t}\right)^2 + \dots \quad (4.1.14)$$

which is similar to the expansion of $K_{\mathbf{X}}(\mathbf{t})$, except for the leading term 1 and coefficients $m_r, r = 1, 2, \dots$. We express joint moments analogously as joint cumulants by

$$\mu^{j_1, \dots, j_k} := E(X_{j_1} \dots X_{j_k}) \quad (4.1.15)$$

where $j_r \in \{1, \dots, J\}$, $r = 1, \dots, k$, allowing repetition of indexes. Then it follows that

$$\begin{aligned}\mu^{j_1, j_2} &= m_1 \varrho(j_1, j_2) \\ \mu^{j_1, j_2, j_3, j_4} &= m_2 \varrho(j_1, \dots, j_4) \\ \mu^{j_1, j_2, j_3, j_4, j_5, j_6} &= m_3 \varrho(j_1, \dots, j_6)\end{aligned}\tag{4.1.16}$$

and so on, where m_r is as in (4.1.13). We see then, for example by setting $c_1 = 1$ and $c_{r>1} = 0$, that we can have non-zero joint moments of orders greater than two, even though no dependence of order greater than two is present in the distribution of \mathbf{X} , according to our definition. This indicates an important difference between joint moments and joint cumulants, in terms of our research.

4.1.2. Some advantages of the archetypal dependence structure

Concerning Spatial Statistics, the archetypal dependence structure introduced above has the following convenient characteristics:

Low dimensionality: With respect to covariance function based Geo-statistics, we need to estimate only one additional parameter per additional order of interaction. For example, assume we want to model interaction among ten variables, as in the example at the outset of subsection 3.1. It was stated that, in principle, we would have to fit a set of $O(10^{10})$ parameters. Under the archetypal dependence structure model, by using a covariance function such as (2.2.8), we need only to estimate four parameters for the covariance matrix, plus five additional ones: c_1, \dots, c_5 .

Geographically Sensible: If we employ a covariance function model for estimating covariances, then covariance between values of every two sites decrease as a function of distance. Since the joint cumulants at (4.1.7) for a given set (j_1, \dots, j_k) of components of field $\mathbf{X} \in \mathbb{R}^J$ are partially a function of covariances between pairs of components, values corresponding to sites closer together will result in higher k -th order interdependence. On the other hand, values corresponding to sites further apart, will result in smaller interdependence parameters, in a way that respects the principle of correlation between pairs of sites-data as a function of distance between sites. This aspect is noticed in the smooth, apparently Gaussian look of the fields generated for the sake of illustration is chapter 5.

Flexibility according to data complexity: In section 2.5 we saw that one of the strengths of the Edgeworth-Sargan distribution was its capacity to be extended, in a natural way, depending on the (subject-matter relevant) features of data to be modeled. We saw as well, that features of this distribution (e.g. location, dispersion, skewness, kurtosis, and moments of order $k > 4$ in general) could be sequentially or “orthogonally” fitted as data complexity required, without altering features already fitted. These characteristics are partially recovered by the archetypal model suggested: joint cumulants of increasing orders can be fitted “orthogonally”, by fitting coefficients c_2, c_3, c_4, \dots . For example, it is possible to have two models with the same covariance matrix, but different fourth order joint cumulants, depending on the values for c_2 fitted for each model.

Handy measure of higher order interdependence: Given a multivariate data set that seems roughly Gaussian (i.e. symmetric, uni-modal, with roughly Normal 1-dimensional marginals), fitting parameters of the model presented in this chapter, allows an immediate measure of interdependence beyond order two interdependence ("gaussianity"). Namely, the following divergence measure:

$$div.measure = \hat{c}_2^2 + \hat{c}_3^2 + \hat{c}_4^2 + \dots \quad (4.1.17)$$

Closeness under marginalization: Marginal distributions of this model, both one-dimensional and multidimensional, belong to the same distribution type. This is a characteristic of distributions whose moment generating functions are of the form (4.1.2), since they are just instances of elliptical distributions for which the form of the characteristic does not change with the dimension of \mathbf{X} . See, for example Hult and Lindskog (2002); Kano (1994). Closeness under marginalization is a sensible requirement for Spatial Statistics models, since data from the environmental variable is often available at a limited number of sites, say J . We seek then estimations of the variable at N extra ungauged sites, or summary statistics from them, even under the possibility of letting $N \rightarrow +\infty$. For preventing inconsistencies in parameters and model interpretations, it is convenient that the distribution model for J sites be of the same type as the model for $J + N$ sites.

4.2. Useful representation of the archetypal model

Working directly with cumulant generating function 4.1.6 can be unwieldy in applications. A more useful representation of the model is necessary, which allows parameters' estimation conveniently. In this section, we introduce a more convenient representation via the standard representation of an elliptical random vector. We use here standard results about the theory of elliptically distributed random variables. An excellent and accessible introduction can be found at Frahm (2004), see also Cambanis et al. (1981); Fang (1990).

Let $\mathbf{X} \in \mathbb{R}^J$ be an elliptically distributed random vector, such as the one having c.g.f. (4.1.6). Then \mathbf{X} admits the following stochastic representation:

$$\mathbf{X} = \mu + R \times \mathbf{U}^{J-1} \times \mathbf{\Gamma}^{\frac{1}{2}} \quad (4.2.1)$$

where,

$\mu \in \mathbb{R}^J$	(Location vector)
$R \geq 0$	(Non-negative r.v.)
\mathbf{U}^{J-1}	(Uniform r.v. on the unit J-dimensional hypersphere)
$\mathbf{\Gamma}^{\frac{1}{2}}$	(Squared root of covariance matrix)

Our model (4.1.6) assumes $\mu = \mathbf{0}$, but a location vector can be added to \mathbf{X} without altering the models properties of interest for this research.

One dimensional random variate R receives the name of "generating variable". It determines important characteristics of random vector \mathbf{X} , such as tail behavior. Actually, the density of

\mathbf{X} , whenever it exists, can be written as:

$$f_{\mathbf{X}}(\mathbf{x}) = \sqrt{\det(\Gamma^{-1})} \frac{\text{Gamma}\left(\frac{J}{2}\right) f_R\left(\left((\mathbf{x} - \mu)' \Gamma^{-1} (\mathbf{x} - \mu)\right)^{1/2}\right)}{2\pi^{\frac{J}{2}} \times \left((\mathbf{x} - \mu)' \Gamma^{-1} (\mathbf{x} - \mu)\right)^{(J-1)/2}} \quad (4.2.2)$$

In order to avoid confusion, we have used *Gamma* for the gamma function Abramowitz (1972). Two examples of generating random variables:

- If \mathbf{X} has a J -dimensional normal distribution, then $R^2 \sim \chi_J^2$.
- If \mathbf{X} has a J -dimensional Student distribution with ν degrees of freedom, then the squared generating variable is proportional to a random variable having Fisher's F distribution, $R^2 \sim J \times F_{J,\nu}$.

In the above two examples one can already see a possible issue with generating variables: They depend on the dimension of the field, J . This issue will be addressed shortly. Given a (flexible enough) model for R , maximum likelihood estimation (say) can be effected on the basis of (4.2.2).

Hence, we shall strive to find the connection between (4.1.6) and R . Note that, since $R \geq 0$, it will be equally good to find a connection between (4.1.6) and R^2 .

4.2.1. Relation between R^2 and the archetypal c.g.f.

Assume that we have random vector $\mathbf{Z} \in \mathbb{R}^J$ with c.g.f (4.1.6), with $\mu = \mathbf{0}$ and covariance matrix equal to the identity matrix, $\Gamma = I_{J \times J}$. For this special case, in agreement with representation (4.2.1), we have

$$\|\mathbf{Z}\|_2 = \sqrt{\langle \mathbf{Z}, \mathbf{Z} \rangle} = \sqrt{\|\mathbf{Z}\| \|\mathbf{Z}\| \cos(0)} = \sqrt{\|R \times \mathbf{U}^{J-1}\| \|R \times \mathbf{U}^{J-1}\|} = R \times 1$$

and then,

$$R^2 = \sum_{j=1}^J Z_j^2 \quad (4.2.3)$$

which in turn means that,

$$E\left((R^2)^k\right) = E\left(\left(\sum_{j_1=1}^J Z_{j_1}^2\right) \times \dots \times \left(\sum_{j_k=1}^J Z_{j_k}^2\right)\right) = \sum_{j_1=1}^J \dots \sum_{j_k=1}^J E(Z_{j_1}^2 \dots Z_{j_k}^2) \quad (4.2.4)$$

Since \mathbf{Z} has c.g.f. given by

$$K_{\mathbf{Z}}(\mathbf{t}) = \frac{c_1}{1!} \left(\frac{1}{2} \mathbf{t}' \mathbf{t}\right) + \frac{c_2}{2!} \left(\frac{1}{2} \mathbf{t}' \mathbf{t}\right)^2 + \frac{c_3}{3!} \left(\frac{1}{2} \mathbf{t}' \mathbf{t}\right)^3 + \dots$$

it follows, as seen in section 4.1.1, that

$$M_{\mathbf{Z}}(\mathbf{t}) = 1 + \frac{m_1}{1!} \left(\frac{1}{2} \mathbf{t}' \mathbf{t} \right) + \frac{m_2}{2!} \left(\frac{1}{2} \mathbf{t}' \mathbf{t} \right)^2 + \frac{m_3}{3!} \left(\frac{1}{2} \mathbf{t}' \mathbf{t} \right)^3 + \dots$$

with coefficients given by

$$\begin{aligned} m_1 &= c_1 \\ m_2 &= c_2 + c_1^2 \\ m_3 &= c_3 + 3c_2c_1 + c_1^3 \\ m_4 &= c_4 + 4c_3c_1 + 3c_2^2 + 6c_2c_1^2 + c_1^4 \end{aligned} \quad (4.2.5)$$

and so on. A particular case of this function is the Gaussian moment generating function, for which all $c_{r>1}$ are set to zero. In particular, for $\xi \sim N_J(\mathbf{0}, I_{J \times J})$,

$$M_{\xi}(\mathbf{t}) = 1 + \frac{c_1}{1!} \left(\frac{1}{2} \mathbf{t}' \mathbf{t} \right) + \frac{c_1^2}{2!} \left(\frac{1}{2} \mathbf{t}' \mathbf{t} \right)^2 + \frac{c_1^3}{3!} \left(\frac{1}{2} \mathbf{t}' \mathbf{t} \right)^3 + \dots \quad (4.2.6)$$

with $c_1 = 1$. Hence joint moments of \mathbf{Z} and ξ are similar, except for what pertains to coefficients c_2, c_3, \dots . In fact, calling

$$h_r(\mathbf{t}) = \left(\frac{1}{2} \mathbf{t}' \mathbf{t} \right)^r$$

one has

$$\begin{cases} \frac{\partial^{r_1+\dots+r_k}}{\partial t_{j_1} \dots \partial t_{j_k}} M_{\xi}(\mathbf{t}) = \frac{c_1}{1!} \frac{\partial^{r_1+\dots+r_k}}{\partial t_{j_1} \dots \partial t_{j_k}} h_1(\mathbf{t}) + \frac{c_1^2}{2!} \frac{\partial^{r_1+\dots+r_k}}{\partial t_{j_1} \dots \partial t_{j_k}} h_2(\mathbf{t}) + \dots \\ \frac{\partial^{r_1+\dots+r_k}}{\partial t_{j_1} \dots \partial t_{j_k}} M_{\mathbf{Z}}(\mathbf{t}) = \frac{m_1}{1!} \frac{\partial^{r_1+\dots+r_k}}{\partial t_{j_1} \dots \partial t_{j_k}} h_1(\mathbf{t}) + \frac{m_2}{2!} \frac{\partial^{r_1+\dots+r_k}}{\partial t_{j_1} \dots \partial t_{j_k}} h_2(\mathbf{t}) + \dots \end{cases}$$

Hence, for odd orders joint moments of both random vectors are zero, and for even orders

$$\begin{aligned} E(\xi_i \xi_j) &= \frac{c_1}{m_1} E(Z_i Z_j) \\ E(\xi_i \xi_j \xi_k \xi_l) &= \frac{c_1^2}{m_2} E(Z_i Z_j Z_k Z_l) \\ &\vdots \\ E(\xi_{j_1}^{r_1} \dots \xi_{j_k}^{r_k}) &= \frac{c_1^{\frac{1}{2} \sum_{j=1}^k r_j}}{m_{\frac{1}{2} \sum_{j=1}^k r_j}} E(Z_{j_1}^{r_1} \dots Z_{j_k}^{r_k}) \end{aligned}$$

whenever $order = \sum_{i=1}^k r_i$ is an even integer. Since $c_1 = 1$, it is clear that the following relation holds, for joint moments of even order:

$$\begin{aligned} m_1 E(\xi_i \xi_j) &= E(Z_i Z_j) \\ m_2 E(\xi_i \xi_j \xi_k \xi_l) &= E(Z_i Z_j Z_k Z_l) \\ &\vdots \\ m_{\frac{1}{2} \sum_{j=1}^k r_j} E(\xi_{j_1}^{r_1} \dots \xi_{j_k}^{r_k}) &= E(Z_{j_1}^{r_1} \dots Z_{j_k}^{r_k}) \end{aligned} \quad (4.2.7)$$

Moments appearing on the left hand side of equation (4.2.7) can be readily found, since they are the moments of a multivariate Gaussian distribution with covariance matrix equal to identity matrix $I_{J \times J}$.

Coefficients $m_1 = 1, m_2, m_3, \dots$ are given in terms of $c_1 = 1, c_2, c_3, \dots$ (and vice versa). Hence we have, by virtue of (4.2.4), identified requirements on all moments of (squared) generating variable R^2 , so that the resulting multivariate distribution \mathbf{X} has cumulant generating function (4.1.6). The task would be now to find a generating random variable which fulfills these moments restrictions.

Summarizing these results: Before proceeding, useful summarizing equations of the preceding analysis are in place. First, since the multivariate Gaussian distribution referred to at equation 4.2.7 has covariance matrix equal to identity, one can write for any set of components (j_1, \dots, j_k) ,

$$m_k E(\xi_{j_1}^2 \dots \xi_{j_k}^2) = E(Z_{j_1}^2 \dots Z_{j_k}^2) \quad (4.2.8)$$

where ξ is a J -dimensional normally distributed vector with mean vector $\mathbf{0}$ and covariance matrix $I_{J \times J}$, the identity matrix on $\mathbb{R}^{J \times J}$. Equation (4.2.4) holds in particular for vector ξ , in which case $R^2 \sim \chi_J^2$, and

$$\sum_{j_1=1}^J \dots \sum_{j_k=1}^J E(\xi_{j_1}^2 \dots \xi_{j_k}^2) = E((\chi_J^2)^k) = \frac{2^k \Gamma(k + \frac{J}{2})}{\Gamma(\frac{J}{2})}$$

Second and more importantly, by virtue of (4.2.8), one can re-write (4.2.4) as

$$E((R^2)^k) = \sum_{j_1=1}^J \dots \sum_{j_k=1}^J m_k E(\xi_{j_1}^2 \dots \xi_{j_k}^2) = m_k \frac{2^k \Gamma(k + \frac{J}{2})}{\Gamma(\frac{J}{2})} \quad (4.2.9)$$

which expresses the moments of R^2 in terms of parameters m_k (hence indirectly of c_k) and the dimension of the random vector \mathbf{X} . Thus, the generating variable can be inferred, via method of moments, for dimensions other than J , for which data is available. This is important, since the generating variable R of random vector \mathbf{X} having an elliptical distribution usually changes with the dimension of \mathbf{X} . In this dissertation work, we use this relation to infer the generating variable of a big random field ($J = 300 \times 300 = 90000$) on the basis of the generating variable estimated from data at $J = 30$ sites, in the context of a synthetic data example. See section 6.

4.2.2. A convenient model for R^2

A convenient model for R^2 should:

1. Be flexible enough, since we do not want to restrict *a priori* the interdependence properties of our multivariate model.
2. Provide means for estimating $E((R^2)^k)$ easily and ideally in closed form, since moments constraints such as 4.2.9 provide means for evaluating the squared generating variable at a different dimension, and spatial interpolation becomes possible.

In this research we use the mixture of gamma distributions introduced by Venturini et al. (2008). The model is given by the mixture

$$R^2 \sim f_{R^2}(x) = \sum_{s=1}^S \pi_s f_s(x | \theta) \quad (4.2.10)$$

where

$$f_s(x | \theta) = \frac{\theta^s}{\Gamma(s)} x^{s-1} e^{-\theta x}$$

The parameters to fit are the weights (π_1, \dots, π_S) and the rate parameter θ . As reported by Venturini et al. (2008), this model allows much flexibility and can model heavy tails of data. The number of components, S , can be safely given a high value (say, 200, as in the authors' application) without incurring in any kind of "under-smoothing". The only disadvantage of increasing S too much, might be that of computational effort or instability of the fitting algorithm. As the authors report, the moments of a random variable represented by a mixture of gammas are given by

$$E\left((R^2)^k\right) = \sum_{s=1}^S \pi_s \frac{\prod_{l=1}^k (s+l-1)}{\theta^k} \quad (4.2.11)$$

Equation (4.2.11) is convenient for two reasons. First, it provides a straightforward means of connecting the parameters of the (squared) generating variable with those of the dependence structure analyzed in this dissertation. For example, after estimating the parameters of the generating variable, an estimator for each coefficient m_k is given by

$$\hat{m}_k = \frac{1}{J^k E(Z^{2k})} \sum_{s=1}^S \hat{\pi}_s \frac{\prod_{l=1}^k (s+l-1)}{\hat{\theta}^k} \quad (4.2.12)$$

Secondly and *most importantly*, for any dimension $J_* \neq J$, the parameters of the squared generating variable $R^2(J_*)$ can be inferred by solving for $\theta, \pi_1, \dots, \pi_S$ the following system of non-linear equations, in a method of moments fashion:

$$\begin{aligned} \sum_{s=1}^S \pi_s \frac{s}{\theta} &= \hat{m}_1 E\left((\chi_{J_*}^2)^1\right) \\ \sum_{s=1}^S \pi_s \frac{s(s+1)}{\theta^2} &= \hat{m}_2 E\left((\chi_{J_*}^2)^2\right) \\ &\vdots \\ \sum_{s=1}^S \pi_s \frac{\prod_{l=1}^S (s+l-1)}{\theta^S} &= \hat{m}_S E\left((\chi_{J_*}^2)^S\right) \\ \sum_{s=1}^S \pi_s &= 1 \end{aligned} \quad (4.2.13)$$

4.2.2.1. The data for estimating f_{R^2}

We recall that

$$f_{\mathbf{X}}(\mathbf{x}) = \sqrt{\det(\Gamma^{-1})} \frac{\text{Gamma}\left(\frac{J}{2}\right) f_R\left(\left((\mathbf{x} - \mu)' \Gamma^{-1} (\mathbf{x} - \mu)\right)^{1/2}\right)}{2\pi^{\frac{J}{2}} \times \left((\mathbf{x} - \mu)' \Gamma^{-1} (\mathbf{x} - \mu)\right)^{(J-1)/2}} \quad (4.2.14)$$

Since random variable R is non-negative, the function $R \mapsto R^2$ is a bijection. Upon application of the change of variables theorem, it can be expressed in terms of the density of R^2 , as

$$f_{\mathbf{X}}(\mathbf{x}) = \sqrt{\det(\Gamma^{-1})} \frac{\text{Gamma}\left(\frac{J}{2}\right) f_{R^2}\left((\mathbf{x} - \mu)' \Gamma^{-1} (\mathbf{x} - \mu)\right)}{\pi^{\frac{J}{2}} \times \left((\mathbf{x} - \mu)' \Gamma^{-1} (\mathbf{x} - \mu)\right)^{(J-2)/2}} \quad (4.2.15)$$

Hence, given a sample $\mathbf{x}_1, \dots, \mathbf{x}_N$ of $\mathbf{X} \in \mathbb{R}^J$, and given Γ and μ , parameter estimation for $f_{\mathbf{X}}$ is effected by maximizing

$$L(\theta, \pi_1, \dots, \pi_S) = \prod_{i=1}^N f_{\mathbf{X}}(\mathbf{x}_i) \propto \prod_{i=1}^N f_{R^2}(t_i) \quad (4.2.16)$$

with respect to θ and (π_1, \dots, π_S) , where $t_i = (\mathbf{x}_i - \mu)' \Gamma^{-1} (\mathbf{x}_i - \mu)$. But given Γ and μ , if $\prod_{i=1}^N f_{\mathbf{X}}(\mathbf{x}_i)$ is to attain its maximum, then $\prod_{i=1}^N f_{R^2}(t_i)$ must also attain its maximum. This means that the sample data necessary for estimating θ and (π_1, \dots, π_S) is given by $t_i = (\mathbf{x}_i - \mu)' \Gamma^{-1} (\mathbf{x}_i - \mu)$, for $i = 1, \dots, N$.

4.3. Parameter estimation

The more convenient representation of our archetypal model at section (4.2) and indeed equation (4.2.15) provide the basis for parameter estimation of the whole model. For example, one can straightforwardly use maximum likelihood estimation.

However, whole model estimation might miss important subject-matter characteristics related to the research questions, such as the probability of groups of components, $(X_{j_1}, \dots, X_{j_k})$, simultaneously trespassing a given quantile. On the other hand, models based on mean and covariance alone, have indeed proved useful in applications. These considerations suggests the following 2-step estimation procedure:

1. Step one: Mean-Covariance determination:
 - a) Fit parameters of model (4.2.15) using maximum likelihood estimation or any other method at hand. This results in estimated values $\hat{\theta}$, $(\hat{\pi}_1, \dots, \hat{\pi}_S)$, $\hat{\mu}$ and $\hat{\Gamma}$.
 - b) Estimate the respective values $\hat{m}_1, \hat{m}_2, \hat{m}_3 \dots$, up to the desired joint cumulant order, by using equation (4.2.9). Note that, implicitly you have estimates $\hat{c}_1, \hat{c}_2, \hat{c}_3, \dots$ via equation 3.2.8.
2. Step Two: Letting $\hat{\Gamma}$, $\hat{\mu}$ and \hat{m}_1 fixed, apply some optimization algorithm on m_2, m_3, \dots in order to make the expected interaction manifestations produced by $\hat{f}_{\mathbf{X}}$ as similar as possible as those observed in data. This second step intends to capture subject-matter specific interaction manifestations and is outlined below.

Note that by letting $\hat{\Gamma}$, $\hat{\mu}$ and \hat{m}_1 fixed at step two above, the mean values and covariances fitted at the first step are not altered. Additionally, note that by optimizing on m_2, m_3, \dots we are indeed optimizing on c_2, c_3, \dots of the dependence structure (4.1.6).

This approach is reminiscent of the approach due to Zheng and Katz (2008) in the context of rainfall modeling, whereby pair-wise covariances among sites data were fitted on a first

step, and then the whole distribution was fitted in a parsimonious manner on a second step. A compromise between overall good fit and adequate covariance modeling was thus attained.

We assume isotropy in the random field to model, so that a covariance function $C_\theta(d)$ is an adequate definition of pair-wise dependence, if necessary, after applying a transformation technique. Concerning the mean of the random field, we assume that it is constantly zero, if necessary, after de-trending by means of a deterministic model on geographic coordinates, or other geographic variables (see, e.g. section 3.6 of Diggle and Ribeiro (2007)). This trend can be added later on for simulation or prediction purposes.

4.3.1. First step: Mean and Covariance determination

Let (s_1, \dots, s_J) denote locations at which data is available. If only one observation per site is available, data can be represented by vector

$$\mathbf{y} = (y_1, \dots, y_J)$$

The model for this data-set is given by probability density

$$f_{\mathbf{X}}(\mathbf{x}) = \sqrt{\det(\Gamma^{-1})} \frac{\Gamma\left(\frac{J}{2}\right) f_{R^2}\left((\mathbf{x} - \mu)' \Gamma^{-1} (\mathbf{x} - \mu)\right)}{\pi^{\frac{J}{2}} \times \left((\mathbf{x} - \mu)' \Gamma^{-1} (\mathbf{x} - \mu)\right)^{(J-2)/2}} \quad (4.3.1)$$

where $\Gamma_{J \times J}$ is covariance matrix determined by $\Gamma_{j_1 j_2} = C_\theta(\text{Dist}(s_{j_1}, s_{j_2}))$, and

$$R^2 \sim f_{R^2}(x) = \sum_{s=1}^S \pi_s f_s(x | \theta) \quad (4.3.2)$$

with

$$f_s(x | \theta) = \frac{\theta^s}{\Gamma(s)} x^{s-1} e^{-\theta x}$$

Maximum likelihood estimation, for example, is effected by maximizing (4.3.1) as a function of the covariance function's vector of parameters ϱ , the mean vector μ (possibly constant or a function of geographical variables), and gamma mixture parameters θ and (π_1, \dots, π_S) .

If at each location a series of I observations are available, so that one has a data matrix

$$\mathbb{Y}_{I \times J} = (\mathbf{y}_1, \dots, \mathbf{y}_I)' \quad (4.3.3)$$

where $\mathbf{y}_i = (y_{i1}, \dots, y_{iJ})$ represents the observations at all sites for time i , with $i = 1, \dots, I$. The Gaussian model for each observation is as before, but now the function to maximize is given by

$$f_{\mathbb{Y}}(\mathbb{Y}) = \prod_{i=1}^I f_{\mathbf{y}}(\mathbf{y}_i) = (\det(\Gamma))^{-\frac{I}{2}} \prod_{i=1}^I \frac{\Gamma\left(\frac{J}{2}\right) f_{R^2}\left((\mathbf{x} - \mu)' \Gamma^{-1} (\mathbf{x} - \mu)\right)}{\pi^{\frac{J}{2}} \times \left((\mathbf{x} - \mu)' \Gamma^{-1} (\mathbf{x} - \mu)\right)^{(J-2)/2}} \quad (4.3.4)$$

which upon maximization effects the estimation of the model parameters.

This latter case poses a problem, since the definition of f_{R^2} as a mixture with S components indicates the term $\prod_{s=1}^S f_{R^2}$ becomes unmanageable. This issue is addressed by Venturini et al. (2008), and their solution is to use a Bayesian estimation approach. They implemented this solution in the R package GMS, which we use for the present work. Given Γ and μ , one can apply the method of Venturini et al. (2008) to obtain (empirical) Bayes estimates of θ and (π_1, \dots, π_S) . In turn, given θ and (π_1, \dots, π_S) , one can apply numerical optimization on 4.3.4 in order to obtain Γ and μ . One can iterate along these two steps until no significant difference between the parameters of one iteration and the other can be observed. We consider the above estimation method, informally, an approximation to the maximum likelihood method.

In case missing data is present, one can implement an approximate version of the Stochastic EM algorithm Gilks et al. (1998); Feodor Nielsen (2000). Briefly stated, one proceeds iteratively imputing missing data at every iteration, and using the complete dataset to estimate all parameters as above. This produces an approximation to the maximum likelihood estimators.

Missing data can be interpreted to be the case in daily rainfall modeling, where one has many zero values. These are then considered as censored values.

Since data represented by (4.3.3) is usually taken chronologically, one should pre-process it by e.g. applying a time-series model to each site's data series. This produces, ideally, a temporally uncorrelated sample of residuals at each site. Data at (4.3.3) would be then the residual process of each time series model, which is Spatially interdependent. Alternatively, one may consider a Spatio-Temporal model Cressie and Wikle (2011); Le and Zidek (2006). Many variants are possible for this step. One can have missing data for periods of time at some sites, gauging stations providing data might have different operations times, or one can employ a Bayesian method of parameter estimation. The reader is referred to Le and Zidek (2006) for details on variants.

The output of this first step is a fitted model for the data at hand, and indeed estimators of all parameters of the archetypal structure: $\hat{\Gamma}$ and $(\hat{c}_1, \hat{c}_2, \hat{c}_3 \dots)$, these last found via $(\hat{m}_1, \hat{m}_2, \hat{m}_3, \dots)$, which in turn are expressed in terms of $\hat{\theta}$ and $(\hat{\pi}_1, \dots, \hat{\pi}_S)$.

4.3.2. Second step: manifestations of higher order

For this step, we begin by computing adequate sample estimates of the interactions manifestations of interest. These may be computed for marginal distributions of dimension not higher than 6, for example, where one can actually estimate them with some stability.

We proceed in this section in a mere conceptual, and rather vague manner. After many attempts, we could not develop a computationally feasible algorithm to implement these ideas, so that further details would be superfluous: We still do not know how to implement these ideas efficiently. For these reasons, this topic must be studied as future research.

For the sake of clearness, we assume we work with interaction among four variables at a time. We shall also use simple estimates of the interaction manifestations, in order to avoid unnecessary distractions.

In the present work, we are interested in three interactions manifestations listed in section

3.1, namely

$$p_{j_1 j_2 j_3 j_4} = \Pr(Y_{j_1} \leq q_{Y_{j_1}, \alpha}, \dots, Y_{j_4} \leq q_{Y_{j_4}, \alpha}) \quad (4.3.5)$$

$$H_{j_1, \dots, j_4} = \int_{-\infty}^{+\infty} \log(f_{j_1, \dots, j_4}) f_{j_1, \dots, j_4} \quad (4.3.6)$$

$$S_{j_1, \dots, j_4}(s_0) = \Pr\left(\sum_{k=1}^4 Y_{j_k} \leq s_0\right) \quad (4.3.7)$$

where $f_{j_1 \dots j_4}$ stands for the marginal density of $(Y_{j_1}, \dots, Y_{j_4})$. We seek to reproduce faithfully these manifestations with our model, or at least, to take them into account.

The distribution of the sum of components (4.3.7) can be attacked either by computing sample estimates of sums' cumulants, and then employing (3.3.12); or directly by using (3.3.15). We employ here the former, for simplicity. Note that for this interaction manifestation, one can consider also higher dimensional marginals, or even all components of $\mathbf{Y} \in \mathbb{R}^J$.

The estimates for the interdependence parameters given above, are

$$\hat{p}_{j_1, j_2, j_3, j_4} = \frac{\#\{(y_{i, j_1}, \dots, y_{i, j_4}) : y_{i, j_k} \leq q_{Y_{j_k}, \alpha}, k = 1, \dots, 4\}}{I} \quad (4.3.8)$$

$$H_{j_1, \dots, j_4} = \frac{1}{I} \sum_{i=1}^I \log(\hat{f}(y_{i, j_1}, \dots, y_{i, j_4})) \hat{f}(y_{i, j_1}, \dots, y_{i, j_4}) \quad (4.3.9)$$

$$\hat{\kappa}_r\left(\sum_{k=1}^4 Y_{j_k}\right) = \hat{\kappa}_r\left(\sum_{k=1}^4 Y_{j_k}\right), r = 1, \dots, 4 \quad (4.3.10)$$

Estimates (4.3.8) and (4.3.10) are just the sample estimates of the parameters. Estimate (4.3.10) can be found by computing sample moments of random variable $S = \sum_{k=1}^4 Y_{j_k}$, and then apply the moment to cumulant formula (3.2.8).

Concerning estimate (4.3.9), $\hat{f}_{j_1, \dots, j_4}$ is a kernel smoothing estimate of the joint density of $(Y_{j_1}, \dots, Y_{j_4})$, see Joe (1989a).

The effect of c_2, c_3, \dots on the joint cumulants of $\mathbf{Y} \in \mathbb{R}^J$ was shown in section 4.1, and the connection of joint cumulants of interdependence manifestations here considered was shown in section 3.3.2.

Thus one can in principle fit c_2, c_3, \dots , so as to minimize objective function

$$\begin{aligned} Z(c_2, c_3, \dots) = & w_1 \sum_{j_1, \dots, j_4} \left\{ \hat{p}_{j_1, \dots, j_4} - \Pr(Y_{j_1} \leq q_{Y_{j_1}, \alpha}, \dots, Y_{j_4} \leq q_{Y_{j_4}, \alpha} \mid c_2, c_3, \dots) \right\}^2 + \\ & w_2 \sum_{j_1, \dots, j_4} \left\{ \hat{H}(f_{j_1, \dots, j_4}) - H(f_{j_1, \dots, j_4} \mid c_2, c_3, \dots) \right\}^2 + \\ & w_3 \sum_{j_1, \dots, j_4} \sum_{r=1}^4 \left\{ \hat{\kappa}_r\left(\sum_{k=1}^4 Y_{j_k}\right) - \kappa_r\left(\sum_{k=1}^4 Y_{j_k} \mid c_2, c_3, \dots\right) \right\}^2 \end{aligned} \quad (4.3.11)$$

where w_1, w_2, w_3 are weights indicating a relative importance of the interaction manifestations, for the problem at hand.

4.4. More flexibility: Transformations on marginals

We have introduced in section 4.1 a dependence structure that provides a probability model which is suitable for Spatial Statistics applications. We have shown the simple structure of its interdependence parameters (join cumulants). With this low dimensional model, interdependence structure, and interdependence parameters, it is possible to consider interdependencies of order greater than 2, along the lines of chapter 3.

We want now to increase the flexibility of our model by considering transformations on its uni-variate marginal random variables. For example, we would like to deal with data sets exhibiting skewed marginal distributions, or a specific degree of kurtosis. The model proposed in 4.1 might not always be directly adequate for these observed characteristics. Alternatively, one might wish to work with the copula of the multivariate distribution, and then non-monotonic transformations on the marginal variables can modify the resulting copula of the distribution, so as to make it more adequate for the research problem at hand. Two cases are considered in this section: the quantile-quantile transformation, which is a monotonically increasing¹ transformation; and the polynomial transformation, which is not necessarily a monotonic transformation.

Throughout, data at hand consists of a realized random sample of size I , represented by

$$\mathbb{Y}_{I \times J} = (\mathbf{y}_1, \dots, \mathbf{y}_I)' \quad (4.4.1)$$

with $\mathbf{y}_i \in \mathbb{R}^J$, for $i = 1, \dots, I$, a row vector representing a multivariate observation. Data is assumed to be representable by a random vector $\mathbf{Y} \in \mathbb{R}^J$,

$$\mathbf{Y} \sim F_{\mathbf{Y}} \quad (4.4.2)$$

having marginal probability distributions $F_{Y_j}(y_j)$, for $j = 1, \dots, J$.

On the other hand, a random vector possessing the archetypal dependence structure will be identified by $\mathbf{X} \in \mathbb{R}^J$,

$$\mathbf{X} \sim F_{\mathbf{X}} \quad (4.4.3)$$

having marginal probability distributions $F_{X_j}(x_j)$, for $j = 1, \dots, J$.

Our aim in this section is to define transformations T_j , on each random variable X_j , such that

$$T_j(X_j) \sim F_{Y_j}$$

where it is assumed that the copula provided by \mathbf{X} is adequate for our modeling purposes, that is,

$$(T_1(X_1), \dots, T_J(X_J)) \sim F_{\mathbf{Y}} \quad (4.4.4)$$

for a convenient selection of the distribution parameters of $F_{\mathbf{X}}$, namely, covariance matrix Γ and coefficients c_1, c_2, c_3, \dots , as in (4.1.6).

¹A function $T : D \subset \mathbb{R} \rightarrow S \subset \mathbb{R}$ is called monotonically increasing, if $x < y$ implies that $T(x) < T(y)$, for every $x, y \in D$.

4.4.1. Quantile-Quantile Transformations

4.4.1.1. One-dimensional marginals of $\mathbf{X} \in \mathbb{R}^J$

It is convenient, for this section, to begin with the identification of the one-dimensional marginal distribution of our archetypal model, defined in section 4.1.

If our model is defined by a cumulant generating function, as in (4.1.6), then the cumulant generating function of every one-dimensional marginal, j , can be found just by setting all other arguments of $K_{\mathbf{X}}(\mathbf{t})$ to zero, namely

$$\begin{aligned} K_{X_j}(t_j) &= \log(E(\exp(X_j t_j))) = \\ &= \log(E(\exp(X_1 \cdot 0 + \dots + X_j t_j + \dots + X_J \cdot 0))) = \\ &= K_{\mathbf{X}}((0, \dots, t_j, \dots, 0)) \end{aligned}$$

Hence it is possible obtain F_{X_j} , the probability distribution function of X_j , by using (3.3.6), the Lugganani and Rice approximation. The inverse function of F_{X_j} , that is $F_{X_j}^{-1}(u)$, can be found for every $u \in (0, 1)$ by means of a root finding algorithm, such as the bisection algorithm, for example.

Alternatively one can apply the techniques of section 4.2 to the specific case of $J_* = 1$ and approximate the 1-dimensional density and, through it, the distribution function of each marginal variable X_j .

4.4.1.2. The Transformation

We assume we have at hand (at least an estimation of) marginal distribution F_{Y_j} , for $j = 1, \dots, J$. This can be available, for example, by fitting a parametric probability distribution to each marginal data set. For each $j = 1, \dots, J$, transformation T_j and its inverse T_j^{-1} are then given, respectively, by

$$T_j(X_j) : = F_{Y_j}^{-1}(F_{X_j}(X_j)) \sim F_{Y_j} \quad (4.4.5)$$

$$T_j^{-1}(Y_j) : = F_{X_j}^{-1}(F_{Y_j}(Y_j)) \sim F_{X_j} \quad (4.4.6)$$

The inverse is well defined, since both F_{X_j} and F_{Y_j} are monotonically increasing functions. That the distribution of the transformed variable is as stated at (4.4.5), follows from the fact that $F_{X_j}(X_j) \sim \text{uniform}(0, 1)$. Then, whenever one applies the inverse of a distribution function F_{Y_j} to an uniformly distributed U , the result is a random variable distributed as F_{Y_j} (see, for example, chapter II of Devroye (1986)). The proof is similar for (4.4.6).

Since $F_{Y_j}(T_j(a)) = F_{X_j}(a)$ from definition at 4.4.5, we have that this transformation preserves quantiles. If $\alpha \in (0, 1)$ is a subject-matter interesting threshold, and $q_{j,\alpha}(X_j)$ is the respective quantile for variable X_j , then

$$\begin{aligned} \alpha &= \Pr(X_j \leq q_{j,\alpha}) = F_{X_j}(q_{j,\alpha}) = \\ &= F_{Y_j}(T_j(q_{j,\alpha})) = \Pr(Y_j \leq T_j(q_{j,\alpha})) \end{aligned} \quad (4.4.7)$$

whereby, $q_{j,\alpha}(Y_j) = T_j(q_{j,\alpha}(X_j))$. For ease of notation, the argument of the quantile is often removed in this work, unless strictly necessary for preventing ambiguity. The argument of the quantile function is understood to be the random variable preceding symbol \leq .

The quantile preservation of transformation T_j is the reason it is called quantile-quantile transformation. It can be used for dealing with one of the three interaction manifestations listed in section 3.1. Namely, the k -dimensional marginal joint probabilities of the form

$$\Pr(X_{j_1} \leq q_{j_1, \alpha}, \dots, X_{j_k} \leq q_{j_k, \alpha})$$

where $(X_{j_1}, \dots, X_{j_k})$ is a subset of the whole random field modeled, (X_1, \dots, X_J) . Values $q_{j_1, \alpha}, \dots, q_{j_k, \alpha}$ are quantiles such that $\Pr(X_{j_i} \leq q_{j_i, \alpha}) = \alpha$, for $1 \leq i \leq k$.

The fitting of a Spatial model that fits *both* covariance structure *and* this type of interdependence manifestations parameters, can be summarized as below. The method resembles the Copula modeling for elliptical distributions (see, for example Demarta and McNeil (2005); Fang et al. (2002)). The objective is to have a model for which rank correlations, marginal distributions, and (multidimensional) marginal joint distributions are fitted consistent with observed data, using relatively few parameters. For the following procedure, the mean vector is fixed to be $\mu = \mathbf{0}$, since this location vector can be modeled by means of the marginal distributions. We shall also set $c_1 = 1$, since the marginals' variances can also be absorbed by the plugged-in marginal distributions.

First Fit data-based marginal distributions F_{Y_1}, \dots, F_{Y_J} . Usually it will be the case that $F_{Y_1} = \dots = F_{Y_J}$.

Second Set an interesting probability threshold α , and compute, for low dimensional marginals (e.g. of dimension 3 or 4) sample estimates

$$\hat{p}_{j_1, j_2, j_3} \approx \Pr(Y_{j_1} \leq \hat{q}_{j_1, \alpha}, \dots, Y_{j_3} \leq \hat{q}_{j_3, \alpha}) \quad (4.4.8)$$

In this case, values \hat{q}_{j_k} stand for the sample quantile of the respective marginal distribution.

Third Fit Kendall's τ_{j_1, j_2} coefficient to every pairs of marginal data, for $1 \leq j_1 < j_2 \leq J$. That is,

$$\tau_{j_1, j_2} = \tau(\mathbb{Y}_{j_1}, \mathbb{Y}_{j_2})$$

Since our dependence structure is a member of the elliptical distributions family, we know (see Lindskog et al. (2003)) that this values provide us with an estimation for correlation matrix Γ , via formula

$$\sin\left(\tau_{j_1, j_2} \times \frac{\pi}{2}\right) \approx \Gamma_{j_1, j_2} \quad (4.4.9)$$

Fourth Fit parameters c_2, c_3, \dots of dependence structure 4.1.6 in order to minimize Z , where

$$Z = \sum_{j_1, j_2, j_3} \left(\hat{p}_{j_1, j_2, j_3} - \Pr\left(X_{j_1} \leq T_{j_1}^{-1}(\hat{q}_{j_1}), \dots, X_{j_3} \leq T_{j_3}^{-1}(\hat{q}_{j_3})\right) \right)^2 \quad (4.4.10)$$

where arguments of the inverse transformation are the sample quantiles found used at (4.4.8). Note that the shape and characteristics imposed by c_2, c_3, \dots affect this objective function via the joint probability $\Pr(*)$ and via the effect of each marginal, F_{X_j} , on each transformation $T_j^{-1} = F_{X_j}^{-1}(F_{Y_j})$.

4.4.1.2.1. Some remarks

To reduce computational effort at (4.4.10), one might wish to work with a small subset of triplets (j_1, j_2, j_3) , perhaps randomly selected.

In case only one observation is available at each location, the third step above cannot be performed directly. Data in this case consists of a vector (y_1, \dots, y_N) with associated location labels (s_1, \dots, s_N) . One possibility is to employ a covariance function, such as (2.2.8) or (2.2.9) and to consider all marginal distributions F_{Y_1}, \dots, F_{Y_J} equal, $F_{Y_1} = \dots = F_{Y_J}$.

Some preliminary computations for this alternative are necessary. Namely, to compute standardized data values, $u_i := F_{Y_1}(y_i)$, for $j = 1, \dots, N$, on the basis of observed data vector (y_1, \dots, y_N) . This results in a new data vector with same dimension as the latter, $(u_1, \dots, u_N)'$, where $u_i \in [0, 1]$.

The alternative third step then consists of fitting, via maximum likelihood or otherwise, a Gaussian copula model with correlation matrix prescribed by a covariance function model, $C(d)$ having $C(0) = 1$.

Third, alternative_variant Let $\Phi(*)$ represent the standard Normal distribution function. Compute transformed data vector

$$x_i = \Phi(u_i), \quad i = 1, \dots, N$$

on the basis of observed data. Then fit the parameters of the covariance function so as to maximize function

$$f(x_1, \dots, x_N) = -\frac{1}{2} \log(\det(\Gamma)) - \frac{1}{2} \sum_{i,j=1}^N x_i x_j \Gamma_{ij}^{-1} \quad (4.4.11)$$

where $\Gamma_{N \times N}$ is a correlation matrix given by $\Gamma_{ij} = C(\text{Dist}(s_i, s_j))$, and Γ_{ij}^{-1} stands for entry i - j of the inverse of Γ . This corresponds to maximum likelihood estimation for a dependence structure of the form (4.1.6), with $c_1 = 1$ and $c_{r>1} = 0$.

Then we can proceed to the fourth step above. Note that, as we fit c_2, c_3, \dots , the correlations of the underlying model are not altered, and thus the ranks correlations of data are properly modeled, as desired.

4.4.2. Polynomial Transformations

In connection with a random vector $\mathbf{X} \in \mathbb{R}^J$, we have seen in section 4.4.1 the usefulness of the quantile-quantile transformation for dealing with interdependence manifestations expressed in the form of quantiles or joint quantiles.

Unfortunately, quantile-quantile transformations do not preserve other interdependence manifestations, such as characteristics of the sum of components and the entropy of (multivariate) marginals. See sections 3.3.2.3 and 3.3.2.2, respectively. If one could obtain the dependence structure of the transformed random vector, then one could in principle model the three types of interaction manifestations presented in this work, and presumably many others. Remember that by dependence structure we mean the cumulant generating function.

In general, it is not possible to know *exactly* into what the dependence structure and dependence parameters are converted, when applying an arbitrary transformation. However, it is sometimes possible to *approximate* them.

In appendix (Taylor_report), we present a method to do so. This method could, in principle, be applied to any function fulfilling the assumptions given there. But there are two reasons for focusing for now on polynomials: Firstly, the dependence structure of the transformed vector can be found exactly, not just approximately. Secondly, they are flexible transformations, the usefulness of which has been documented (see Fleishman (1978); Headrick (2010); Headrick and Zumbo (2008); Headrick (2002)).

Joint cumulants and cumulant generating functions of random vectors \mathbf{Y} being the result of applying polynomial transformations for (multivariate) marginals of a random vector \mathbf{X} , are topics addressed by McCullagh (1987, 1984); Barndorff-Nielsen and Cox (1990); wa Binyavanga (2009).

4.4.2.1. Polynomial transformations

The following method is useful rather for simulation of random vectors with prescribed joint cumulants.

Estimation on the basis of \mathbf{Y} is also possible, but in the case of non-monotonic transformations (a most important case), it becomes cumbersome, and would represent a large section for a topic not dealt with in the illustrations section: If our polynomial should have k different roots, then one has to include for each observed \mathbf{y}_i a latent indicator vector (ψ_1, \dots, ψ_J) , with $\psi_j \in \{1, \dots, k\}$, which determines which of the different possible pre-images \mathbf{x}_i corresponds to the observed \mathbf{y}_i . This latent variable should then be either integrated out, or incorporated into an MCMC algorithm that samples from it at each iteration. Hence we omit here the estimation part.

For each marginal random variable X_j of $\mathbf{X} \in \mathbb{R}^J$, we consider functions of the form

$$T_j(X_j) \rightarrow \sum_{r_j=1}^{R_j} a_{r_j} X_j^{r_j} := Y_j \quad (4.4.12)$$

where $R_j \in \mathbb{N}$ is a specified order, and coefficients $a_{r_1}, \dots, a_{r_{R_j}}$ are to be fitted on the basis of available data, in such a way that the resulting random vector $\mathbf{Y} = (Y_1, \dots, Y_J)$ presumably constitutes an adequate multivariate model for data.

For the sake of simplicity, and since it is sensible to assume common marginal distributions when dealing with many environmental variables, we assume the same order and coefficients for each transformation. Hence

$$R_1 = \dots = R_J = R$$

and all marginal transformations are defined in terms of a common set of coefficients, a_1, \dots, a_R .

4.4.2.2. Dependence structure of the transformed vector

Let random vector $\mathbf{Y} = (Y_1, \dots, Y_J)$ be defined by (4.4.12). Following appendix (Taylor_report), its moment generating function is given by

$$M_{\mathbf{Y}}(\mathbf{t}) = \sum_{r_1 + \dots + r_J = 0}^{\infty} \frac{t_1^{r_1} \dots t_J^{r_J}}{r_1! \dots r_J!} \left(\sum_{s_1^1, \dots, s_1^{r_1} = 0}^R \dots \sum_{s_J^1, \dots, s_J^{r_J} = 0}^R \frac{a_{s_1^1} \dots a_{s_1^{r_1}}}{s_1^1! \dots s_1^{r_1}!} \times \dots \right. \\ \left. \dots \times \frac{a_{s_J^1} \dots a_{s_J^{r_J}}}{s_J^1! \dots s_J^{r_J}!} E \left(X_1^{\sum_{i=1}^{r_1} s_1^i} \dots X_J^{\sum_{i=1}^{r_J} s_J^i} \right) \right) \quad (4.4.13)$$

Whence each joint moment is given by

$$E(Y_1^{r_1} \dots Y_J^{r_J}) = \sum_{s_1^1, \dots, s_1^{r_1} = 0}^R \dots \sum_{s_J^1, \dots, s_J^{r_J} = 0}^R \frac{a_{s_1^1} \dots a_{s_1^{r_1}}}{s_1^1! \dots s_1^{r_1}!} \times \dots \\ \dots \times \frac{a_{s_J^1} \dots a_{s_J^{r_J}}}{s_J^1! \dots s_J^{r_J}!} E \left(X_1^{\sum_{i=1}^{r_1} s_1^i} \dots X_J^{\sum_{i=1}^{r_J} s_J^i} \right) \quad (4.4.14)$$

Finally, using the relation described by (3.2.18), one can use these joint moments to obtain joint cumulants. In turn, with the joint cumulants at hand, κ_{r_1, \dots, r_J} , one can write down the dependence structure of \mathbf{Y} in the form of its Taylor expansion, that is, as in equation (3.2.10). Computation becomes quickly unmanageable, as J , the number of random variables for which interdependence parameters are modeled, increases. Hence one can simulate fields with interactions of orders 4 or 5, at the most. By means of polynomial transformations, one can enrich considerably the spectrum of possible simulated fields characteristics. For example, one could include diverse degrees of asymmetry into the modeling random vector distribution by means of this technique.

Example 3. Polynomial of order three, four-wise dependence parameters sought.

For example, let $\mathbf{X} \in \mathbb{R}^4$ and define $\mathbf{Y} \in \mathbb{R}^4$ by,

$$Y_j := a_0 + a_1 X_j + a_2 X_j^2 + a_3 X_j^3$$

for $j = 1, \dots, 4$.

We are interested in four-wise dependence parameters such as

$$\text{cum}(Y_{j_1}, Y_{j_2}, Y_{j_3}, Y_{j_4}) \quad (4.4.15)$$

One finds by (4.4.14) that, for example,

$$\mu_{1,1,2,0,0} = E(Y_1 Y_2 Y_3^2) = \sum_{s_1=0}^3 \sum_{s_2=0}^3 \sum_{s_3^1=0}^3 \sum_{s_3^2=0}^3 \frac{a_{s_1} a_{s_2} a_{s_3^1} a_{s_3^2}}{s_1! s_2! s_3^1! s_3^2!} E(X_1^{s_1} X_2^{s_2} X_3^{s_3^1 + s_3^2})$$

Then joint cumulants for \mathbf{Y} can be found using the moments to cumulants conversion formula (3.2.18),

$$\text{cum}(Y_{j_1}, Y_{j_2}, Y_{j_3}, Y_{j_4}) = \sum_{\pi} \left\{ \left((-1)^{|\pi|-1} (|\pi| - 1)! \right) J_{\pi}^* \right\} E(Y_{j_1} \dots Y_{j_4}) \quad (4.4.16)$$

4.4.2.3. Fitting parameters “orthogonally”

We would like to keep for the transformed model \mathbf{Y} the desirable property, originally inspired by the Edgeworth-Sargan distribution (section 2.5), of allowing the fitting of higher order dependence parameters “orthogonally”. That is, that the fitting of order r joint cumulants of \mathbf{Y} , should not alter the value of joint cumulants of order $s < r$.

Covariances of the transformed model, \mathbf{Y} , are

$$\begin{aligned} \text{cum}(Y_{j_1}, Y_{j_2}) &= E(Y_{j_1} Y_{j_2}) - E(Y_{j_1}) E(Y_{j_2}) = \\ &= \sum_{s_1=0}^R \sum_{s_2=0}^R \frac{a_{s_1} a_{s_2}}{s_1! s_2!} E(X_{j_1}^{s_1} X_{j_2}^{s_2}) - \left(\sum_{s_1=0}^R \frac{a_{s_1}}{s_1!} E(X_{j_1}^{s_1}) \right) \left(\sum_{s_2=0}^R \frac{a_{s_2}}{s_2!} E(X_{j_2}^{s_2}) \right) = \\ &= \sum_{s_1=0}^R \sum_{s_2=0}^R \frac{a_{s_1} a_{s_2}}{s_1! s_2!} E(X_{j_1}^{s_1} X_{j_2}^{s_2}) - \sum_{s_1=0}^R \sum_{s_2=0}^R \frac{a_{s_1} a_{s_2}}{s_1! s_2!} E(X_{j_1}^{s_1}) E(X_{j_2}^{s_2}) = \\ &= \sum_{s_1=0}^R \sum_{s_2=0}^R \frac{a_{s_1} a_{s_2}}{s_1! s_2!} \left\{ E(X_{j_1}^{s_1} X_{j_2}^{s_2}) - E(X_{j_1}^{s_1}) E(X_{j_2}^{s_2}) \right\} \quad (4.4.17) \end{aligned}$$

We see that covariances of \mathbf{Y} are affected by moments of the original variable, \mathbf{X} , of order up to $2R$.

Since, according to section 4.1.1, one has

$$\begin{aligned} \mu^{j_1, j_2} &= m_1 \varrho(j_1, j_2) \\ \mu^{j_1, j_2, j_3, j_4} &= m_2 \varrho(j_1, \dots, j_4) \\ \mu^{j_1, j_2, j_3, j_4, j_5, j_6} &= m_3 \varrho(j_1, \dots, j_6) \\ &\vdots \end{aligned}$$

and zero for odd orders. Additionally,

$$\begin{aligned} m_1 &= c_1 \\ m_2 &= c_2 + c_1^2 \\ m_3 &= c_3 + 3c_2 c_1 + c_1^3 \\ m_4 &= c_4 + 4c_3 c_1 + 3c_2^2 + 6c_2 c_1^2 + c_1^4 \\ &\vdots \end{aligned}$$

Then, one notices that changing the value of $c_{k/2}$ will affect, via $m_{k/2}$, the order k joint moments of \mathbf{X} . If $k \leq 2R$, these joint moments, in turn, will affect covariances between components of the transformed variable, \mathbf{Y} . But if $k > 2R$, then covariances of \mathbf{Y} are *not* affected. In general, one can show the following relation between moment orders k of variable \mathbf{X} and joint moments of transformed vector \mathbf{Y} ,

$$\begin{aligned} k \leq R &\rightarrow E(Y_j) \\ k \leq 2R &\rightarrow E(Y_{j_1} Y_{j_2}) \\ k \leq 3R &\rightarrow E(Y_{j_1} Y_{j_2} Y_{j_3}) \\ k \leq 4R &\rightarrow E(Y_{j_1} \dots Y_{j_4}) \end{aligned}$$

where symbol \rightarrow above must be read “affects”. Since the k -th joint moment of original vector \mathbf{X} is connected with coefficient $m_{k/2}$, which in turn can be written in terms of

$$c_1, c_2, \dots, c_{k/2}$$

one would have, in principle, to extend the number of these coefficients.

For example, in order to have flexibility in fitting covariances and third order joint cumulants of \mathbf{Y} , one would have to allow c_1, c_2, \dots, c_{2R} in the structure of \mathbf{X} for fitting covariance, and then c_{2R+1}, \dots, c_{3R} for providing additional flexibility to third order joint cumulants. In fact, we shall use c_1, c_2, \dots, c_{2R} and coefficients a_0, \dots, a_R for fitting first and second order moments.

Alternatively, one could impose that only a subset of each group of coefficients be non-zero. For example, $c_1, 0, \dots, 0$ for fitting covariances of \mathbf{Y} , then $c_{2R+1}, 0, \dots, 0$ for fitting third order joint cumulants, then $c_{3R+1}, 0, \dots, 0$ for fitting fourth order joint cumulants, and so on. Each of these groups has only one non-zero coefficient, and $R - 1$ zeros.

4.4.3. Combinations of both types of transformations

A third variant of the use of transformations is combining both types. When dealing with joint-quantiles related dependence manifestations, one can add more flexibility to the underlying dependence structure of section 4.4.1 by allowing it to be itself the transformation of a random vector having the archetypal dependence structure.

Namely, as in section 4.4.1, we have multivariate data to be modeled by random vector $\mathbf{Y} \sim F_{\mathbf{Y}}$; marginal distributions F_{Y_1}, \dots, F_{Y_J} ; an underlying random vector $\mathbf{X} \in \mathbb{R}^J$; and marginal transformations

$$T_j(X_j) := F_{Y_j}^{-1}(F_{X_j}(X_j))$$

such that

$$(T_J(X_J), \dots, T_J(X_J)) \sim F_{\mathbf{Y}}$$

But in this section, random vector $\mathbf{X} \in \mathbb{R}^J$ is assumed to be the result of polynomially transforming random vector $\mathbf{Z} \in \mathbb{R}^J$, which \mathbf{Z} possesses an archetypal dependence structure,

$$K_{\mathbf{Z}}(\mathbf{t}) = c_1 \frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t} + \frac{1}{2!} c_2 \left[\frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t} \right]^2 + \frac{1}{3!} c_3 \left[\frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t} \right]^3 + \dots$$

In this way, additional flexibility is given to the quantile-quantile transformation approach.

4.5. Dealing with censored/truncated data

It is often the case that we are in need to model censored or truncated data, in Spatial Statistics. Censored data arises, for example, when the precision of a gauging device under a certain threshold is questionable. Rainfall modeling provides a typical example of the need to model truncated data: one can just not have “negative” rainfall.

The approach taken in the present research work, and applied in section 6, is to use MCMC simulation to “fill in” the censored or truncated values with simulated latent variables consistent with observed data. The approach is called “data augmentation” and was introduced

by Tanner and Wong (1987); see also Gilks et al. (1998); van Dyk and Meng (2001). Two resources on this approach in the context of rainfall modeling are Sanso and Guenni (2000, 1999), where an underlying multivariate Gaussian model.

4.6. Simulation

One can sample realizations from a model with dependence structure (4.1.6) by means of Approximate Gibbs Sampler / Sequential simulation, both in the context of a random field, and in the general context of a multivariate random variable. A second approach relies on the "generating variable" of the model, which model is an instance of the family of Elliptically Contoured distributions. Both methods are briefly introduced in this section.

4.6.1. Gibbs Sampler / Sequential simulation

The objective is to sample $\mathbf{X} \sim F_{\mathbf{X}}$. This can be approximately attained by sampling each component from the conditional distributions

$$X_j^{(t)} \sim \Pr \left(X_j \mid X_1^{(t)}, \dots, X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, \dots, X_J^{(t-1)} \right) \quad (4.6.1)$$

one at a time, or block-wise. The super-index indicates the iteration number. After sufficiently many iterations, realizations obtained, $(X_1^{(t)}, \dots, X_J^{(t)})$ are approximately distributed as $F_{\mathbf{X}}$. Details can be found at chapter 5 of Gilks et al. (1998).

In the context of Spatial Statistics, each component represents the value of the random field at a specific location. Since the dimension of the fields to simulate makes specification of full conditionals of the form (4.6.1) unfeasible, the conditioning components must be limited to those representing a set of locations that are as close as possible to the new to-simulate location, its so-called neighbors. Thus, a random field simulated with this method would follow the next steps

1. Select, either randomly or in a systematic way, a new location on the plane s_j at which to sample your random quantity of interest $X_j = Z(s_j)$.
2. Identify a set of k neighbors, on which a realization of the random variable is already available. The number of neighbors to select depends on the computational resources at hand, but some playing around with this parameter indicates that it should not be smaller than 5.
3. Sample X_j from its conditional distribution, given the values of its selected neighbors.

The reader is referred to chapter 3 of Diggle and Ribeiro (2007), and the references therein for sequential simulation. The random fields presented in sections 5.1 and 5.2 of this dissertation were simulated using this scheme, with $k = 5$ and the sampling technique presented below.

In terms of this work, the most important question is how to simulate from the conditional distributions, given that all one has is the cumulant generating function of the full distribution. To this end we employ the formula due to Skovgaard (1987) (see also Kolassa (2006);

Barndorff-Nielsen and Cox (1990)), which produces a precise approximation to the distribution function of conditional distributions of a multivariate random variable, when only the cumulant generating function is available.

Namely, given $\mathbf{x}_{-j} = (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_J)$, one has

$$F_{\mathbf{X}_{-j}}(x_j) = \Pr(X_j \leq x_j \mid X_i = x_i, i \neq j) \approx \Phi(r) + \phi(r) \left(\frac{1}{r} - q \right) \quad (4.6.2)$$

where

$$r = \text{sign}(\hat{\lambda}_j) \sqrt{2 \left\{ \hat{\lambda}^T \mathbf{x} - \hat{\lambda}_{-j}^T \mathbf{x}_{-j} - K_{\mathbf{X}}(\hat{\lambda}) + K_{\mathbf{X}_{-j}}(\hat{\lambda}_{-j}) \right\}} \quad (4.6.3)$$

$$q = \frac{1}{\hat{\lambda}_j} \det(K''_{\mathbf{X}_{-j}}(\hat{\lambda}_{-j})) \det(K''_{\mathbf{X}}(\hat{\lambda}))^{-\frac{1}{2}} \quad (4.6.4)$$

and $\hat{\lambda} \in \mathbb{R}^J$, $\hat{\lambda}_{-j} \in \mathbb{R}^{J-1}$ are the solutions to equations

$$\begin{aligned} \nabla K_{\mathbf{X}}(\hat{\lambda}) &= (x_1, \dots, x_J) \\ \nabla K_{\mathbf{X}_{-j}}(\hat{\lambda}_{-j}) &= \mathbf{x}_{-j} \end{aligned}$$

Additionally, $\hat{\lambda}_j$ is the corresponding component of $\hat{\lambda}$, and $K''_{\mathbf{X}}(\hat{\lambda})$ stands for the matrix of second derivatives on the c.g.f. evaluated at $\hat{\lambda}$. Finally, $K_{\mathbf{X}_{-j}}$ is the c.g.f. of components of \mathbf{X} not including j .

With the aid of this approximation, one can readily sample from 4.6.1 by sampling $u^* \sim \text{Unif}(0, 1)$ and then solving numerically equation

$$F_{\mathbf{X}_{-j}}(x_j^*) = u^*$$

for x_j^* . This value x_j^* constitutes an approximate simulation from the conditional distribution in question.

4.6.2. Generating Variable method

The representation of elliptically distributed random vector $\mathbf{X} \in \mathbb{R}^J$ given by (4.2.1) is very useful for the sake of simulation. In order to sample from \mathbf{X} having cumulant generating function

$$K_{\mathbf{X}}(\mathbf{t}) = c_1 \frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t} + \frac{1}{2!} c_2 \left[\frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t} \right]^2 + \frac{1}{3!} c_3 \left[\frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t} \right]^3 + \dots \quad (4.6.5)$$

do the following:

1. Identify the connection parameters m_1, m_2, m_3, \dots corresponding to the prescribed c_1, c_2, c_3, \dots by using (4.1.12).
2. Use equation (4.2.11) to identify the distribution of the squared generating variable R^2 through its moments, $E\left((R^2)^k\right)$. This results in parameters θ and (π_1, \dots, π_S) for probability density $f_{R^2}(t) = \sum_{s=1}^S \pi_s \frac{\theta^s}{\Gamma(s)} t^{s-1} \exp(-t\theta)$.

3. Sample $R^2 \sim f_{R^2}$ and set $R = \sqrt{R^2}$. Sampling can be performed by sampling $s \in \{1, \dots, S\}$ from a multinomial distribution with class probabilities (π_1, \dots, π_S) , and then using this value to sample $R^2 \sim \text{Gamma}(s, \theta)$.
4. Sample \mathbf{u}^{J-1} , a random vector uniformly distributed on the unit hypersphere on \mathbb{R}^J . This can be attained by sampling J i.i.d. standard normal variates, $\xi = (\xi_1, \dots, \xi_J)$, and then dividing by the euclidean norm: $\mathbf{u}^{J-1} := \xi / \|\xi\|_2$.
5. Your realization \mathbf{x} of \mathbf{X} is given by setting $\mathbf{x} := R \times \mathbf{u}^{J-1} \times \Gamma^{1/2}$. A vector of non-zero means, $\mu \in \mathbb{R}^J$, can be added if necessary at this stage: $\mathbf{x} := \mu + R \times \mathbf{u}^{J-1} \times \Gamma^{1/2}$.

One can then generate easily realizations of random fields using this approach, provided one has matrix $\Gamma^{\frac{1}{2}}$. As dimension J increases, this becomes unfeasible or unpractical. This is a well-known issue in spatial statistics and several algorithms are available for producing (approximate) realizations from Gaussian random fields without having to compute $\Gamma^{\frac{1}{2}}$, two of which are the *turning bands* method (see, for example Ripley (1981)) and the *Spectral Method* (see Cressie (1991)). The following section shows how we can exploit such techniques to simulate random fields having a dependence structure prescribed by our model.

4.6.2.1. Deviance from Normality

It will be convenient, e.g. for simulating big random fields of dimension $J \gg 1$, to be able to express the squared generating variable R^2 of our dependence model as a product of two generating variables:

$$R^2 = R_*^2 \times \chi_J^2 \quad (4.6.6)$$

Then, simulation from random vector $\mathbf{X} \in \mathbb{R}^J$ with location vector μ and dispersion matrix Γ , can proceed by simulating a realization \mathbf{z}_i from a random vector $\mathbf{Z} \sim N_J(\mathbf{0}, \Gamma)$ and then setting

$$\mathbf{x}_i := \mu + \sqrt{R_*^2} \times \mathbf{z}_i \quad (4.6.7)$$

As already mentioned, there are algorithms available for simulation of big Gaussian random fields, hence we assume that the sampling of \mathbf{z}_i is not an issue.

In order to find the distribution of scaling variable R_*^2 , assume the density of the squared generating variable R^2 of $\mathbf{X} \in \mathbb{R}^J$ has been fitted as a mixture of gamma distributions,

$$f_{R^2}(z) = \sum_{s=1}^S \pi_s \frac{\theta^s}{\Gamma(s)} z^{s-1} e^{-\theta z}$$

The idea is now to find a random variable R_*^2 such that $R_*^2 \times \chi_J^2 = R^2$, where R^2 is the squared generating variable of the model. To avoid cumbersome notation we label:

$$\begin{aligned} R_*^2 &:= \xi \\ \chi_J^2 &:= x \end{aligned}$$

Now, if we had x , application of the change of variables theorem dictates that the density of ξ is:

$$f_{\xi|x}(\xi) = f_{R^2}(\xi x) x$$

We just have to integrate out this variable x , which is a χ_J^2 random variable. Specifically,

$$\begin{aligned}
 f_\xi(\xi) &= \int_0^{+\infty} f_{\xi|x}(\xi) f_x(x) dx = \\
 &= \int_0^{+\infty} \left\{ \sum_{s=1}^S \pi_s \frac{\theta^s}{\Gamma(s)} (\xi x)^{s-1} e^{-\theta \xi x} \right\} \frac{\left(\frac{1}{2}\right)^{\frac{J}{2}}}{\Gamma\left(\frac{J}{2}\right)} x^{\frac{J}{2}-1} e^{-\frac{1}{2}x} dx = \\
 &= \sum_{s=1}^S \pi_s \frac{\theta^s \left(\frac{1}{2}\right)^{\frac{J}{2}}}{\Gamma(s) \Gamma\left(\frac{J}{2}\right)} \xi^{s-1} \int_0^{+\infty} x^{(s+\frac{J}{2})-1} e^{-(\theta\xi+\frac{1}{2})x} dx = \\
 &= \sum_s^S \pi_s \frac{\theta^s \left(\frac{1}{2}\right)^{\frac{J}{2}} \Gamma\left(s+\frac{J}{2}\right)}{\Gamma(s) \Gamma\left(\frac{J}{2}\right) (\theta\xi+\frac{1}{2})^{s+\frac{J}{2}}} \xi^{s-1} \quad (4.6.8)
 \end{aligned}$$

Hence the density of $\xi = R_*^2$ is given by (4.6.8), using the same estimated parameters $\theta, \pi_1, \dots, \pi_S$ for the R^2 variable. One can then sample random vector $\mathbf{X} \in \mathbb{R}^J$ possessing the required dependence structure by using relation (4.6.7).

The usefulness of the formulas presented in this section will become evident in section 6, when dealing with a big random field, simulated under the dependence structure studied in this dissertation.

Part II.

Examples and Illustrations

5. Two Random fields

We shall present in this chapter the analysis of two random fields having non-zero interdependence parameters (i.e. joint cumulants) of order higher than two. It will be noted that they possess virtually the same variogram function as the Gaussian fields presented for comparison. Still, they present non-Normal behavior with respect to other measures of dependence.

The fields were simulated using sequential simulation with the aid of formula (4.6.2). Five neighbors were used for each realization. For each example we generated 2650 independent, uniformly distributed random variables, $u_1, u_2, \dots \sim Unif(0, 1)$. Using this "random path", both the Gaussian field and the non-Gaussian fields were simulated, in order to prevent confusion of the effect of the random path employed with the effect of the joint cumulants used.

The 2650 random values correspond to a 50×53 grid on the plane. Simulation proceeded along the rows of this grid, one row after the other (no random selection of the next location to sample from). The values obtained for the first three rows of each field, that is, the first 150 simulated values, were discarded for analysis. The reason for this is to avoid a "one dimensional" effect for those first simulation that had no neighbors in the Y-axis. Hence the examples are concerned with a 50×50 grid.

5.1. Random Fields Set 1

A powered exponential covariance model, as in (2.2.8), was used with parameter $(\theta_1, \theta_2, \sigma_0^2, \sigma_1^2) = (3, 1, 0, 1)$. Five random fields were simulated with the same random path, the difference being only in the coefficients of the dependence structure (4.1.6). Meaningful names were assigned to each field for convenience. So, field 2-D refers to a field with only c_1 set to non-zero; field 4-D refers to a field with non-zero c_2 , and so on. The values used for each field are presented on table 5.1.1. Although more combinations are obviously possible, simple configuration may help discriminate the effects of the different coefficients.

Additionally, a Gaussian random field simulated using the SVD method was also generated, for comparisons purposes. This correspond to a theoretically correct random field, with no artifacts produced by the sequential simulation.

Plots of the random fields appear on figures 5.1.1 through 5.1.3. These fields look very similar.

5.1.1. Empirical Variograms

In order to analyze the dependence structure, we present a series of empirical variograms on figure 5.1.4.

Coeff / F. Name	c_1	c_2	c_3	c_4	c_5
2-D	1	0	0	0	0
4-D	1	2	0	0	0
6-D	1	0	2	0	0
8-D	1	0	0	2	0
10-D	1	0	0	0	2

Table 5.1.1.: Random fields c.g.f coefficients configurations

First, the empirical variograms of the ranks, on the upper left plot. We notice that the sequentially simulated fields have virtually the same empirical variogram for their ranks. This is not the case for the SVD simulated Gaussian field. This indicates an effect of the sequential simulation method employed, and points to the desirability in the future of using an exact simulation method for the non-Gaussian fields also, e.g. by means of the "generating variable" method presented in section 4.6.2.

Second, on the upper right plot, the empirical variograms of the fields indicate that the inclusion of coefficients $c_{r>1}$ have altered the variance of the fields. This is most clear for field 4-D.

Third, empirical variograms resulting from scaling all variables are presented on the lower left plot. By scaling it is meant to subtract the mean of all values and divide by their standard deviation. The matching is better.

Fourth, by scaling only the values of the non-Gaussian fields, the variograms of all sequentially simulated random fields are virtually the same. That is to say, with respect to this bi-variate dependence parameter, they are almost indistinguishable. For the rest of this section, we work with the unscaled 2-D field and the scaled non-Gaussian fields.

5.1.2. Marginal Distributions

The next step consists in testing whether the marginal distributions of all fields are the same. To explore the matter, we present quantile-quantile plots of the fields values. In figure 5.1.5 the quantile-quantile plots that compare the marginal distribution of each of the non-Gaussian fields with that of field 2-D are shown. Apart from some deviance for the lower quantiles, the marginal distributions look quite similar. Moreover, one might not possess data from all the fields in practice. Q-Q plots of randomly selected samples of size $n=200$ are also presented for comparison: it is difficult to state from these plots that data come from different distributions.

Further, two tests are applied for comparing data from the simulated fields; the Anderson-Darling test and the Kolmogorov-Smirnov tests for equality in distributions. Since using all data might result in spurious rejection of the equality hypothesis, a Monte Carlo procedure was followed instead:

1. Randomly select a sample of size $n=200$ from each field's data.
2. Perform the Anderson-Darling and the Kolmogorov-Smirnov tests on each pair of samples. Check whether the attained p-value for each pairwise comparison was or not greater than 0.05, and store this information.

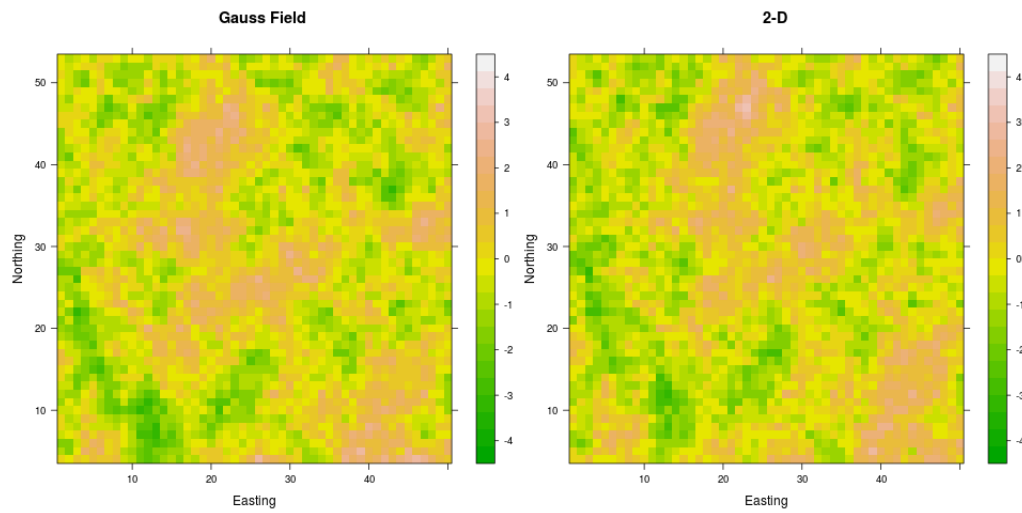


Figure 5.1.1.: Perfect Gaussian Random field (left). Sequentially simulated random field (right)

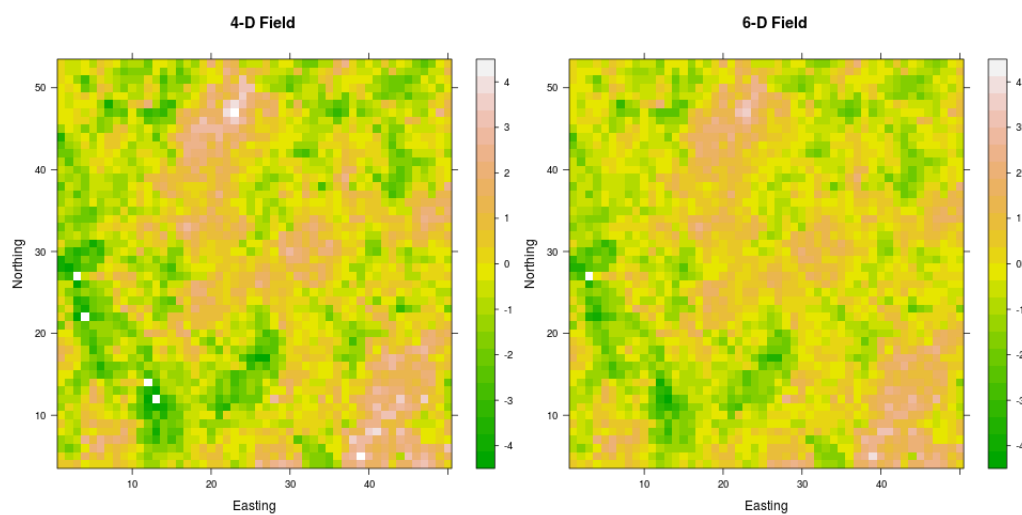


Figure 5.1.2.: Field with 4-th (left) and 6-th (right) order non-zero joint cumulants

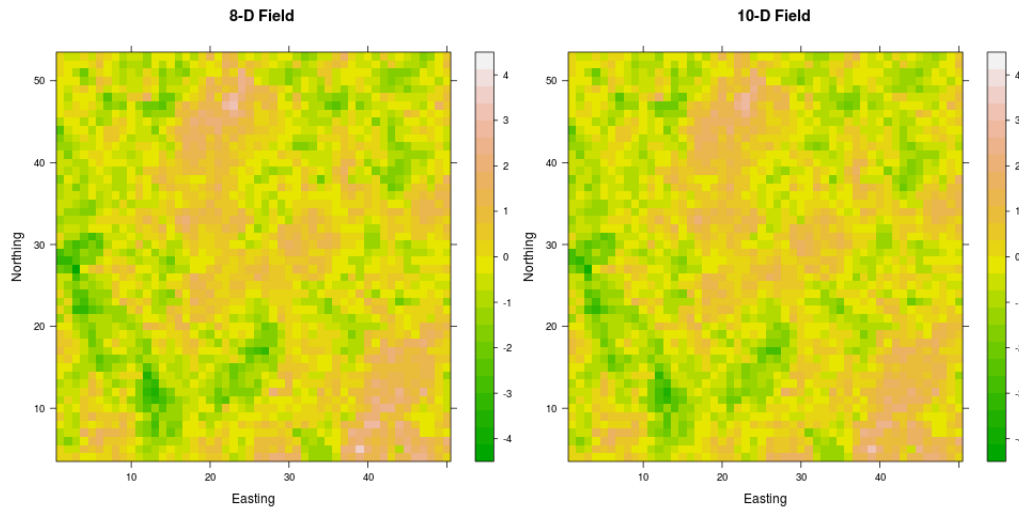


Figure 5.1.3.: Field with 8-th (left) and 10-th (right) order non-zero joint cumulants

3. Repeat steps one and two $B = 1000$ times.
4. Check the proportion of times that the equality in distribution was rejected.

The results of this experiment are presented in table 5.1.2. The theoretically correct Gaussian field was included as reference; an independent Normally distributed random sample was simulated each iteration for comparison. As it is seen, the proportion of times that a truly Gaussian variable is rejected to be equal in distribution to the Gaussian field data, is for both tests greater than the same proportion for the non-Gaussian fields data. This points to the near-normality of the marginal distributions of all fields. Additionally, non-Gaussian fields have virtually indistinguishable marginals.

5.1.3. Two, Three and Four dimensional Marginals

The following step is to test for the normality of samples of two, three and four dimensional marginal distributions of vectors formed from the fields. For this part of the analysis, "interesting distances" were selected by visual analysis of the empirical variograms of figure 5.1.4. Distances used appear on table 5.1.3.

For the two dimensional marginals, pairs of values from locations lying at distance approximately two (2) were formed into samples from the 2-dimensional marginals.

In order to form three dimensional samples, triangles with sides approximately the selected interesting distances were identified on the field grid and their values taken. In this manner, samples of the three dimensional marginal were taken for each field.

For the four-dimensional marginals, one needs six distance categories in order to create a 4 by 4 covariance matrix. Interesting distances were again selected, but this time no straightforward representation on the plane is available. Hence we used non-metrical multi-dimensional scaling in order to obtain an approximate 2-dimensional representation of the six locations. This representation serves as a "mold", as the one seen in figure 5.1.6. This mold was then randomly rotated and its center randomly assigned to locations on the field's

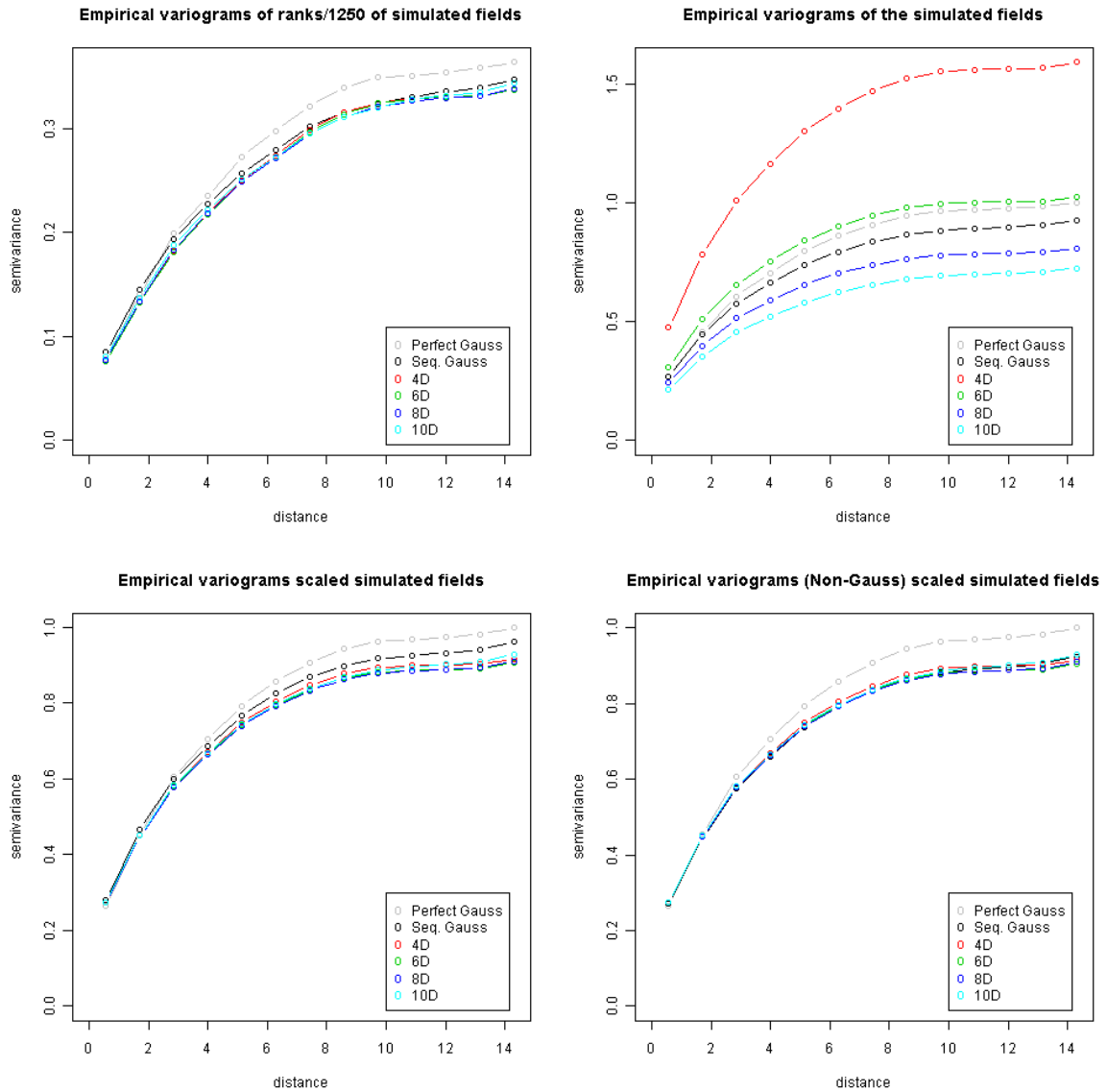


Figure 5.1.4.: Empirical variograms from simulated fields. More convenient variant corresponds to scaling fields 4-D, 6-D, 8-D and 10-D. In this way all sequentially simulated fields possess the same empirical variogram.

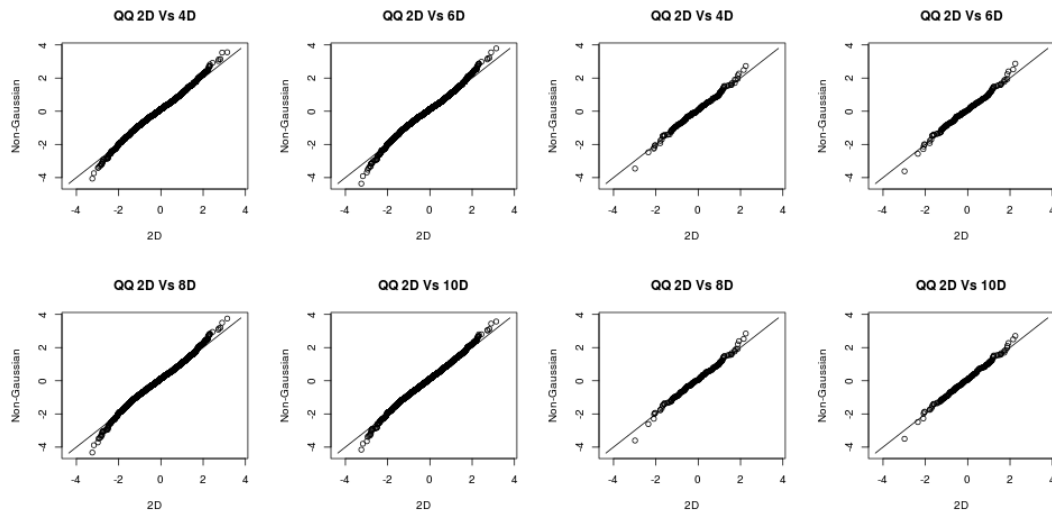


Figure 5.1.5.: Quantile-Quantile plots of values of simulated fields: all data (left), a randomly selected sample of size $n=200$ (right).

		Anderson-Darlin Test					
		Gauss	2D	4D	6D	8D	10D
Gauss		0.197	0.016	0.116	0.137	0.124	0.103
2D		NA	NA	0.003	0.004	0.002	0.002
4D		NA	NA	NA	0.000	0.000	0.000
6D		NA	NA	NA	NA	0.000	0.000
8D		NA	NA	NA	NA	NA	0.000
10D		NA	NA	NA	NA	NA	NA
		Kolmogorov-Smirnov Test					
		Gauss	2D	4D	6D	8D	10D
Gauss		0.197	0.016	0.116	0.137	0.124	0.103
2D		NA	NA	0.003	0.004	0.002	0.002
4D		NA	NA	NA	0.000	0.000	0.000
6D		NA	NA	NA	NA	0.000	0.000
8D		NA	NA	NA	NA	NA	0.000
10D		NA	NA	NA	NA	NA	NA

Table 5.1.2.: Tests for equality in marginal distributions of the different fields analyzed. Proportion of times, out of 1000, in which p-value of the test was smaller than 0.05.

grid. Values corresponding to locations on which the "arms" of the mold fell, were formed into samples from the 4-dimensional marginals of the field's data.

The testing for marginal normality was performed as follows for each dimension d of the marginal: 2, 3 and 4.

1. Select a d -dimensional sample of size $n = 500$ from the field.
2. Apply the multivariate Shapiro test proposed by Villasenor Alva and Estrada (2009), which test as alternative hypothesis non-Normality. This is a test specialized for Normality testing.
3. Perform steps one and two $B = 100$ times, and store the p-value obtained.

Since the Shapiro Test is very sensible, even with small sample sizes, a visual exploration of the p-values produces by the tests will provide a quick view of the situation for the different marginals. We see the results in figures 5.1.7, 5.1.8 and 5.1.9 the histograms of the p-values obtained for marginal distributions of dimension 2, 3 and 4, respectively. In general, the test rightly rejects the hypothesis of Normality. This rejection seems more marked as dimension of the marginal increases.

5.1.4. Interaction Manifestation: Sums of components

Sums of values corresponding to sets of locations on the field are important statistics for hydrological applications (for example, if the field represents precipitation values).

We analyze again the behavior of sums components of 2, 3 and 4 dimensional marginals. Data samples are collected using the same interesting distances as in last section, and with the same procedure for the 4-dimensional marginal.

Specifically, comparison was performed for each dimension d of the marginal, as follows:

1. Collect $n = 2000$ samples from the d -dimensional marginal of the field into a $n \times d$ data matrix A .
2. Set to zero all negative values in A (other positive thresholds are also interesting). This is to keep the analogy to precipitation modeling, and to prevent canceling out by negative values with a great absolute value.
3. Compute the sums of the components, by adding along the rows of A . This results in a vector a of size n containing the sums of components for this sample.
4. Compute the sample 75%, 90% and 99.5% quantiles of sums in vector a , and store this information.

Marginals	Approx. distances considered
2-dim	2
3-dim	1,3,7
4-dim	1,3,5,7,9,11

Table 5.1.3.: Interesting distances used for the multivariate marginals analysis

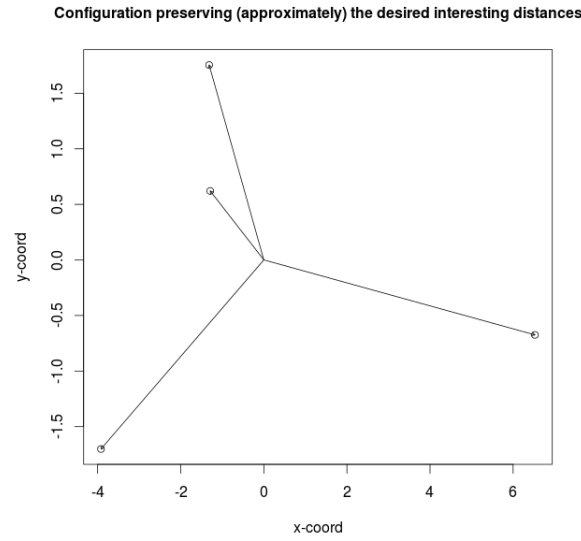


Figure 5.1.6.: 2-dimensional approximate representation of the six distances involved in the 4-dimensional marginal analysis

5. Repeat steps 1 through 4 a total of $B = 100$ times.

This procedure provides a means of comparing the sums of components, particularly at their uppermost tails. The results are presented in figures 5.1.10, 5.1.11 and 5.1.12, for marginal distributions of dimension 2, 3 and 4, respectively.

A pattern can be observed, regardless of marginal dimension: quantiles of the sums corresponding to non-Gaussian fields are in general lower than those corresponding to the Gaussian field for the 75% and 90% quantiles. But 99.5% quantiles are appreciably higher for sums from non-Gaussian fields.

5.1.5. Interaction manifestations: A statistic built on the marginal joint probability distributions

To finish this comparisons session, we employ the congregation measure used by Bárdossy and Pegram (2009) for the sake of model validation. This measure is now briefly introduced. The object of analysis is a random vector \mathbf{X} of dimension J , which can correspond to a multidimensional marginal vector of a higher dimensional random vector (as in the example below). Set a threshold quantile, say $a = 90\%$, and define binary random variables for each $j = 1, \dots, J$

$$V_j = \begin{cases} 1, & F_j(X_j) > a \\ 0, & F_j(X_j) \leq a \end{cases} \quad (5.1.1)$$

This results in a discrete random vector $\mathbf{V} = (V_1, \dots, V_J)$. The congregation measure referred to is defined to be the entropy of \mathbf{V} ,

$$\text{congr}(\mathbf{X}) = - \sum \Pr(V_1, \dots, V_J) \log(\Pr(V_1, \dots, V_J)) \quad (5.1.2)$$

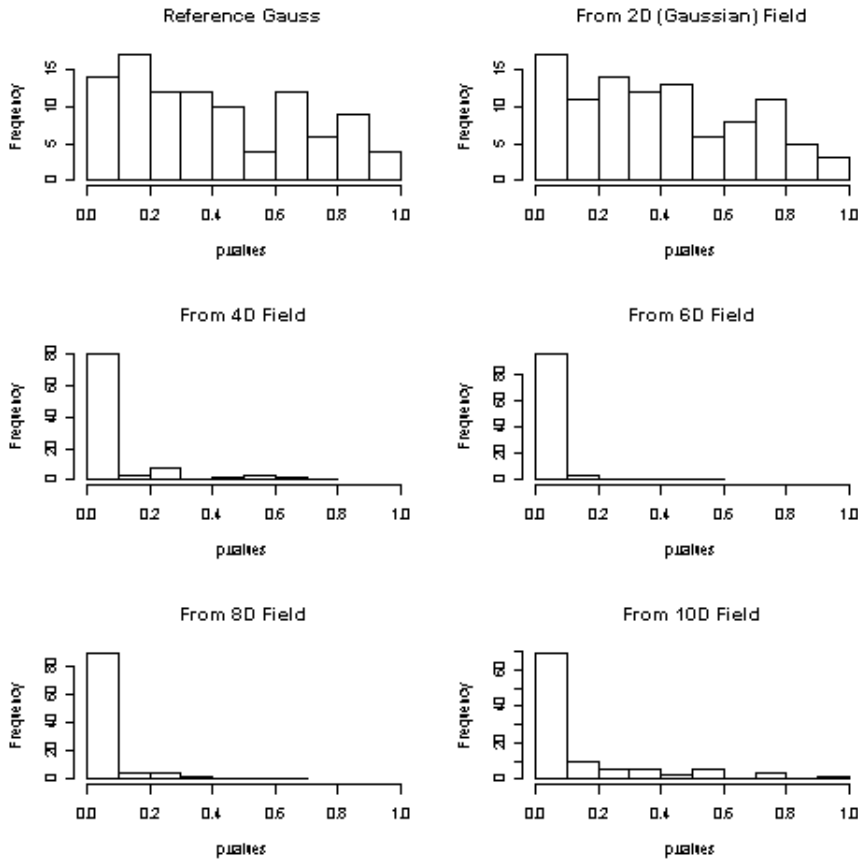


Figure 5.1.7.: Histograms for Multivariate Shapiro's test for Normality applied to the 2-dimensional marginal distributions. The test rightly rejects most of the time the hypothesis of normality.

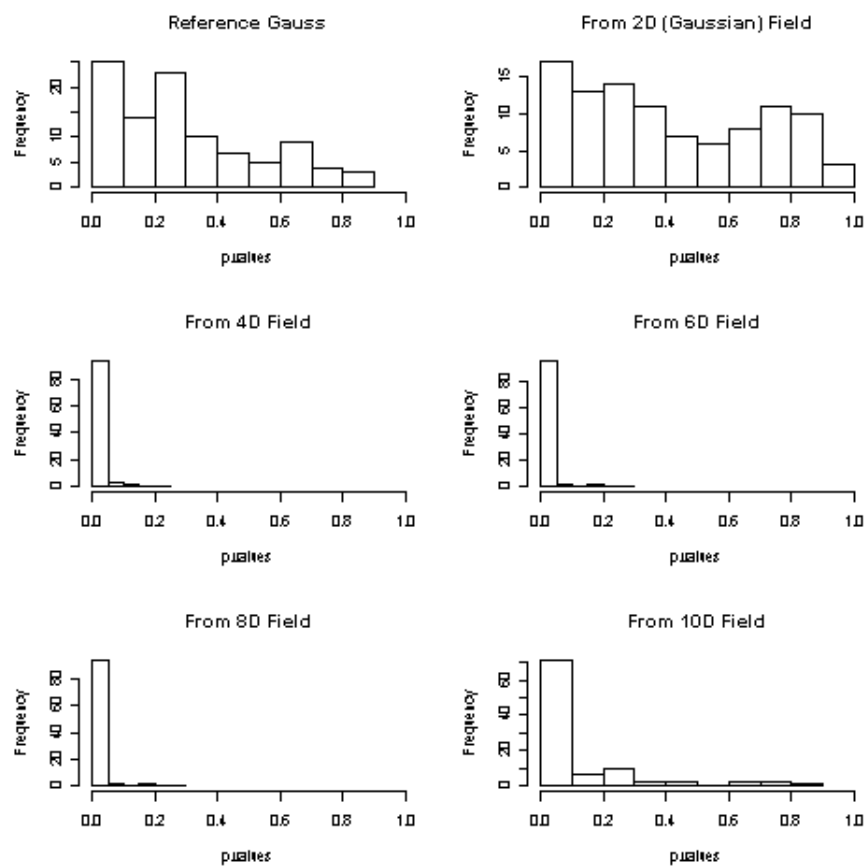


Figure 5.1.8.: Histograms for Multivariate Shapiro's test for Normality applied to the 3-dimensional marginal distributions. The test rightly rejects most of the time the hypothesis of normality.

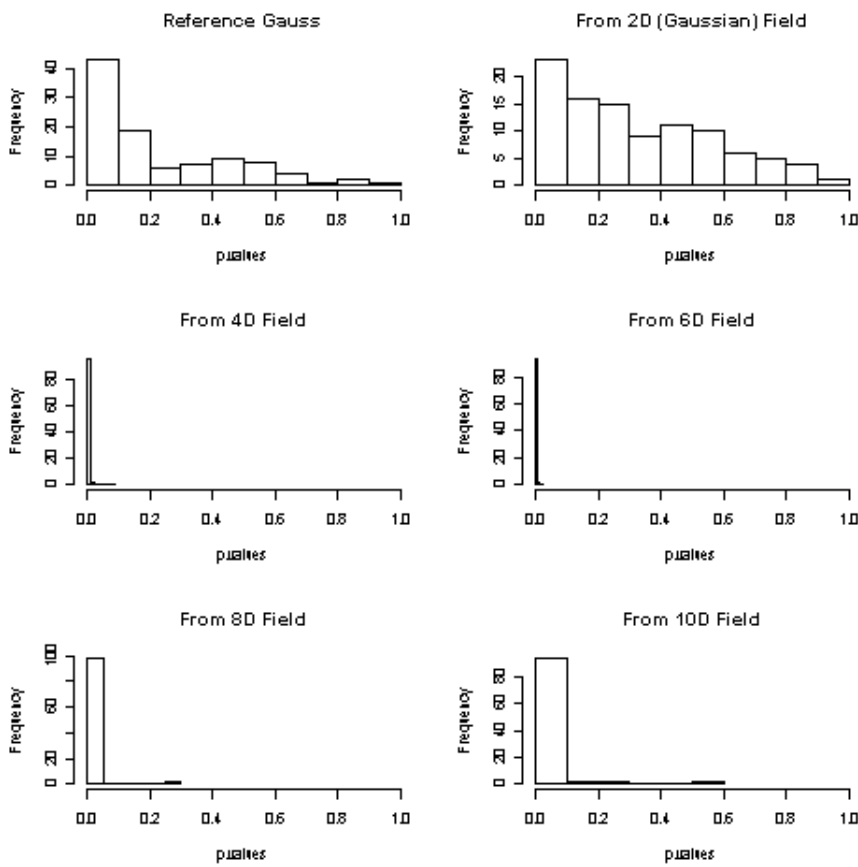


Figure 5.1.9.: Histograms for Multivariate Shapiro's test for Normality applied to the 4-dimensional marginal distributions. The test rightly rejects most of the time the hypothesis of normality.

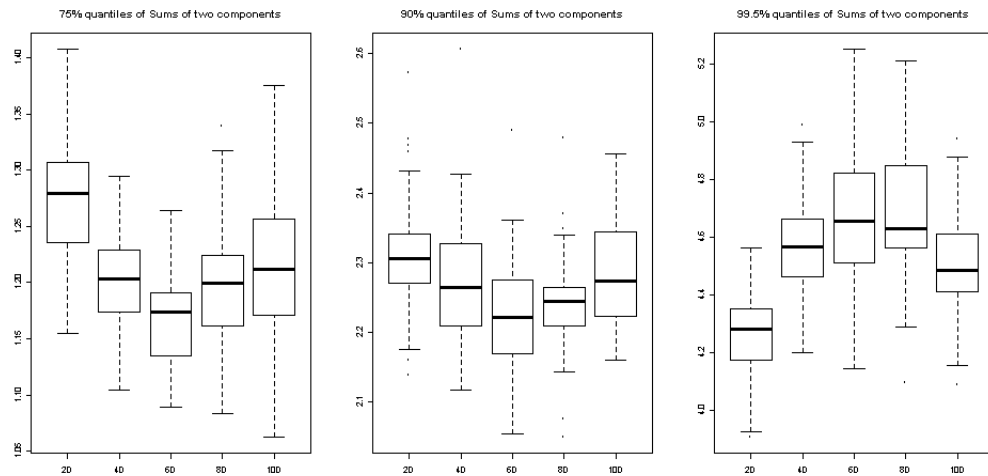


Figure 5.1.10.: Comparison of sums of components for 2-dimensional marginals. Increase is appreciable for sums of components of non-Gaussian fields at the uppermost quantile.

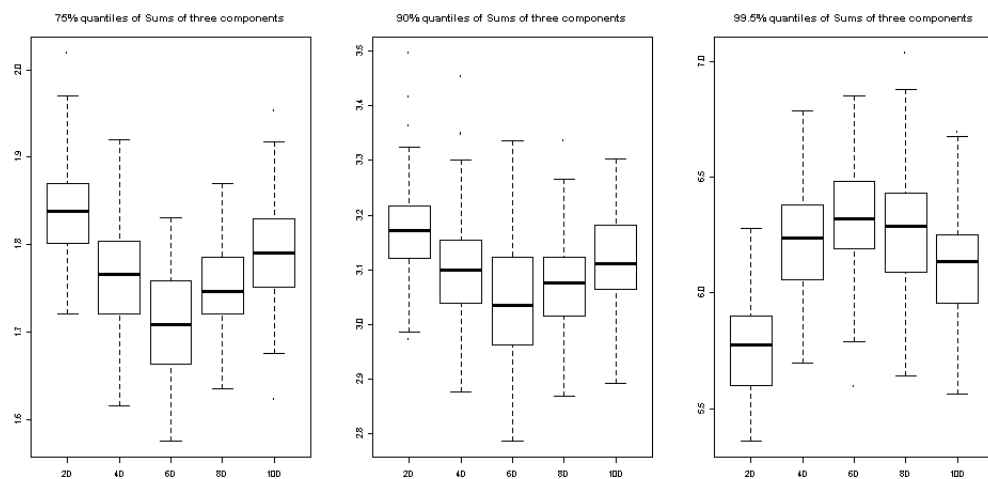


Figure 5.1.11.: Comparison of sums of components for 3-dimensional marginals. Increase is appreciable for sums of components of non-Gaussian fields at the uppermost quantile.

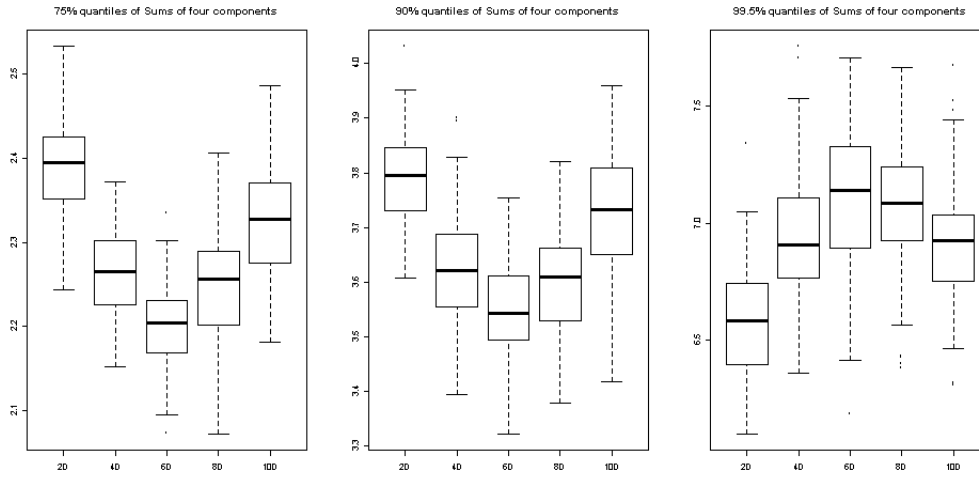


Figure 5.1.12.: Comparison of sums of components for 4-dimensional marginals. Increase is appreciable for sums of components of non-Gaussian fields at the uppermost quantile.

summed over all possible values of (V_1, \dots, V_J) . That is, the measure is defined as the entropy of the joint distribution of the binary variables just defined. A higher value of this measure indicates less congregation.

One characteristic of this entropy or congregation measure, is that it is not affected by monotonic transformations on random variables X_j , since each one is investigated as to whether it trespasses its own quantile.

The comparison procedure is similar to the procedure used for the sums of components, and uses the same interesting distances:

1. Collect $n = 2000$ J -dimensional samples from the field. This sample is considered to be n -realizations of \mathbf{X} .
2. Compute the sample estimate of $\text{congr}(\mathbf{X})$, and store this information.
3. Repeat steps one and two a total of $B = 100$ times.

The results are presented in figure 5.1.13 for marginal distributions of dimension 2, 3 and 4. The pattern observed is: congregation is greater for data from non-Gaussian fields, regardless of the dimension of the marginal under analysis.

Summary

We have seen that it is possible to simulate a random field having identical 1-dimensional marginal distribution and variogram as a Gaussian field, but with different responses in terms of the two interdependence manifestations considered: sums of higher dimensional marginal components, and a congregation measure based on the joint probability of the multivariate marginals.

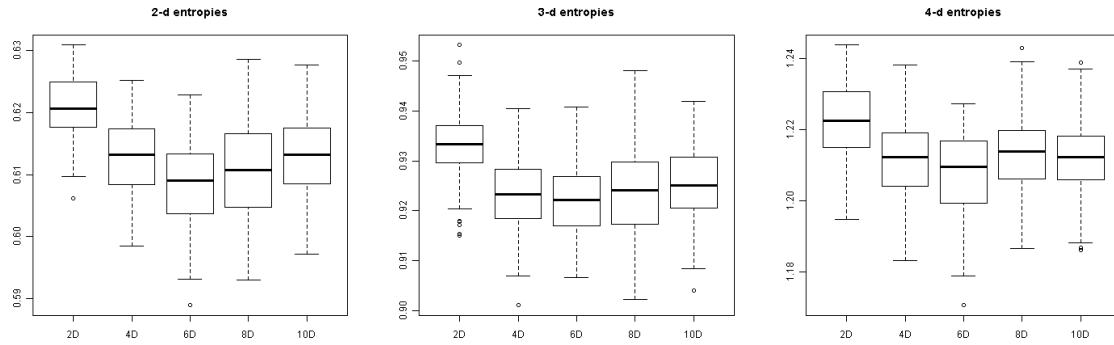


Figure 5.1.13.: Comparison of congregation measure for marginals of different dimensions and data from the different fields. From left to right results of the comparison procedure are given for marginals of dimension 2, 3 and 4. Congregation seems to be always greater for the non-Gaussian fields.

For the next example, only the results of the analysis are presented, unless a comment is strictly necessary. The procedures followed for comparison are identical with those of this section.

5.2. Random Fields Set 2

We used the same covariance model as in last section for simulation of the fields. Namely a powered exponential model with $(\theta_1, \theta_2, \sigma_0^2, \sigma_1^2) = (3, 1, 0, 1)$. Coefficients of the dependence structure used for this experiment are shown in table 5.2.1.

Plots of the simulated fields, including the theoretically correct Gaussian field, are presented in figures 5.2.1, 5.2.2 and 5.2.3.

It is found, by observation of the empirical variograms, that if all sequentially simulated fields are scaled, then the empirical variograms are virtually the same. The analysis proceeds with data of all sequentially simulated fields scaled.

Quantile-Quantile plots, presented in figure 5.2.5, indicate that 1-dimensional marginal distributions of all fields are similar. Additionally, the Anderson-Darling and Kolmogorov-Smirnov tests very seldom reject the hypothesis of equality at level $\alpha = 0.05$, as shown in table 5.2.2.

Coeff / F. Name	c_1	c_2	c_3	c_4	c_5
2-D	1	0	0	0	0
4-D	1	1	0	0	0
6-D	1	0	1	0	0
8-D	1	0	0	1	0
10-D	1	0	0	0	1

Table 5.2.1.: Random fields c.g.f coefficients configurations

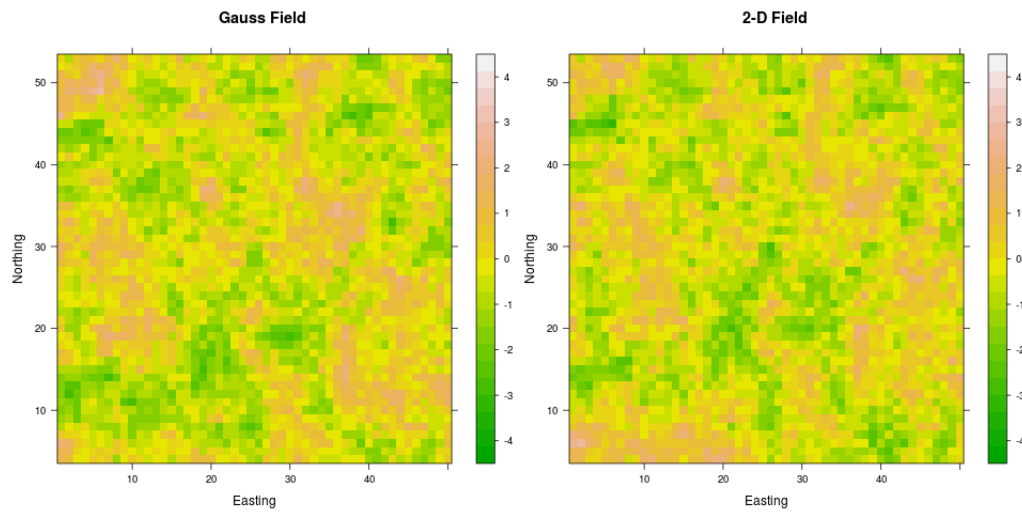


Figure 5.2.1.: Perfect Gaussian Random field (left). Sequentially simulated random field (right)

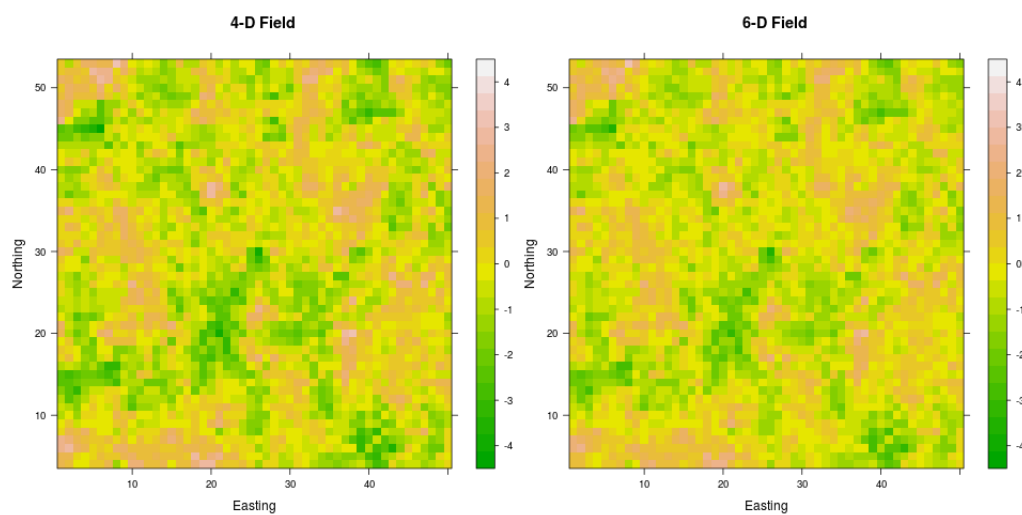


Figure 5.2.2.: Field with 4-th (left) and 6-th (right) order non-zero joint cumulants

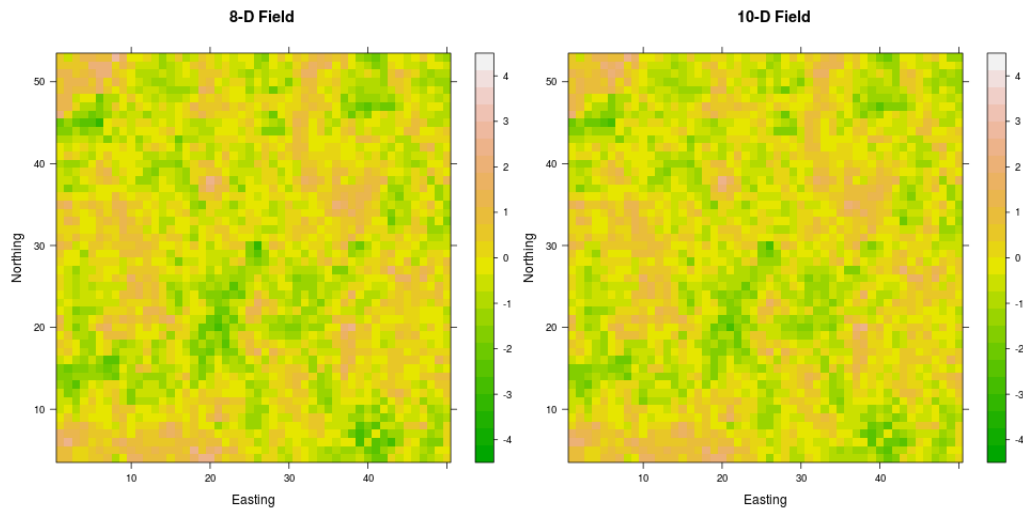


Figure 5.2.3.: Field with 8-th (left) and 10-th (right) order non-zero joint cumulants

Anderson-Darlin Test						
	Gauss	2D	4D	6D	8D	10D
Gauss	0.043	0.000	0.003	0.004	0.003	0.002
2D	NA	NA	0.005	0.005	0.001	0.001
4D	NA	NA	NA	0.000	0.000	0.000
6D	NA	NA	NA	NA	0.000	0.000
8D	NA	NA	NA	NA	NA	0.000
10D	NA	NA	NA	NA	NA	NA
Kolmogorov-Smirnov Test						
	Gauss	2D	4D	6D	8D	10D
Gauss	0.029	0.001	0.005	0.005	0.003	0.002
2D	NA	NA	0.001	0.000	0.000	0.000
4D	NA	NA	NA	0.000	0.000	0.000
6D	NA	NA	NA	NA	0.000	0.000
8D	NA	NA	NA	NA	NA	0.000
10D	NA	NA	NA	NA	NA	NA

Table 5.2.2.: Tests for equality in marginal distributions of the different fields analyzed. Proportion of times, out of 1000, in which p-value of the test was smaller than 0.05.

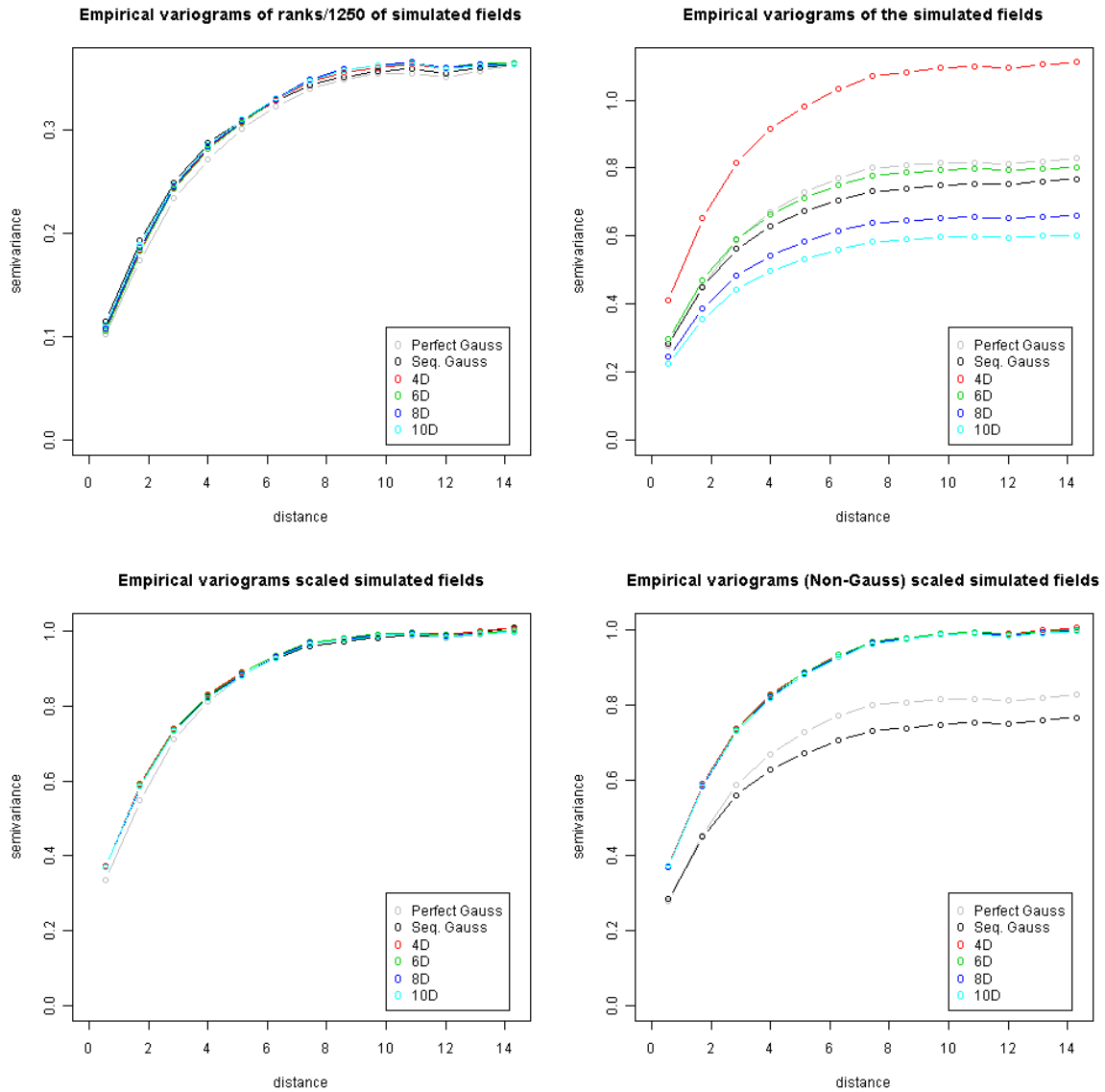


Figure 5.2.4.: Empirical variograms from simulated fields. More convenient variant corresponds to scaling all fields: 2-D, 4-D, 6-D, 8-D and 10-D. In this way all sequentially simulated fields possess the same empirical variogram.

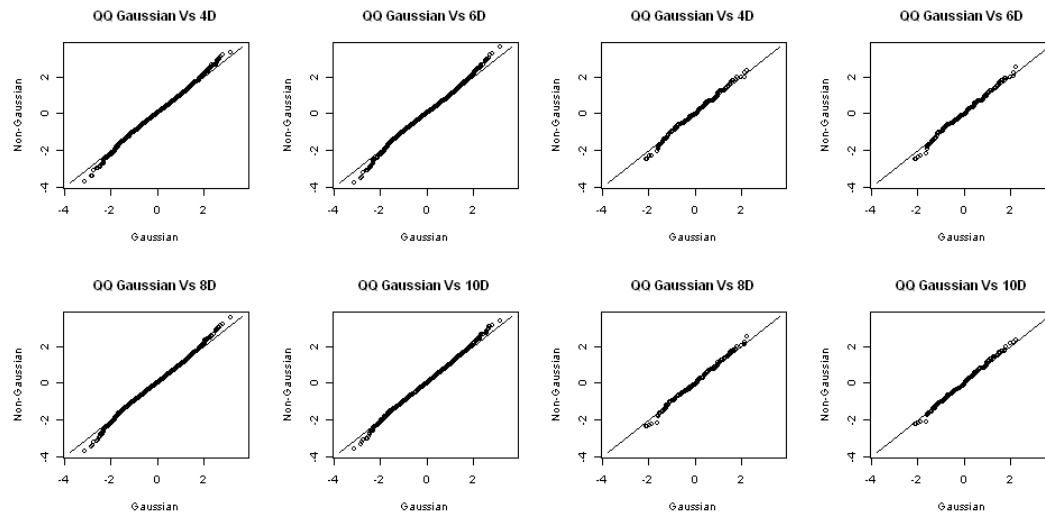


Figure 5.2.5.: Quantile-Quantile plots of values of simulated fields: all data (left), a randomly selected sample of size $n=200$ (right).

Concerning Shapiro-type multivariate tests for Normality, the non-Gaussian simulated fields are closer to Normality, as can be inferred from figures 5.2.6, 5.2.7 and 5.2.8. Field 10D has two-dimensional marginals that are very similar to those of a Gaussian distribution (figure 5.2.6).

Regarding the sums of components, non-Gaussian fields present a systematic increase in the quantiles, as compared with the Gaussian field. This is true even for field 10D, which is very similar to a Gaussian field in its 2-dimensional marginals. See figures 5.2.9, 5.2.10 and 5.2.11. The congregation measure presents a less clear picture, with respect to this criterion the fields simulated are not visibly different.

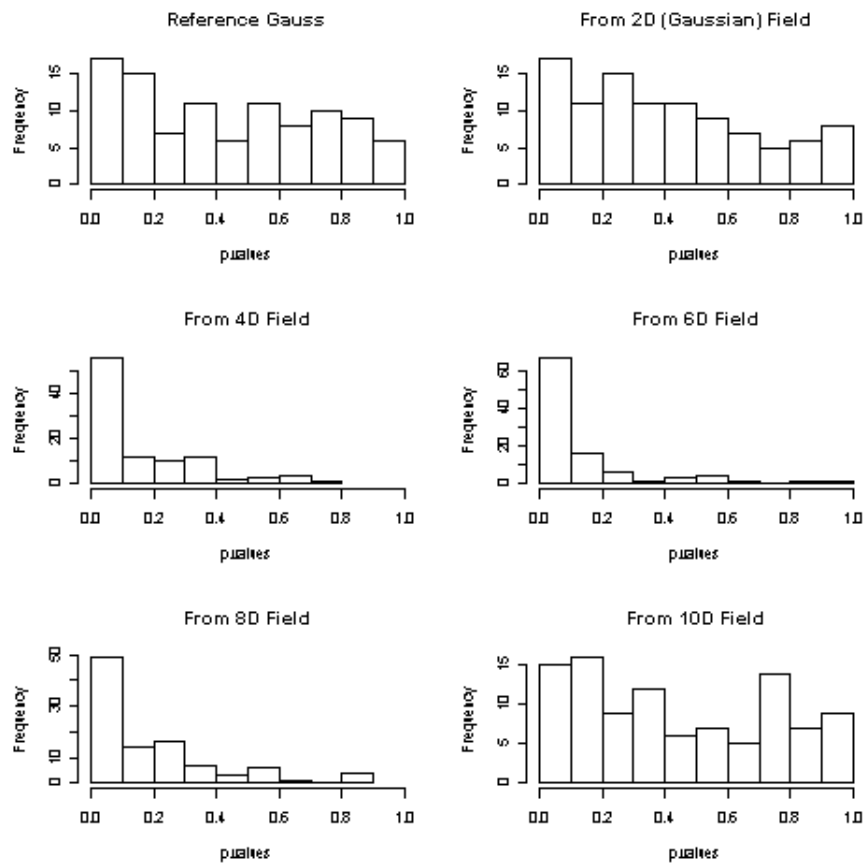


Figure 5.2.6.: Histograms for Multivariate Shapiro's test for Normality applied to the 2-dimensional marginal distributions. The test rightly rejects most of the time the hypothesis of normality, except for field 10D, which exhibits a similar rejection pattern as the 2D and the true Gaussian fields.

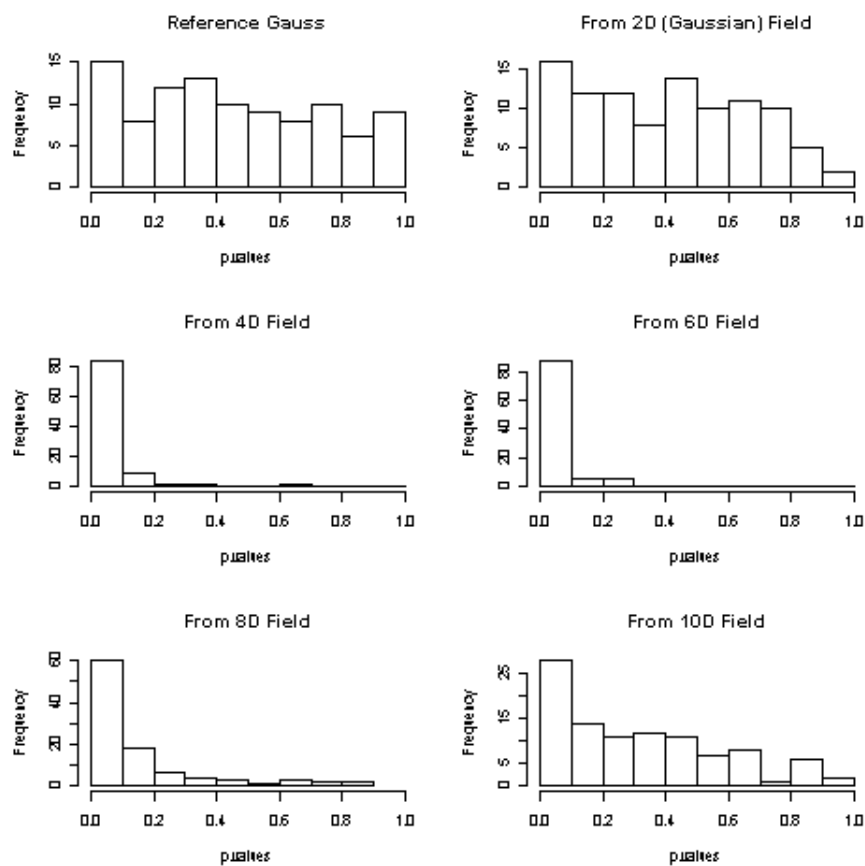


Figure 5.2.7.: Histograms for Multivariate Shapiro's test for Normality applied to the 3-dimensional marginal distributions. The test rightly rejects most of the time the hypothesis of normality. Random field 10D is roughly Gaussian, however.

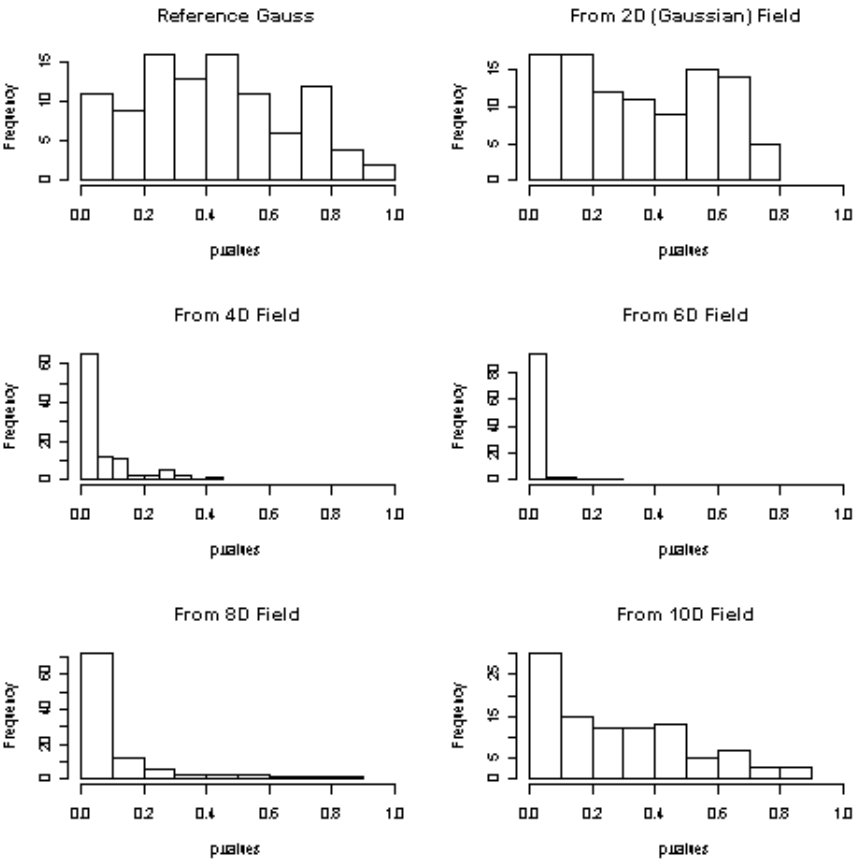


Figure 5.2.8.: Histograms for Multivariate Shapiro's test for Normality applied to the 4-dimensional marginal distributions. The test rightly rejects most of the time the hypothesis of normality. Random field 10D is roughly Gaussian, however.

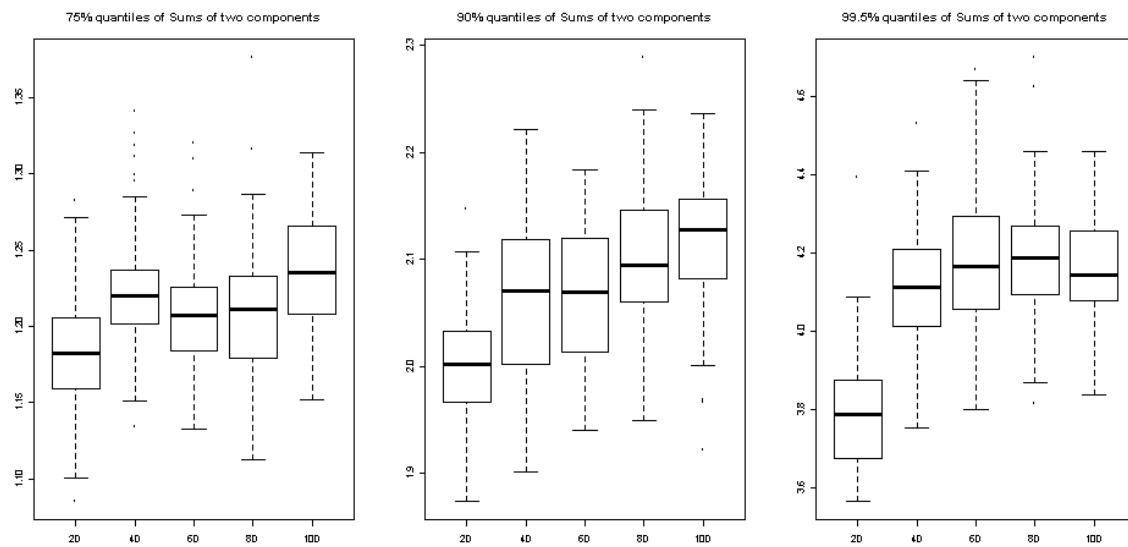


Figure 5.2.9.: Comparison of sums of components for 2-dimensional marginals. Increase is appreciable for sums of components of non-Gaussian fields at the uppermost quantile.

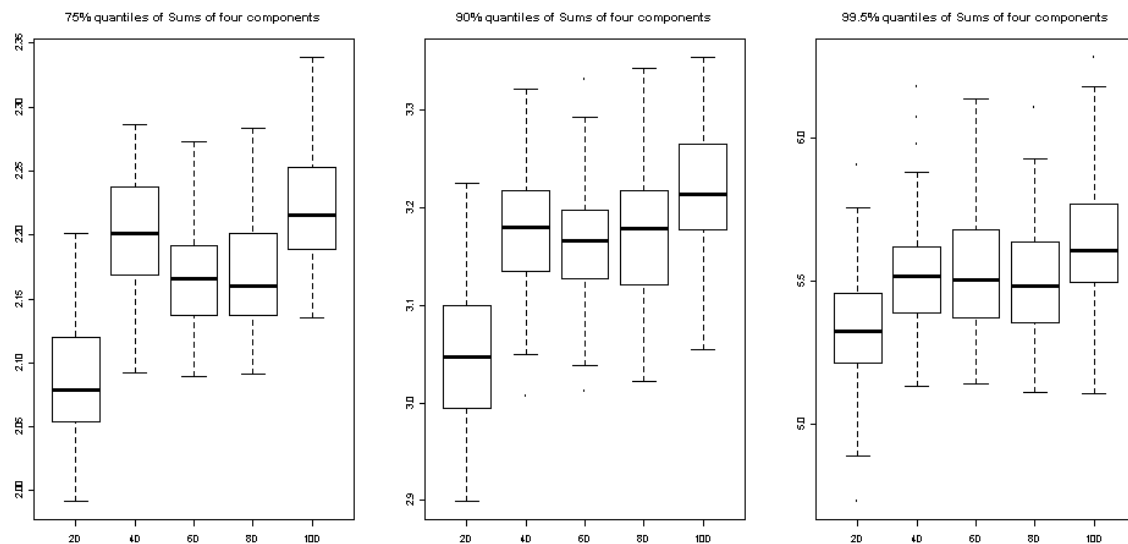


Figure 5.2.10.: Comparison of sums of components for 4-dimensional marginals. Increase is appreciable for sums of components of non-Gaussian fields already at the 75% quantile.

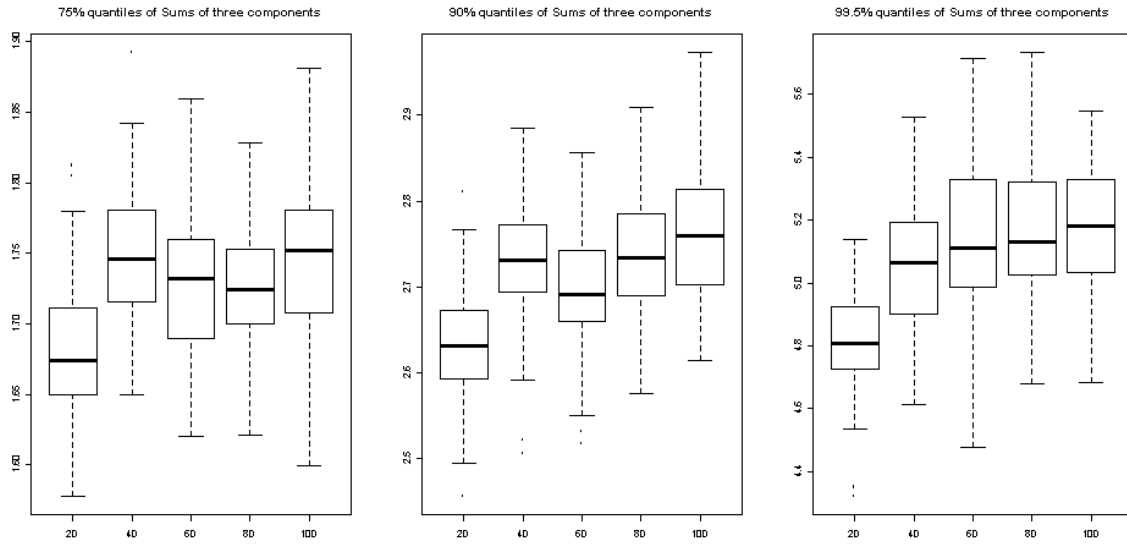


Figure 5.2.11.: Comparison of sums of components for 3-dimensional marginals. Increase is appreciable for sums of components of non-Gaussian fields at the uppermost quantile.

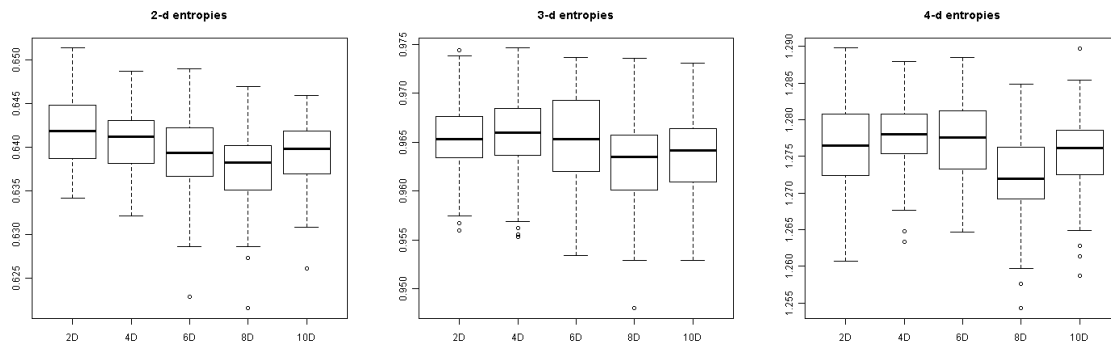


Figure 5.2.12.: Comparison of congregation measure for marginals of different dimensions and data from the different fields. From left to right results of the comparison procedure are given for marginals of dimension 2, 3 and 4. Congregation seems to be always greater for the non-Gaussian fields.

6. Inference in a quasi-real setting

Our aim is now to study the consequences of higher order interdependence in an extended period of time, and try to recover faithfully the characteristics (as expressed in parameters) of a random field we can observe only at a very limited number of sites. We use a total of $n = 3650$ realizations of the field. The reader might think of daily rainfall modeling as represented by the following example, where the sites correspond to gaging stations, and the study period spans ten years.

We examine the consequences of interdependence as expressed in the following three interactions manifestations:

1. The sum of positive values of the field. The mean of the field is set to zero throughout the simulation time.
2. The total number of components of the field (i.e. the number of locations or "pixels" in a map) above a given threshold.
3. The entropy measure used by Bárdossy and Pegram (2009), as explained in section 5.1.5, applied to triples of variables. Each triple comprises three sites forming a right-angled triangle.

We shall see the difference between a Normal field and a field with non-zero higher order (> 2) joint cumulants, regarding these measures. The parameters of the non-Normal random vector have been selected to make it look very similar to a Normal one, up to six order joint cumulants (interdependence parameters).

6.1. The simulated fields

The J -dimensional random fields simulated for this example, where $J = 300 \times 300 = 90000$, correspond to a J -dimensional vector with dependence structure

$$K_{\mathbf{X}}(\mathbf{t}) = c_1 \frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t} + \frac{1}{2!} c_2 \left[\frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t} \right]^2 + \frac{1}{3!} c_3 \left[\frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t} \right]^3 + \dots \quad (6.1.1)$$

where

$$\begin{aligned} c_1 &= 1 \\ c_2 &= 0 \\ c_3 &= 10 \\ c_4 &= 495 \end{aligned}$$

A total of $n = 3650$ fields were simulated as follows:

1. Simulate $n = 3650$ Gaussian random fields \mathbf{y}_i , $i = 1, \dots, n$, of size $J = 300 \times 300 = 90000$ using a fast algorithm. In this dissertation we used the method of circulant embedding (see Chan and Wood (1997)), as implemented in the statistical Software R. The fields are simulated with mean 0, and using an exponential covariogram as dependence model with nugget effect 0, variance 1 and range parameter 20.
2. Simulate $n = 3650$ realizations of R_*^2 , a random variable such that the moments of $R^2 = R_*^2 \times \chi_J^2$ are given by

$$E\left((R^2)^k\right) = J^k m_k E\left(\xi^{2k}\right) \iff E\left((R_*^2)^k\right) = \frac{J^k m_k E\left(\xi^{2k}\right)}{E\left((\chi_J^2)^k\right)}$$

where ξ is a standard Normal Random variable. We used for the specific example below representation:

$$R_*^2 = \left(\sum_{s=0}^5 a_s \xi^s \right)^2$$

with $\xi \sim N(0, 1)$ and coefficients given in table (6.1.1). The resulting moments correspond to parameters (m_1, m_2, m_3, m_4) that, on application of relation (4.1.13) result in the desired coefficients: $c_1 = 1$, $c_2 = 0$, $c_3 = 10$ and $c_4 = 495$.

3. Each of the $n = 3650$ fields $\mathbf{x}_i \in \mathbb{R}^J$ employed for this example are given by

$$\mathbf{x}_i = \sqrt{R_*^2} \times \mathbf{y}_i \quad (6.1.2)$$

In order to analyze to what extent the differences between the Gaussian and non-Gaussian fields are altered by the marginal distributions of \mathbf{X} , which are slightly non-Normal (they are the same up to the 5th cumulant, being almost indistinguishable for most Normality tests), we also consider the random fields obtained by applying a Quantile-Quantile transformation to each of the 90000 components of \mathbf{X} . Hence, each component of field $\mathbf{X} \in \mathbb{R}^J$ is exactly normally distributed with mean 0 and variance 1. As we shall see, most of the consequences on the interaction manifestations introduced by the scaling variable R_*^2 are not altered by this Q-Q transformation.

6.2. Analysis

In Figure (6.2.1) the simulation mechanism, together with an approximation to the density of $\sqrt{R_*^2}$, are shown. One can see that the distribution is highly concentrated around 1. However, values such as $\sqrt{R_*^2} = 6$ or $\sqrt{R_*^2} = 7$ are also possible, though with relatively low

a_0	a_1	a_2	a_3	a_4	a_5
0.918076529	0.023806437	0.004666416	0.059499895	0.010020358	0.001326858

Table 6.1.1.: Coefficients producing R_*^2 for the fields example. Coefficients were fitted by the method of moments.

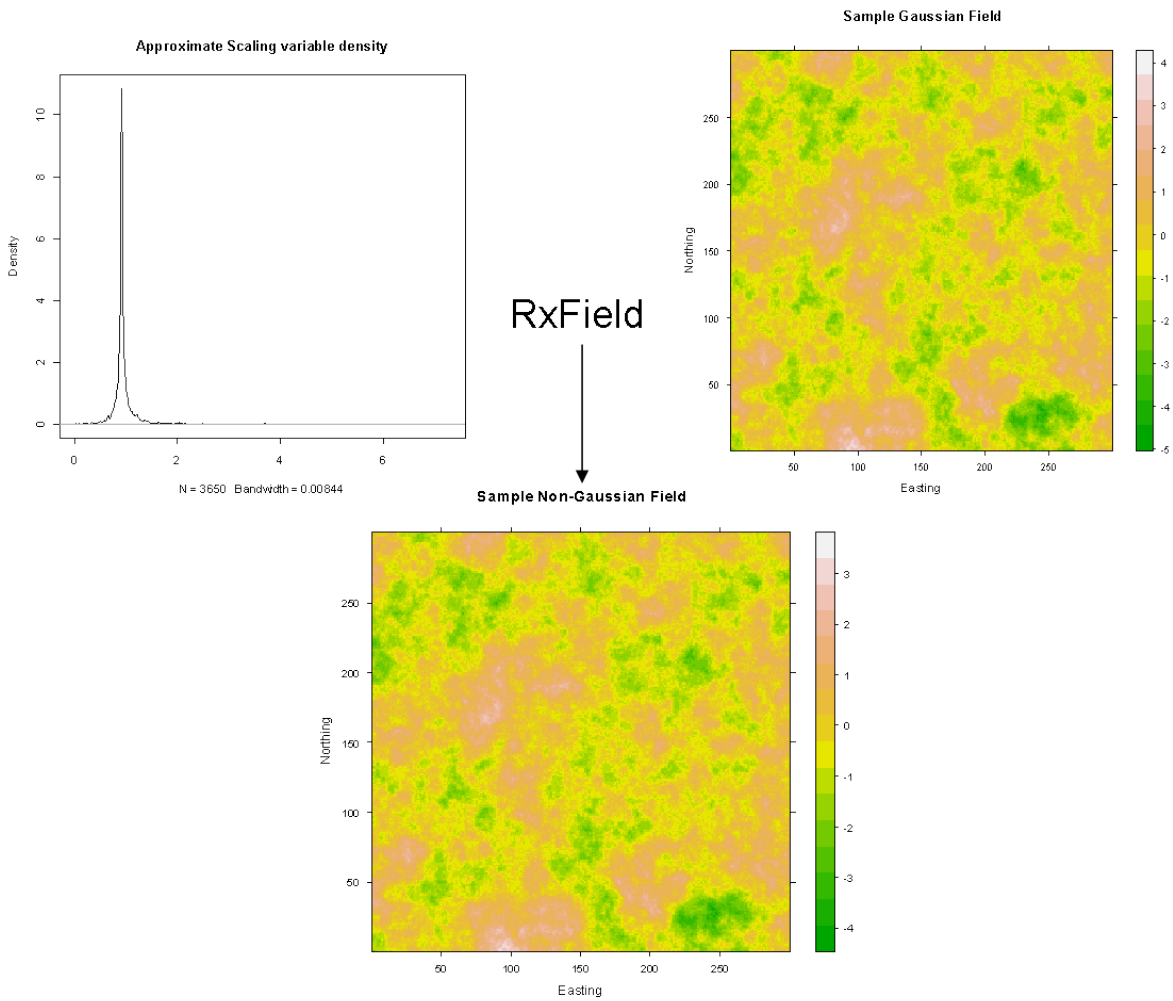


Figure 6.2.1.: Simulation Scheme and kernel smoothing approximation to the density of $\sqrt{R_*^2}$

probability. As the size of the field increases, the consequences for an interaction manifestation such as the sum of components become huge.

Each field coming from the non-Gaussian mechanism is actually a Gaussian field that has been scaled, but the characteristics of the fields in the long run are quite different. This is most evidently observed in the QQ-transformed fields, where the QQ transformation for each component ("pixel" of the map) is made on the basis of all the observations for that pixel, along the 3650 realizations. As an example, see figure (6.2.2). We see a considerable clustering of high values.

Hence it is possible, if necessary with the aid of QQ transforms, to model mechanisms leading to such clusterings. Actually, the scaling variable represented in figure (6.2.1) might be identifiable, in the context of rainfall modeling, with some large scale atmospheric process. In figure (6.2.3) we can see box-plots outlining the distributions of the sums of positive values for the fields analyzed. The effect of the scaling variable (that is, of the high order joint cumulants) is clearly manifested on the non-Gaussian and QQ-Transformed fields. This effect is non attributable to scaling on the 1-D marginal distributions, as indicated by the distribution of the sums for the QQ-transformed fields; the dependence structures of the random fields are different.

If the random fields were to represent daily rainfall for a period of 10 years, a model based on the Gaussian dependence structure would clearly prescribe a smaller total rainfall over large areas. An increase in the total rainfall such as the one observed in the non-Gaussian field might, for example, be due to a large scale atmospheric process. This process would then be modeled with the aid of $\sqrt{R_*^2}$.

Even more impressive is the effect on the probability of simultaneously trespassing a given threshold, in the course of $n = 3650$ realizations. In the remaining plots of figure (6.2.3), we present box-plots giving an idea of the distribution of the number of components with values above thresholds 1.04, 1.28 and 2.5, in the period given. These thresholds correspond to the 85%, 90% and 99.38% quantiles of a Standard Normal distribution, which is approximately the distribution of the Gaussian field here simulated. Note that the divergence between the responses of the Gaussian and non-Gaussian fields grows more and more as one moves towards the uppermost quantile of the marginal distribution.

Again, in the context of rainfall modeling, this plot gives an idea of the total area over which one might expect intense rainfall, both under the Gaussian dependence and under the non-Gaussian dependence assumption. A single event with 40000 locations (44% of the total area, in this example) receiving extreme rainfall above the 99.38% historical quantile, might cause huge, totally unexpected losses in the course of only ten years. The left plot at figure (6.2.2) provides an idea of the kind of field producing such an extreme response.

Note that the response, regarding this interaction manifestation, of the QQ-transformed random vector, which has Standard Normal marginals, is very similar to that of the untransformed, non-Gaussian field.

One can infer that by manipulation of the 6th and 8th joint cumulants of a random vector, as introduced via the scaling variable R_*^2 , one can model important features in its dependence structure. This was to be expected, in view of the conceptualization of joint cumulants as extensions of correlation in section 3.2. Concerning coefficients c_2, c_3, c_4, \dots of the dependence structure (6.1.1), one can set more coefficients to zero, not just c_2 as in this example, making the distribution look more and more like a Gaussian multivariate distribution with respect to

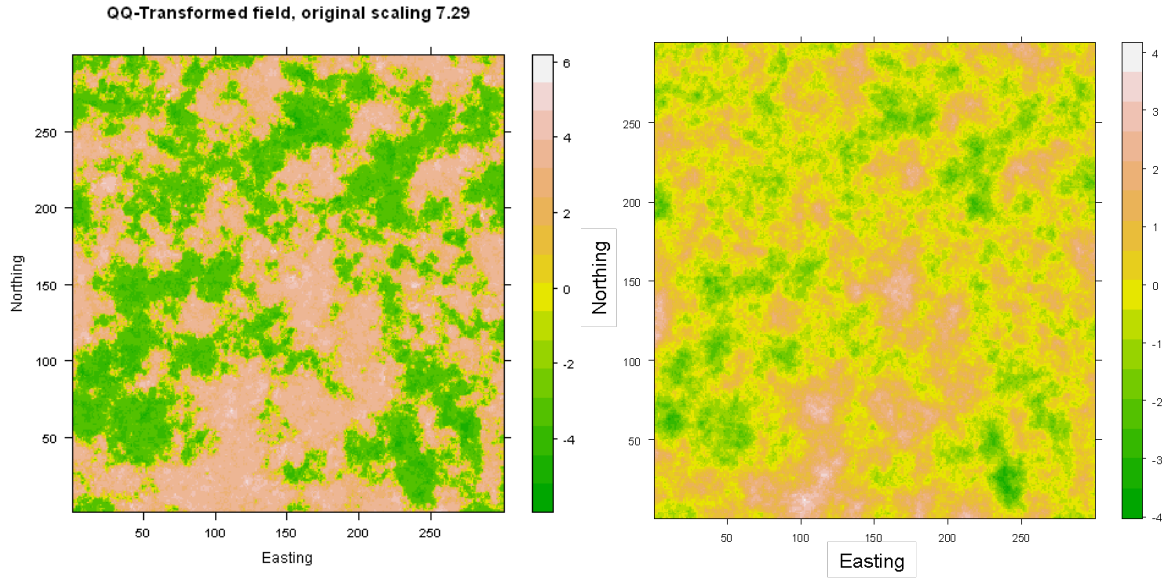


Figure 6.2.2.: Original Gaussian and Scaled sample Fields, after QQ transformation

its low dimensional marginal distributions. However, as the dimension of the random vector under analysis increases, the closeness to normality of these low dimensional marginals tells less and less about the interdependence among big sets of components.

Already by letting the fourth joint cumulants be zero (i.e. setting $c_2 = 0$) the resulting variable is close to Normality in the two dimensional marginals. In Figure (6.2.4), the empirical copulas formed of data from the first three locations of the random fields are presented. On the basis of this similarity, it is not easy foresee great dissimilarity for higher dimensional marginals distributions.

And still, even in terms of triplets of components, the distributions present a different behavior, particularly in their upper quantiles. To see this, we apply the congregation or entropy measure due to Bárdossy and Pegram (2009) to data from triplets of locations, randomly selected from the fields locations. Specifically, we selected right-angled triangles with two equal catheti. Several lengths for these catheti were selected and the entropy measure applied to all data ($n = 3650$) of the locations on which the triangles' vertices fell. For each distance, we selected randomly fifty triplets. Results are expressed in the form of box-plots of the entropies obtained, and presented in figure (6.2.5). Since a lower value of this measure indicates more association, we notice that the association is systematically higher for the non-Gaussian random variable. This result is similar to what Bárdossy and Pegram (2009) observed in their analysis of rainfall data.

Taking as illustration data from locations 1, 3 and 5, we see in figure (6.2.6) that the increase in entropy (disaggregation) of the data obtained from the Gaussian field increases considerably at the uppermost corner of the distribution.

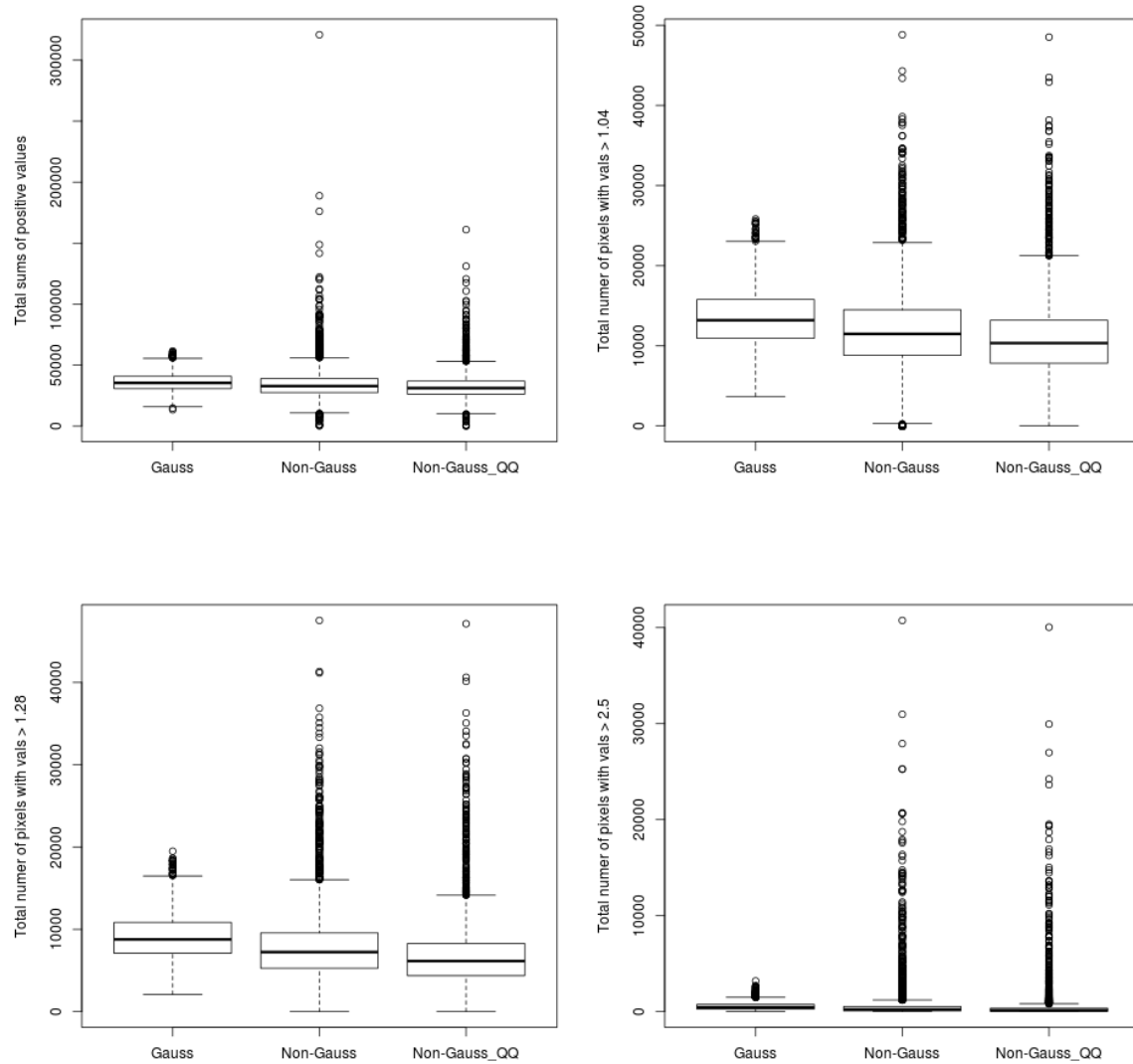


Figure 6.2.3.: From left to right and downwards: Box-plots of the sum of positive values and of number of components above 1.04, 1.28 and 2.5 on each field's realization. Regarding the number of components (locations) above the given thresholds, divergence between the Gaussian and non-Gaussian fields become more and more apparent as one moves towards the uppermost part of the marginal distribution.

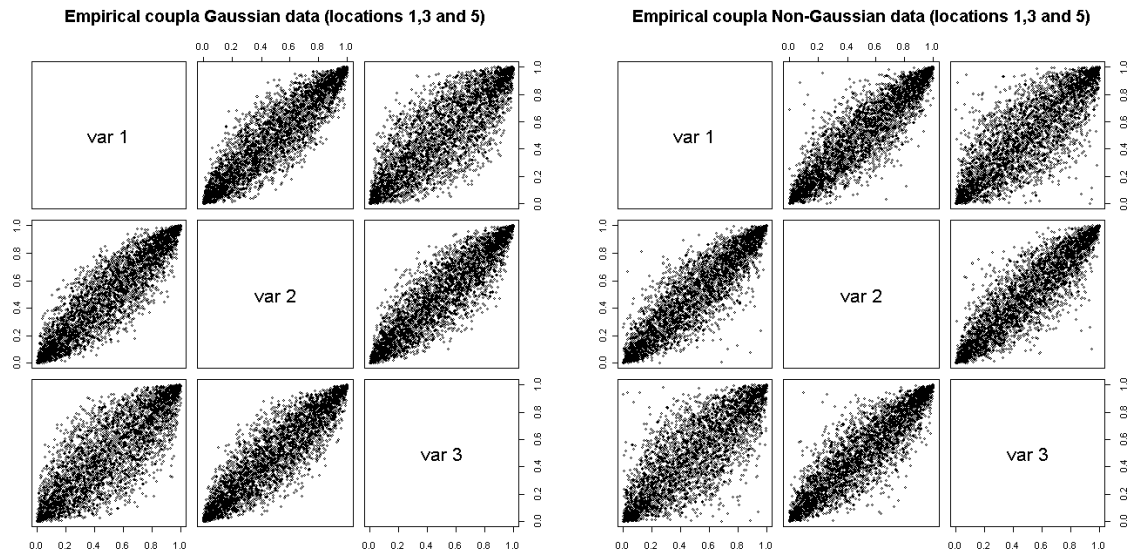


Figure 6.2.4.: Empirical copulas of data from locations 1,3 and 5 from the random fields simulated.

6.3. Estimating the parameters of the field

Now we intend to reproduce the situation often found in practice, where one has data from a set of gaging stations taking measurements of a random field (e.g. rainfall), and one attempts to infer characteristics of the whole field. We focus on the non-Gaussian field without any transformation. A good command of this basic model can ensure fitting with little difficulty copula models with given marginals, as well.

For this part of the example, we selected 30 locations on the plane and took the data of the fields corresponding to those locations as given data. Values under zero were truncated at zero and considered to be no-rain values. These will be considered upon model fitting as censored values, and imputed in the course of the "Stochastic EM algorithm", henceforth St-EM (see, for example Feodor Nielsen (2000); Gilks et al. (1998)). In appen, we explain briefly the estimation method used for this example.

The locations selected are illustrated in figure (6.3.1), where negative values have been removed.

To provide an idea of the data at hand, data from five of the sites are presented in figure (6.3.2).

Estimated parameters

One hundred and twenty (120) iterations of the St-EM algorithm were run, and the first sixty (60) discarded. The chains can be seen in figure 6.3.3 for the mean, and the covariance model parameters. An average of the last 60 iterations produces our estimators for these parameters, as presented and compared with the parameters used for the simulation in table:

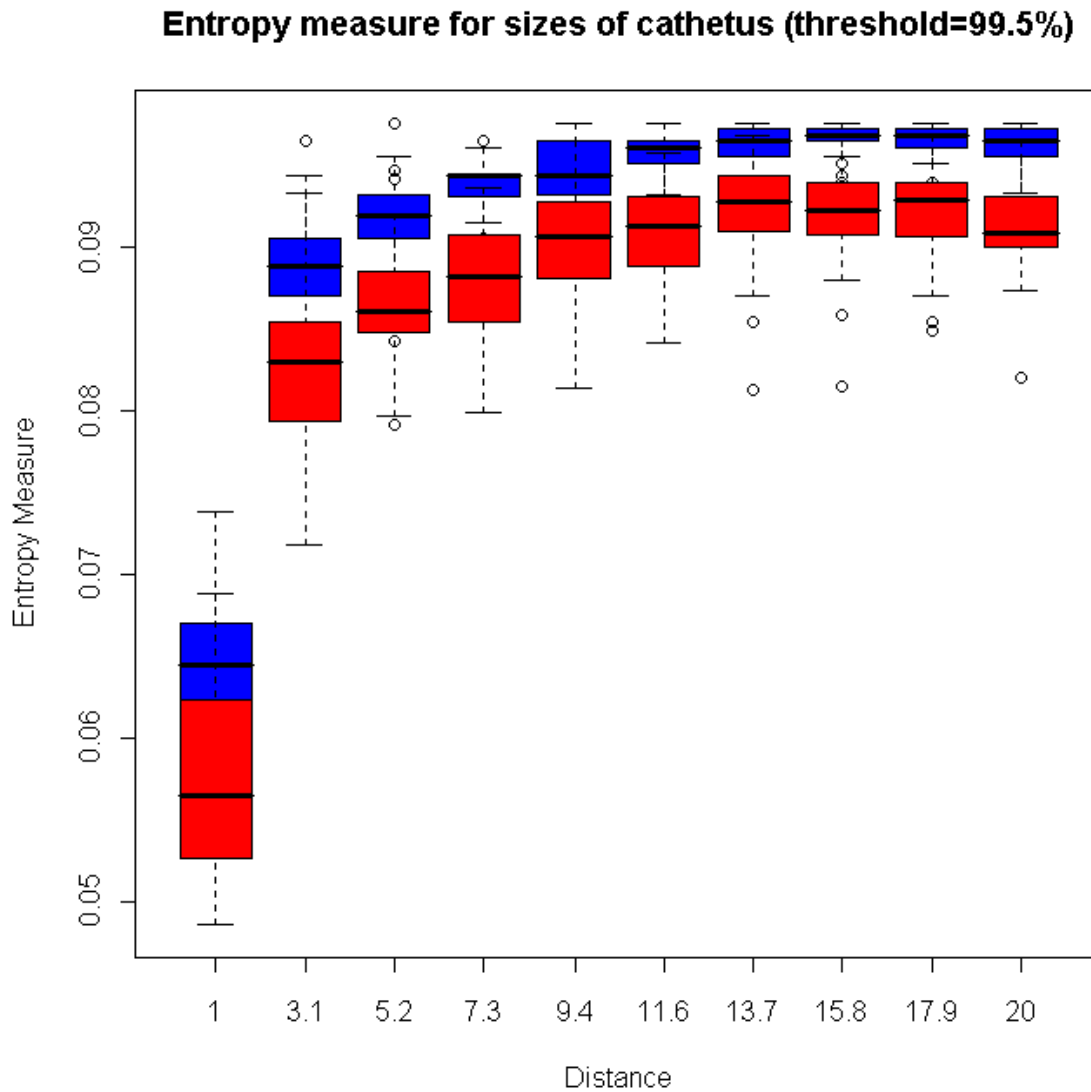


Figure 6.2.5.: Entropy congregation measure applied to data from randomly selected triplets of locations. Box-plots are organized in terms of the size of the catheti of the right-angled triangles constituting the triplets. Fifty triplets were selected per distance category and their entropies computed. *Blue boxes* represent the results for the Gaussian field data. Selected quantile threshold was 99.5%.

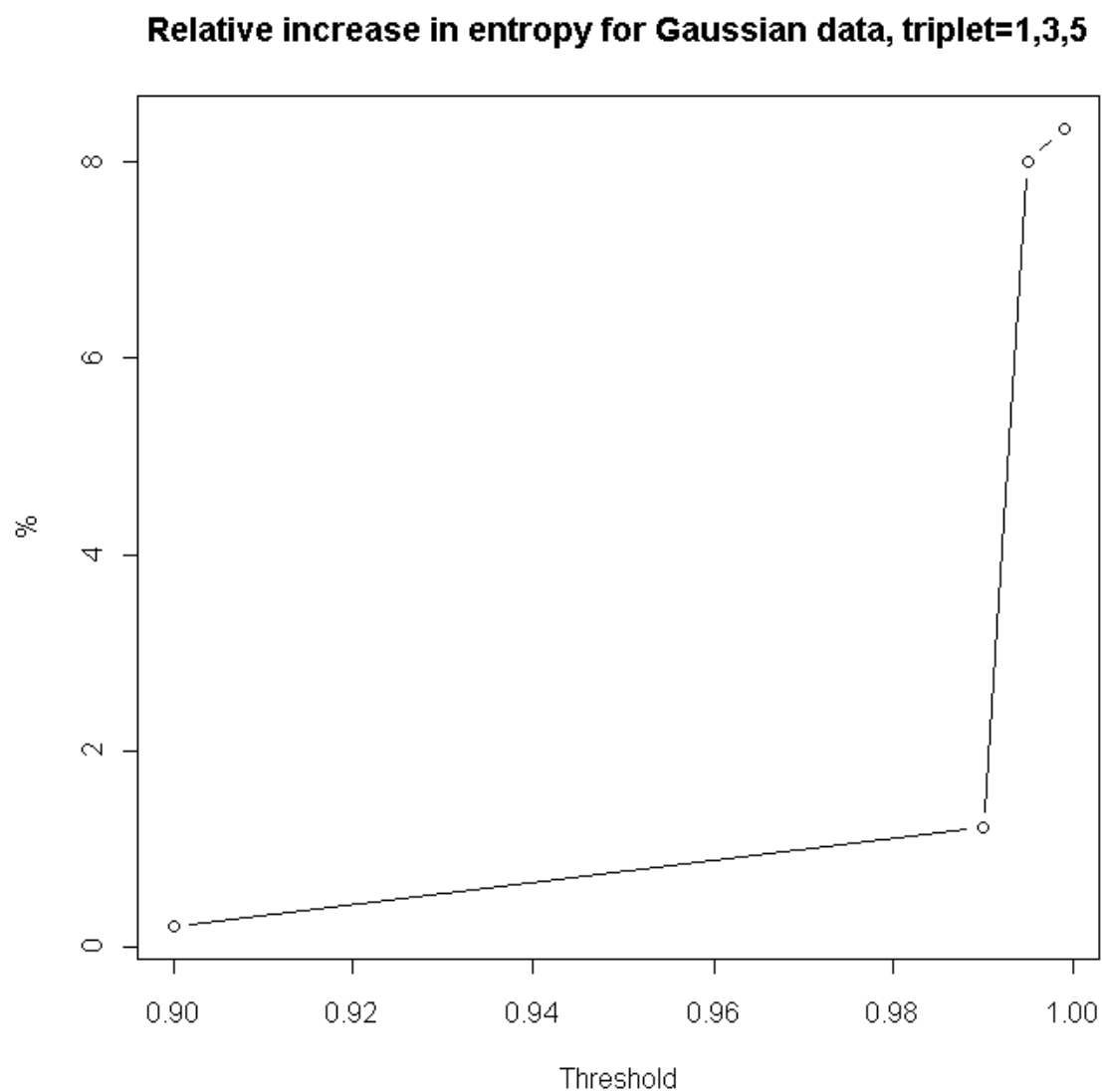


Figure 6.2.6.: Ratio of the entropy measure computed for data from Gaussian to the entropy computed from Non-Gaussian data, at thresholds 90%, 99%, 99.5% and 99.9%. Locations for the triplet are 1,3 and 5. The increase is considerable above the 99% quantile.

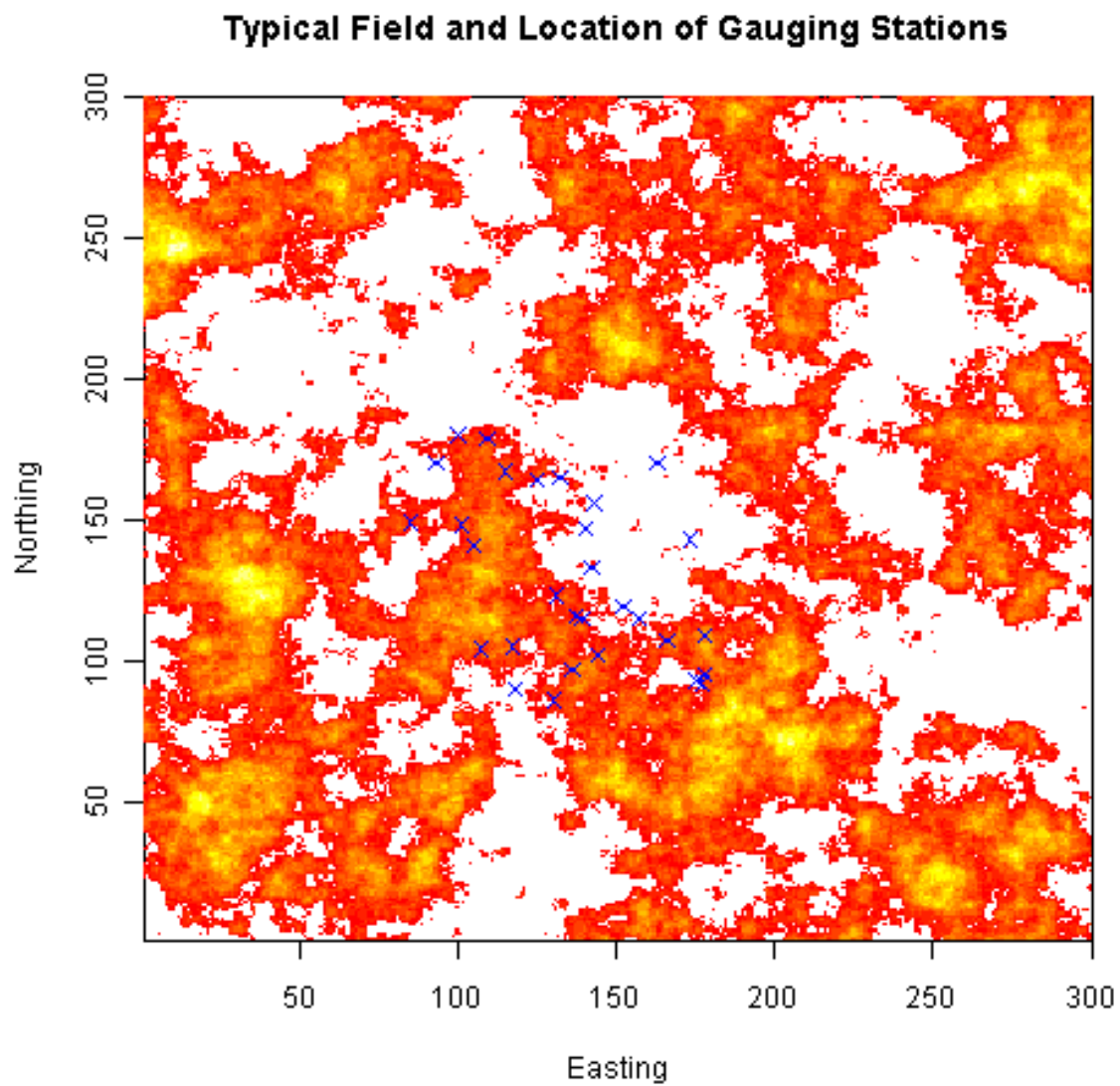


Figure 6.3.1.: Locations with "gaging stations" are marked with an x . Negative values were regarded as zero.

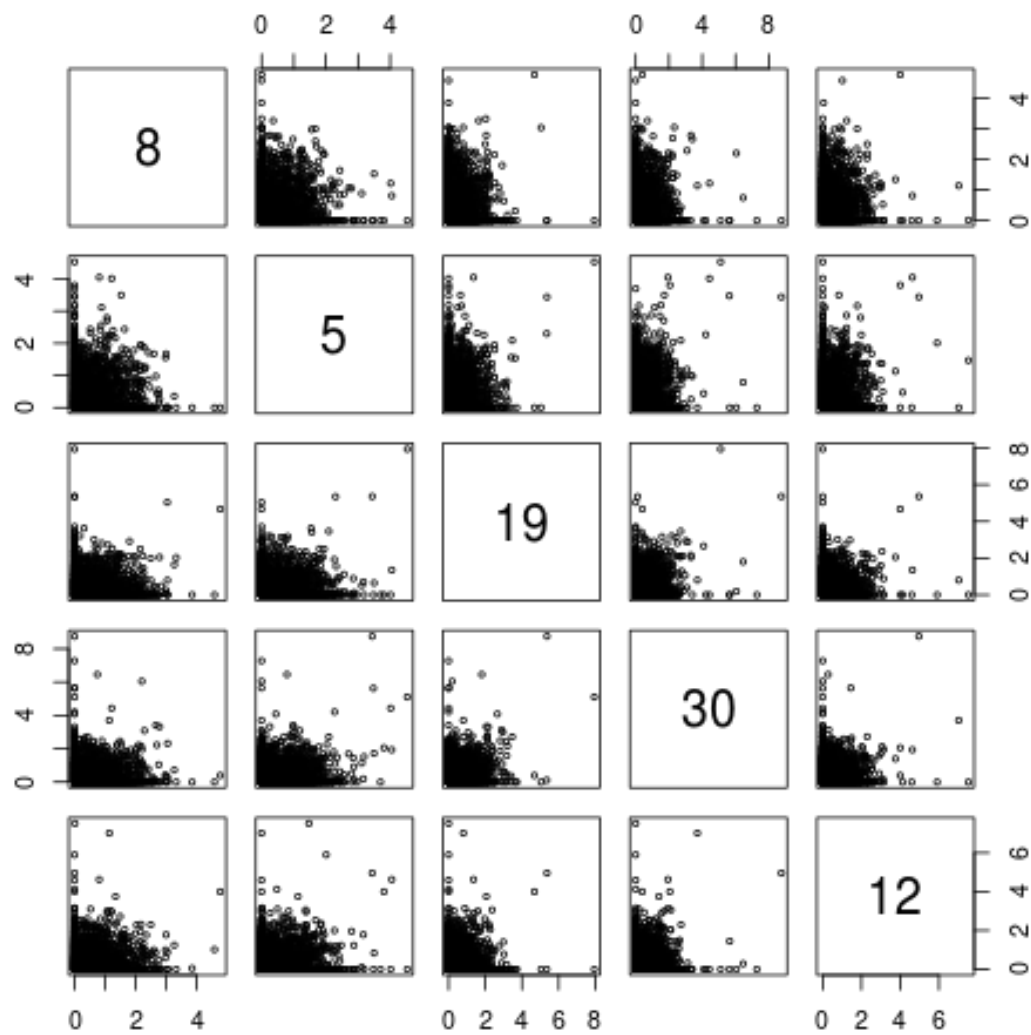


Figure 6.3.2.: Data at five of the sites from which the field's characteristics are to be estimated.

- Estimated mean $\hat{\mu} = 0.0036$, true parameter $\mu = 0$.
- Estimated Nugget effect $\hat{s}_0 = 0.00003$, true parameter $s_0 = 0$.
- Estimated Range parameter $\hat{R}g = 20.29$, true parameter $Rg = 20$.

We can say that the estimation is acceptable for these parameter.

Concerning the distribution of the generating variable R^2 , we used a mixture of $S = 100$ gamma distributions. A plot of the fitted density can be seen in figure 6.3.4. We notice that values much higher than those typical of a Chi-squared distribution with 30 degrees of freedom (which corresponds to a 30-D Gaussian distribution) appear.

The squared "deviance from normality" or scaling variable R_*^2 was also computed, using the techniques of section 4.6.2.1. A plot of its density and probability distribution function can be found in figure 6.3.5. The square root of this variable will be used for simulation of validation fields.

6.3.0.0.1. Parameters fitted to the Squared Generating Variable: The parameters estimated for the mixture of $S = 100$ Gamma distributions are given at table D.0.1 of the appendix. Via equation (4.2.12), it is possible to find the parameters m_k , $k = 1, 2, 3, \dots, K$, that connect the generating variable of the random field with the dependence structure. We computed values up to $K = 49$, but results are very similar already for $K = 9$. The first nine coefficients are given in table (6.3.1).

Using (4.2.5) one can also have estimates for the coefficients of the dependence structure. These are presented in table (6.3.1) as well. The interdependence coefficients, c_2, c_3, c_4 are clearly under-estimated. This is not surprising, since the approximate maximum likelihood estimation effected by the St-EM algorithm attempts to fit *the whole* distribution. Hence a second step is recommendable after estimating model parameters by maximum likelihood or a similar method.

At this second step, interaction manifestations such as those listed in section (3.1) or the ones used in this section are quantified for the available data and incorporated into the estimation procedure. This two step procedure was suggested in section (4.3).

6.4. Inference for the whole field

We generate now $n = 3650$ fields of the same dimension as in the previous section using the "dimension adaptation" technique, explained in section 4.6.2, for the scaling variable R_*^2 . We shall then be able to simulate big Non-Gaussian fields by simulating Gaussian fields via a fast method (e. g. turning bands), and then multiplying each of these fields times a realization from $\sqrt{R_*^2}$.

The dimension adaptation method consists in:

1. Find an estimate for the squared generating variable R^2 at dimension $J = 300 \times 300$, using relations given by (4.2.13) and estimates \hat{m}_k . Parameters estimates $\hat{\theta}$ and $(\hat{\pi}_1, \dots, \hat{\pi}_{100})$ are thus found, which fulfill the moments equations prescribed for a model with c.g.f as in (6.1.1). These estimated parameters define a (squared) generating variable suitable for the new dimension, $J = 300 \times 300$.

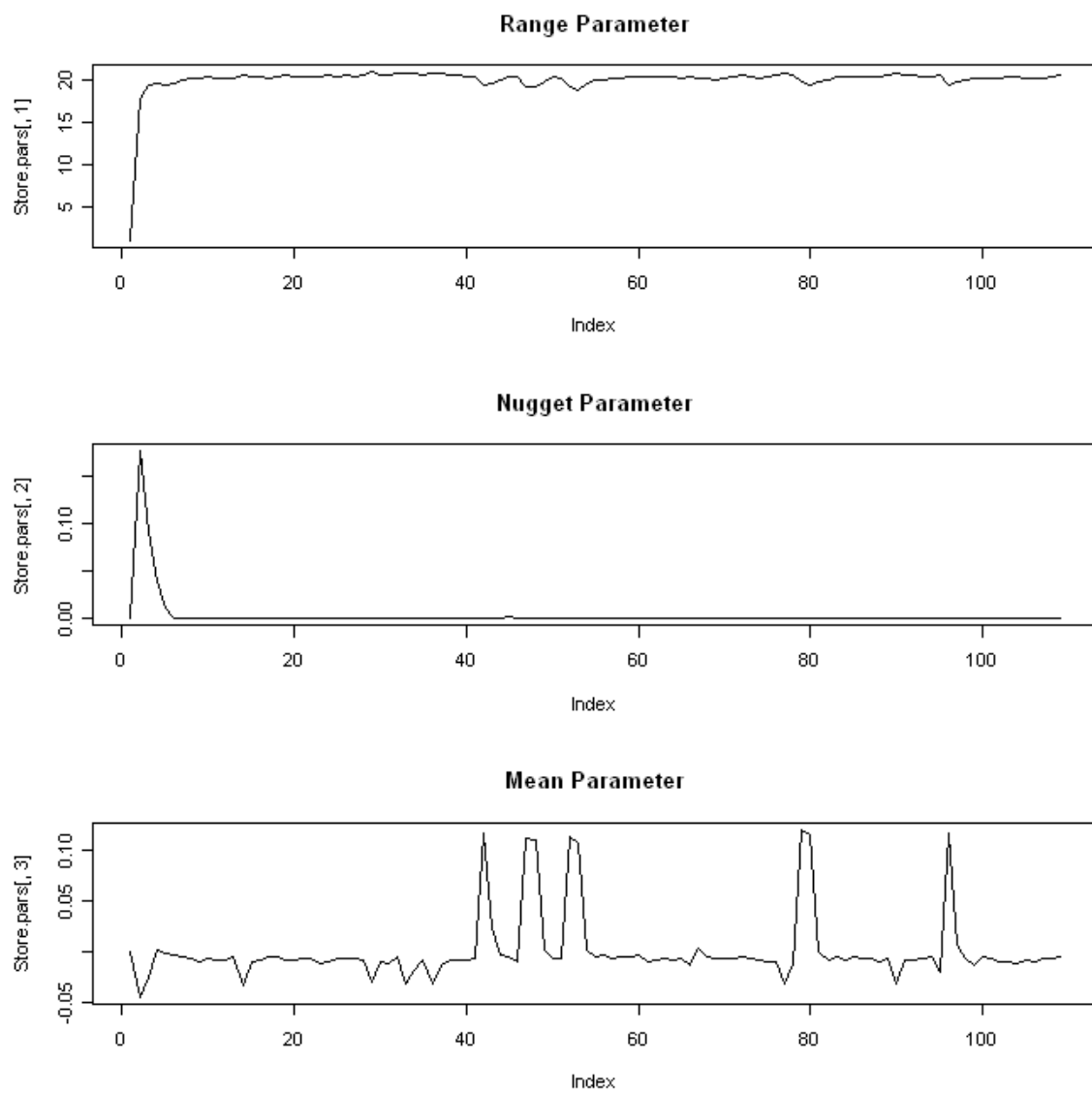


Figure 6.3.3.: St-EM produced chains for the mean, Nugget and range parameters.

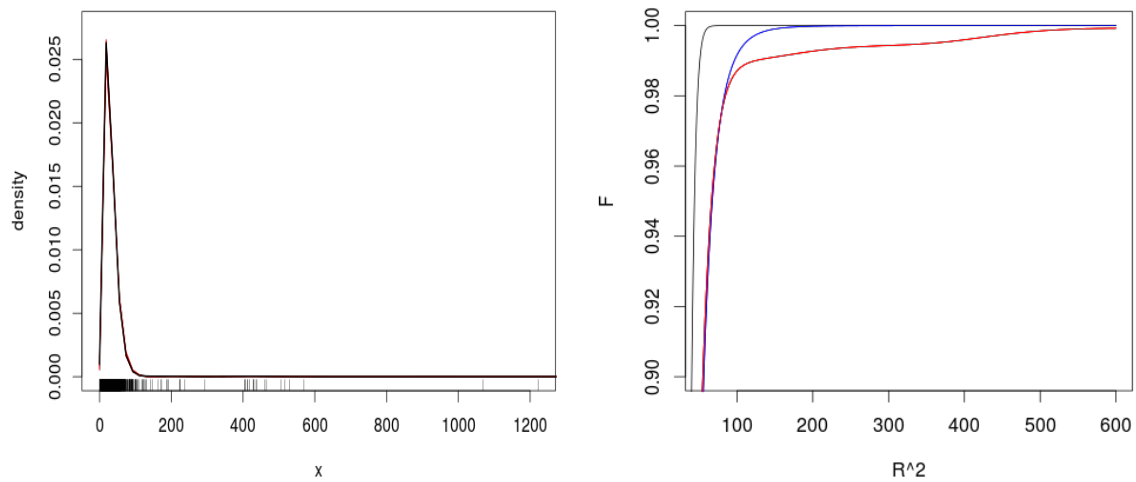


Figure 6.3.4.: Probability density (left) and distribution (right) functions of the squared generating variable, as estimated from the 30 Stations' dataset. Estimated distribution for R^2 appears in red as compared in its upper quantiles with the squared generating variables of a multivariate normal distribution (black) and that of a mult. Student r.v. with 15 degrees of freedom (blue).

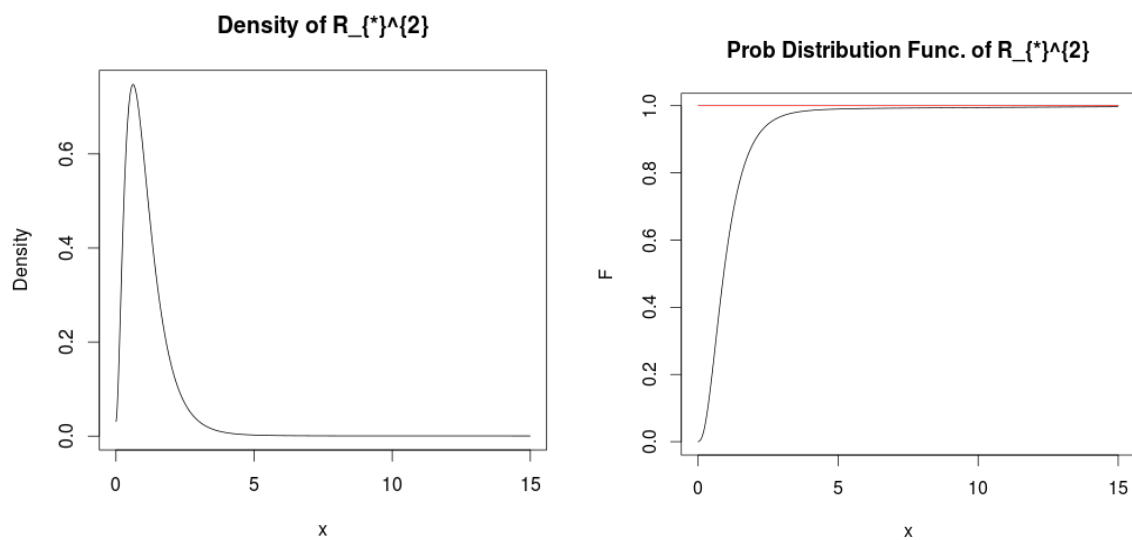


Figure 6.3.5.: Density (left) and probability distribution (right) of the squared deviance from normality random variable R_*^2

k	\hat{m}_k	\hat{c}_k	c_k
1	1.1068	1.1068	1
2	1.0775	-0.1475	0
3	2.7323	1.8662	10
4	9.1687	0.4248	495
5	28.8510	6.3693	
6	81.5158	-34.4707	
7	205.6870	-12.9141	
8	464.8952	-674.6965	
9	947.6988	3279.9002	

Table 6.3.1.: First nine m_k and c_k coefficients fitted on the basis of the 30 sites data. Original c_k interdependence coefficients of the c.g.f are also presented for comparison. These coefficients are clearly underestimated for $k \geq 2$, which calls for a second estimation step addressing specifically interdependence manifestations.

2. Apply (4.6.8) to derive scaling variable R_*^2 from $\hat{\theta}$ and $(\hat{\pi}_1, \dots, \hat{\pi}_{100})$ fitted on step 1.

One can simulate random fields having the desired dependence structure, by sampling a Gaussian field with covariance matrix prescribed by the estimated covariance function parameters, $\mathbf{Y} \in \mathbb{R}^J$, and setting $\mathbf{X} := \hat{\mu} + R_* \mathbf{Y}$.

Squared Generating variable, R^2 , for the new dimension

One ideally would solve, for step 1 above, the system of non-linear equations (4.2.13) to obtain unequivocal estimates $\hat{\theta}$ and $(\hat{\pi}_1, \dots, \hat{\pi}_{100})$. However, the huge numbers involved when dealing with moments of very high order make the system unstable when more than $K = 50$ equations are considered. Hence one must content oneself with solving the following least squares problem for some moment order $K < S$:

$$\min_{(\theta, \pi_1, \dots, \pi_{100})} \sum_{k=1}^K \left(\sum_{s=1}^S \pi_s \frac{\prod_{l=1}^k (s+l-1)}{\theta^k} - J^k m_k E(Z^{2k}) \right)^2$$

subject to:

$$\begin{aligned} \sum_{s=1}^S \pi_s \frac{s}{\theta} &= J \hat{m}_1 E(Z^2) \\ \sum_{s=1}^S \pi_s \frac{s(s+1)}{\theta^2} &= J^2 \hat{m}_2 E(Z^4) \\ &\vdots \\ \sum_{s=1}^K \pi_s \frac{\prod_{l=1}^K (s+l-1)}{\theta^K} &= J^K \hat{m}_K E(Z^{2K}) \\ \sum_{s=1}^S \pi_s &= 1 \end{aligned} \tag{6.4.1}$$

where $S = 100$ is the number of mixture components for the squared generating variable. We used for this example $K = 49$. Additionally, in order to avoid sub-optimal local minima, the algorithm implemented performed two steps per iteration i , namely:

1. Define some initial $\theta^{(0)}$. For example, $\theta^{(0)} = 1$.
2. Given $\theta^{(i)}$, solve the optimization problem in $(\pi_1, \dots, \pi_{100})$ as a quadratic problem with linear constraints. For this problem very efficient methods are available for its solution. Set $(\pi_1, \dots, \pi_{100})^{(i)}$ to the solution of this problem.
3. Given $(\pi_1, \dots, \pi_{100})^{(i)}$ minimize the objective function subject to $\theta > 0$. This is a one-dimensional problem, and efficient algorithms are available. Set $\theta^{(i+1)}$ to the solution of this problem.
4. Return to step 2 until some optimality criterion is reached, such as $|\theta^{(i)} - \theta^{(i+1)}| < tol$, for a given small positive constant tol .

Estimated parameters $\hat{\theta}$ and $(\hat{\pi}_1, \dots, \hat{\pi}_{100})$ are given at table (6.4.1).

A plot of the estimated distribution of the squared generating variable, R^2 , appears in red to the right of figure 6.4.1, where those of a Gaussian distribution, χ^2_{90000} , and a Student with 15 degrees of freedom, $90000 \times F_{90000,15}$, are also presented for comparison. We see that our estimated variable lies somewhere in between for the new dimension of the field.

Scaling variable, R^2_* , for the new dimension

The second step necessary for inference on the whole field, consists in estimating scaling variable R^2_* adequately for the target dimension of the field, $J = 300 \times 300$. This was done applying equation (4.6.8), using the parameters just estimated for R^2 at $J = 300 \times 300$. The estimated density for R^2_* and its estimated probability distribution can be seen in figure 6.4.2. Now it is possible to simulate big Non-Gaussian fields with characteristics inferred from the 30 stations' data.

$\hat{\theta}, \hat{\pi}_{87} - \hat{\pi}_{93}$	$\hat{\pi}_{94} - \hat{\pi}_{100}$
$\hat{\theta} = 0.0009605027$	0.076672074
0.003262554	0.087159170
0.013749613	0.097646270
0.024236676	0.108133376
0.034723745	0.118620488
0.045210820	0.129107604
0.055697899	0.139594726
0.066184984	

Table 6.4.1.: Estimated parameters for the squared generating variable, R^2 , of the 90000-dimensional non-Gaussian field. Estimated weights $\hat{\pi}_1$ through $\hat{\pi}_{86}$ are 0.00.

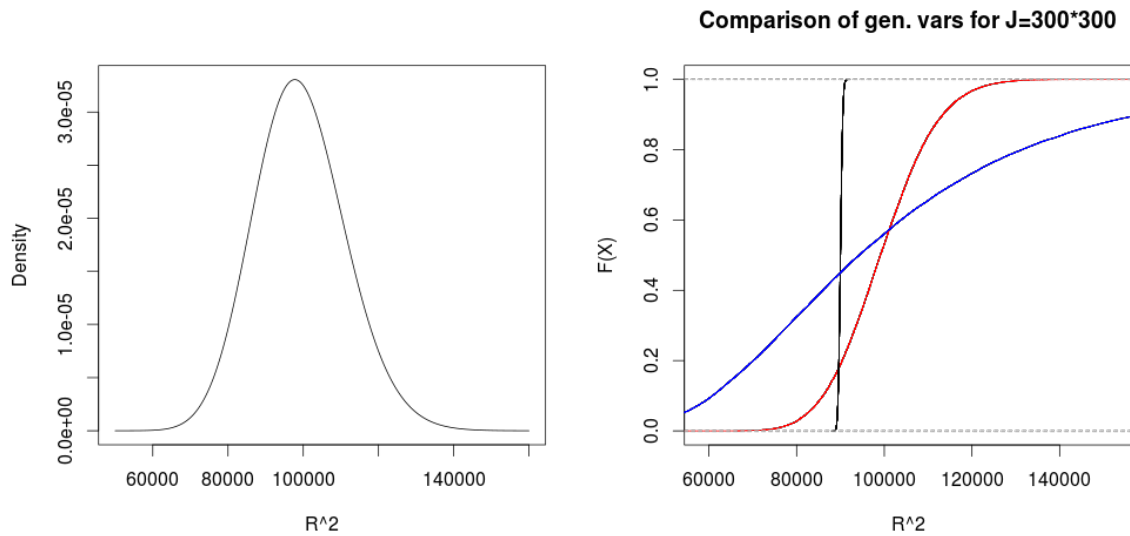


Figure 6.4.1.: Left: Probability density function of the squared generating variable estimated for dimension $J = 90000$. Right: squared generating variables for a Gaussian (black), a Student-t with 15 degrees of freedom (blue), and the fitted generating variable (red). These variables are adapted to dimension $J = 90000$ of the random field.

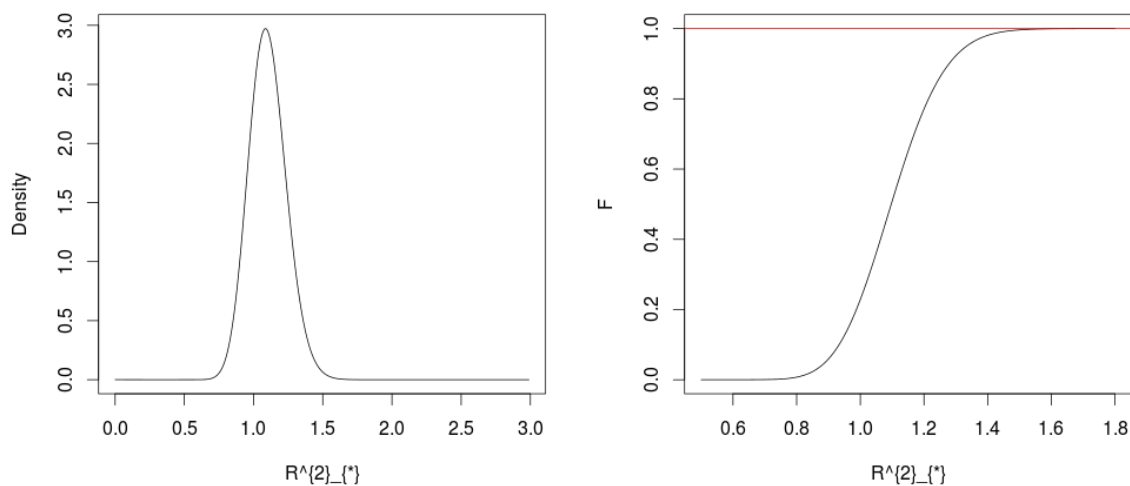


Figure 6.4.2.: Estimated density (left) and probability distribution (right) of the squared scaling variable for dimension $J = 90000$

6.4.1. (Partial) Inferential results

Parameter estimation was performed, as formerly explained, by applying an approximation to the Maximum likelihood method. Interaction manifestations such as those illustrated in figures 6.2.5 or 6.2.3 were not considered explicitly for estimation, which in this work is recommended as a second estimation step. Still, the estimated generating variable has captured some of the non-Gaussian interactions present in the dataset. In this section we shall see the implications for the complete field ($J = 300 \times 300$).

We first analyze the distribution of the sum of positive components, and the number of locations having values greater than thresholds $a \in \{1.04, 1.28, 2.5\}$.

In figure (6.4.3), from left to right and downwards, observed values of the sum of positive components, together with observed values of the number of components above the given thresholds, are presented. This figure corresponds to figure 6.2.3. One can see that the interaction manifestations studied in this example are only partially recovered from the simulated fields.

The components' sums of the non-Gaussian fields realizations have been pulled up as compared to those of the Gaussian one, with an increase of 12% and 10% for the maximum values of the non-Gaussian and QQ-transformed fields, respectively.

The observed numbers of components above the given thresholds of the non-Gaussian and QQ-transformed fields also differ sensibly from those of the Gaussian field. For thresholds 1.04, 1.28 and 2.5, the maximum values observed are 16%, 38% and 63% higher for data of the non-Gaussian field, and 13%, 19% and 30% for data of the QQ-transformed field, respectively. Of course, this difference is not even remotely as big as the ones inferred from figure 6.2.3.

Concerning the entropy criterion, the newly simulated non-Gaussian field recovers little of the behavior present on the original fields. In figure 6.4.4, one observes box-plots of the obtained entropy criterion applied to the same locations used in figure 6.2.5. Blue box-plots correspond to data from the Gaussian field. The congregation measure for each distance class is virtually the same for the Gaussian and non-Gaussian cases. This supports once again the necessity for the consideration during estimation of this type of interdependence manifestation.

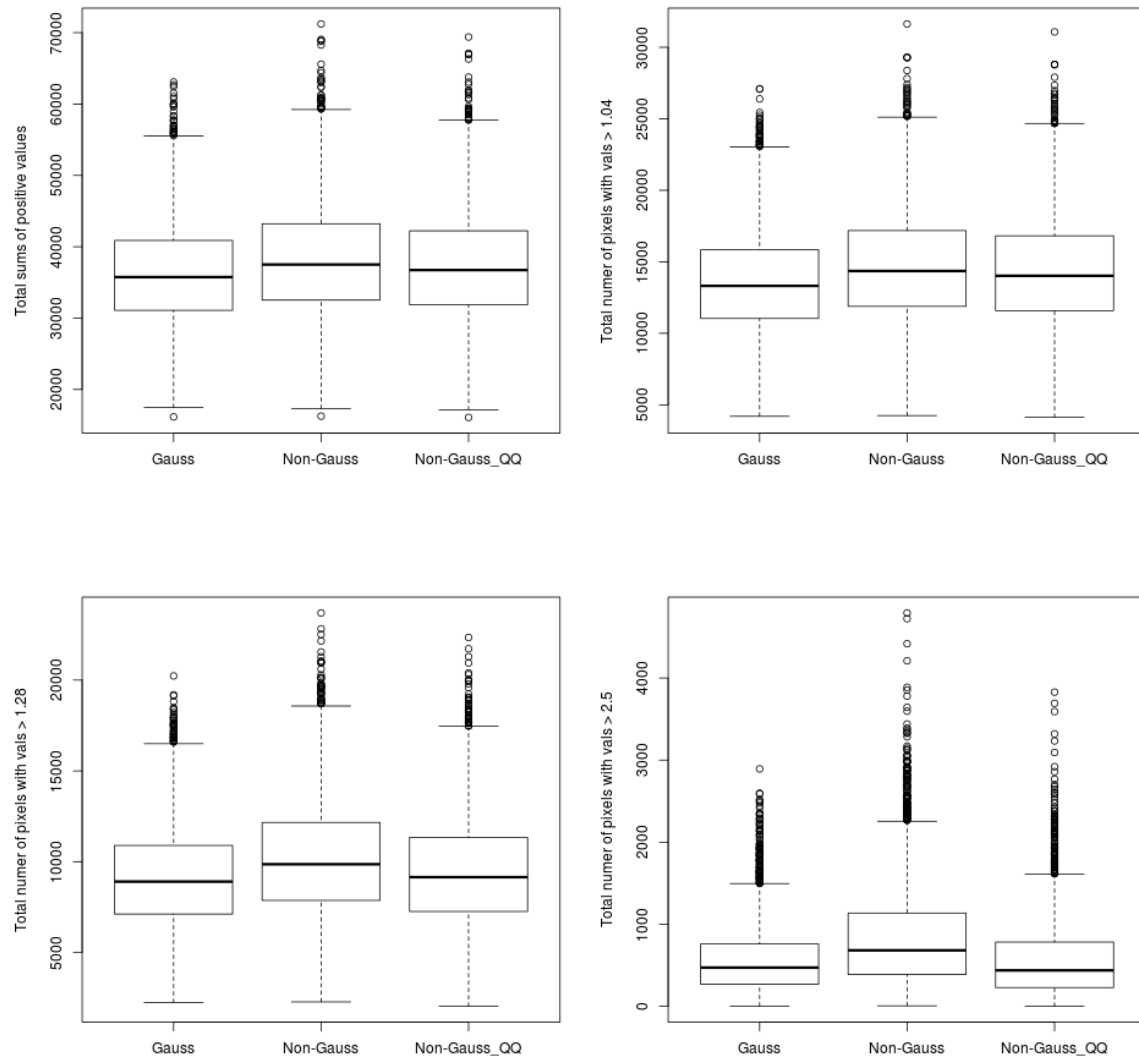


Figure 6.4.3.: From left to right and downwards: Box-plots of the sum of positive values and of number of components above 1.04, 1.28 and 2.5 on each field's realization. Regarding the number of components (locations) above the given thresholds, divergence between the Gaussian and non-Gaussian fields become more and more apparent as one moves towards the uppermost part of the marginal distribution. This figure corresponds to figure (6.2.3) of the original fields.

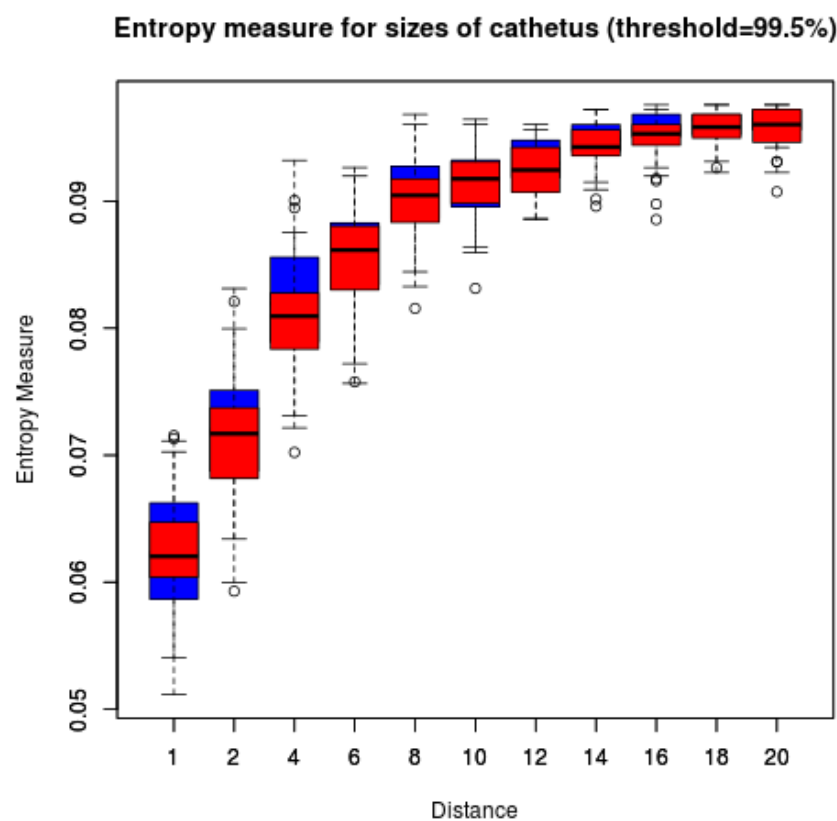


Figure 6.4.4.: Entropy congregation measure applied to data from the same selected triplets of locations as in figure 6.2.5. Box-plots are organized in terms of the size of the catheti of the right-angled triangles constituting the triplets. *Blue boxes* represent the results for the Gaussian field data. Selected quantile threshold was 99.5%.

7. Summary and outlook

Summary

What Bárdossy and Pegram (2009, 2012) found in the course of their research, and the illustrative examples presented at this dissertation, both point out that there is a need to quantify and model explicitly interactions of more than two variables at a time. This is what we call higher order interactions in this research work.

Consequences of higher order interactions can be huge as dimension of the field increases, as exemplified in the example at chapter 6. Neglecting higher order interactions in modeling can lead to important underestimation of subject-matter relevant features of the mechanism generating the data at hand. This would have in turn implications for the predictive capacity of the statistical model fitted, one of its most important characteristics Dawid (1984).

We have taken an initial step in the direction of addressing the issue of higher order interdependence in the context of Spatial Statistics, where low dimensionality of the model, and ease to extend the model are important requirements.

Joint cumulants were shown to be sensible building blocks for models that allow explicitly for higher order interdependencies. Joint cumulants have a reasonable interpretation as measures of interdependence, are natural extensions of the covariance coefficient, and one can impose conditions on them, so as to reduce the number of parameters to estimate in the model.

An example of a convenient interdependence structure, in the form of a cumulant generating function, was presented in this work. This model was seen to be a reasonable extension to the Gaussian model, as its cumulant generating function is a particular case of our model.

However, we suggest that interaction quantification must proceed on an application-specific basis. The specific aspect of interdependence (interaction manifestation) that is relevant to the problem at hand must be the departure point for interaction quantification. A model can then be built, on the basis of joint cumulants, such that data simulated from such a model produces similar interaction manifestations as the observed data. For example, a similar joint (empirical) distribution of low dimensional marginals above a given quantile, or a similar distribution for the sums of the vector's components.

We saw how to connect a number of such interaction manifestations to joint cumulants, our building blocks. Inference can in principle proceed in a method of "moments" fashion. We could produce interesting interaction manifestations at the examples in chapters 5 and 6, by manipulating the joint cumulants of order greater than two.

We presented a method for performing approximate Maximum Likelihood Estimation for the model proposed. However, we are not yet able to recover the joint cumulants in such a way, that interactions manifestations are properly reproduced (see final part of chapter 6). A second estimation step is required in which, leaving fixed joint cumulants of order one and two (mean and covariance), one fits higher order cumulants so that observed interaction

manifestations are better matched by those of the fitted model.

Outlook

There is a lot of future work to perfect and extend the methodology here presented.

One immediate step is to implement the second estimation step referred to above, whereby the interesting interaction manifestations may be explicitly taken into account on parameter estimation. This is very likely to enhance the power of the methodology, and make it capable of faithfully reproducing the interesting interaction manifestations observed in data. Predictive capacity can then be increased, but also new insight into the data generating mechanism can be aided, if one can identify model parameters as mostly responsible for specific interaction manifestations. Eventually such parameters or sets of parameters may receive a physical interpretation.

In order to increase the flexibility of the methodology, one may work with the copula of the model, rather than with the model itself. The copula characteristics themselves may be altered by means of polynomial (non-monotonic) transformations on the marginal components. These polynomial transformations were partially studied in this dissertation. This opens up another course of future research, in order to make the methodology more applicable.

A future course of action is to apply extended versions of the methodology to the research areas in which the issue of higher order interdependence was made clear. Namely, Downscaling and Daily precipitation modeling. The question whether the method is able to reproduce conveniently or not the application-specific interaction manifestations, will be elucidated by applying it to real world problems. Our opinion is that the method has a lot of potential to recover the relevant interaction characteristics and hence to provide better forecasts of the variables analyzed (e.g. rainfall). This opinion is based on the illustrations presented in this dissertation.

Additional research areas, in which considering higher order interactions can result in substantial model improvement are: Time Series Analysis, where diagnostics of a good model could go beyond verifying lack of temporal autocorrelation of the residual process (linear modeling) or some function of it (e.g. ARCH modeling). Empirical finance and economic analysis, where our archetypal model can enlarge the spectrum of copulas at hand for joint variable modeling (cf. Patton (2012)), and the consequences of dependence beyond correlation can be better isolated.

Finally, the methodology provides the possibility of connecting joint cumulants of order ≥ 2 straightforwardly to summary interdependence statistics, such as addition of the random vector's components. Hence we conjecture that our methodology has a role to play in the ongoing search for statistics that can better explain (re-)insurance losses. Recent research due to Kousky and R.M. (2011) indicates that micro-correlations combined with tail dependence can produce (in terms of current models) unexpected huge losses, provided one aggregates sufficiently many loss-cases. This stands in direct connection with what we observed for the example in chapter 6. More research in this connection is desirable.

A. Joint cumulants derivation

Our object of study is the cumulant generating function of a random variable $\mathbf{X} \in \mathbb{R}^J$. We shall be interested in joint cumulants such as

$$\text{cum}(X_{j_1}, \dots, X_{j_r}) \quad (\text{A.0.1})$$

where some, or all, of the indexes can be repeated. Hence it is convenient to refer to a random vector $\mathbf{X}^* \in \mathbb{R}^{J^*}$ having the components of \mathbf{X} , even repeated, and then find the joint cumulants that appear with degree at most one, of this “new” random vector. Thus we can, without loss of generality, focus on finding the joint cumulants with degree not greater than one, given by

$$\frac{\partial^r}{\partial t_{j_r} \dots \partial t_{j_1}} K_{\mathbf{X}^*}(\mathbf{t}) \big|_{\mathbf{t}=\mathbf{0}} := \text{cum}(X_{j_1}, \dots, X_{j_r}) \quad (\text{A.0.2})$$

where no t_j , for $j \in \{j_1, \dots, j_r\}$, is repeated.

For example, when computing the variance of a component, X_j , of \mathbf{X} , one would rather compute the covariance of vector $\mathbf{X}^* = (X_j, X_j)$, namely $\kappa_{11}(\mathbf{X}^*)$.

The archetypal dependence structure advocated for in this work is given by

$$K_{\mathbf{X}^*}(\mathbf{t}) = c_1 \frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t} + \frac{1}{2!} c_2 \left[\frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t} \right]^2 + \frac{1}{3!} c_3 \left[\frac{1}{2} \mathbf{t}^T \mathbf{\Gamma} \mathbf{t} \right]^3 + \dots \quad (\text{A.0.3})$$

for some coefficients c_1, c_2, c_3, \dots and covariance matrix $\mathbf{\Gamma}_{J^* \times J^*}$, and $\mathbf{t} \in \mathbb{R}^{J^*}$.

By expansion, the above expression can be written as

$$K_{\mathbf{X}^*}(\mathbf{t}) = \frac{c_1}{1!} \frac{1}{2} \sum_{j_1, j_2=1}^J t_{j_1} t_{j_2} \Gamma_{j_1 j_2} + \frac{c_2}{2!} \frac{1}{2^2} \sum_{j_1, \dots, j_4=1}^J t_{j_1} \dots t_{j_4} \Gamma_{j_1 j_2} \Gamma_{j_3 j_4} + \frac{c_3}{3!} \frac{1}{2^3} \sum_{j_1, \dots, j_6=1}^J t_{j_1} \dots t_{j_6} \Gamma_{j_1 j_2} \Gamma_{j_3 j_4} \Gamma_{j_5 j_6} + \dots \quad (\text{A.0.4})$$

For each coefficient $c_{\frac{r}{2}}$, for r even, there appears a sum of the form

$$\frac{c_{\frac{r}{2}}}{\frac{r}{2}!} \frac{1}{2^{\frac{r}{2}}} \sum_{j_1=1}^J \dots \sum_{j_{2r}=1}^J t_{j_1} \dots t_{j_r} \Gamma_{j_1 j_2} \dots \Gamma_{j_{r-1} j_r} \quad (\text{A.0.5})$$

This is the only block-summand of (A.0.4) that does not vanish upon differentiation with respect to each variable and equation to zero, as in (A.0.2). Other blocks will vanish either upon differentiation with respect to a variable that does not appear in them, or upon equation to zero, since such blocks become a sum of zeroes. So, it suffices to focus on this block, to differentiate it and equate it with zero.

Let each member of the (A.0.5) be labeled

$$s_{j_1, \dots, j_r} = t_{j_1} \dots t_{j_r} \Gamma_{j_1 j_2} \dots \Gamma_{j_{r-1} j_r}$$

then, we have stated that,

$$\frac{\partial^r}{\partial t_{j_r} \dots \partial t_{j_1}} K_{\mathbf{X}^*}(\mathbf{t}) \Big|_{\mathbf{t}=\mathbf{0}} = \frac{c_{\frac{r}{2}}}{\frac{r}{2}!} \frac{1}{2^{\frac{r}{2}}} \sum_{j_1=1}^J \dots \sum_{j_{2r}=1}^J \frac{\partial^r}{\partial t_{j_r} \dots \partial t_{j_1}} s_{j_1, \dots, j_r} \quad (\text{A.0.6})$$

Partial differentiation of s_{j_1, \dots, j_r} is readily found to be

$$\frac{\partial^r}{\partial t_{j_r} \dots \partial t_{j_1}} s_{j_1, \dots, j_r} = \Gamma_{j_1 j_2} \dots \Gamma_{j_{r-1} j_r} \quad (\text{A.0.7})$$

Sub-indexes appearing in the factors, $\Gamma_{j_1 j_2}, \Gamma_{j_3 j_4}, \dots$ constitute a partition of size $\frac{r}{2}$ of the set $A = \{j_1, j_2, \dots, j_r\}$. That is, the union of the $\frac{r}{2}$ non-overlapping sets

$$\{j_1, j_2\}, \{j_3, j_4\}, \dots, \{j_{r-1}, j_r\}$$

formed with elements of set $A = \{j_1, j_2, \dots, j_r\}$, is equal to that set:

$$\{j_1, j_2\} \cup \{j_3, j_4\} \cup \dots \cup \{j_{r-1}, j_r\} = A$$

Since the sum at (A.0.6) runs over all indexes in A , the sum returning the joint cumulant in question comprises all partitions of size two of A . How many different partitions of size two can be obtained for A , by forming sets out of different combinations of indexes? In general, a set with n elements, n even, can be seen to have

$$1 \times 3 \times \dots \times (n-1)$$

such partitions.

We have shown that joint cumulants of the archetypal dependence structure are given by

$$cum(X_{j_1}, \dots, X_{j_r}) = \frac{c_{\frac{r}{2}}}{\frac{r}{2}!} \frac{1}{2^{\frac{r}{2}}} \sum_{j_1, \dots, j_r=1}^J \Gamma_{j_1 j_2} \dots \Gamma_{j_{r-1} j_r} \quad (\text{A.0.8})$$

B. Joint cumulants of transformed vectors

Let $\mathbf{X} \in \mathbb{R}^J$ have cumulant generating function as in (A.0.3). We are interested in joint cumulants and cumulant generating function of random vector $\mathbf{Y} \in \mathbb{R}^J$, obtained by

$$Y_j = T_j(X_j) \quad (\text{B.0.1})$$

for some function T_j , and $j = 1, \dots, J$.

Our strategy is to find the joint moments of $\mathbf{Y} \in \mathbb{R}^J$ so constructed, and then use moment to cumulants formula (3.2.18) to get the joint cumulants.

One dimensional case

The technique employed here is usually called the “delta method” Casella and Berger (2001); Hurt (1976); Oehlert (1992). For the sake of clarity, let us begin with $X \in \mathbb{R}$ and $Y = T(X)$. Having ideas so illustrated will help to better follow the notationally cumbersome multi-variate case.

Assume that $T(*)$ has a valid Taylor expansion around $a = \mu_X$

$$T(x) = T(a) + \frac{1}{1!}T'(a) \times (x - a) + \frac{1}{2!}T''(a) \times (x - a)^2 + \frac{1}{3!}T^{(3)}(a) \times (x - a)^3 + \dots \quad (\text{B.0.2})$$

Then, by using the linearity property of the expected value, one has,

$$\begin{aligned} E(T(X)) &= T(\mu_X) + \frac{1}{1!}T'(\mu_X) \times E(X - \mu_X) + \frac{1}{2!}T''(\mu_X) \times E[(X - \mu_X)^2] + \\ &\quad + \frac{1}{3!}T^{(3)}(\mu_X) \times E[(X - \mu_X)^3] + \dots = \\ &= T(\mu_X) + \frac{1}{2!}T''(\mu_X) \times \sigma_X^2 + \frac{1}{3!}T^{(3)}(\mu_X) \times \tilde{\mu}_3 + \dots \quad (\text{B.0.3}) \end{aligned}$$

where $\tilde{\mu}_3$ is the third moment of X around its mean. We consider, with no loss of generality, that $\mu_X = 0$. Then the general expression is

$$E(T(X)) = E\left(\sum_{r=0}^{\infty} \frac{T^{(r)}(0)}{r!} X^r\right) = \sum_{r=0}^{\infty} \frac{T^{(r)}(0)}{r!} \mu_r \quad (\text{B.0.4})$$

where $\mu_r = E(X^r)$.

Thus, the expected value of a function of a random variable can be approximated in terms of the moments around the mean of the original random variable. In the literature, the series is taken up to the second term.

In order to find the moment generating function of Y ,

$$\begin{aligned} M_Y(t) &= E(\exp(tY)) = E\left(1 + \frac{tY}{1!} + \frac{t^2 Y^2}{2!} + \frac{t^3 Y^3}{3!} + \dots\right) \\ &= 1 + \frac{t}{1!} E(Y) + \frac{t^2}{2!} E(Y^2) + \frac{t^3}{3!} E(Y^3) + \dots \quad (\text{B.0.5}) \end{aligned}$$

one can apply the delta method to each summand of (B.0.5), namely, to each factor $E(Y^s)$, $s = 1, 2, \dots$, up to a practically useful order r .

Then one has for each factor $E(Y^s)$,

$$\begin{aligned} E(Y^s) &= E\left(\left(\sum_{r_1=0}^{\infty} \frac{T^{(r_1)}(0)}{r_1!} X^{r_1}\right) \times \dots \times \left(\sum_{r_s=0}^{\infty} \frac{T^{(r_s)}(0)}{r_s!} X^{r_s}\right)\right) \\ &= E\left(\sum_{r_1=0}^{\infty} \dots \sum_{r_s=0}^{\infty} \left(\frac{T^{(r_1)}(0)}{r_1!} \dots \frac{T^{(r_s)}(0)}{r_s!} X^{r_1} \dots X^{r_s}\right)\right) \\ &= \sum_{r_1=0}^{\infty} \dots \sum_{r_s=0}^{\infty} \frac{T^{(r_1)}(0)}{r_1!} \dots \frac{T^{(r_s)}(0)}{r_s!} \mu_{r_1+\dots+r_s} \quad (\text{B.0.6}) \end{aligned}$$

In practice, one truncates the summations at some useful order R , and so we can express (B.0.5) as

$$\begin{aligned} M_Y(t) &= 1 + \frac{t}{1!} \sum_{r_1=0}^R \frac{T^{(r_1)}(0)}{r_1!} \mu_{r_1} + \frac{t^2}{2!} \sum_{r_1, r_2=0}^R \frac{T^{(r_1)}(0) T^{(r_2)}(0)}{r_1! r_2!} \mu_{r_1+r_2} \\ &\quad + \frac{t^3}{3!} \sum_{r_1, r_2, r_3=0}^R \frac{T^{(r_1)}(0) T^{(r_2)}(0) T^{(r_3)}(0)}{r_1! r_2! r_3!} \mu_{r_1+r_2+r_3} + \dots \quad (\text{B.0.7}) \end{aligned}$$

where evaluation at zero of $T^{(r)}$ has been omitted to simplify notation. In this way, we can express the moments and the moment generating function of transformed variable Y in terms moments of X . To find cumulants of Y , one can use a moments to cumulants inversion formula, such as explained in section 3.2.1.

Multidimensional case

Let $\mathbf{Y} \in \mathbb{R}^J$ be defined in terms of another random vector, as in (B.0.1). Its moment generating function is

$$M_{\mathbf{Y}}(\mathbf{t}) = \sum_{s_1=0}^{\infty} \dots \sum_{s_J=0}^{\infty} \frac{t_1^{s_1} \dots t_J^{s_J}}{s_1! \dots s_J!} m_{s_1 \dots s_J} \quad (\text{B.0.8})$$

where

$$m_{s_1 \dots s_J} = E(Y_1^{s_1} \times \dots \times Y_J^{s_J}) \quad (\text{B.0.9})$$

Joint moments of $\mathbf{Y} \in \mathbb{R}^J$ of the form (B.0.9) are now the object of analysis. We set R as truncation order for the Taylor expansion of each function T_j . One has, in analogy to the

one-dimensional case,

$$E(Y_1^{s_1} \times \dots \times Y_J^{s_J}) = E\left(\left(\sum_{r_1=0}^R \frac{T_1^{(r_1)}}{r_1!} X_1^{r_1}\right)^{s_1} \times \dots \times \left(\sum_{r_J=0}^R \frac{T_J^{(r_J)}}{r_J!} X_J^{r_J}\right)^{s_J}\right) \quad (\text{B.0.10})$$

Each exponentiated term will be conveniently expanded as illustrated below for first component,

$$\begin{aligned} \left(\sum_{r_1=0}^R \frac{T_1^{(r_1)}}{r_1!} X_1^{r_1}\right)^{s_1} &= \sum_{r_1^1=0}^R \dots \sum_{r_1^{s_1}=0}^R \frac{T_1^{(r_1^1)} \dots T_1^{(r_1^{s_1})}}{r_1^1! \dots r_1^{s_1}!} X_1^{r_1^1 + \dots + r_1^{s_1}} \\ &= \sum_{r_1^1, \dots, r_1^{s_1}=0}^R \frac{T_1^{(r_1^1)} \dots T_1^{(r_1^{s_1})}}{r_1^1! \dots r_1^{s_1}!} X_1^{\sum_{i=1}^{s_1} r_1^i} \end{aligned}$$

Thus, one has

$$\begin{aligned} E(Y_1^{s_1} \times \dots \times Y_J^{s_J}) &= E\left\{ \left[\sum_{r_1^1, \dots, r_1^{s_1}=0}^R \frac{T_1^{(r_1^1)} \dots T_1^{(r_1^{s_1})}}{r_1^1! \dots r_1^{s_1}!} X_1^{\sum_{i=1}^{s_1} r_1^i} \right] \times \dots \right. \\ &\quad \left. \dots \times \left[\sum_{r_J^1, \dots, r_J^{s_J}=0}^R \frac{T_J^{(r_J^1)} \dots T_J^{(r_J^{s_J})}}{r_J^1! \dots r_J^{s_J}!} X_J^{\sum_{i=1}^{s_J} r_J^i} \right] \right\} \quad (\text{B.0.11}) \end{aligned}$$

which can be written

$$\begin{aligned} E(Y_1^{s_1} \times \dots \times Y_J^{s_J}) &= E\left\{ \sum_{r_1^1, \dots, r_1^{s_1}=0}^R \dots \sum_{r_J^1, \dots, r_J^{s_J}=0}^R \left[\frac{T_1^{(r_1^1)} \dots T_1^{(r_1^{s_1})}}{r_1^1! \dots r_1^{s_1}!} X_1^{\sum_{i=1}^{s_1} r_1^i} \times \dots \right. \right. \\ &\quad \left. \dots \times \frac{T_J^{(r_J^1)} \dots T_J^{(r_J^{s_J})}}{r_J^1! \dots r_J^{s_J}!} X_J^{\sum_{i=1}^{s_J} r_J^i} \right] \Big\} = \\ &\quad \sum_{r_1^1, \dots, r_1^{s_1}=0}^R \dots \sum_{r_J^1, \dots, r_J^{s_J}=0}^R \frac{T_1^{(r_1^1)} \dots T_1^{(r_1^{s_1})}}{r_1^1! \dots r_1^{s_1}!} \dots \frac{T_J^{(r_J^1)} \dots T_J^{(r_J^{s_J})}}{r_J^1! \dots r_J^{s_J}!} E\left(X_1^{\sum_{i=1}^{s_1} r_1^i} \dots X_J^{\sum_{i=1}^{s_J} r_J^i}\right) \quad (\text{B.0.12}) \end{aligned}$$

So, with the aid of equation (B.0.12) one can find both the joint moments and the moment generating function of random vector $\mathbf{Y} \in \mathbb{R}^J$ in terms of the moments of the original vector $\mathbf{X} \in \mathbb{R}^J$. Joint cumulants, and cumulant generating function of \mathbf{Y} can then be obtained by moments to cumulants formula (3.2.18).

Conditions on T_j

We deal now briefly with conditions that function $T_j(\cdot)$, for $j = 1, \dots, J$, has to fulfil in order to have a valid Taylor expansion, as in (B.0.2). This is a well studied topic in mathematical

analysis. More details, and proofs of the statements here presented, can be found on chapter 9 of Apostol (1974).

For the sake of this research, we assume that $T_j \in C^\infty$ on interval $[-b, b]$, where $b \in \mathbb{R}$ will be assumed to be as big as practically necessary.

Derivatives of polynomial growth If there exists a positive constant M , such that $|T_j^{(n)}(x_j)| \leq M^n$, for all $n \in \mathbb{N}$, then the Taylor series can be shown to converge to function $T_j(*)$. Compare this with Oehlert (1992).

Monotonically increasing transformations In the case T_j is a monotonically increasing transformation, then the Taylor series also converges by virtue of the Bernstein theorem for Taylor series.

Finally, in the useful case, explored in this work,

$$T_j(x_j) = \sum_{r=0}^R a_r x_j^r$$

the Taylor expansion is in fact function T_j , if the truncation order is set to R , or grater.

C. Outline of Estimation Procedure at section 6.3

We present now an outline of the estimation procedure used at section 6.3.

Data consists of a matrix $\mathbb{X} \in \mathbb{R}^{I \times J}$, where I represents the number of realizations of the field, and J the dimension of the field. Each row \mathbf{X}_i represents then a field a can be written

$$\mathbf{X}_i = (\mathbf{X}_i^{\text{obs}}, \mathbf{X}_i^{\text{NA}}) \quad (\text{C.0.1})$$

where $\mathbf{X}_i^{\text{obs}}$ and \mathbf{X}_i^{NA} represent the observed and unobserved (censored) part of the field. As the censored part we take all components having negative values. These values will be imputed in the course of the Stochastic EM algorithm outlined below.

If we had no censored data, each row of \mathbb{X} would have likelihood

$$L_{\mathbf{X}_i}(\Theta) = \sqrt{\det(\Gamma^{-1})} \frac{\text{Gamma}\left(\frac{J}{2}\right) f_{R^2}\left((\mathbf{x}_i - \mu)' \Gamma^{-1} (\mathbf{x}_i - \mu)\right)}{\pi^{\frac{J}{2}} \times \left((\mathbf{x}_i - \mu)' \Gamma^{-1} (\mathbf{x}_i - \mu)\right)^{(J-2)/2}} \quad (\text{C.0.2})$$

and the likelihood of the model would be

$$L_{\mathbb{X}}(\Theta) = \prod_{i=1}^I \sqrt{\det(\Gamma^{-1})} \frac{\text{Gamma}\left(\frac{J}{2}\right) f_{R^2}\left((\mathbf{x}_i - \mu)' \Gamma^{-1} (\mathbf{x}_i - \mu)\right)}{\pi^{\frac{J}{2}} \times \left((\mathbf{x}_i - \mu)' \Gamma^{-1} (\mathbf{x}_i - \mu)\right)^{(J-2)/2}} \quad (\text{C.0.3})$$

where $\mu \in \mathbb{R}^J$ is a vector of means, $\Gamma_{J \times J}$ is a positive definite correlation matrix (defined in terms of a covariance function), and f_{R^2} is the density of the squared generating variable of \mathbf{X} . All parameters are considered in the parameter vector Θ .

The density of R^2 is given by the mixture

$$f_{R^2}(x) = \sum_{s=1}^S \pi_s \frac{\theta^j x^{j-1} e^{-x\theta}}{\text{Gamma}(j)} \quad (\text{C.0.4})$$

We can maximize $L_{\mathbb{X}}(\Theta)$ as follows:

1. Assign some initial negative values to the censored values at \mathbb{X} , and to parameters θ and (π_1, \dots, π_S) of f_{R^2} .
2. Given the available \mathbb{X} , θ and (π_1, \dots, π_S) , maximize $L_{\mathbb{X}}(\Theta)$ as a function of μ and correlation matrix Γ . Since for Γ we are using a covariance model with $\sigma^2 := 1$, optimization will comprise the parameters of such covariance model: nuggets effect, range parameter, etc.

3. Given the available values for \mathbb{X} , μ and correlation matrix Γ , maximize $L_{\mathbb{X}}(\Theta)$ as a function of θ and (π_1, \dots, π_S) . For this end, we use the Bayes estimate of Venturini et al. (2008), giving a total weight to the prior distribution of 20%. We simulate 500 iterations of the MCMC algorithm as implemented by the authors for the R statistical software, and use the mean of the last 400 iterations as estimator, for each parameter.
4. Given the available values for μ , Γ , θ and (π_1, \dots, π_S) , run 100 iterations of the Metropolis-Hastings algorithm in order to sample each \mathbf{X}_i^{NA} . The vector \mathbf{X}_i^{NA} takes here the place of the unknown parameters to sample from in an MCMC algorithm. The "density" of which \mathbf{X}_i^{NA} is the parameter is given by $L_{\mathbf{X}_i}$. We used for this application a Normal transition kernel with standard deviation 0.5.
5. Return to item 2.

As explained in the text, we run 120 iteration of this algorithm for parameter estimation at section 6.3.

Steps 2 and 3 represent an approximation to the maximization step of the stochastic EM algorithm. Step 4 is an approximation to the simulation step.

Hence the above algorithm is an approximation to the well-known Stochastic Expectation Maximization algorithm. At each iteration a new value for each parameter is obtained, but after some iterations these parameters stabilize each around a mean value. Such mean value is close to the maximum likelihood estimator.

D. Parameters fitted

The parameters of the fitted squared generating variable for the 30-dimensional data set of section 6 are given below.

The estimated scale parameter is $\hat{\theta} = 0.1001$. The estimated weights for the mixture are shown in table (D.0.1).

$\pi_1 - \pi_{25}$	$\pi_{26} - \pi_{50}$	$\pi_{51} - \pi_{75}$	$\pi_{76} - \pi_{100}$
0.0102839943275482	6.4481467940787e-07	4.25248158706497e-06	9.2449300955788e-07
0.0194490400357542	8.99802080083257e-06	6.93153347633369e-05	9.48032987468127e-06
0.932763676114771	2.44297216708648e-05	2.46362171317745e-06	7.24961905082507e-07
0.0235093688707954	3.69539031868027e-05	3.4419397069178e-05	1.75589427501765e-06
0.00260751029388689	1.17208659431384e-05	3.21018232687351e-06	3.63718280516155e-06
0.000639273897128121	1.62040701813094e-06	1.03605807541582e-05	4.0838859054133e-06
0.000623981538329266	4.88543962800929e-06	1.33457899271894e-05	5.26231404637288e-06
9.86118056206257e-05	2.31109821479957e-06	4.43100804553691e-06	9.8716770049286e-06
5.19981182935072e-05	2.28990253896776e-06	1.75114481178519e-05	7.67959883100292e-06
2.16309695383026e-05	2.01547567163614e-06	8.16646486322687e-07	2.43599672513347e-06
3.19858463128171e-05	8.19651986368629e-07	2.43161303910772e-06	5.50503936874e-06
1.45261959615697e-05	3.9617751617422e-06	2.4786914449827e-06	1.43167643093757e-05
1.64103157085842e-05	5.27226668244265e-06	1.53768902610807e-05	2.53889853278942e-06
2.7092981227538e-05	4.93257685427866e-07	5.82088224063488e-06	5.20528055986649e-06
1.69635677596202e-05	2.80721785583637e-05	1.54985060460838e-06	2.17271609096933e-06
3.01403598010169e-06	7.52259418256762e-06	2.07247067963341e-06	3.80223511506013e-06
0.00274416870585206	2.1629007404713e-06	9.57845335436306e-07	5.50298810152373e-05
1.6515232058326e-05	0.00422370145635887	2.42803353978282e-06	3.38023758549003e-07
3.73285129947467e-06	9.02322580558261e-05	4.27195604921782e-06	3.02559221589016e-06
3.45703680969391e-06	3.95578223857611e-06	2.28222858308876e-06	8.75052068116695e-05
3.25414341475162e-05	4.4198280079391e-05	1.07792266856996e-06	5.64570039945599e-05
0.000130413762878289	0.000109373309893208	1.93668978119392e-06	1.95835208627243e-06
0.00122196647479606	0.000188241570657547	7.77133862265519e-07	0.000100843042477168
2.37386248440859e-05	3.6621641606569e-05	1.62941240328692e-06	5.11156959231813e-05
6.66869935770567e-06	1.78313484952466e-06	4.40808262633827e-07	0.000174107569063703

Table D.0.1.: Estimated weights for Gamma mixture representing squared generating variable of 30-D model.

Bibliography

- (2006). Curse of dimensionality. In Kotz, S., Read, C. B., Balakrishnan, N., Vidakovic, B., and Johnson, N. L., editors, *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc., Hoboken, NJ, USA.
- Aas, K., Czado, C., Frigessi, A., and Bakken, H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198.
- Abramowitz, M. (1972). *Handbook of Mathematical Functions: with formulas, graphs, and mathematical tables*. Dover Publications, New York.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.
- Apostol, T. M. (1974). *Mathematical Analysis, Second Edition*. Pearson.
- Bárdossy, A. and Li, J. (2008). Geostatistical interpolation using copulas. *Water Resources Research*, 44(W07412):doi:10.1029/2007WR006115.
- Bárdossy, A. and Pegram, G. (2009). Copula based multisite model for daily precipitation simulation. *Hydrology and Earth System Sciences Discussions*, 6(3):4485–4534.
- Bárdossy, A. and Pegram, G. (2011). Downscaling precipitation using regional climate models and circulation patterns toward hydrology. *Water Resources Research*, 47(4):n/a–n/a.
- Bárdossy, A. and Pegram, G. (2012). Multiscale spatial recorrelation of RCM precipitation to produce unbiased climate change scenarios over large areas and small. *Water Resources Research*, 48(9).
- Barndorff-Nielsen, O. E. and Cox, D. R. (1990). *Asymptotic techniques for use in statistics*. Chapman and Hall, London [u.a.].
- Bedford, T. and Cooke, R. M. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial intelligence*, 32(1-4):245–268.
- Billingsley, P. (1986). *Probability and measure*. Wiley series in probability and mathematical statistics. Wiley, New York, 2nd ed edition.
- Block, H. W. and Fang, Z. (1988). A multivariate extension of hoeffding’s lemma. *The Annals of Probability*, 16(4):1803–1820.
- Blomqvist, N. (1950). On a measure of dependence between two random variables. *The Annals of Mathematical Statistics*, 21(4):593–600.

- Brillinger, D. R. (1974). *Time series: data analysis and theory*. Holt, Rinehart, and Winston, New York.
- Cambanis, S., Huang, S., and Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368 – 385.
- Casella, G. and Berger, R. L. (2001). *Statistical Inference, Second Edition*. Duxbury Press.
- Chan, G. and Wood, A. T. (1997). Algorithm AS 312: An algorithm for simulating stationary gaussian random fields. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(1):171–181.
- Charpentier, A., Fermanian, J.-D., and Scaillet, O. (2006). The estimation of copulas: Theory and practice. In Rank, J., editor, *Copulas: From theory to application in finance*, pages 35–60. Risk Publications.
- Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, 31(2):131–145.
- Cherubini, U., Luciano, E., and Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons, Hoboken, NJ.
- Cox, D. R. (2006). *Principles of statistical inference*. Cambridge University Press.
- Cressie, N. and Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Wiley.
- Cressie, N. A. C. (1991). *Statistics for spatial data*. Wiley series in probability and mathematical statistics. Wiley, New York.
- Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2):pp. 278–292.
- Del Brio, E. B., Níguez, T.-M., and Perote, J. (2009). Gram-Charlier densities: a multivariate approach. *Quantitative Finance*, 9(7):855–868.
- Demarta, S. and McNeil, A. J. (2005). The t copula and related copulas. *International Statistical Review*, 73(1):111–129.
- Devroye, L. (1986). *Non-uniform random variate generation*. Springer-Verlag, New York.
- Diggle, P. J. and Ribeiro, P. J. (2007). *Model-based Geostatistics*. Springer.
- Embrechts, P., Mcneil, A., and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. In Dempster, M. H. A., editor, *Risk Management: Value at Risk and Beyond*, pages 176–223. Cambridge University Press, Cambridge.
- Fang, H.-B., Fang, K.-T., and Kotz, S. (2002). The meta-elliptica distribution with given marginals. *Journal of Multivariate Analysis*, 82:1–16.
- Fang, K.-T. (1990). *Symmetric multivariate and related distributions*. Number 36 in Monographs on statistics and applied probability. Chapman and Hall, London ; New York.

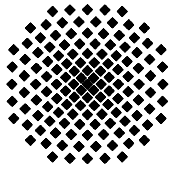
- Feodor Nielsen, S. (2000). The stochastic em algorithm: estimation and asymptotic results. *Bernoulli*, 6(3):457–489.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43(4):521–532.
- Frahm, G. (2004). *Generalized elliptical distributions: theory and applications*. PhD thesis, Universität zu Köln.
- Gallant, A. R. and Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, 55(2):pp. 363–390.
- Gebelein, H. (1941). Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM - Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379.
- Genest, C., Masiello, E., and Tribouley, K. (2009). Estimating copula densities through wavelets. *Insurance: Mathematics and Economics*, 44(2):170–181.
- Genz, A. (1993). Comparison of methods for the computation of multivariate normal probabilities. *Computing Sciences and Statistics*, 25:400–405.
- Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1998). *Markov chain Monte Carlo in practice*. Chapman & Hall, Boca Raton, Fla.
- Gräler, B. and Pebesma, E. (2011). The pair-copula construction for spatial data: a new approach to model spatial dependency. *Procedia Environmental Sciences*, 7:206–211.
- Grün, S. and Rotter, S. (2010). *Analysis of parallel spike trains*. Springer, New York.
- Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics & Data Analysis*, 40(4):685 – 711.
- Headrick, T. C. (2010). *Statistical simulation: power method polynomials and other transformations*. Chapman & Hall/CRC, Boca Raton.
- Headrick, T. C. and Zumbo, B. D. (2008). A method for simulating multivariate non normal distributions with specified standardized cumulants and intraclass correlation coefficients. *Communications in Statistics - Simulation and Computation*, 37(3):617–628.
- Hebb, D. O. (2002). *The organization of behavior : a neuropsychological theory*. L. Erlbaum Associates, Mahwah, N.J.
- Hoeffding, W. (1940). Masstabinvariante korrelations-theorie. *Schriften Math. Inst. Univ. Berlin*, 5:181–233.
- Hulle, M. M. V. (2005). Edgeworth approximation of multivariate differential entropy. *Neural Computation*, 17(9):1903–1910.

- Hult, H. and Lindskog, F. (2002). Multivariate extremes, aggregation and dependence in elliptical distributions. *Advances in Applied Probability*, 34(3):587–608.
- Hurt, J. (1976). Asymptotic expansions of functions of statistics. *Aplikace matematiky*, 21(6):444–456.
- Huzurbazar, S. (1999). Practical saddlepoint approximations. *The American Statistician*, 53(3):pp. 225–232.
- Jaworski, P. (2010). *Copula theory and its applications: proceedings of the workshop held in Warsaw, 25-26 September 2009*. Number 198 in Lecture notes in statistics–proceedings. Springer, Heidelberg ; New York.
- Joe, H. (1989a). Estimation of entropy and other functionals of a multivariate density. *Annals of the Institute of Statistical Mathematics*, 41(4):683–697.
- Joe, H. (1989b). Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association*, 84(405):pp. 157–164.
- Joe, H. (1996). Families of m -variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. *Lecture Notes-Monograph Series*, 28:pp. 120–141.
- Joe, H. (1997). *Multivariate models and dependence concepts*. Chapman Hall, Boca Raton.
- Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419.
- Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1/2):pp. 149–176.
- Kano, Y. (1994). Consistency property of elliptic probability density functions. *Journal of Multivariate Analysis*, 51(1):139 – 147.
- Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):pp. 81–93.
- Kendall, M. G. and Stuart, A. (1969). *The advanced theory of statistics Vol. 1, Distribution theory*. Griffin, London.
- Kitanidis, P. (1997). *Introduction to geostatistics: applications in hydrogeology*. Cambridge University Press.
- Kolassa, J. (2006). *Series approximation methods in statistics*, volume 88. Springer.
- Kolassa, J. and Li, J. (2010). Multivariate saddlepoint approximations in tail probability and conditional inference. *Bernoulli*, 16(4):1191–1207.
- Kousky, C. and R.M., C. (2011). The limits of securitization: Micro-correlations, fat tails and tail dependence. In Böcker, K., editor, *Rethinking Risk Measurement and Reporting: Vol. I. Risk Books*.
- Kruskal, J. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2):115–129.

- Lancaster, H. O. (1969). *The chi-squared distribution*. Wiley, New York.
- Le, N. D. and Zidek, J. V. (2006). *Statistical analysis of environmental space-time processes*. Springer series in statistics. Springer, New York.
- Lehmann, E. L. (1966). Some concepts of dependence. *The Annals of Mathematical Statistics*, 37(5):1137–1153.
- Lindskog, F., McNeil, A., and Schmock, U. (2003). Kendall's tau for elliptical distributions. In Müller, W. A., Bihn, M., Bol, G., Nakhaeizadeh, G., Rachev, S. T., Ridder, T., and Vollmer, K.-H., editors, *Credit Risk*, pages 149–156. Physica-Verlag HD, Heidelberg.
- Long, J. D. (2006). Kendall's tau-II. In Kotz, S., Read, C. B., Balakrishnan, N., Vidakovic, B., and Johnson, N. L., editors, *Encyclopedia of Statistical Sciences*. John Wiley & Sons, Inc., Hoboken, NJ, USA.
- Lugannani, R. and Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability*, 12(2):pp. 475–490.
- Marron, J. S. and Ruppert, D. (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(4):pp. 653–671.
- Mathai, A. and Moschopoulos, P. (1991). On a multivariate gamma. *Journal of Multivariate Analysis*, 39(1):135–153.
- Mauleon, I. and Perote, J. (2000). Testing densities with financial data: an empirical comparison of the Edgeworth–Sargan density to the student t. *The European Journal of Finance*, 6(2):225–239.
- McCullagh, P. (1984). Tensor notation and cumulants of polynomials. *Biometrika*, 71(3):pp. 461–476.
- McCullagh, P. (1987). *Tensor methods in statistics*. Chapman and Hall, London; New York.
- Mendel, J. (1991). Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications. *Proceedings of the IEEE*, 79(3):278–305.
- Muirhead, R. (1982). *Aspects of multivariate statistical theory*. John Wiley & Sons, New York.
- Nelsen, R. (1999). *An introduction to copulas*. Springer Verlag, New York.
- Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, 46(1):27–29.
- Patton, A. J. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110(0):4 – 18. <ce:title>Special Issue on Copula Modeling and Dependence</ce:title>.
- Pearson, K. (2011). *The life, letters and labours of Francis Galton*. Cambridge University Press, Cambridge.

- Perote, J. (2004). The multivariate Edgeworth–Sargan density. *Spanish Economic Review*, 6(1):77–96.
- Rényi, A. (1959). On measures of dependence. *Acta Mathematica Hungarica*, 10(3):441–451.
- Ripley, B. D. (1981). *Spatial statistics*. Wiley series in probability and mathematical statistics. Wiley, New York.
- Rota, G.-C. (1964). The number of partitions of a set. *The American Mathematical Monthly*, 71(5):498.
- Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119.
- Sanso, B. and Guenni, L. (1999). Venezuelan rainfall data analysed by using a bayesian space-time model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):345–362.
- Sanso, B. and Guenni, L. (2000). A nonstationary multisite model for rainfall. *Journal of the American Statistical Association*, 95(452):pp. 1089–1100.
- Sargan, J. D. (1976). Econometric estimators and the edgeworth approximation. *Econometrica*, 44(3):pp. 421–448.
- Schmid, F., Schmidt, R., Blumentritt, T., Gaißer, S., and Ruppert, M. (2010). Copula-based measures of multivariate association. In Jaworski, P., Durante, F., Härdle, W. K., and Rychlik, T., editors, *Copula Theory and Its Applications*, volume 198, pages 209–236. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Schmidt, A. M. and O’Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):743–758.
- Schweizer, B. and Wolff, E. F. (1981). On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9(4):879–885.
- Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization*. Wiley series in probability and mathematical statistics. Wiley, New York.
- Singh, V. P. and (ed.), W. G. S. (2007). Special issue: Copulas in hydrology. *Journal of Hydrologic Engineering*, 12(4).
- Skovgaard, I. (1987). Saddlepoint expansions for conditional distributions. *Journal of Applied Probability*, pages 875–887.
- Slifker, J. F. and Shapiro, S. S. (1980). The johnson system: Selection and parameter estimation. *Technometrics*, 22(2):pp. 239–246.
- Smith, P. J. (1995). A recursive formulation of the old problem of obtaining moments from cumulants and vice versa. *The American Statistician*, 49(2):217–218.

- Spearman, C. (1904). The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101.
- Staude, B., Grün, S., and Rotter, S. (2010). Higher-order correlations and cumulants. In Grün, S. and Rotter, S., editors, *Analysis of Parallel Spike Trains*, pages 253–280. Springer US, Boston, MA.
- Steyn, H. (1993). On the problem of more than one kurtosis parameter in multivariate analysis. *Journal of Multivariate Analysis*, 44(1):1 – 22.
- Streitberg, B. (1990). Lancaster interactions revisited. *The Annals of Statistics*, 18(4):1878–1885.
- Streitberg, B. (1999). Exploring interactions in high-dimensional tables: a bootstrap alternative to log-linear models. *The Annals of Statistics*, 27(1):405–413.
- Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540.
- van Dyk, D. A. and Meng, X.-L. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50.
- Venturini, S., Dominici, F., and Parmigiani, G. (2008). Gamma shape mixtures for heavy-tailed distributions. *The Annals of Applied Statistics*, 2(2):756–776.
- Villasenor Alva, J. A. and Estrada, E. G. (2009). A generalization of shapiro–wilk’s test for multivariate normality. *Communications in Statistics Theory and Methods*, 38(11):1870–1883.
- wa Binyavanga, K. (2009). Calculating cumulants of a taylor expansion of a multivariate function. *International Statistical Review*, 77(2):212–221.
- Zheng, X. and Katz, R. W. (2008). Simulation of spatial dependence in daily rainfall using multisite generators. *Water Resources Research*, 44(9):n/a–n/a.



**Institut für Wasser- und
Umweltsystemmodellierung
Universität Stuttgart**

Pfaffenwaldring 61
70569 Stuttgart (Vaihingen)
Telefon (0711) 685 - 64717/64749/64752/64679
Telefax (0711) 685 - 67020 o. 64746 o. 64681
E-Mail: iws@iws.uni-stuttgart.de
<http://www.iws.uni-stuttgart.de>

Direktoren

Prof. Dr. rer. nat. Dr.-Ing. András Bárdossy
Prof. Dr.-Ing. Rainer Helmig
Prof. Dr.-Ing. Silke Wieprecht

Vorstand (Stand 19.08.2013)

Prof. Dr. rer. nat. Dr.-Ing. A. Bárdossy
Prof. Dr.-Ing. R. Helmig
Prof. Dr.-Ing. S. Wieprecht
Prof. Dr. J.A. Sander Huisman
Jürgen Braun, PhD
apl. Prof. Dr.-Ing. H. Class
Dr.-Ing. H.-P. Koschitzky
Dr.-Ing. M. Noack
Jun.-Prof. Dr.-Ing. W. Nowak, M.Sc.
Dr. rer. nat. J. Seidel
Dr.-Ing. K. Terheiden

Emeriti

Prof. Dr.-Ing. habil. Dr.-Ing. E.h. Jürgen Giesecke
Prof. Dr.h.c. Dr.-Ing. E.h. Helmut Kobus, PhD

**Lehrstuhl für Wasserbau und
Wassermengenwirtschaft**

Leiter: Prof. Dr.-Ing. Silke Wieprecht
Stellv.: Dr.-Ing. Kristina Terheiden
Versuchsanstalt für Wasserbau
Leiter: Dr.-Ing. Markus Noack

**Lehrstuhl für Hydromechanik
und Hydrosystemmodellierung**

Leiter: Prof. Dr.-Ing. Rainer Helmig
Stellv.: apl. Prof. Dr.-Ing. Holger Class
**Jungwissenschaftlergruppe: Stochastische
Modellierung von Hydrosystemen**
Leiter: Jun.-Prof. Dr.-Ing. Wolfgang Nowak, M.Sc.

Lehrstuhl für Hydrologie und Geohydrologie

Leiter: Prof. Dr. rer. nat. Dr.-Ing. András Bárdossy
Stellv.: Dr. rer. nat. Jochen Seidel
Hydrogeophysik der Vadosen Zone
(mit Forschungszentrum Jülich)
Leiter: Prof. Dr. J.A. Sander Huisman

**VEGAS, Versuchseinrichtung zur
Grundwasser- und Altlastensanierung**

Leitung: Jürgen Braun, PhD, AD
Dr.-Ing. Hans-Peter Koschitzky, AD

Verzeichnis der Mitteilungshefte

- 1 Röhnisch, Arthur: *Die Bemühungen um eine Wasserbauliche Versuchsanstalt an der Technischen Hochschule Stuttgart*, und Fattah Abouleid, Abdel: *Beitrag zur Berechnung einer in lockeren Sand gerammten, zweifach verankerten Spundwand*, 1963
- 2 Marotz, Günter: *Beitrag zur Frage der Standfestigkeit von dichten Asphaltbelägen im Großwasserbau*, 1964
- 3 Gurr, Siegfried: *Beitrag zur Berechnung zusammengesetzter ebener Flächen-tragwerke unter besonderer Berücksichtigung ebener Stauwände, mit Hilfe von Randwert- und Lastwertmatrizen*, 1965
- 4 Plica, Peter: *Ein Beitrag zur Anwendung von Schalenkonstruktionen im Stahlwasserbau*, und Petrikat, Kurt: *Möglichkeiten und Grenzen des wasserbaulichen Versuchswesens*, 1966

- 5 Plate, Erich: *Beitrag zur Bestimmung der Windgeschwindigkeitsverteilung in der durch eine Wand gestörten bodennahen Luftschicht*, und
Röhnisch, Arthur; Marotz, Günter: *Neue Baustoffe und Bauausführungen für den Schutz der Böschungen und der Sohle von Kanälen, Flüssen und Häfen; Gesteigungskosten und jeweilige Vorteile*, sowie Unny, T.E.: *Schwingungsuntersuchungen am Kegelstrahlschieber*, 1967
- 6 Seiler, Erich: *Die Ermittlung des Anlagenwertes der bundeseigenen Binnenschiffahrtsstraßen und Talsperren und des Anteils der Binnenschiffahrt an diesem Wert*, 1967
- 7 *Sonderheft anlässlich des 65. Geburtstages von Prof. Arthur Röhnisch mit Beiträgen von* Benk, Dieter; Breitling, J.; Gurr, Siegfried; Haberhauer, Robert; Honekamp, Hermann; Kuz, Klaus Dieter; Marotz, Günter; Mayer-Vorfelder, Hans-Jörg; Miller, Rudolf; Plate, Erich J.; Radomski, Helge; Schwarz, Helmut; Vollmer, Ernst; Wildenhahn, Eberhard; 1967
- 8 Jumikis, Alfred: *Beitrag zur experimentellen Untersuchung des Wassernachschubs in einem gefrierenden Boden und die Beurteilung der Ergebnisse*, 1968
- 9 Marotz, Günter: *Technische Grundlagen einer Wasserspeicherung im natürlichen Untergrund*, 1968
- 10 Radomski, Helge: *Untersuchungen über den Einfluß der Querschnittsform wellenförmiger Spundwände auf die statischen und rammtechnischen Eigenschaften*, 1968
- 11 Schwarz, Helmut: *Die Grenztragfähigkeit des Baugrundes bei Einwirkung vertikal gezogener Ankerplatten als zweidimensionales Bruchproblem*, 1969
- 12 Erbel, Klaus: *Ein Beitrag zur Untersuchung der Metamorphose von Mittelgebirgsschneedecken unter besonderer Berücksichtigung eines Verfahrens zur Bestimmung der thermischen Schneequalität*, 1969
- 13 Westhaus, Karl-Heinz: *Der Strukturwandel in der Binnenschiffahrt und sein Einfluß auf den Ausbau der Binnenschiffskanäle*, 1969
- 14 Mayer-Vorfelder, Hans-Jörg: *Ein Beitrag zur Berechnung des Erdwiderstandes unter Ansatz der logarithmischen Spirale als Gleitflächenfunktion*, 1970
- 15 Schulz, Manfred: *Berechnung des räumlichen Erddruckes auf die Wandung kreiszylindrischer Körper*, 1970
- 16 Mobasseri, Manoutschehr: *Die Rippenstützmauer. Konstruktion und Grenzen ihrer Standsicherheit*, 1970
- 17 Benk, Dieter: *Ein Beitrag zum Betrieb und zur Bemessung von Hochwasserrückhaltebecken*, 1970

- 18 Gàl, Attila: *Bestimmung der mitschwingenden Wassermasse bei überströmten Fischbauchklappen mit kreiszylindrischem Staublech*, 1971, vergriffen
- 19 Kuz, Klaus Dieter: *Ein Beitrag zur Frage des Einsetzens von Kavitationerscheinungen in einer Düsenströmung bei Berücksichtigung der im Wasser gelösten Gase*, 1971, vergriffen
- 20 Schaak, Hartmut: *Verteilleitungen von Wasserkraftanlagen*, 1971
- 21 *Sonderheft zur Eröffnung der neuen Versuchsanstalt des Instituts für Wasserbau der Universität Stuttgart mit Beiträgen von* Brombach, Hansjörg; Dirksen, Wolfram; Gàl, Attila; Gerlach, Reinhard; Giesecke, Jürgen; Holthoff, Franz-Josef; Kuz, Klaus Dieter; Marotz, Günter; Minor, Hans-Erwin; Petrikat, Kurt; Röhnisch, Arthur; Rueff, Helge; Schwarz, Helmut; Vollmer, Ernst; Wildenhahn, Eberhard; 1972
- 22 Wang, Chung-su: *Ein Beitrag zur Berechnung der Schwingungen an Kegelstrahlschiebern*, 1972
- 23 Mayer-Vorfelder, Hans-Jörg: *Erdwiderstandsbeiwerte nach dem Ohde-Variationsverfahren*, 1972
- 24 Minor, Hans-Erwin: *Beitrag zur Bestimmung der Schwingungsanfachungsfunktionen überströmter Stauklappen*, 1972, vergriffen
- 25 Brombach, Hansjörg: *Untersuchung strömungsmechanischer Elemente (Fluidik) und die Möglichkeit der Anwendung von Wirbelkammerelementen im Wasserbau*, 1972, vergriffen
- 26 Wildenhahn, Eberhard: *Beitrag zur Berechnung von Horizontalfilterbrunnen*, 1972
- 27 Steinlein, Helmut: *Die Eliminierung der Schwebstoffe aus Flußwasser zum Zweck der unterirdischen Wasserspeicherung, gezeigt am Beispiel der Iller*, 1972
- 28 Holthoff, Franz Josef: *Die Überwindung großer Hubhöhen in der Binnenschifffahrt durch Schwimmerhebwerke*, 1973
- 29 Röder, Karl: *Einwirkungen aus Baugrundbewegungen auf trog- und kastenförmige Konstruktionen des Wasser- und Tunnelbaues*, 1973
- 30 Kretschmer, Heinz: *Die Bemessung von Bogenstaumauern in Abhängigkeit von der Talform*, 1973
- 31 Honekamp, Hermann: *Beitrag zur Berechnung der Montage von Unterwasserpipelines*, 1973
- 32 Giesecke, Jürgen: *Die Wirbelkammertriode als neuartiges Steuerorgan im Wasserbau*, und Brombach, Hansjörg: *Entwicklung, Bauformen, Wirkungsweise und Steuereigenschaften von Wirbelkammerverstärkern*, 1974

- 33 Rueff, Helge: *Untersuchung der schwingungserregenden Kräfte an zwei hintereinander angeordneten Tiefschützen unter besonderer Berücksichtigung von Kavitation*, 1974
- 34 Röhnisch, Arthur: *Einpreßversuche mit Zementmörtel für Spannbeton - Vergleich der Ergebnisse von Modellversuchen mit Ausführungen in Hüllwellrohren*, 1975
- 35 *Sonderheft anlässlich des 65. Geburtstages von Prof. Dr.-Ing. Kurt Petrikat mit Beiträgen von:* Brombach, Hansjörg; Erbel, Klaus; Flinspach, Dieter; Fischer jr., Richard; Gàl, Attila; Gerlach, Reinhard; Giesecke, Jürgen; Haberhauer, Robert; Hafner Edzard; Hausenblas, Bernhard; Horlacher, Hans-Burkhard; Hutarew, Andreas; Knoll, Manfred; Krummet, Ralph; Marotz, Günter; Merkle, Theodor; Miller, Christoph; Minor, Hans-Erwin; Neumayer, Hans; Rao, Syamala; Rath, Paul; Rueff, Helge; Ruppert, Jürgen; Schwarz, Wolfgang; Topal-Gökceli, Mehmet; Vollmer, Ernst; Wang, Chung-su; Weber, Hans-Georg; 1975
- 36 Berger, Jochum: *Beitrag zur Berechnung des Spannungszustandes in rotations-symmetrisch belasteten Kugelschalen veränderlicher Wandstärke unter Gas- und Flüssigkeitsdruck durch Integration schwach singulärer Differentialgleichungen*, 1975
- 37 Dirksen, Wolfram: *Berechnung instationärer Abflußvorgänge in gestauten Gerinnen mittels Differenzenverfahren und die Anwendung auf Hochwasserrückhaltebecken*, 1976
- 38 Horlacher, Hans-Burkhard: *Berechnung instationärer Temperatur- und Spannungsfelder in langen mehrschichtigen Hohlzylindern*, 1976
- 39 Hafner, Edzard: *Untersuchung der hydrodynamischen Kräfte auf Baukörper im Tiefwasserbereich des Meeres*, 1977, ISBN 3-921694-39-6
- 40 Ruppert, Jürgen: *Über den Axialwirbelkammverstärker für den Einsatz im Wasserbau*, 1977, ISBN 3-921694-40-X
- 41 Hutarew, Andreas: *Beitrag zur Beeinflußbarkeit des Sauerstoffgehalts in Fließgewässern an Abstürzen und Wehren*, 1977, ISBN 3-921694-41-8, vergriffen
- 42 Miller, Christoph: *Ein Beitrag zur Bestimmung der schwingungserregenden Kräfte an unterströmten Wehren*, 1977, ISBN 3-921694-42-6
- 43 Schwarz, Wolfgang: *Druckstoßberechnung unter Berücksichtigung der Radial- und Längsverschiebungen der Rohrwandung*, 1978, ISBN 3-921694-43-4
- 44 Kinzelbach, Wolfgang: *Numerische Untersuchungen über den optimalen Einsatz variabler Kühlsysteme einer Kraftwerkskette am Beispiel Oberrhein*, 1978, ISBN 3-921694-44-2
- 45 Barczewski, Baldur: *Neue Meßmethoden für Wasser-Luftgemische und deren Anwendung auf zweiphasige Auftriebsstrahlen*, 1979, ISBN 3-921694-45-0

- 46 Neumayer, Hans: *Untersuchung der Strömungsvorgänge in radialen Wirbelkammerverstärkern*, 1979, ISBN 3-921694-46-9
- 47 Elalfy, Youssef-Elhassan: *Untersuchung der Strömungsvorgänge in Wirbelkammerdioden und -drosseln*, 1979, ISBN 3-921694-47-7
- 48 Brombach, Hansjörg: *Automatisierung der Bewirtschaftung von Wasserspeichern*, 1981, ISBN 3-921694-48-5
- 49 Geldner, Peter: *Deterministische und stochastische Methoden zur Bestimmung der Selbstdichtung von Gewässern*, 1981, ISBN 3-921694-49-3, vergriffen
- 50 Mehlhorn, Hans: *Temperaturveränderungen im Grundwasser durch Brauchwassereinleitungen*, 1982, ISBN 3-921694-50-7, vergriffen
- 51 Hafner, Edzard: *Rohrleitungen und Behälter im Meer*, 1983, ISBN 3-921694-51-5
- 52 Rinnert, Bernd: *Hydrodynamische Dispersion in porösen Medien: Einfluß von Dichteunterschieden auf die Vertikalvermischung in horizontaler Strömung*, 1983, ISBN 3-921694-52-3, vergriffen
- 53 Lindner, Wulf: *Steuerung von Grundwasserentnahmen unter Einhaltung ökologischer Kriterien*, 1983, ISBN 3-921694-53-1, vergriffen
- 54 Herr, Michael; Herzer, Jörg; Kinzelbach, Wolfgang; Kobus, Helmut; Rinnert, Bernd: *Methoden zur rechnerischen Erfassung und hydraulischen Sanierung von Grundwasserkontaminationen*, 1983, ISBN 3-921694-54-X
- 55 Schmitt, Paul: *Wege zur Automatisierung der Niederschlagsermittlung*, 1984, ISBN 3-921694-55-8, vergriffen
- 56 Müller, Peter: *Transport und selektive Sedimentation von Schwebstoffen bei gestautem Abfluß*, 1985, ISBN 3-921694-56-6
- 57 El-Qawasmeh, Fuad: *Möglichkeiten und Grenzen der Tropfbewässerung unter besonderer Berücksichtigung der Verstopfungsanfälligkeit der Tropfelemente*, 1985, ISBN 3-921694-57-4, vergriffen
- 58 Kirchenbaur, Klaus: *Mikroprozessorgesteuerte Erfassung instationärer Druckfelder am Beispiel seegangsbelasteter Baukörper*, 1985, ISBN 3-921694-58-2
- 59 Kobus, Helmut (Hrsg.): *Modellierung des großräumigen Wärme- und Schadstofftransports im Grundwasser*, Tätigkeitsbericht 1984/85 (DFG-Forschergruppe an den Universitäten Hohenheim, Karlsruhe und Stuttgart), 1985, ISBN 3-921694-59-0, vergriffen
- 60 Spitz, Karlheinz: *Dispersion in porösen Medien: Einfluß von Inhomogenitäten und Dichteunterschieden*, 1985, ISBN 3-921694-60-4, vergriffen
- 61 Kobus, Helmut: *An Introduction to Air-Water Flows in Hydraulics*, 1985, ISBN 3-921694-61-2

- 62 Kaleris, Vassilios: *Erfassung des Austausches von Oberflächen- und Grundwasser in horizontalebene Grundwassermodellen*, 1986, ISBN 3-921694-62-0
- 63 Herr, Michael: *Grundlagen der hydraulischen Sanierung verunreinigter Porengrundwasserleiter*, 1987, ISBN 3-921694-63-9
- 64 Marx, Walter: *Berechnung von Temperatur und Spannung in Massengestein infolge Hydratation*, 1987, ISBN 3-921694-64-7
- 65 Koschitzky, Hans-Peter: *Dimensionierungskonzept für Sohlbelüfter in Schußbrinnen zur Vermeidung von Kavitationsschäden*, 1987, ISBN 3-921694-65-5
- 66 Kobus, Helmut (Hrsg.): *Modellierung des großräumigen Wärme- und Schadstofftransports im Grundwasser*, Tätigkeitsbericht 1986/87 (DFG-Forschergruppe an den Universitäten Hohenheim, Karlsruhe und Stuttgart) 1987, ISBN 3-921694-66-3
- 67 Söll, Thomas: *Berechnungsverfahren zur Abschätzung anthropogener Temperaturanomalien im Grundwasser*, 1988, ISBN 3-921694-67-1
- 68 Dittrich, Andreas; Westrich, Bernd: *Bodenseeufenerosion, Bestandsaufnahme und Bewertung*, 1988, ISBN 3-921694-68-X, vergriffen
- 69 Huwe, Bernd; van der Ploeg, Rienk R.: *Modelle zur Simulation des Stickstoffhaushaltes von Standorten mit unterschiedlicher landwirtschaftlicher Nutzung*, 1988, ISBN 3-921694-69-8, vergriffen
- 70 Stephan, Karl: *Integration elliptischer Funktionen*, 1988, ISBN 3-921694-70-1
- 71 Kobus, Helmut; Zilliox, Lothaire (Hrsg.): *Nitratbelastung des Grundwassers, Auswirkungen der Landwirtschaft auf die Grundwasser- und Rohwasserbeschaffenheit und Maßnahmen zum Schutz des Grundwassers*. Vorträge des deutsch-französischen Kolloquiums am 6. Oktober 1988, Universitäten Stuttgart und Louis Pasteur Strasbourg (Vorträge in deutsch oder französisch, Kurzfassungen zweisprachig), 1988, ISBN 3-921694-71-X
- 72 Soyeaux, Renald: *Unterströmung von Stauanlagen auf klüftigem Untergrund unter Berücksichtigung laminarer und turbulenter Fließzustände*, 1991, ISBN 3-921694-72-8
- 73 Kohane, Roberto: *Berechnungsmethoden für Hochwasserabfluß in Fließgewässern mit überströmten Vorländern*, 1991, ISBN 3-921694-73-6
- 74 Hassinger, Reinhard: *Beitrag zur Hydraulik und Bemessung von Blocksteinrampen in flexibler Bauweise*, 1991, ISBN 3-921694-74-4, vergriffen
- 75 Schäfer, Gerhard: *Einfluß von Schichtenstrukturen und lokalen Einlagerungen auf die Längsdispersion in Porengrundwasserleitern*, 1991, ISBN 3-921694-75-2
- 76 Giesecke, Jürgen: *Vorträge, Wasserwirtschaft in stark besiedelten Regionen; Umweltforschung mit Schwerpunkt Wasserwirtschaft*, 1991, ISBN 3-921694-76-0

- 77 Huwe, Bernd: *Deterministische und stochastische Ansätze zur Modellierung des Stickstoffhaushalts landwirtschaftlich genutzter Flächen auf unterschiedlichem Skalenniveau*, 1992, ISBN 3-921694-77-9, vergriffen
- 78 Rommel, Michael: *Verwendung von Kluftdaten zur realitätsnahen Generierung von Kluftnetzen mit anschließender laminar-turbulenter Strömungsberechnung*, 1993, ISBN 3-92 1694-78-7
- 79 Marschall, Paul: *Die Ermittlung lokaler Stofffrachten im Grundwasser mit Hilfe von Einbohrloch-Meßverfahren*, 1993, ISBN 3-921694-79-5, vergriffen
- 80 Ptak, Thomas: *Stofftransport in heterogenen Porenaquiferen: Felduntersuchungen und stochastische Modellierung*, 1993, ISBN 3-921694-80-9, vergriffen
- 81 Haakh, Frieder: *Transientes Strömungsverhalten in Wirbelkammern*, 1993, ISBN 3-921694-81-7
- 82 Kobus, Helmut; Cirpka, Olaf; Barczewski, Baldur; Koschitzky, Hans-Peter: *Versucheinrichtung zur Grundwasser und Altlastensanierung VEGAS, Konzeption und Programmrahmen*, 1993, ISBN 3-921694-82-5
- 83 Zang, Weidong: *Optimaler Echtzeit-Betrieb eines Speichers mit aktueller Abflußregenerierung*, 1994, ISBN 3-921694-83-3, vergriffen
- 84 Franke, Hans-Jörg: *Stochastische Modellierung eines flächenhaften Stoffeintrages und Transports in Grundwasser am Beispiel der Pflanzenschutzmittelproblematik*, 1995, ISBN 3-921694-84-1
- 85 Lang, Ulrich: *Simulation regionaler Strömungs- und Transportvorgänge in Karst-aquiferen mit Hilfe des Doppelkontinuum-Ansatzes: Methodenentwicklung und Parameteridentifikation*, 1995, ISBN 3-921694-85-X, vergriffen
- 86 Helmig, Rainer: *Einführung in die Numerischen Methoden der Hydromechanik*, 1996, ISBN 3-921694-86-8, vergriffen
- 87 Cirpka, Olaf: *CONTRACT: A Numerical Tool for Contaminant Transport and Chemical Transformations - Theory and Program Documentation -*, 1996, ISBN 3-921694-87-6
- 88 Haberlandt, Uwe: *Stochastische Synthese und Regionalisierung des Niederschlages für Schmutzfrachtberechnungen*, 1996, ISBN 3-921694-88-4
- 89 Croisé, Jean: *Extraktion von flüchtigen Chemikalien aus natürlichen Lockergesteinen mittels erzwungener Luftströmung*, 1996, ISBN 3-921694-89-2, vergriffen
- 90 Jorde, Klaus: *Ökologisch begründete, dynamische Mindestwasserregelungen bei Ausleitungskraftwerken*, 1997, ISBN 3-921694-90-6, vergriffen
- 91 Helmig, Rainer: *Gekoppelte Strömungs- und Transportprozesse im Untergrund - Ein Beitrag zur Hydrosystemmodellierung-*, 1998, ISBN 3-921694-91-4, vergriffen

- 92 Emmert, Martin: *Numerische Modellierung nichtisothermer Gas-Wasser Systeme in porösen Medien*, 1997, ISBN 3-921694-92-2
- 93 Kern, Ulrich: *Transport von Schweb- und Schadstoffen in staugeregelten Fließgewässern am Beispiel des Neckars*, 1997, ISBN 3-921694-93-0, vergriffen
- 94 Förster, Georg: *Druckstoßdämpfung durch große Luftblasen in Hochpunkten von Rohrleitungen* 1997, ISBN 3-921694-94-9
- 95 Cirpka, Olaf: *Numerische Methoden zur Simulation des reaktiven Mehrkomponententransports im Grundwasser*, 1997, ISBN 3-921694-95-7, vergriffen
- 96 Färber, Arne: *Wärmetransport in der ungesättigten Bodenzone: Entwicklung einer thermischen In-situ-Sanierungstechnologie*, 1997, ISBN 3-921694-96-5
- 97 Betz, Christoph: *Wasserdampfdestillation von Schadstoffen im porösen Medium: Entwicklung einer thermischen In-situ-Sanierungstechnologie*, 1998, ISBN 3-921694-97-3
- 98 Xu, Yichun: *Numerical Modeling of Suspended Sediment Transport in Rivers*, 1998, ISBN 3-921694-98-1, vergriffen
- 99 Wüst, Wolfgang: *Geochemische Untersuchungen zur Sanierung CKW-kontaminierter Aquifere mit Fe(0)-Reaktionswänden*, 2000, ISBN 3-933761-02-2
- 100 Sheta, Hussam: *Simulation von Mehrphasenvorgängen in porösen Medien unter Einbeziehung von Hysterese-Effekten*, 2000, ISBN 3-933761-03-4
- 101 Ayros, Edwin: *Regionalisierung extremer Abflüsse auf der Grundlage statistischer Verfahren*, 2000, ISBN 3-933761-04-2, vergriffen
- 102 Huber, Ralf: *Compositional Multiphase Flow and Transport in Heterogeneous Porous Media*, 2000, ISBN 3-933761-05-0
- 103 Braun, Christopherus: *Ein Upscaling-Verfahren für Mehrphasenströmungen in porösen Medien*, 2000, ISBN 3-933761-06-9
- 104 Hofmann, Bernd: *Entwicklung eines rechnergestützten Managementsystems zur Beurteilung von Grundwasserschadensfällen*, 2000, ISBN 3-933761-07-7
- 105 Class, Holger: *Theorie und numerische Modellierung nichtisothermer Mehrphasenprozesse in NAPL-kontaminierten porösen Medien*, 2001, ISBN 3-933761-08-5
- 106 Schmidt, Reinhard: *Wasserdampf- und Heißluftinjektion zur thermischen Sanierung kontaminierter Standorte*, 2001, ISBN 3-933761-09-3
- 107 Josef, Reinhold: *Schadstoffextraktion mit hydraulischen Sanierungsverfahren unter Anwendung von grenzflächenaktiven Stoffen*, 2001, ISBN 3-933761-10-7

- 108 Schneider, Matthias: *Habitat- und Abflussmodellierung für Fließgewässer mit unscharfen Berechnungsansätzen*, 2001, ISBN 3-933761-11-5
- 109 Rathgeb, Andreas: *Hydrodynamische Bemessungsgrundlagen für Lockerdeckwerke an überströmbaren Erddämmen*, 2001, ISBN 3-933761-12-3
- 110 Lang, Stefan: *Parallele numerische Simulation instationärer Probleme mit adaptiven Methoden auf unstrukturierten Gittern*, 2001, ISBN 3-933761-13-1
- 111 Appt, Jochen; Stumpp Simone: *Die Bodensee-Messkampagne 2001, IWS/CWR Lake Constance Measurement Program 2001*, 2002, ISBN 3-933761-14-X
- 112 Heimerl, Stephan: *Systematische Beurteilung von Wasserkraftprojekten*, 2002, ISBN 3-933761-15-8, vergriffen
- 113 Iqbal, Amin: *On the Management and Salinity Control of Drip Irrigation*, 2002, ISBN 3-933761-16-6
- 114 Silberhorn-Hemminger, Annette: *Modellierung von Kluftaquifersystemen: Geostatistische Analyse und deterministisch-stochastische Kluftgenerierung*, 2002, ISBN 3-933761-17-4
- 115 Winkler, Angela: *Prozesse des Wärme- und Stofftransports bei der In-situ-Sanierung mit festen Wärmequellen*, 2003, ISBN 3-933761-18-2
- 116 Marx, Walter: *Wasserkraft, Bewässerung, Umwelt - Planungs- und Bewertungsschwerpunkte der Wasserbewirtschaftung*, 2003, ISBN 3-933761-19-0
- 117 Hinkelmann, Reinhard: *Efficient Numerical Methods and Information-Processing Techniques in Environment Water*, 2003, ISBN 3-933761-20-4
- 118 Samaniego-Eguiguren, Luis Eduardo: *Hydrological Consequences of Land Use / Land Cover and Climatic Changes in Mesoscale Catchments*, 2003, ISBN 3-933761-21-2
- 119 Neunhäuserer, Lina: *Diskretisierungsansätze zur Modellierung von Strömungs- und Transportprozessen in geklüftet-porösen Medien*, 2003, ISBN 3-933761-22-0
- 120 Paul, Maren: *Simulation of Two-Phase Flow in Heterogeneous Poros Media with Adaptive Methods*, 2003, ISBN 3-933761-23-9
- 121 Ehret, Uwe: *Rainfall and Flood Nowcasting in Small Catchments using Weather Radar*, 2003, ISBN 3-933761-24-7
- 122 Haag, Ingo: *Der Sauerstoffhaushalt staugeregelter Flüsse am Beispiel des Neckars - Analysen, Experimente, Simulationen -*, 2003, ISBN 3-933761-25-5
- 123 Appt, Jochen: *Analysis of Basin-Scale Internal Waves in Upper Lake Constance*, 2003, ISBN 3-933761-26-3

- 124 Hrsg.: Schrenk, Volker; Batereau, Katrin; Barczewski, Baldur; Weber, Karolin und Koschitzky, Hans-Peter: *Symposium Ressource Fläche und VEGAS - Statuskolloquium 2003, 30. September und 1. Oktober 2003*, 2003, ISBN 3-933761-27-1
- 125 Omar Khalil Ouda: *Optimisation of Agricultural Water Use: A Decision Support System for the Gaza Strip*, 2003, ISBN 3-933761-28-0
- 126 Batereau, Katrin: *Sensorbasierte Bodenluftmessung zur Vor-Ort-Erkundung von Schadensherden im Untergrund*, 2004, ISBN 3-933761-29-8
- 127 Witt, Oliver: *Erosionsstabilität von Gewässersedimenten mit Auswirkung auf den Stofftransport bei Hochwasser am Beispiel ausgewählter Stauhaltungen des Oberrheins*, 2004, ISBN 3-933761-30-1
- 128 Jakobs, Hartmut: *Simulation nicht-isothermer Gas-Wasser-Prozesse in komplexen Kluft-Matrix-Systemen*, 2004, ISBN 3-933761-31-X
- 129 Li, Chen-Chien: *Deterministisch-stochastisches Berechnungskonzept zur Beurteilung der Auswirkungen erosiver Hochwasserereignisse in Flusstauhaltungen*, 2004, ISBN 3-933761-32-8
- 130 Reichenberger, Volker; Helmig, Rainer; Jakobs, Hartmut; Bastian, Peter; Niessner, Jennifer: *Complex Gas-Water Processes in Discrete Fracture-Matrix Systems: Upscaling, Mass-Conservative Discretization and Efficient Multilevel Solution*, 2004, ISBN 3-933761-33-6
- 131 Hrsg.: Barczewski, Baldur; Koschitzky, Hans-Peter; Weber, Karolin; Wege, Ralf: *VEGAS - Statuskolloquium 2004*, Tagungsband zur Veranstaltung am 05. Oktober 2004 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2004, ISBN 3-933761-34-4
- 132 Asie, Kemal Jabir: *Finite Volume Models for Multiphase Multicomponent Flow through Porous Media*. 2005, ISBN 3-933761-35-2
- 133 Jacoub, George: *Development of a 2-D Numerical Module for Particulate Contaminant Transport in Flood Retention Reservoirs and Impounded Rivers*, 2004, ISBN 3-933761-36-0
- 134 Nowak, Wolfgang: *Geostatistical Methods for the Identification of Flow and Transport Parameters in the Subsurface*, 2005, ISBN 3-933761-37-9
- 135 Süß, Mia: *Analysis of the influence of structures and boundaries on flow and transport processes in fractured porous media*, 2005, ISBN 3-933761-38-7
- 136 Jose, Surabhin Chackiath: *Experimental Investigations on Longitudinal Dispersive Mixing in Heterogeneous Aquifers*, 2005, ISBN: 3-933761-39-5
- 137 Filiz, Fulya: *Linking Large-Scale Meteorological Conditions to Floods in Mesoscale Catchments*, 2005, ISBN 3-933761-40-9

- 138 Qin, Minghao: *Wirklichkeitsnahe und recheneffiziente Ermittlung von Temperatur und Spannungen bei großen RCC-Staumauern*, 2005, ISBN 3-933761-41-7
- 139 Kobayashi, Kenichiro: *Optimization Methods for Multiphase Systems in the Sub-surface - Application to Methane Migration in Coal Mining Areas*, 2005, ISBN 3-933761-42-5
- 140 Rahman, Md. Arifur: *Experimental Investigations on Transverse Dispersive Mixing in Heterogeneous Porous Media*, 2005, ISBN 3-933761-43-3
- 141 Schrenk, Volker: *Ökobilanzen zur Bewertung von Altlastensanierungsmaßnahmen*, 2005, ISBN 3-933761-44-1
- 142 Hundecha, Hirpa Yesheatesfa: *Regionalization of Parameters of a Conceptual Rainfall-Runoff Model*, 2005, ISBN: 3-933761-45-X
- 143 Wege, Ralf: *Untersuchungs- und Überwachungsmethoden für die Beurteilung natürlicher Selbstreinigungsprozesse im Grundwasser*, 2005, ISBN 3-933761-46-8
- 144 Breiting, Thomas: *Techniken und Methoden der Hydroinformatik - Modellierung von komplexen Hydrosystemen im Untergrund*, 2006, 3-933761-47-6
- 145 Hrsg.: Braun, Jürgen; Koschitzky, Hans-Peter; Müller, Martin: *Ressource Untergrund: 10 Jahre VEGAS: Forschung und Technologieentwicklung zum Schutz von Grundwasser und Boden*, Tagungsband zur Veranstaltung am 28. und 29. September 2005 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2005, ISBN 3-933761-48-4
- 146 Rojanschi, Vlad: *Abflusskonzentration in mesoskaligen Einzugsgebieten unter Berücksichtigung des Sickerraumes*, 2006, ISBN 3-933761-49-2
- 147 Winkler, Nina Simone: *Optimierung der Steuerung von Hochwasserrückhaltebecken-systemen*, 2006, ISBN 3-933761-50-6
- 148 Wolf, Jens: *Räumlich differenzierte Modellierung der Grundwasserströmung alluvialer Aquifere für mesoskalige Einzugsgebiete*, 2006, ISBN: 3-933761-51-4
- 149 Kohler, Beate: *Externe Effekte der Laufwasserkraftnutzung*, 2006, ISBN 3-933761-52-2
- 150 Hrsg.: Braun, Jürgen; Koschitzky, Hans-Peter; Stuhmann, Matthias: *VEGAS-Statuskolloquium 2006*, Tagungsband zur Veranstaltung am 28. September 2006 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2006, ISBN 3-933761-53-0
- 151 Niessner, Jennifer: *Multi-Scale Modeling of Multi-Phase - Multi-Component Processes in Heterogeneous Porous Media*, 2006, ISBN 3-933761-54-9
- 152 Fischer, Markus: *Beanspruchung eingeeerdeter Rohrleitungen infolge Austrocknung bindiger Böden*, 2006, ISBN 3-933761-55-7

- 153 Schneck, Alexander: *Optimierung der Grundwasserbewirtschaftung unter Berücksichtigung der Belange der Wasserversorgung, der Landwirtschaft und des Naturschutzes*, 2006, ISBN 3-933761-56-5
- 154 Das, Tapash: *The Impact of Spatial Variability of Precipitation on the Predictive Uncertainty of Hydrological Models*, 2006, ISBN 3-933761-57-3
- 155 Bielinski, Andreas: *Numerical Simulation of CO₂ sequestration in geological formations*, 2007, ISBN 3-933761-58-1
- 156 Mödinger, Jens: *Entwicklung eines Bewertungs- und Entscheidungsunterstützungssystems für eine nachhaltige regionale Grundwasserbewirtschaftung*, 2006, ISBN 3-933761-60-3
- 157 Manthey, Sabine: *Two-phase flow processes with dynamic effects in porous media - parameter estimation and simulation*, 2007, ISBN 3-933761-61-1
- 158 Pozos Estrada, Oscar: *Investigation on the Effects of Entrained Air in Pipelines*, 2007, ISBN 3-933761-62-X
- 159 Ochs, Steffen Oliver: *Steam injection into saturated porous media – process analysis including experimental and numerical investigations*, 2007, ISBN 3-933761-63-8
- 160 Marx, Andreas: *Einsatz gekoppelter Modelle und Wetterradar zur Abschätzung von Niederschlagsintensitäten und zur Abflussvorhersage*, 2007, ISBN 3-933761-64-6
- 161 Hartmann, Gabriele Maria: *Investigation of Evapotranspiration Concepts in Hydrological Modelling for Climate Change Impact Assessment*, 2007, ISBN 3-933761-65-4
- 162 Kebede Gurmessa, Tesfaye: *Numerical Investigation on Flow and Transport Characteristics to Improve Long-Term Simulation of Reservoir Sedimentation*, 2007, ISBN 3-933761-66-2
- 163 Trifković, Aleksandar: *Multi-objective and Risk-based Modelling Methodology for Planning, Design and Operation of Water Supply Systems*, 2007, ISBN 3-933761-67-0
- 164 Götzing, Jens: *Distributed Conceptual Hydrological Modelling - Simulation of Climate, Land Use Change Impact and Uncertainty Analysis*, 2007, ISBN 3-933761-68-9
- 165 Hrsg.: Braun, Jürgen; Koschitzky, Hans-Peter; Stuhmann, Matthias: *VEGAS – Kolloquium 2007*, Tagungsband zur Veranstaltung am 26. September 2007 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2007, ISBN 3-933761-69-7
- 166 Freeman, Beau: *Modernization Criteria Assessment for Water Resources Planning; Klamath Irrigation Project, U.S.*, 2008, ISBN 3-933761-70-0

- 167 Dreher, Thomas: *Selektive Sedimentation von Feinstschwebstoffen in Wechselwirkung mit wandnahen turbulenten Strömungsbedingungen*, 2008, ISBN 3-933761-71-9
- 168 Yang, Wei: *Discrete-Continuous Downscaling Model for Generating Daily Precipitation Time Series*, 2008, ISBN 3-933761-72-7
- 169 Kopecki, Ianina: *Calculational Approach to FST-Hemispheres for Multiparametrical Benthos Habitat Modelling*, 2008, ISBN 3-933761-73-5
- 170 Brommundt, Jürgen: *Stochastische Generierung räumlich zusammenhängender Niederschlagszeitreihen*, 2008, ISBN 3-933761-74-3
- 171 Papafotiou, Alexandros: *Numerical Investigations of the Role of Hysteresis in Heterogeneous Two-Phase Flow Systems*, 2008, ISBN 3-933761-75-1
- 172 He, Yi: *Application of a Non-Parametric Classification Scheme to Catchment Hydrology*, 2008, ISBN 978-3-933761-76-7
- 173 Wagner, Sven: *Water Balance in a Poorly Gauged Basin in West Africa Using Atmospheric Modelling and Remote Sensing Information*, 2008, ISBN 978-3-933761-77-4
- 174 Hrsg.: Braun, Jürgen; Koschitzky, Hans-Peter; Stuhmann, Matthias; Schrenk, Volker: *VEGAS-Kolloquium 2008 Ressource Fläche III*, Tagungsband zur Veranstaltung am 01. Oktober 2008 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2008, ISBN 978-3-933761-78-1
- 175 Patil, Sachin: *Regionalization of an Event Based Nash Cascade Model for Flood Predictions in Ungauged Basins*, 2008, ISBN 978-3-933761-79-8
- 176 Assteerawatt, Anongnart: *Flow and Transport Modelling of Fractured Aquifers based on a Geostatistical Approach*, 2008, ISBN 978-3-933761-80-4
- 177 Karnahl, Joachim Alexander: *2D numerische Modellierung von multifraktionalem Schwebstoff- und Schadstofftransport in Flüssen*, 2008, ISBN 978-3-933761-81-1
- 178 Hiester, Uwe: *Technologieentwicklung zur In-situ-Sanierung der ungesättigten Bodenzone mit festen Wärmequellen*, 2009, ISBN 978-3-933761-82-8
- 179 Laux, Patrick: *Statistical Modeling of Precipitation for Agricultural Planning in the Volta Basin of West Africa*, 2009, ISBN 978-3-933761-83-5
- 180 Ehsan, Saqib: *Evaluation of Life Safety Risks Related to Severe Flooding*, 2009, ISBN 978-3-933761-84-2
- 181 Prohaska, Sandra: *Development and Application of a 1D Multi-Strip Fine Sediment Transport Model for Regulated Rivers*, 2009, ISBN 978-3-933761-85-9

- 182 Kopp, Andreas: *Evaluation of CO₂ Injection Processes in Geological Formations for Site Screening*, 2009, ISBN 978-3-933761-86-6
- 183 Ebigbo, Anozie: *Modelling of biofilm growth and its influence on CO₂ and water (two-phase) flow in porous media*, 2009, ISBN 978-3-933761-87-3
- 184 Freiboth, Sandra: *A phenomenological model for the numerical simulation of multiphase multicomponent processes considering structural alterations of porous media*, 2009, ISBN 978-3-933761-88-0
- 185 Zöllner, Frank: *Implementierung und Anwendung netzfreier Methoden im Konstruktiven Wasserbau und in der Hydromechanik*, 2009, ISBN 978-3-933761-89-7
- 186 Vasin, Milos: *Influence of the soil structure and property contrast on flow and transport in the unsaturated zone*, 2010, ISBN 978-3-933761-90-3
- 187 Li, Jing: *Application of Copulas as a New Geostatistical Tool*, 2010, ISBN 978-3-933761-91-0
- 188 AghaKouchak, Amir: *Simulation of Remotely Sensed Rainfall Fields Using Copulas*, 2010, ISBN 978-3-933761-92-7
- 189 Thapa, Pawan Kumar: *Physically-based spatially distributed rainfall runoff modelling for soil erosion estimation*, 2010, ISBN 978-3-933761-93-4
- 190 Wurms, Sven: *Numerische Modellierung der Sedimentationsprozesse in Retentionsanlagen zur Steuerung von Stoffströmen bei extremen Hochwasserabflussergebnissen*, 2011, ISBN 978-3-933761-94-1
- 191 Merkel, Uwe: *Unsicherheitsanalyse hydraulischer Einwirkungen auf Hochwasserschutzdeiche und Steigerung der Leistungsfähigkeit durch adaptive Strömungsmodellierung*, 2011, ISBN 978-3-933761-95-8
- 192 Fritz, Jochen: *A Decoupled Model for Compositional Non-Isothermal Multiphase Flow in Porous Media and Multiphysics Approaches for Two-Phase Flow*, 2010, ISBN 978-3-933761-96-5
- 193 Weber, Karolin (Hrsg.): *12. Treffen junger WissenschaftlerInnen an Wasserbauinstituten*, 2010, ISBN 978-3-933761-97-2
- 194 Bliefernicht, Jan-Geert: *Probability Forecasts of Daily Areal Precipitation for Small River Basins*, 2011, ISBN 978-3-933761-98-9
- 195 Hrsg.: Koschitzky, Hans-Peter; Braun, Jürgen: *VEGAS-Kolloquium 2010 In-situ-Sanierung - Stand und Entwicklung Nano und ISCO -*, Tagungsband zur Veranstaltung am 07. Oktober 2010 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2010, ISBN 978-3-933761-99-6

- 196 Gafurov, Abror: *Water Balance Modeling Using Remote Sensing Information - Focus on Central Asia*, 2010, ISBN 978-3-942036-00-9
- 197 Mackenberg, Sylvia: *Die Quellstärke in der Sickerwasserprognose: Möglichkeiten und Grenzen von Labor- und Freilanduntersuchungen*, 2010, ISBN 978-3-942036-01-6
- 198 Singh, Shailesh Kumar: *Robust Parameter Estimation in Gauged and Ungauged Basins*, 2010, ISBN 978-3-942036-02-3
- 199 Doğan, Mehmet Onur: *Coupling of porous media flow with pipe flow*, 2011, ISBN 978-3-942036-03-0
- 200 Liu, Min: *Study of Topographic Effects on Hydrological Patterns and the Implication on Hydrological Modeling and Data Interpolation*, 2011, ISBN 978-3-942036-04-7
- 201 Geleta, Habtamu Itafa: *Watershed Sediment Yield Modeling for Data Scarce Areas*, 2011, ISBN 978-3-942036-05-4
- 202 Franke, Jörg: *Einfluss der Überwachung auf die Versagenswahrscheinlichkeit von Staustufen*, 2011, ISBN 978-3-942036-06-1
- 203 Bakimchandra, Oinam: *Integrated Fuzzy-GIS approach for assessing regional soil erosion risks*, 2011, ISBN 978-3-942036-07-8
- 204 Alam, Muhammad Mahboob: *Statistical Downscaling of Extremes of Precipitation in Mesoscale Catchments from Different RCMs and Their Effects on Local Hydrology*, 2011, ISBN 978-3-942036-08-5
- 205 Hrsg.: Koschitzky, Hans-Peter; Braun, Jürgen: *VEGAS-Kolloquium 2011 Flache Geothermie - Perspektiven und Risiken*, Tagungsband zur Veranstaltung am 06. Oktober 2011 an der Universität Stuttgart, Campus Stuttgart-Vaihingen, 2011, ISBN 978-3-933761-09-2
- 206 Haslauer, Claus: *Analysis of Real-World Spatial Dependence of Subsurface Hydraulic Properties Using Copulas with a Focus on Solute Transport Behaviour*, 2011, ISBN 978-3-942036-10-8
- 207 Dung, Nguyen Viet: *Multi-objective automatic calibration of hydrodynamic models – development of the concept and an application in the Mekong Delta*, 2011, ISBN 978-3-942036-11-5
- 208 Hung, Nguyen Nghia: *Sediment dynamics in the floodplain of the Mekong Delta, Vietnam*, 2011, ISBN 978-3-942036-12-2
- 209 Kuhlmann, Anna: *Influence of soil structure and root water uptake on flow in the unsaturated zone*, 2012, ISBN 978-3-942036-13-9

- 210 Tuhtan, Jeffrey Andrew: *Including the Second Law Inequality in Aquatic Ecodynamics: A Modeling Approach for Alpine Rivers Impacted by Hydropeaking*, 2012, ISBN 978-3-942036-14-6
- 211 Tolossa, Habtamu: *Sediment Transport Computation Using a Data-Driven Adaptive Neuro-Fuzzy Modelling Approach*, 2012, ISBN 978-3-942036-15-3
- 212 Tatomir, Alexandru-Bodgan: *From Discrete to Continuum Concepts of Flow in Fractured Porous Media*, 2012, ISBN 978-3-942036-16-0
- 213 Erbertseder, Karin: *A Multi-Scale Model for Describing Cancer-Therapeutic Transport in the Human Lung*, 2012, ISBN 978-3-942036-17-7
- 214 Noack, Markus: *Modelling Approach for Interstitial Sediment Dynamics and Reproduction of Gravel Spawning Fish*, 2012, ISBN 978-3-942036-18-4
- 215 De Boer, Cjestmir Volkert: *Transport of Nano Sized Zero Valent Iron Colloids during Injection into the Subsurface*, 2012, ISBN 978-3-942036-19-1
- 216 Pfaff, Thomas: *Processing and Analysis of Weather Radar Data for Use in Hydrology*, 2013, ISBN 978-3-942036-20-7
- 217 Lebreinz, Hans-Henning: *Addressing the Input Uncertainty for Hydrological Modeling by a New Geostatistical Method*, 2013, ISBN 978-3-942036-21-4
- 218 Darcis, Melanie Yvonne: *Coupling Models of Different Complexity for the Simulation of CO₂ Storage in Deep Saline Aquifers*, 2013, ISBN 978-3-942036-22-1
- 219 Beck, Ferdinand: *Generation of Spatially Correlated Synthetic Rainfall Time Series in High Temporal Resolution - A Data Driven Approach*, 2013, ISBN 978-3-942036-23-8
- 220 Guthke, Philipp: *Non-multi-Gaussian spatial structures: Process-driven natural genesis, manifestation, modeling approaches, and influences on dependent processes*, 2013, ISBN 978-3-942036-24-5
- 221 Walter, Lena: *Uncertainty studies and risk assessment for CO₂ storage in geological formations*, 2013, ISBN 978-3-942036-25-2
- 222 Wolff, Markus: *Multi-scale modeling of two-phase flow in porous media including capillary pressure effects*, 2013, ISBN 978-3-942036-26-9
- 223 Mosthaf, Klaus Roland: *Modeling and analysis of coupled porous-medium and free flow with application to evaporation processes*, 2013, ISBN 978-3-942036-27-6
- 224 Leube, Philipp Christoph: *Methods for Physically-Based Model Reduction in Time: Analysis, Comparison of Methods and Application*, 2013, ISBN 978-3-942036-28-3
- 225 Rodríguez Fernández, Jhan Ignacio: *High Order Interactions among environmental variables: Diagnostics and initial steps towards modeling*, 2013, ISBN 978-3-942036-29-0

Die Mitteilungshefte ab der Nr. 134 (Jg. 2005) stehen als pdf-Datei über die Homepage des Instituts: www.iws.uni-stuttgart.de zur Verfügung.