

Analysis and Modelling of Mining Induced Seismicity

by

Ben Bredenkamp

Thesis submitted in partial fulfillment
of the requirements for the degree

of

Master of Science in Engineering Sciences

in the Department of Process Engineering
at the University of Stellenbosch

Supervised by

Professor C. Aldrich

Stellenbosch

Date December 2006

Declaration



I, the undersigned, hereby declare that the work contained in this thesis is my own original work and that I have not previously, in its entirety, or in part submitted it at any university for a degree.

Signature:

Date:

Abstract

Earthquakes and other seismic events are known to have catastrophic effects on people and property. These large-scale events are almost always preceded by smaller-scale seismic events called precursors, such as tremors or other vibrations. The use of precursor data to predict the realization of seismic hazards has been a long-standing technical problem in different disciplines. For example, blasting or other mining activities have the potential to induce the collapse of rock surfaces, or the occurrence of other dangerous seismic events in large volumes of rock. In this study, seismic data (T4) obtained from a mining concern in South Africa were considered using a nonlinear time series approach. In particular, the method of surrogate analysis was used to characterize the deterministic structure in the data, prior to fitting a predictive model.

The seismic data set (T4) is a set of seismic events for a small volume of rock in a mine observed over a period of 12 days. The surrogate data were generated to have structure similar to that of T4 according to some basic seismic laws. In particular, the surrogate data sets were generated to have the same autocorrelation structure and amplitude distributions of the underlying data set T4. The surrogate data derived from T4 allow for the assessment of some basic hypotheses regarding both types of data sets.

The structure in both types of data (i.e. the relationship between the past behavior and the future realization of components) was investigated by means of three test statistics, each of which provided partial information on the structure in the data. The first is the *average mutual information* between the reconstructed past and futures states of T4. The second is a *correlation dimension* estimate, D_c which gives an indication of the deterministic structure (predictability) of the reconstructed states of T4. The final statistic is the *correlation coefficients* which gives an indication of the predictability of the future behavior of T4 based on the past states of T4.

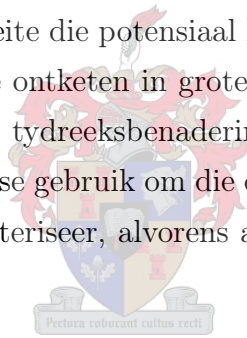
The past states of T4 was reconstructed by reducing the dimension of a delay coordinate embedding of the components of T4. The map from past states to future realization of T4 values was estimated using Long Short-Term Recurrent Memory (LSTM) neural networks. The application of LSTM Recurrent Neural Networks on point processes has not been reported before in literature.

Comparison of the stochastic surrogate data with the measured structure in the T4 data set showed that the structure in T4 differed significantly from that of the surrogate data sets. However, the relationship between the past states and the future realization of components for both T4 and surrogate data did not appear to be deterministic. The application of LSTM in the modeling of T4 shows that the approach could model point processes at least as well or even better than previously reported applications on time series data.



Ekserp

Die katastrofiese gevolge van aardbewings of ander seismiese gebeurtenisse op mens en eiendom is welbekend. Hierdie grootskaalse gebeurtenisse word amper altyd voorafgegaan deur kleiner-skaalse seismiese gebeurtenisse, soos trillings of ander vibrasies en die benutting van sulke data om die gerealiseerde seismiese risiko te voorspel, is 'n ou probleem sonder 'n duidelike oplossing. In hierdie studie was die seismiese data vanaf mynbedrywighede in Suid-Afrika oorweeg, waar die gevolge van uitgrawings en ander aktiwiteite die potensiaal het om rotswande te laat verbrokkel of ander seismiese gebeure te ontketen in groter rotsvolumes. Die studie is gedoen aan die hand van nie-lineêre tydreeksbenaderings. In die besonder is die tegniek bekend as surrogaatdataanalise gebruik om die deterministiese struktuur in die seismiese data stel (T4) te karakteriseer, alvorens a voorspellende model daarop gepas is.



Die seismiese datastel, T4, is a waargenome stel seismiese gebeure vir 'n beperkte rotsvolume in 'n myn, gemeet oor 'n periode van 12 dae. Surrogaatdatastelle was gegenerer om 'n struktuur soortgelyk aan T4 te hê, in ooreenstemming met 'n paar basiese seismiese wette. Surrogaatdatastelle is gegenerer om dieselfde autokorrelasiestruktuur en amplitudeverdeling te hê as die onderliggende datastel, T4. Met die surrogaatdata afgelei vanaf T4 kon 'n aantal basiese hipotetese aangaande die data getoets word.

Die struktuur in albei soorte data (i.e. die verwantskap tussen die pas afgelope gedrag en die daaropvolgende gedrag) is bestudeer aan die hand van drie toetsstatistieke, waarvan elkeen gedeeltelike inligting verskaf oor die struktuur in die data. Die eerste is die *gemiddelde wedersydse inligting* tussen die herwinde afgelope toestande en daaropvolgende herwinde toestande van 'n datastel. Die tweede toetsstatistiek is 'n beraming van die *korrelasiedimensie*, D_c , wat inligting verskaf oor die deterministiese struktuur (voorspelbaarheid) van herwinde toestande van 'n datas-

tel. Die laaste statistiek is die *korrelasie koëffisiënt* wat 'n aanduiding gee van die voorspelbaarheid van daaropvolgende gedrag waargeneem in T4 en die pas afgelope toestande. Die vloei van toestande in T4 is herwin met 'n dimensievermindering van die tydverwante komponentverpakking van T4. Die passing van die afbeelding van die afgelope toestande na die daaropvolgende gedrag van T4 is beraam met lang kort-termyn terugvoer-neurale netwerke (LKTT).

In die vergelyking tussen die stochastiese surrogaat- en seismiese data, skei die seismiese data betekenisvol van die surrogaatdatastelle. Die skeiding dui daarop dat verdere ondersoek na die strukture in T4 geregverdig is. Desnieteenstaande wys die verwantskap tussen die afgelope herwinde toestande en die daaropvolgende gedrag nie die kenmerkende tekens wat met 'n deterministiese struktuur verwag word nie.

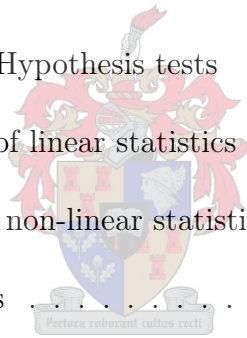
Die toepassing van LSTM Recurrent Neural Networks op puntprosesse is tot op hede nie gerapporteer in die literatuur nie. Die toepassing van LKTT in die modellering van T4 wys dat LKTT puntprosesse kan modelleer. Die toepassing van LKTT op T4 het trouens beter resultate behaal as wat voorheen vir tydreeksdata gerapporteer is.



Contents

1	Introduction	2
1.1	Motivation and Problem Statement	2
1.2	Layout of the thesis	4
2	Literature Review: Modelling Mining-Induced Seismicity	6
2.1	Background	6
2.2	The nature of mining-induced seismicity	7
2.3	The relationship between mining and mining-related seismicity	10
2.4	A framework for modelling crust-scale seismicity	12
2.5	Modelling of mining-induced seismicity from a crust-scale perspective	17
2.6	Concluding Remarks	18
3	Analysis of seismic data using the method of surrogates	20
3.1	Introduction	20
3.2	Definitions and measurements	21
3.2.1	Sampling the seismic dataset, T4	21
3.2.2	Previous work - ISSI's investigation	23
3.2.3	Seismic Laws	25
3.2.4	Linear statistics	31

3.2.5	Embedding	35
3.2.6	Independent component analysis	37
3.2.7	Discriminating statistics	38
3.2.8	Surrogate data	41
3.2.9	Hypothesis testing	45
3.2.10	Generating synthetic seismic data	47
3.2.11	ISSI's attractor analysis	48
3.3	The seismic behavior of T4	51
3.3.1	Seismic laws and stationarity	51
3.3.2	Nonlinear measures on attractors	59
3.4	Surrogate data and Hypothesis tests	65
3.4.1	Comparison of linear statistics	67
3.4.2	Comparative non-linear statistics and hypothesis tests	69
3.5	Concluding Remarks	77
4	Modelling T4 using Long Short-Term Memory	80
4.1	Introduction	80
4.2	Modelling Using Mean LSTM Networks	82
4.3	Data set training and validation	84
4.4	Estimating the components of T4	88
4.5	Hypothesis tests on correlation coefficients	91
4.6	Concluding Remarks	96
5	Conclusions	97
A	The rule derivations for the LSTM network	99



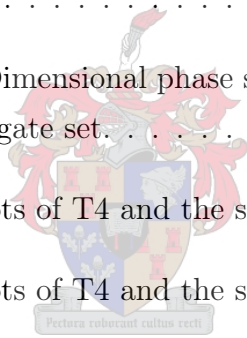
A.1 LSTM forward pass	99
A.2 LSTM backward pass	104
A.3 Functional LSTM correctness	107
B Algorithm listing	109
B.1 Code for ICA transforms	109
B.2 Code for the surrogate computation	113



List of Figures

1.1	A coarse mind map of the study	5
3.1	The seismic moment and energy point process of the T4 data set as a function of time.	23
3.2	The location of seismic events in the T4 data set.	24
3.3	The interaction between the cumulative apparent volume and the energy index prior to a large seismic event.	26
3.4	An explanation of the frequency - size relationship in a seismic data set.	28
3.5	Selecting a sequence of events in plotting the trend for seismic event clustering.	29
3.6	An explanation of plotting the trend for seismic event clustering. . . .	30
3.7	A representation of a comparison between the surrogate data and the observations using a quantile-quantile plot.	33
3.8	An explanation of the autocorrelation vs. autocorrelation comparison. . .	34
3.9	A graphical representation of two ICA attractors.	36
3.10	A graphical representation of what the average mutual information measures.	40
3.11	An schematic explanation for a hypothesis test	46
3.12	The time interval between two consecutive events, for all the events in the T4 Data set.	49

3.13	The D_c estimates of a set of surrogate data sets and the T4 data set's inter-event interval-times.	50
3.14	The GR relationships in the components of the T4 data set.	55
3.15	T4, clustering of large events for event sequences of length 10.	56
3.16	The lack of event clustering in the event sequences of length 10.	57
3.17	The cross correlations between the components of the T4 dataset.	58
3.18	The AMI_J scores for the norm of the 10 dimensional ICs of the T4 dataset as a delay lag in J	61
3.19	$D_c(\epsilon/\epsilon_o)$ estimates of the components of the T4 dataset.	64
3.20	A 3-dimensional plot of the components of T4 and the components of a surrogate set.	66
3.21	The norms of a 10-Dimensional phase space reconstruction of the T4 data set and a surrogate set.	67
3.22	q-q plot and Acc plots of T4 and the surrogates for Dt_i	68
3.23	q-q plot and Acc plots of T4 and the surrogates for the $\log(M_i)$	69
3.24	q-q plot and the Acc plot for T4 and its surrogates for $\log(E_i)$	70
3.25	A comparison between the cross-correlations of the T4 data set and its surrogates	71
3.26	Surrogate analysis, hypothesis test of Dt_i	72
3.27	Surrogate analysis, hypothesis test of $\log(M_i)$	73
3.28	Surrogate analysis, hypothesis test of $\log(E_i)$	74
3.29	Surrogate analysis, hypothesis test of $(Dt_i, \log(M_i))$	75
3.30	Surrogate analysis, hypothesis test of $(Dt_i, \log(E_i))$	76
3.31	Surrogate Hypothesis test for $(Dt_i, \log(M_i), \log(E_i))$	77
3.32	Surrogate Hypothesis test for $(Dt_i, \log(M_i), \log(E_i))$	78



4.1 A schematic representation of the 7 steps for modeling and evaluating the components of T4. 83

4.2 The input-output map for modeling T4 with LSTM. 85

4.3 An error per epoch plot for an LSTM training sequence. 86

4.4 Training and validation errors for the selected LSTM networks. 87

4.5 The number of times a \bar{Y}_i vector was selected for training or for validation. 88

4.6 A plot of the estimate of $S(Dt_{i+1})$ in comparison to the actual values. 89

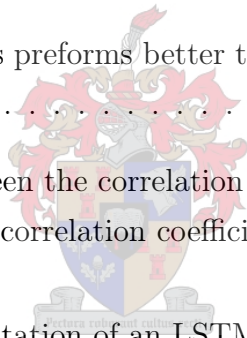
4.7 The autocorrelation coefficient estimates for the modeled components of the T4 dataset. 90

4.8 The cross correlation coefficients for the modeled components. 91

4.9 The estimator of \bar{Y}_i s preforms better than the autocorrelation structure. 93

4.10 The difference between the correlation coefficients of the LSTM estimators and the autocorrelation coefficients of the T4 dataset. 95

A.1 A graphical representation of an LSTM network. 102



Acknowledgements

*“ ‘I tell you what!’ said the Donkey brightly.
‘Perhaps it’s an animal that can’t talk but thinks it can.’ ”*
The Magician’s Nephew, C.S. Lewis

I would like to thank:

- God the Father, for giving me and helping me to complete this task.
- Mum and Dad, friends and relatives for their support and encouragement.
- Chris Aldrich for his patience and wisdom.
- ISSI for their data and help at the onset of the analysis.
- The NRF for their part in funding the study.
- Gordon Jemwa, Astrid Sindle, Miss L. Overbeek, Reg Dodds and Gabriel Cillié for their help in writing the thesis.
- Steve Kroon, Prof. De Villiers and Tannie Amanda Botha for residence on Stellenbosch.
- Andrew M. Inggs, for the use of his Tex thesis layout.

I would also like to express my gratitude to the University of Stellenbosch where I conducted the study and whose support provided the means.

Nomenclature

μ	The rigidity of a volume of rock.
AC_n^{data}	The range of autocorrelation coefficient for the data set in lag n .
AC_n^i	The range of autocorrelation coefficient for surrogate set i in lag n .
α	A portion of probability in the tail of a distribution, the alpha error in a hypothesis test.
s_i^{10}	A set of 10-dimensional ICA scores, iterated in event order by i .
$AMI(.,.)$	The average mutual information measure between two data sets.
AMI_J	The AMI score of a time ordered variable with itself, separated by a lag of J .
D_c	A shorthand for $D_c(\epsilon/\epsilon_0)$ if specifying the scales is not required.
$D_c(\epsilon/\epsilon_0)$	The Judd correlation dimension at the standardized distance of the inter-point distance distribution of a sampled attractor
Dt_k	A time interval between events k and $k+1$ in T4, after the small event submission.
E_i	Seismic energy released, i 'th in the sequence of seismic events.
ϵ	A small length in the inter-point distance distribution of a sampled attractor
ϵ/ϵ_0	A length in the inter-point distance distribution of a sampled attractor standardized to length ϵ_0 .
f	The range of frequencies
f_0	The zero frequency in a range of frequencies

$\overline{(Loc_i)}$	The location estimate for a seismic event i .
$\xrightarrow{LSTM_k}$	An LSTM map, optimized on a training set for the k 'th training and validation set deviation.
M_i	Seismic moment, i^{th} in the sequence of seismic events.
NPE_k	The normalized prediction error for a set of outputs for an LSTM network over a sequence of desired outputs.
$P_D(.,.)$	A discrete, unspecified, two-variable probability distribution function.
$\overline{P_i}$	The observed event in a point process of indexed events.
$P_M(.,.)$	A discrete, unspecified, one-variable probability distribution function.
$P_S(.)$	A discrete, unspecified, one-variable probability distribution function.
R	A correlation coefficient.
ρ	A specified correlation coefficient.
R_{si}	The correlation coefficients for components in T4 i with a mean LSTM estimator of the training or validation set, s .
$S(.)$	The function used to scale the desired vectors of the LSTM map, to with in the range of the network's output units.
$\overline{Seis_i}$	A set of embedded T4 components, iterated in event order by i .
s_n	A sample from a surrogate population, red colored noise.
Sur	The population of sampled stochastic systems, each system adhering to a specified CDF and autocorrelation structure.
sur_i	A sample from a surrogate population, Sur , random variable.
t	A test statistic with a Student T-distribution.
W	A short and for $W_{10,D_{embed}}$ when the size of the matrix is unimportant.
$W_{10,D_{embed}}$	A matrix of 10 rows and D_{embed} columns, separating $\overline{Seis_i}$ into principal components.

- $\overline{X}_{sk}(t)$ An input vector for LSTM, t 'th in the sequence of inputs, for the k 'th selection of a training and validation set samples and an element of the s 'th one of the two.
- \overline{Y}_i A desired output for LSTM, i 'th in the time ordered sequence of events.
- $\overline{Y}_{sk}(t)$ An output vector for LSTM, t 'th in the sequence of inputs, for the k 'th selection of a training and validation set samples and an element of the s 'th one of the two.
- $z(.,.)$ A random variable sampled on the space of correlation coefficient pairs. Normal distributed if the pair are the same.



Chapter 1

Introduction

Seismic events, for example earth tremors induced by plate tectonic movements in the earth's crust or blasting in mining activities, pose catastrophic hazards to people and property. An accurate assessment of seismic hazards associated with a volume of rock is therefore an important problem. Since these hazards are realized through seismic events, the problem entails the location of the volume of rock in which a seismic event will occur, as well as the timing of the events associated with the hazard. In the case of mining activities, it must be established when a volume of rock becomes potentially unsafe, as all the seismic events above a certain threshold may realize the hazard. To this end, mining operations use seismic monitoring systems to predict the probability of a seismic event occurring in rock mass.

1.1 Motivation and Problem Statement

Seismic signals are inherently difficult to interpret and there is as yet no universal approach that can be used to successfully analyse these signals. Hence, research is being undertaken in different directions in an effort to explore plausible interpretations of seismic data. In this thesis, the feasibility of using an approach inspired by nonlinear dynamics theory in the analysis of seismic data was investigated. In particular, state space reconstruction methods are proposed to quantify predictive structure underlying the occurrence of seismic events in a seismic event sequence (Kantz and Schreiber, 1997).

In nonlinear dynamics theory, the evolution of a system is given by a functional relationship among the rates of change of the underlying variables of the system with

respect to time, commonly referred to as the state space. However, in most cases, the governing variables of a system are either unknown or difficult to access. Instead, what is usually observed are other variables that contain information on the hidden variables. The state space reconstruction method uses measured observations to derive a functional relationship governing the evolution of the system. The success of state space reconstruction methods depends on the levels of uncertainties associated with the measurement process and how well the observations sample the state space. To assess how good a reconstruction is, nonlinear statistical quantities including among other, average mutual information, correlation dimension and predictability are used.

An implicit assumption in state space reconstruction is that the observations fully capture the dynamics driving the system. However, if the observations do not fully reflect the state space of the system, the reconstruction may yield an incorrect state space. In that case, the only structures observed are the distribution and correlations in the data. Therefore, incorrect conclusions may be reached on the basis of such a state space reconstruction. To minimize this risk, the method of surrogate data analysis is used as a mechanism for testing the presence of non-trivial structures in the data (Schreiber and Schmitz, 2000).

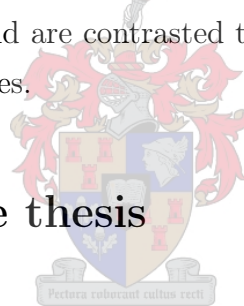
Surrogate data analysis is a technique for sampling a number of reconstructed state spaces similar to that of the data set, under the assumption that the worst-case scenario (that is, only trivial structures are present in data) is true. If the worst-case scenario is not applicable to the data set its nonlinear statistics would differ from those of the surrogate set. State space reconstruction can then be used for further analysis of the observed data. In particular, if a state space driving the system can be observed, the future behavior of the system can be modelled using the current observations from the system.

Predictability is a natural test of evidence of determinism in observed data. Fitting of reconstructed state space data to a model requires appropriate modelling tools. In this study long short-term memory recurrent neural networks (LSTM) are used (Hochreiter and Schmidhuber, 1991). LSTM is an experimental recurrent neural network which has been shown to give improved performance over other types of recurrent neural networks. The LSTM network has been demonstrated mostly on benchmark examples, but little work has been done in extending to real-world problems, especially those that can exploit the special properties of LSTM.

In the case of mining-induced seismicity, the observable functional relationships are still unknown (Helmstetter and Sornette, 2002; Mendecki et al., 1997). Therefore, it is not yet possible to define a state space using analytical approaches (Gere and Shan, 1984; Rikitake, 1976). Although an exact state space is yet to be established, this study approximates the state-space driving mining induced seismicity via reconstruction methods. ISSI (ISSI©, 2003) is a company based in South Africa that has established itself as a centre of excellence in monitoring and analysis of seismic activity in mines. It is involved in the development and maintenance of seismic monitoring systems. Additionally, they also provide analytical software to assist in the interpretation of signals obtained from the monitoring systems.

Figure 1.1 shows a mind map for visualizing the different steps in the study outlined above. At the top of Figure 1.1 are the mining-induced seismic data and corresponding surrogate data generated according to some specified hypothesis. In this case, the hypothesis is that state space reconstruction captures similar structure in the seismic data as the surrogate data. Both sets are passed through the state space reconstruction mechanism and are contrasted through the observed flow properties of the constructed state spaces.

1.2 Layout of the thesis



The rest of the thesis is organized as follows. In Chapter 2, a literature review of predictive modelling of mining-induced and crust-scale seismicity is presented. To reconstruct a state space from a seismic system requires the quantification and separation of the stochastic and structural components in the system. The predictive modeling of seismicity likewise requires separating the two components. The literature overview provides the motivation of this study and demonstrates its contribution to the field. An attempt is made in establishing a common ground between the mining-induced seismicity and crust-scale seismicity, since these two areas of study appear to have developed rather separately. The method of surrogate data analysis for the testing of hypotheses is discussed and applied to the seismic data in Chapter 3. Chapter 4 considers the accuracy of the functional map realized by the state space flow as expressed by a mean estimator based on long short-term memory (LSTM) recurrent neural networks. The predictability in T4 modelled by LSTM is compared to the optimal theoretical bound that can be expected from red colored noise. Finally, Chapter 5 concludes and highlights the contributions of the study.

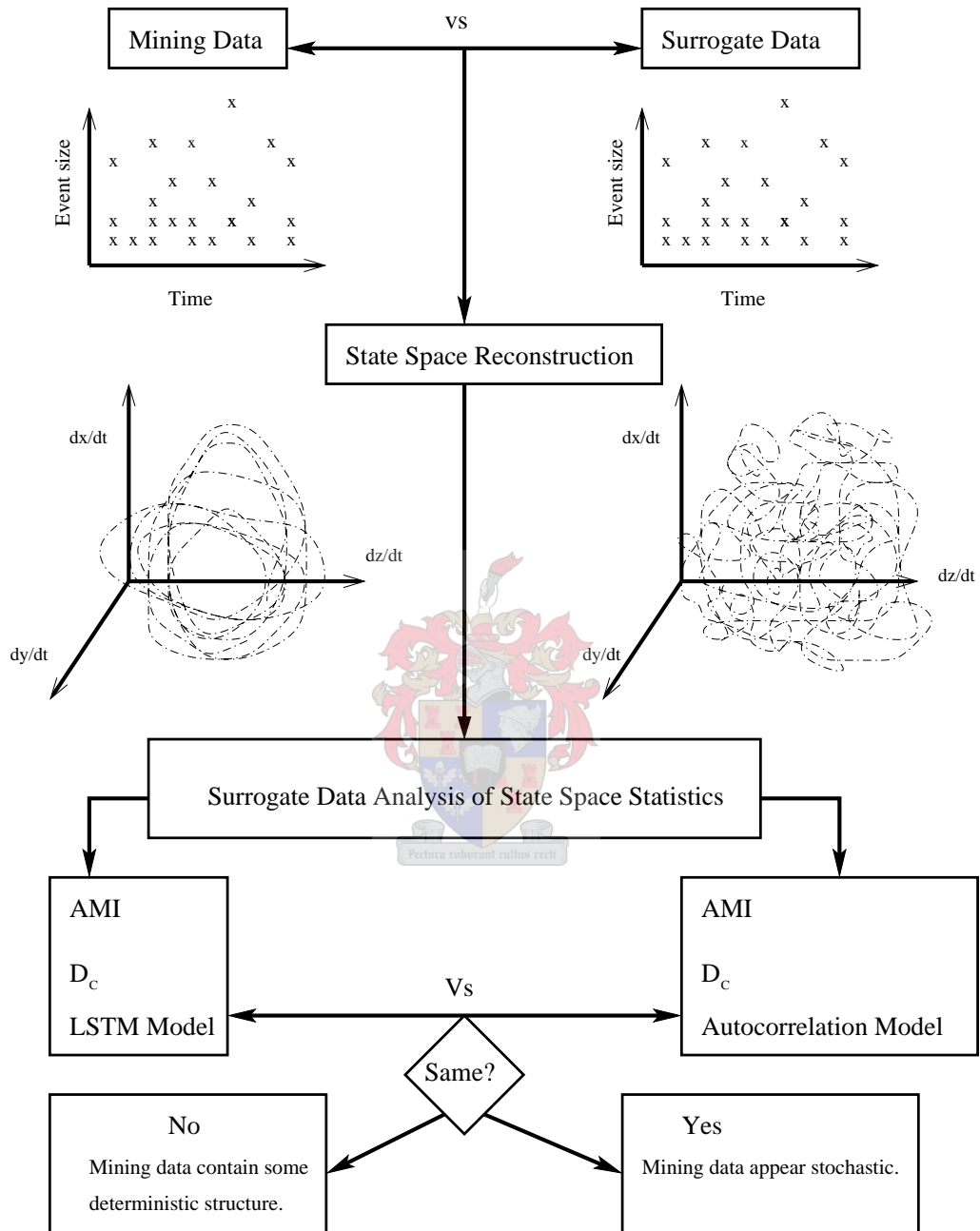


Figure 1.1: A mind map of the study. Starting at the top are the mining induced seismic data in contrast with the point process of the surrogate data. The surrogate data are autocorrelated and scaled noise, sampled from a continuous distribution.

Chapter 2

Literature Review: Modelling Mining-Induced Seismicity

2.1 Background

The prediction of significant seismic events produced by the earth's crust is a major challenge that has been facing geologists for quite some time (Gere and Shan, 1984; Rikitake, 1976). Solving the problem for short time scales has so far proved very difficult. Some researchers have even concluded that the initiation of seismic events is inherently random and, therefore, unpredictable (Barriere and Turcotte, 1994; Geller et al., 1997; Hooge et al., 1994). Several authors have noted the appearance of precursors preceding large crust-generated seismic events over long-time observation periods (Eneva and Ben-Zion, 1997b; Habermann, 1981; Kanamori, 1981; Rikitake, 1976; Vorobieva, 1999). This has led to speculations of alternative approaches to the prediction problem for long- to medium-time scales (Sornette, 2000). In the closely related field of modelling seismic generating processes, some advances have been made and ballpark estimates of the location, time (Ben-Zion, 1996) and hazard estimates of seismic events have improved (Darpahi-Noubary, 2002). It has been argued that proper seismic hazard estimation during the development of infrastructure requires only knowledge of the statistical nature of earthquakes (Gere and Shan, 1984). One school of thought further argues that most of the resources used in the study of earthquakes should be spent on minimizing seismic hazards – something which is necessary despite the poor predicatibility of earthquakes (Geller et al., 1997). Minimizing seismic hazard is apparently an easier task than solving the prediction problem, especially since the prediction problem is intractable.

Precursors to large mining-induced seismic events and rock burst have been predicted through turbulence in estimated rock flow parameters (Mendecki et al., 1997). Another study reported that estimates of at least five (5) different rock flow parameters are necessary to give a decisive indicator (Poplawski, 1997). However, establishing reliable indicators is made difficult by the inconsistency of trends in the precursors to dangerous seismic events because similar types of flow in observed seismic data do not always result in the same behavior (Eneva, 1998). These issues, together with the inherent difficulties of interpreting seismicity, result in a very low accurate prediction rate. On the other hand, modelling failure process that occur in the mining environment has proved to be fruitful (McGarr, 2000). These investigations have resulted in, among other, better establishment of seismic hazards during mining (Cai et al., 2001; Fujii et al., 1997; Mansurov, 2001) as well as improvements in mine engineering techniques (Pytel, 2003).

Modelling and establishing precursors in mining-related seismicity lies between the controlled environment of rock-breaking experiments and the problem of large-scale earthquakes. In mining-induced seismicity similar parameters as in global or crust-scale seismicity are monitored, and similar interpretation problems are encountered (Kagan and Vere-Jones, 1996). Consequently, these parallels provide an opportunity to formulate and test hypotheses for crust-scale seismicity using what has been learned in mining-induced seismicity (Eneva, 1998; Pollard, 2000).



2.2 The nature of mining-induced seismicity

Characteristics or plausible governing physical laws exhibited by observed mining-induced seismicity have been proposed in the literature. These physical laws form the basis of modelling or hypothesis testing regarding the system since they dictate specific types of correlations between the variables representing the system as well as invariance of the test statistics. Establishing the cause-effect relationships is an important subject in the literature and helps to pin-point further areas of investigation (Helmstetter and Sornette, 2002; Pollard, 2000).

Mining-induced seismicity is typically represented by a point process in combination with a location parameter. Each point in space and time represents a transition from elastic strain to inelastic strain, with an observed seismic energy release and seismic moment. An elastic solid described by a tensor field relating its stress and

strain fields fails when the stress exceeds a given threshold, with the strain field subsequently describing a relatively large acceleration. In a homogeneous rock sample a failure is a repeatable experiment. In a heterogeneous sample set-up, the stress-strain relationship behaves differently from the homogeneous case (Feynman et al., 1989; Rikitake, 1976).

Although moment and energy estimates associated with a seismic event are inherently inaccurate, they are important for comparing different events in the system. Locating the source of the event may also be inaccurate as a result of the triangulation process. Non-stationarity invariably occurs in the observed data since the seismic monitoring system can only detect events above a set minimum lower bound and, also requires frequent adjustments due to drift in its accuracy. Many approaches for dealing with the sampling noise problem have been proposed literature (Bodri, 2001; Darpahi-Noubary, 2002).

If a volume of rock has been mined out of a rock mass and the deformed rock mass can only reach equilibrium through seismic activity, the sum of the seismic moments (M_i) will be proportional to the mined out volume (V_m), i.e

$$\sum_{t_0}^{t_{\text{inf}}} M_i \propto \mu V_m \quad (2.1)$$



for the time interval ($t_0 - t_{\text{inf}}$), and where μ is normalizing constant known as rigidity.

The seismic events are driven by the closure of nearby stopes, which accumulate energy in proportion to the product of the volume of closure and the overhead burden stress normal to the plane of the stope. In mining most of the pent-up energy is used for creating the fault necessary for the seismic moment to occur. Seismic energy release is considered inefficient (McGarr, 2000; Pytel, 2003).

The apparent volume is an estimate of the volume of rock that was deformed to generate the occurrence of the seismic moment and is given by

$$V_{Ai} = \frac{M_i^2}{2\mu E_i} \quad (2.2)$$

indexed over the sequence of events. The accumulated apparent volume for the seismic events generated in a given volume of rock is used to estimate the rate

of deformation that occurs in that volume due to the smaller seismic events. A positive correlation exists between co-seismic and a-seismic deformation. As a result, cumulative apparent volume over a given volume provides insight into the rate and distribution of co-seismic deformation.

No precursory behavior is demonstrated by the cumulative plots of scalar terms, if the scalar terms provide only a one-dimensional description (viz. E_i and M_i used in this study) of each of the the seismic event sources in the cumulative sum.

A log-log plot of seismic moment versus the energy release demonstrates a linear relationship. The parameters for the relationship do not have universal values. The ratio between the energy released and the expected energy release given the event's seismic moment, forms part of the precursory behavior of the system.

The time independent size distribution of the mining seismicity follows the Gutenberg-Richter power law (Section 3.2.3). In the time order event sequence, foreshock and aftershocks to large events are demonstrated. Both of the sequences obey a power law and the frequency of aftershocks decays according to Omori's law with an exponent slightly larger than unity (see below). Four parameters have been reported to be independent: the average time between seismic events; the average distance between consecutive seismic events; the sum of the seismic moments; and the sum of the seismic energies (Mendecki et al., 1997; Poplawski, 1997).

Mining induced seismic events are localized in both time and space. Space localization is characterized by a process of nucleation in the region of the impending event. Such Clusters of seismic events are characterized by their correlation dimension. A correlation dimension larger than two is an indicator of a space-filling nucleation while a correlation dimension of less than one is related to a plane-filling nucleation (Eneva, 1998).

In the preceding paragraphs, many relationships between parameters and associated invariant properties have been highlighted. These properties have been used and applied in the mining environment to monitor seismic hazards. These relationships have also been included in the body of knowledge representing seismicity in general.

2.3 The relationship between mining and mining-related seismicity

Realized seismic hazards are a major risk on mining operations. As mining activities become more extensive and deeper, the risk of seismic hazards also increases. Therefore, it is essential to have a better understanding of the underlying seismic dynamics to minimise risk associated with the hazards (Vieira et al., 2001). In the following sections a summary of different approaches to estimating current failure properties and/or projections of future failure properties of the rock mass in mining operations is given.

The system generating seismic events can be divided into different scales depending on the type of questions being asked. On a small scale, questions arise about specific pillars or rock faces. Is this pillar building up stress to create a violent rock burst? How do pillars fail gracefully? Is a specific rock face preparing for a rock burst? How is this fault being influenced by the oncoming stope? The next scale of influence on the seismic hazard lies in the combination of pillars and stopes. What is the best strategy to mine this pocket of ore? Or, what strategy will lead to successful rock mass softening and stope closure? On a large scale, the interest is in how the mining activity interacts with the regional rock mass. Large events could be the result of horizontal layers separating as the stope is closed (Pytel, 2003). McGarr (2000) postulate that even though a surrounding rock mass maybe close to failure, mining activity initially lowers the pore pressure and has a stabilizing effect on the surrounding rock mass. Continuing excavations concentrate pressure gradients around the stope, resulting in instabilities. To resolve these and related questions, different research directions centered around the interpretation of seismic events observed on the different scales have been discussed in literature.

The knowledge-based approach relates fundamental rock mass and source models to observations. Subsequently, these source models are related to large scale fundamental properties of current or future states of failure. Using this approach, Mendecki et al. (1997) established the rate of deformation in a volume of rock or pressure gradient. The pressure gradient is inferred through the notion of dilatancy in Hanson et al. (2002) using computer aided tomography (CAT). McGarr (2000) divided mining seismicity into events associated with stope closure and events on nearby faults due to changes in the regional stress-strain relationship.

In some cases the observed system may lack sufficient information to fit all the degrees of freedom in the model. One can then resort to bootstrap statistical methods that combine the knowledge-based approach with numerical simulation. The parametric space used in the model is approximated and discretized for the numerical simulation, resulting in units of interaction. These units interact with each other according to some set of laws, creating a conglomerate of units. The behaviour of the conglomeration of units is investigated as the conglomerate is protruded in some way. Failure models of this kind have a range of applications (Newman, 1995). For example, these simulations were used to estimate the statistical difference which a change in excavation strategy will have on the expected seismicity (Vieira et al., 2001). Kaiser and Tang (1998) investigated failure properties of a pillar in different conditions, derived from a fundamentally different approach to modelling failure. Cai et al. (2001) related the observed seismicity to the extent of failure experienced by a pillar. Beck and Brady (2002) estimated the expected seismicity to a planned excavation.

Data mining techniques have also been applied to seismic data to establish probabilistic rules. Also, in combination with other hypotheses, one can test for the significance of the data set belonging to some class of processes, or for process invariance properties exhibited by the data. Eneva (1998) used three parameters to describe spatial and temporal characteristics of mining-induced seismicity and found rules to relate trends in the test statistics to the occurrence of large events. An seismic warning signal index was proposed in Poplawski (1997) as a weighted sum of a combinations of parameters. In Eneva (1994) the spatial distribution of mining-induced seismicity was distinguished from a multi-fractal system as a mono-fractal system, characterizing invariant properties of the spatial distribution. Spacial distribution tendencies can be used in combination with probabilistic rules to estimate the future behavior of the mining induced seismicity more reliably than without it.

Mining-induced seismicity studies have improved the understanding and management of the related hazards through source parameter estimation and extrapolation. This understanding paid for the installation of a seismic observation system within 10 months of operation (Poplawski, 1997). The essence of the relationship between the failure process and its related seismicity can be validated by the accuracy of the simulated failure models evaluated in terms of the simulated seismicity they produce. Statistics based on seismicity exhibit precursory behaviour providing indications of the scale and nature of possible determinism and independence exhibited

by the system (Eneva, 1998). All these connections demonstrate the importance and validity of studying the mining-induced seismicity, in particular to gain insight into the failure processes involved.

2.4 A framework for modelling crust-scale seismicity

Large earthquakes have a recurrence time in the same place. Within such an area, the smaller the earthquake the more frequently it occurs. Once it has occurred, an earthquake diffuses into more earthquakes in the surrounding vicinity that decrease exponentially in size (Helmstetter and Sornette, 2002; Rikitake, 1976). The large time lapse between successive large earthquakes and the short time span of the actual event leads to a very small independent probability of occurrence of any given large event, based on the Poisson distribution. Because of the need to pinpoint the probability of a seismic event's occurrence location and (possibly) time, the search is on to establish the dependent probability and finding improvements on the independent probability (Aki, 1981). Establishing exactly what earthquakes depend on, given the size or rather precursors to earthquakes, has proved inconclusive. Some researchers have proposed that the short time scale occurrence of earthquakes are rather arbitrary and, thus independent of the observed short term precursors (Abercrombie et al., 1995; Abercrombie and Mori, 1994; Barriere and Turcotte, 1994).

Seismicity is a result of the inhomogeneity of the earth's crust, which one school of thought assumes to be on the point of failure, commonly referred to as a state of self-organized criticality (SOC) (Barriere and Turcotte, 1994). The stress in the crust is induced by a series of forces exerted in and on the crust. The global system of plate tectonics, the mid-ocean ridges, subduction zones, and faults form one part of the exerted forces. Other forces include gravity, the resultant pressure of liquids trapped in porous material in the crust and outward pressure from the mantle. Earthquakes are triggered by natural and man-made activities. The natural inducers include volcanic activity, other earthquakes, tides and atmospheric pressure changes. Man-made triggers include reservoir water level fluctuations, deep well liquid-waste disposal, mining and nuclear bomb detonations. It has been observed that triggers "leading to incremental deviatoric stresses of less than 10 atmospheric pressure are enough to induce events in the uppermost crust ranging in magnitude

up to 7" (Sornette, 2000).

Earthquakes are related to each other in apparent size by estimating the amount of deformation and energy released at the source in terms of the P and S waves arriving at the measurement station. The deformation and energy observations are model dependent (Gere and Shan, 1984). Due to the propagation qualities of the S and P waves, the point processes documented in the seismic catalogs are the most extensive and accurate observations of seismicity available. The energy release at the source has been found to be in a log-linear relationship with the source deformation. A consequence of the log-linear relationship is that the regional and global rates of seismic energy release can be monitored. These have been found to exhibit stationary properties, that is, the process is not in a transient state. Another consequence is that no realistic number of small earthquakes can make up for the energy released by a large seismic event. An ultimate point of strain has been estimated for the earth's crust and the largest recorded events demonstrated to be near that point (Rikitake, 1976). The type and style of seismic source deformation has been studied and related to the type and style of source deformation observed during rock breaking experiments (Abercrombie et al., 1995). On the other side of the seismicity size spectrum are the smallest events which are a source of non-stationarity in the observed seismicity. Not all events of the same small size energy release or deformation are recorded and any modelling based on the seismic source parameters should take this source of nonstationarity into account (Li et al., 2002).

Theories on the long-term drivers of seismicity have been labeled as stationary. Some of the largest forces involved have been identified and now generally accepted. Theories on the medium- to short-term behavior of large seismic events were originally dominated by observations of fault zones, rock-breaking experiments and the notions of wet and dry dilatancy. The model for dilatancy was developed using rock-breaking experiments. In rock-breaking experiments, the failure of the sample is directly related to the loading history of the sample. Failure in heterogeneous samples is preceded by a fault preparation process. In general, buckling causes dilatancy fracturing on the outside of the buckle, and compression on the inside of the buckle. The specific point of failure is preceded by a nucleation process and a resultant fractal fault area displayed as well as a series of foreshocks. The notions of dilatancy contributed to the hypothesised existence of signatures to large seismic events.

The idea that earthquakes are preceded by foreshocks was strengthened by observations that reservoir-induced seismic events and mining-induced events are usually also preceded by foreshocks. These ideas culminated in theories of stress build-up, a fault preparing for a large seismic event and, finally failure setting in, resulting in a large seismic event. This set of theories includes the notion of asperities and the onset and end of seismic gaps (Habermann, 1981). The body of theories on the medium- to short-term behaviour of the earthquake-generating process have been questioned as they fail to predict or even uniquely identify foreshocks. The independent empirically observed laws still considered valid explained by these theories are the characteristic frequency-size power law distribution of events, the inverse power law in the decreasing rate of aftershocks after a large event and the relationship between the size of the subsequent event and the log of the radius of area affected by abnormality (Helmstetter and Sornette, 2002; Sornette, 2000).

The emergence of the concept of self-organized criticality led to development of new theories on medium- to short-term earthquake behaviour. Firstly, self-organized criticality implied that the crust is on the verge of failure most of the time and, secondly, the short-term time predictability of self-organized criticality systems in general has been questioned. Studies into the nature of the earthquake generating mechanism are still ongoing. It has been known for some time that observed seismic catalogs alone do not provide enough information on the medium- and short-term physical earthquake driving mechanism (Kagan and Vere-Jones, 1996; Kanamori, 1981; Rikitake, 1976). A number of issues relating to the physical mechanism of earthquake generation are currently under investigation. These include (a) the effect and distribution of subtle pressure changes in the crust caused by and resulting in earthquakes (Stein, 2003); (b) the question of why some events grow into large events and others do not (Abercrombie and Mori, 1994); and, (c) the long-term temporal evolution of fault regions (Eneva and Ben-Zion, 1997a).

In general, the dependence exhibited by the observed seismic parameters can be used to understand the underlying physical mechanism responsible for the seismic events as well as hypothesis testing within a rigorous statistical framework (Habermann, 1981; Kanamori, 1981). Parallel to the study of the physical process, statistical models and machine learning approaches have been used to establish and quantify the extent of the dependence exhibited between observed parameters and the location in time and space of an event of a given size. In all cases, pre-processing of the seismic data forms an important part of the rigorous statistical framework because

of the qualities of the observed seismicity. The interested reader is referred to the cited works herewith and in Brillinger (1997).

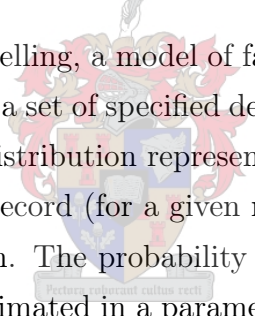
An inference procedure for a rule set used in medium to large-scale seismic event prediction is discussed in Eneva and Ben-Zion (1997b). Furthermore, two other similar historical inference procedures based on the seismicity of a region are discussed, and some of their weaknesses addressed. To create a rule, sequential portions of seismicity are divided into seismicity exhibiting some quality, typically a large event, and seismicity preceding such behavior, typically precursory behavior. A rule associates intervals of precursory seismicity with each other, and counts the empirical probability of a given quality related to that association. A rule set is then a collection of these rules, representing the probability of occurrence of the required quality, given the associated description of seismicity, as expressed in the data (Russell and Norvig, 1995; Witten and Frank, 1999). The non-parametric approach needs sufficient data to provide adequate error bounds on the rules. The success of the adapted rule-inferencing method is measured on synthetic seismic catalogs. The identification of foreshocks based solely on seismicity, or *seismic prediction*, is discussed in Yamashina (1981) and Vorobieva (1999). The probability that a given seismic event will be followed by a larger seismic event is estimated by a rule set, derived in a similar fashion to the prediction algorithms for the preceding long-term estimation.

Artificial neural networks (Haykin, 1999) are used in Bodri (2001) to give an estimate of the long-term time interval to a future large seismic event, given the rate of preceding smaller events, partitioned into classes of different sizes. We refer to this class of models as weakly parameteric models. An artificial neural network is a universal function approximator with an adjustable, but unknown bias-variance relationship (Fan and Gijbels, 1996). The functional mapping from the independent to the dependent variables is done with no prior assumptions, being dependent purely on the optimization of the network of weights. As a general modelling method this puts similar constraints on the data as a non-parametric approach.

Colombo et al. (1997) used expert systems to estimate a maximum event magnitude field for a region in Costa Rica. One of the motivations for the use of expert systems was the large number of parameters being monitored. More parameters than the seismicity of the region are used by the system. Non-linear regression is used to model the functional form of the active seismic regions and the related maximum observed seismic events. This style of regression represents a weakly parametrized

approach, using a piecewise linear function as a general approximator. Maximum moment field estimation is revisited below.

In contrast to the empirical prediction rules of the preceding sections, Kagan and Knopoff (1987) present a well-defined stochastic model that a given event is followed by a larger event within a relatively small time interval at a pre-described offset in location. The time interval is small compared to the recurrence time of the predicted event. The model is derived from a quasi-static fracture growth model and has only three adjustable parameters. The optimal values for these parameters are found using a maximum likelihood estimator. The model is used to identify foreshock sequences in progress and provides a method of quantifying the success of the model. A rule list expressing the relationship between preceding events and following event sizes is extracted from the model with the maximum likelihood estimated parameters. Independent models governed by a few parameters optimized for a given system and used to express a most likely future behaviour are examples of parametric modelling, or non-empirical modelling (Kagan and Knopoff, 1987).

In parametric stochastic modelling, a model of failure is defined representing a family of distributions exhibiting a set of specified dependencies between the parameters of the system. The specific distribution representing the parameters most likely expressed in the seismic track record (for a given region) is used to give the expected future behavior of the system. The probability of occurrence for a large event, i.e. a maximum event field, is estimated in a parametric approach in Akkaya and Yüce-men (2000); Akkaya and Yüce-
men (2002). The model uses space, time and location correlations of the preceding seismicity on a given fault. Darpahi-Noubary (2002) discuss a parametric estimation of the probability of occurrence for a large event in a relatively stable seismic region. As an alternative to filtering the smaller portion of non-stationary seismic events, they use a generalization of the exponential distribution for the occurrence of events in the region known as the family of generalized Pareto distributions. The generalization demonstrates that the assumption of the exponential distribution for the arrival of a size class of seismic events is biased.

At the heart of all parametric stochastic modelling is a model of failure applicable to the seismic scenario in question, with no more than adequate variance and as small a bias as possible. In Kagan and Vere-Jones (1996) an attempt is made to express the fundamental structure exhibited by the seismic failure process as measured by its seismicity. An empirical failure model is presented capable of expressing the *a pri-*

ori described structure. Several authors have also proposed dependence structures and corresponding stochastic models capable of expressing these laws (Abinante and Knopoff, 1995; Helmstetter and Sornette, 2002; Newman, 1995; Sornette et al., 1992). With the emergence of self-organized criticality, generalized stochastic models expressing the self-organized earthquake style laws have also been investigated (Blanter and Shnirman, 1996; Hooge et al., 1994).

In the search for precursors of large earthquakes, different contributing factors have been proposed and advocated as theories. Consequently, many models of dependence have been proposed, extended or rejected. Observations from the crustal and related systems are systemized into different independent components. The statistical significance of the temporal-dependent components have been tested. It has been found that the crustal system exhibits self-similar dependencies over different temporal and spacial scales. Some authors have argued that precursory behavior to large seismic events lies only in the eye of the beholder, and that the study of minimizing seismic hazard should not put much focus on the short term prediction problem. On the other hand, rule-based algorithmic approaches have successfully minimized the error on the expected time interval to the next large seismic event. Hence, stochastic models are being developed that closely mimic the crustal system's observations resulting in better hazard estimation. New statistical procedures are establishing probabilities of seismic hazards which has not been possible before.



2.5 Modelling of mining-induced seismicity from a crust-scale perspective

Mining-induced seismicity and crust-scale seismicity correspond in some aspects and differ in others. Similarities in the system represent scale invariant properties in the crustal failure process while differences reflect system variant properties and beg explanations since "the exception tests the rule".

One of the methods of monitoring the failure process is the creation of an expected maximum event field representing the seismic hazard in a given area. The field can be expressed in static or dynamic terms. The dynamical description is commonly known as prediction (adding a time parameter to the hazard term). In both crust-scale and mining-induced seismicity, the static description of the failure process plays a dominant role. A static description looks at a description of the current

state of the rock and associates a hazard term with that state. In the study of crust-scale seismicity dynamical descriptions of seismic hazard have shown some success (Bodri, 2001; Sornette, 2000). However, in the mining environment the subject is not discussed often, nor are seismic rates parameters often used. This lack of use of seismic rates is interesting since similarities between rock flow and turbulent flow have been established (Mendecki et al., 1997). In reconstructing the system responsible for turbulence (from a point process) the role and specification of time intervals between points have been discussed in literature (Castro and Sauer, 1999; Pavlov et al., 2001; Sauer, 1994). Similarly, the time intervals between large seismic events can be expressed as the realization of the rate at which that large event is arriving.

Mining-induced seismic events have been put forward as evidence for the self-organized criticality of the crust. Shortcomings in the statistical framework for the analysis of seismicity have been highlighted, e.g. Kagan and Vere-Jones (1996). The statistical modelling approach to seismicity requires a well-defined framework to make reliable conclusions. A similar method and test statistics developed for crust-scale seismicity can be used in mining-induced seismicity. Due to the higher sampling rate in a smaller area, deterministic test statistics have been used to quantify some of the temporal qualities of the seismicity (Eneva, 1998). Correspondences and differences in the statistical models for failure resulting in seismicity can be tested for mining-induced seismicity, and deviations explained for a better understanding of the failure model in general.

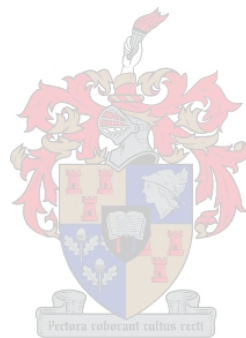
2.6 Concluding Remarks

The available literature on seismicity presents many perspectives on mining-induced seismicity and its relationship to crust-scale seismicity. One group of these is related to the similarities and differences between the two types of seismicity and the corresponding systems generating them. Another viewpoint is related to the predictability of seismicity, whether it is on a mining-induced or on a tectonic-induced scale.

The similarities reported in the literature between the two types of seismicity are striking considering their differences. Mining-induced seismicity is generated by a man-induced, transient system on a very small scale. In contrast, crust-scale seis-

micity is generated by a natural, non-transient large scale system. All the major seismic laws developed for crust-scale seismicity have been reported for mining-induced seismicity as well. Despite the similarities in the apparently different systems, no generally applicable theory exists that fully explains the nature of the system that generates these two types of seismicity.

The system responsible for generating seismicity is still under investigation in the hope of determining the extent to which it can be predicted and why. The self-organized criticality perspective, if successful, may provide answers to the how and why. However, seismicity as generated by a self-organized criticality system has not yet been fully formalized. Prediction studies using empirically gathered information are also being pursued. There is as yet no single method that offers a full explanation of seismicity and its behaviour, nor has a concise summary of relevant facts on the seismic laws been established.

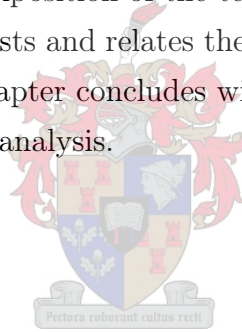


Chapter 3

Analysis of seismic data using the method of surrogates

In this chapter hypothesis tests conducted on the T4 seismic data set are discussed. The chapter starts with an exposition of the terms and definitions used in the construction of the hypothesis tests and relates the hypothesis test conducted on T4 to seismicity in general. The chapter concludes with a discussion and summary of the results of the surrogate data analysis.

3.1 Introduction



Surrogate data analysis is essentially a statistical hypothesis testing technique for probing the nature of structure in observed data. A null hypothesis is assumed for the data and tested against an alternative hypothesis using bootstrap sampling methods. An appropriate test statistic is sampled from realizations of the data, also known as *surrogate data*, generated according to the null hypothesis. If the distribution of test statistic of the surrogate data is significantly different from the same statistic computed on the observed data the null hypothesis is rejected, otherwise it is accepted.

For the T4 seismic data, the hypothesis test is constructed by sampling two sets of test statistics on the real data. The surrogate data are generated according to some stochastic model (null hypothesis) against which the observed data are tested to decide if the two sets belong to the same underlying population. The test statistics are sampled from a state space reconstruction of the system producing each of the

data sets (Sauer, 1991, 1994).

3.2 Definitions and measurements

3.2.1 Sampling the seismic dataset, T4

We define a seismic event as a sudden fracture in a rock mass described by a set of seismic source parameters satisfying

$$\overline{P}_i = (\overline{Loc}_i, M_i, E_i) \quad (3.1)$$

observed at specified time stamps t_i . Here, the seismic event (\overline{P}_i, t_i) is estimated by the time of occurrence, its location in the rock $\overline{Loc}_i = (x_i, y_i, z_i)$, the deformation M_i , and energy E_i released at that location. The T4 seismic data set is a set of N estimates of seismic events that occurred in a mine over time for a fixed volume of rock, that is

$$T4 = \{(\overline{P}_i, t_i) \bullet i = 1 \dots N\} \quad (3.2)$$

with i indexing the events in order of occurrence. It will be assumed that the errors made in the measurements of the location, deformation and energy in T4 are negligible.

A seismic monitoring system consists of a fixed array of instruments that measure ground velocity and acceleration. These instruments enable so called primary (P) and secondary (S) waves (Weisstein, 2004) propagated in the surrounding rock mass to be sampled within a specified frequency range. The derivative and integral with respect to time of both P and S waves give estimates of the ground acceleration and movement respectively.

S and P waves originate from seismic events. Sudden rock fracture is associated with two orthogonal components – friction and compression. The two components are mostly orthogonal. In the volume of rock, the compression component is mostly along one of the basis vectors, the friction component along the other two. The P wave is initiated by the compression component, while the S wave is initiated by the frictional component. In mines, S waves travel slightly faster than P waves. Using

this information, combined with knowledge of the locations of the instruments, the location of an event can be triangulated.

The measured ground movement at a sensor is divided into frequency components. The ground movement measurements for all the sensors that pick up an event are stacked on the same frequency components. The S and the P components are sampled on separate frequency components. The power components over the stacked spectra are adjusted to form a single power estimate for each of the observed frequencies to compensate for random fluctuations in each of the individual measurements. A Brune model (see below) is then fitted to the re-sampled spectra (Brune, 1990). The model-dependent seismic source parameters, E_i and M_i , are derived using the Brune model.

The Brune model ($D(f)$) is a function that relates a frequency f to the power associated with that frequency $\Omega(f)$. It is parameterized by two parameters: the corner frequency f_0 , and the power at the zero frequency Ω_0 . Deriving the seismic source parameters from the S and P wave power spectra requires an infinite range of frequencies. The Brune model, fitted to the finite frequencies of the stacked spectra, can be extended to infinite frequencies under the required adjustment. The Brune model is defined as

$$\Omega(f) = D(f) = \frac{\Omega_0}{1 + (f/f_0)^2}. \quad (3.3)$$

The seismic moment M_i is estimated from the power at the zero frequency of the stacked spectra by extrapolating the Brune model. The seismic energy E_i is estimated from the integral of the Brune model over the infinite spectrum of the frequencies. The errors in the seismic moment and the seismic energy are derived from the residuals after the Brune model has been removed from the stacked spectra. The scale of the seismic source parameters E_i and M_i and their absolute errors are rather large and, therefore logarithmic transformations are used for numerical stability.

The sequence of recorded seismic moment changes M_i and energy releases E_i in the T4 data set are shown in Figure 3.1 as a function of each event's time stamp. T4 consists of 3835 events ranging in Richter magnitude from -1.5 to 2.9 sampled over a period of 288 hours for a mass of rock approximately 0.2150 km^3 volume. The distribution of the location of events in the rock mass \overline{Loc}_i of T4 is depicted in Figure 3.2. In this study the location of an event is not taken into consideration

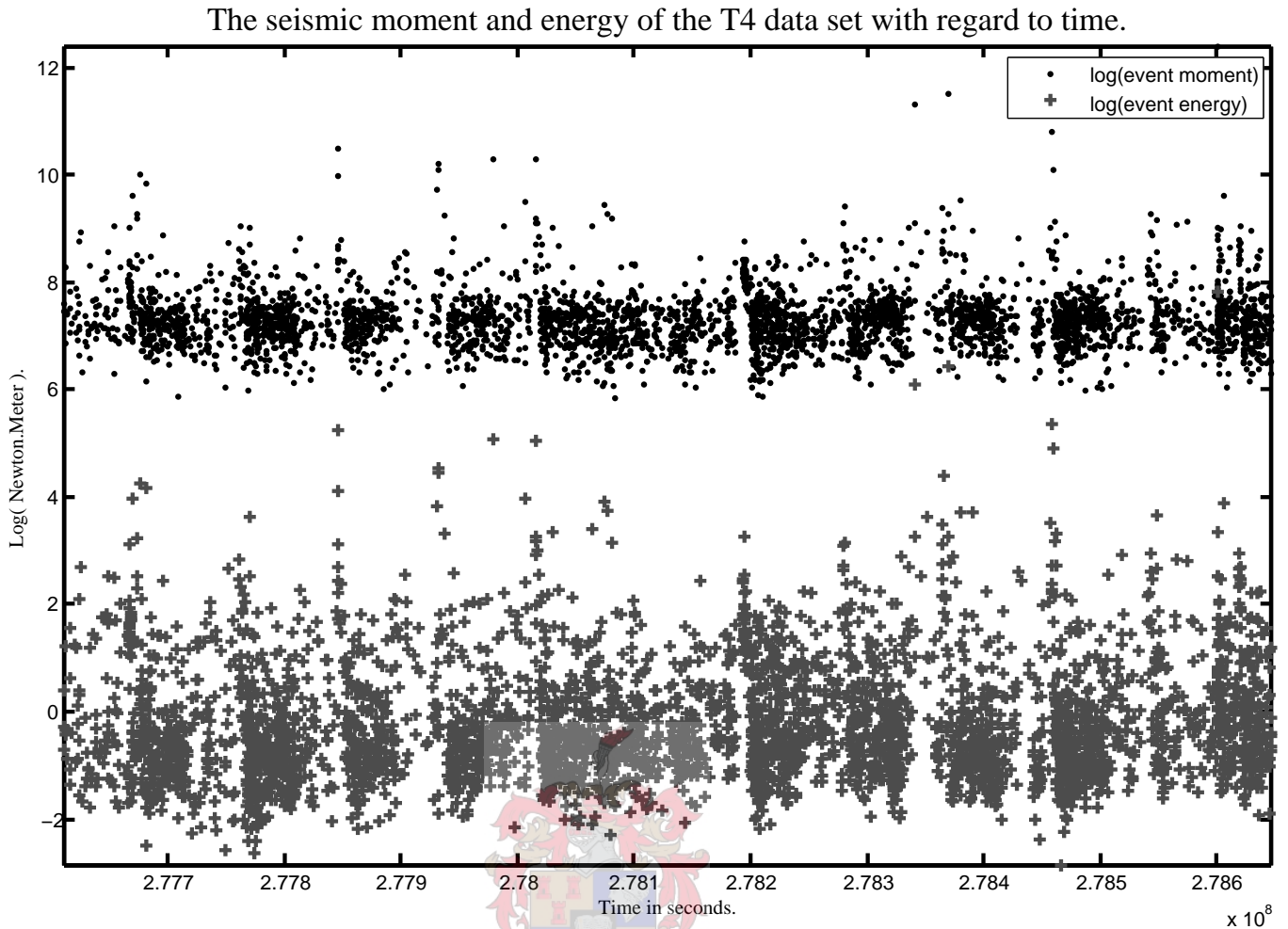


Figure 3.1: The T4 data set's seismic moment (.) and seismic energy (+) measurements plotted on a logarithmic scale as a function of time.

and the data set is considered to represent a single seismic source. T4 is considered in more detail in Section 3.3.

3.2.2 Previous work - ISSI's investigation

ISSI have published a book explaining the source and interpretation of their seismic data sets (Mendecki et al., 1997). They did an analysis on the attractors formed by the process generating the seismic events to establish determinism in the measurements. A fundamental model of the fracture of solids was used to derive a method for the prediction of large seismic events, and estimation of the pressure gradients in the rock causing the fractures. The resulting model had a moderate 33.3% success rate in predicting large seismic events.

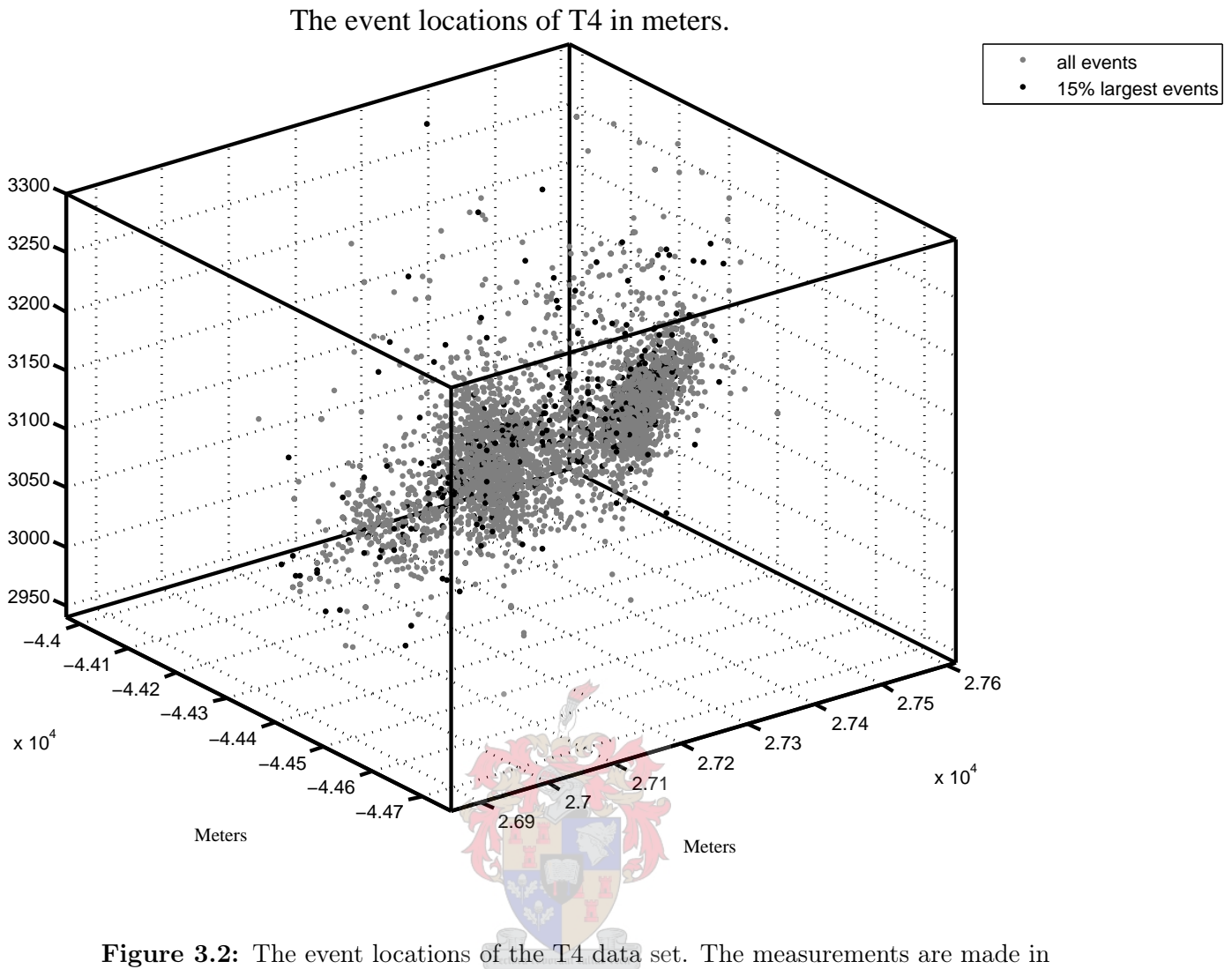


Figure 3.2: The event locations of the T4 data set. The measurements are made in meters but the orientation of the axes is unknown.

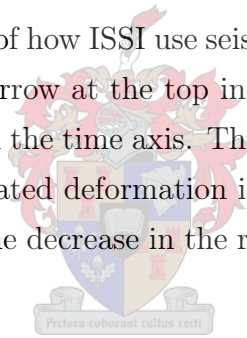
Mining deforms the rock mass of a shaft mine out of its natural balanced form. The shift from balance induces a pressure gradient in the rock so that the rock becomes unbalanced or unstable. The pressure gradient, or stress, is described with a tensor field. The rock flows along this gradient and is turbulent at fracture points in the rock. The strain is also described with a tensor field. The turbulence is then described along the interaction between these two tensor fields.

Stress and strain are related through Hooke's linear law for solid materials given the rigidity of a material (Feynman et al., 1989). Fracture points in the rock are preceded by a nonlinear deviation from Hooke's law in the tensor field's stress-strain relationship. The nonlinear deviation consist of an acceleration of rock deformation for the same amount of work done by the deformation process. This signature occurs irrespective of the size of the fracture.

The energy index and the accumulated apparent volume can be computed from the seismic moments and energies of consecutive events in a given portion of rock. The energy index of an event is its observed energy in proportion to its expected energy, given its moment, as exhibited by the moment-energy relationship in the rest of the data set. The energy index is usually filtered using a moving median with respect to time. Apparent volume is a function that relates the seismic moment (in Joules) to the volume of rock deformed during the seismic event. The accumulated apparent volume is the sum of the apparent volumes with respect to time.

Using Hooke's law, a large fracture in a given portion of rock can be detected when the accumulated apparent volume accelerates and the energy index decreases simultaneously. ISSI uses these two quantities to predict the occurrence of large seismic events. The occurrence of this pattern depends on correct selection of the volume of rock, set of seismic events, filters and time frame for the two quantities. The context selection and pattern recognition is done manually.

Figure 3.3 shows an example of how ISSI use seismic data to establish the occurrence of big seismic events. Each arrow at the top indicates when a large event occurred at the corresponding point on the time axis. The signature of the large events at the end of day 21 are the accelerated deformation in the accumulated apparent seismic volume during day 21, and the decrease in the released energy per unit deformation (days 19 to 22).



3.2.3 Seismic Laws

A seismic data set is a set of measurements of a point process that consists of a series of event measurements. The three basic measures of a seismic event are (a) the time at which it occurs t_i , (b) the estimated amount of rock deformation $\log(M_i)$ and, (c) work done during the event $\log(E_i)$. Seismic laws describe empirical statistical dependencies between the variables of the seismic event generating process.

Typical features, or seismic laws, observed in crust-scale seismicity have been related to mining-induced seismicity (Chapter 2). A number of stochastic systems have similarly been reported to demonstrate dependencies comparable with the seismic laws. Observing systems demonstrating these 'seismic laws' does not necessarily lead to the conclusion that these systems are therefore stochastic. Instead, the correct assertion would rather be that the system cannot be distinguished from a stochastic

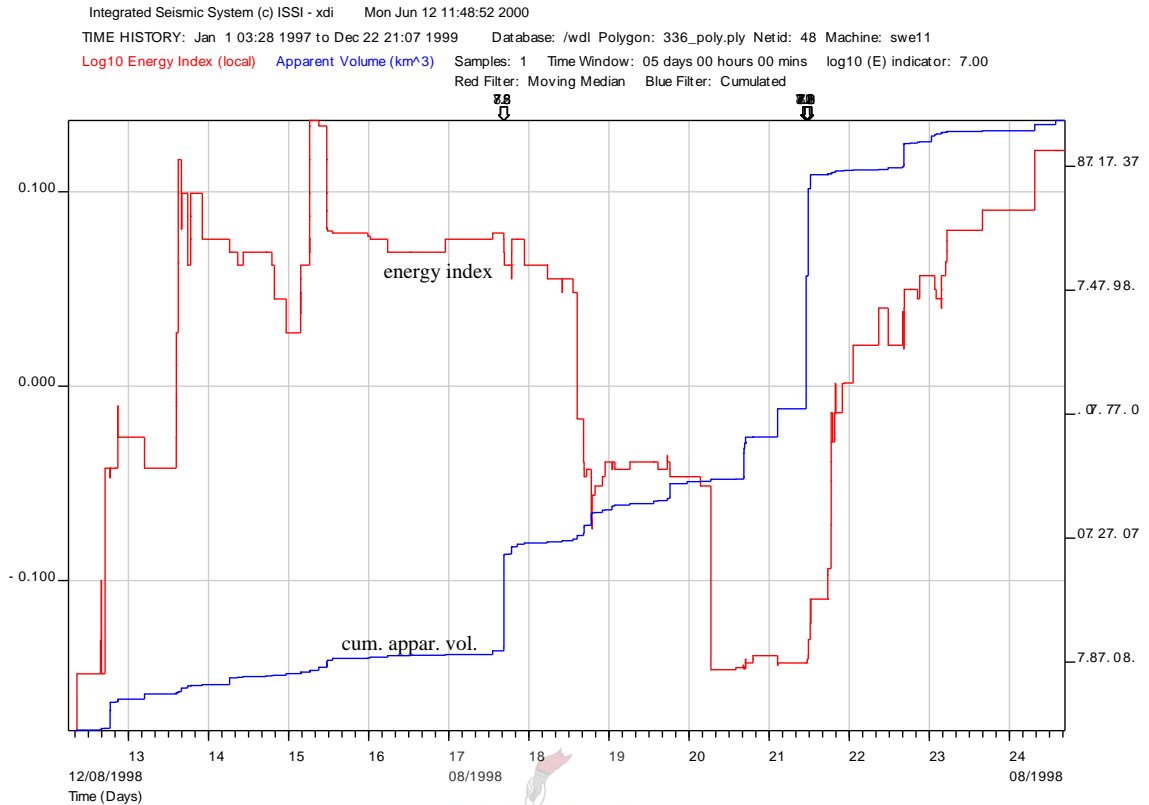


Figure 3.3: A method of estimating the time to fracture for a large seismic event. The cumulative apparent volume (left-hand, bottom corner) and the filtered energy index (left-hand, top corner) plotted versus a time scale. The two curves exhibit the expected signature of a failure for solid materials. The graph was kindly provided by ISSI and plotted using their seismic interpretation software.

system on the basis of these dependencies only. In other words, the expected behavior of the stochastic systems exhibiting these dependencies will provide an indication of the expected behavior of a seismic event generating system under assumption of these dependencies alone.

Consideration of seismic laws is important for a number of reasons. Firstly, these place the study into the larger context of investigation into seismicity. Secondly, a number of existing relationships can result in an emergent behavior influencing new possible measures of the system. A reported result on a system can either be explained by the existing laws reported for the system, or demonstrate a property that needs to be included into known laws for that system. Finally, the method of surrogate data analysis can be used as a mechanism for generating synthetic seismic activity if the structure in the surrogates can account for the observed seismic laws.

Law I - The Gutenberg-Richter relationship

The Gutenberg-Richter relationship describes a linear functional relation between the event's moment size and the logarithm of the frequency of occurrences of the event (Figure 3.4). It can be expressed as

$$Y_{M^L} = \log_{10}\left(\sum_i I(M_i \geq M^L)\right) \quad (3.4)$$

where $I(M_i \geq M^L)$ is an indicator function that is equal to one if $M_i \geq M^L$ and equal to zero if $M_i < M^L$. The law is used to establish the stationary range of observed seismic events (Chapter 2). It has been shown that in the case of mining-induced seismicity the law deviates from the linear form observed in crust-scale seismicity (McGarr, 2000).

The slope ($-b$) is an indication of the amount energy the seismic system can exert in principle, as a consequence of the third seismic law discussed below. A large b indicates relatively few large events while a small b indicates a high number of large seismic events generated by the system and larger maximum event sizes. The size of b is not universally fixed in seismic systems.

Law II - Clustering of large events in time

The second law relates to the clustering of large events in time and is considered a combination of seismic after-shocks (Omori's law) and seismic foreshocks. Seismic foreshocks only occur in a third of mining-induced large events (Chapter 2). The law is not explicitly stated in a functional form, but only observed as a trend that is induced by the clustering of large events around hazardous events due to fore- and after-shock activity.

The trend can be depicted graphically through the following steps:

1. Select the seismic events, (t_i^L, X_i^L) , larger than a given threshold (X^L) , from the data set $X_i^L \geq X^L$.
2. Stagger the large seismic event sequence into sequences of length $n+1$, $\{(t_k^L, X_k^L), (t_{k+1}^L, X_{k+1}^L) \dots (t_{k+n}^L, X_{k+n}^L)\}$.

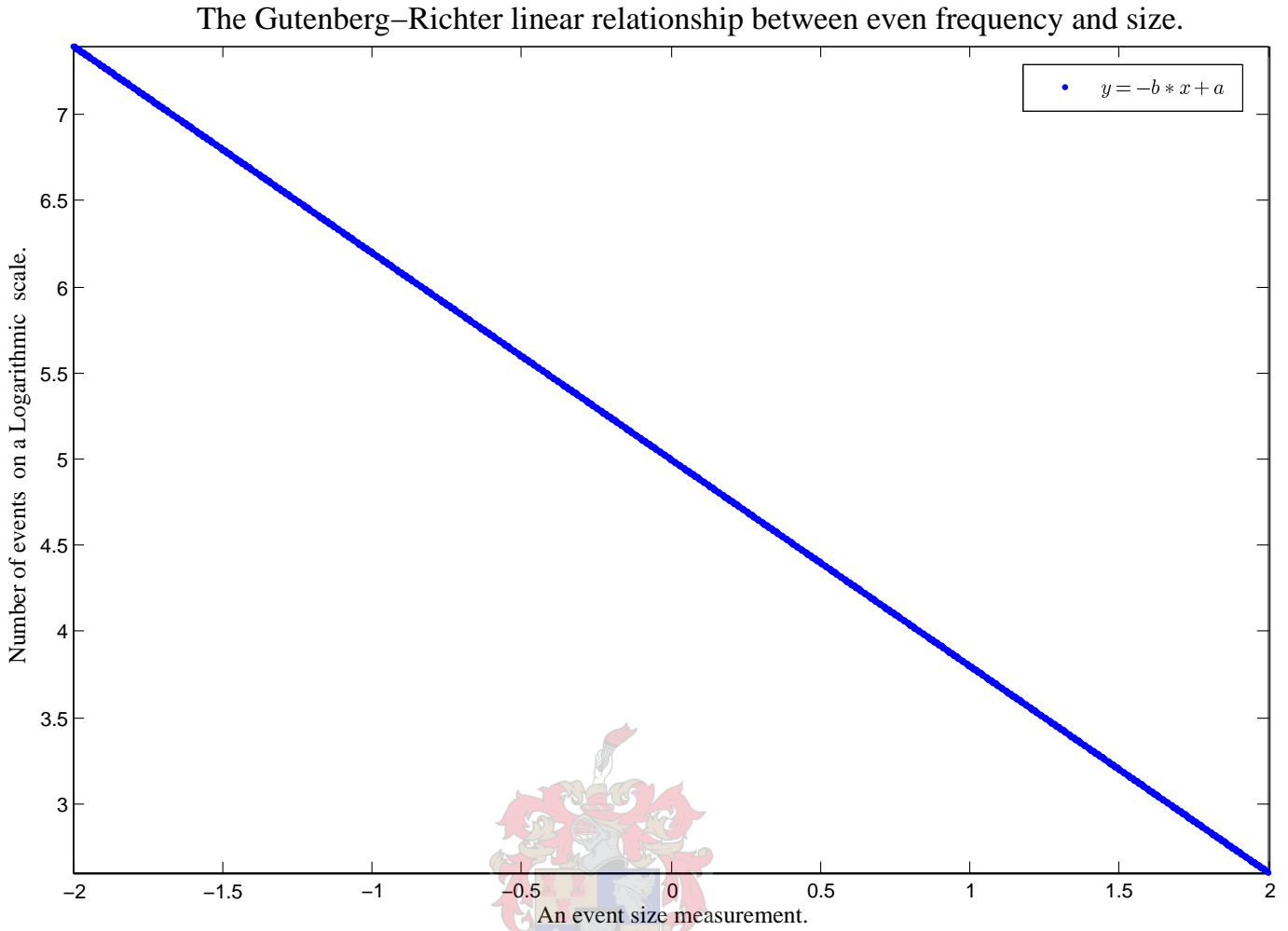


Figure 3.4: The size vs. the log of cumulative frequency, relationship. $Y(M^L) = \log_{10}(\sum_i I(M_i \geq M^L))$. According to literature the G-R relationship for seismic activity follows a linear scaling law: $Y(M^L) = -b * M^L + a$.

3. Plot the average arrival interval between the events,

$$E(Dt_i^L) = 1/(n+1) \sum_{k=i}^{i+n} (t_{k+1}^L - t_k^L) \text{ as a function of the average event size,}$$

$$E(X_i^L) = 1/(n+1) \sum_{k=i}^{i+n} (X_{k+1}^L) \text{ for each sequence.}$$

4. Plot the trend for two different thresholds, firstly for large events, $X_i^L \geq 85\%$ of X_i and secondly for all the events in the data set, $X_i^L > -\infty$.

The selection of a large event sequence is depicted in Figure 3.5 which shows a number of event values, $X_{101}, X_{102}, \dots, X_{109}$ sampled at the corresponding time stamps. Each of the events is larger than the threshold value X^L . The value pair for the sequence plotted schematically in Figure 3.6 is then $(E(X_i^L), E(Dt_i^L))$.

The average inter-event arrival intervals for large events plotted as a function of the

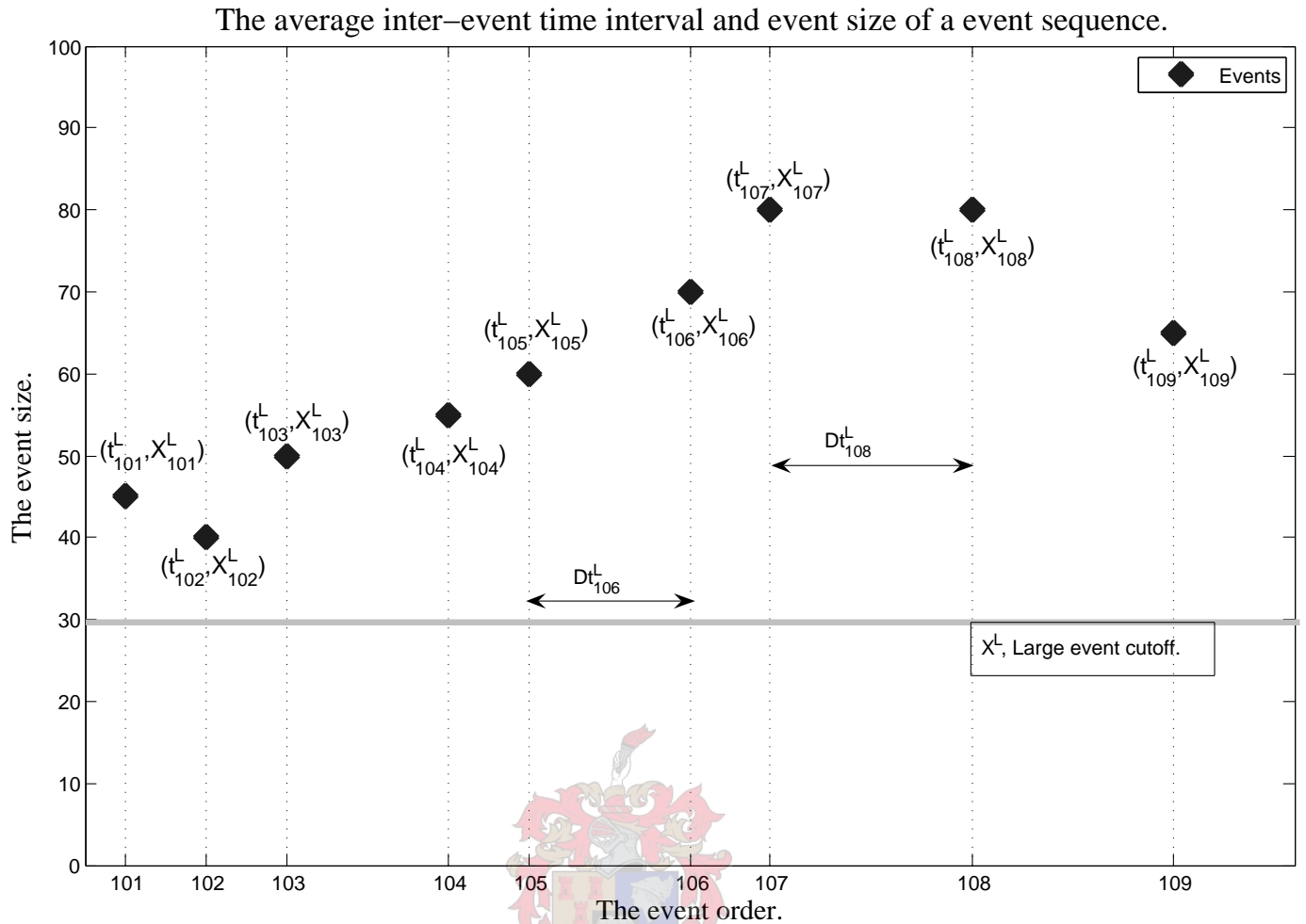


Figure 3.5: Selecting a sequence of events in plotting the trend for seismic event clustering.

average event size are depicted in Figure 3.6 for the case where large events cluster in time. Clustering of larger events around hazardous seismic events will cause the average inter-event arrival intervals to be smaller for sequences demonstrating hazardous events. The lower part of the average event sizes will show a whole range of average inter-event arrival intervals of the large event set. Conversely, if the same test is conducted taking the whole data set as large events, the trend of faster event arrival rates for a sequence containing larger events should not be visible since fore-shocks are not clearly visible prior to every event. Fore- and after-shocks are considered only as the clustering in time of larger events around large events. Fore- and after-shocks are not observed as a general increase of seismic events around the occurrence of large events.

The trend for the average inter-event time interval vs. the average event size for clustering events.

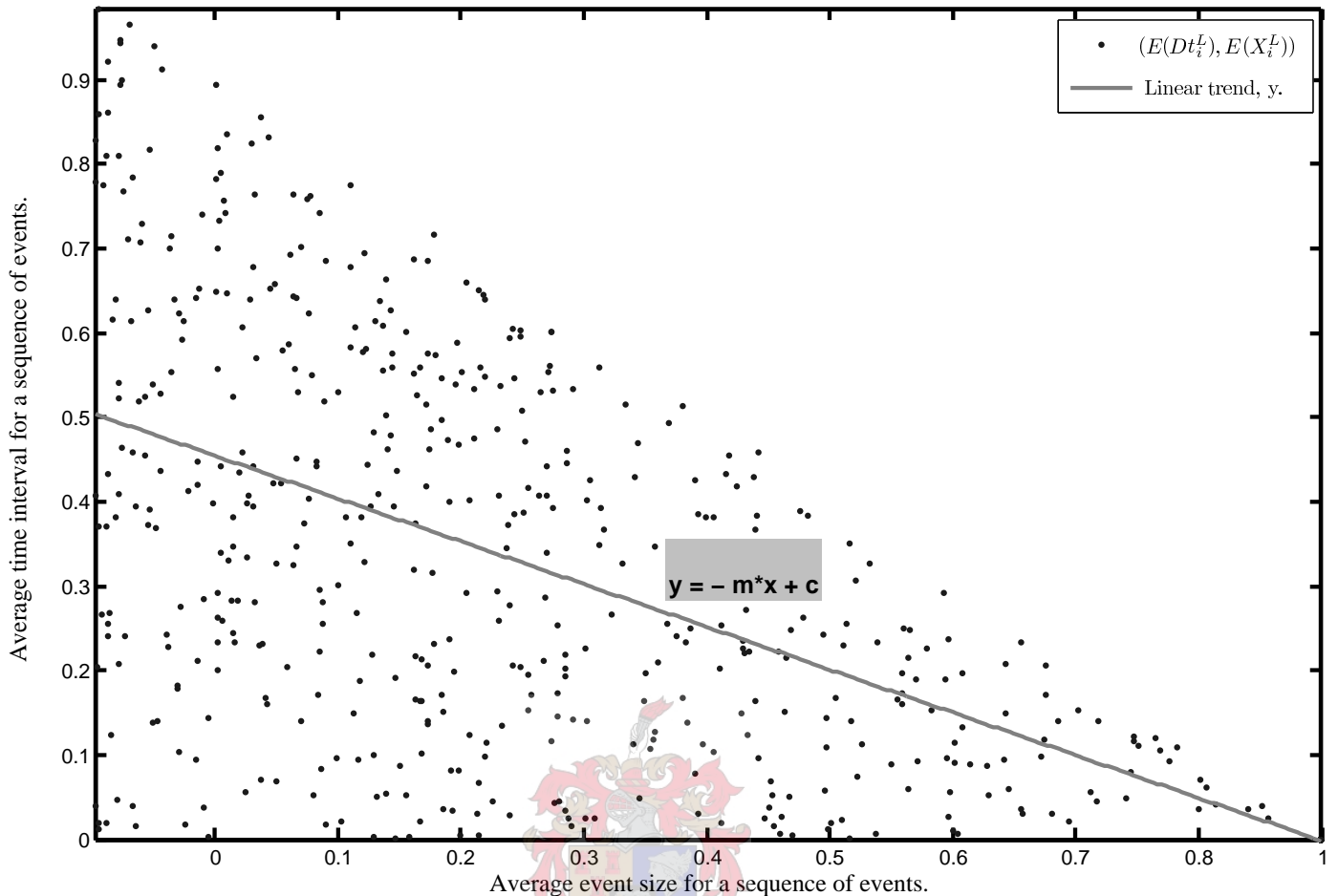


Figure 3.6: An explanation of plotting a schematic of the trend for seismic event clustering.

Law III - Seismic deformation and energy relationship

The third law is the log-linear relationship between seismic deformation and seismic energy release. The linear relationship is depicted in Figure 3.17 on the top right side axes. To demonstrate the importance of the relationship consider the functional form of the relationship,

$$y = f(M_i) = a * M_i - b. \quad (3.5)$$

The functional relationship dictates that $10^{1.7}$ events of $\log_{10}(M) = 9$ are required to release the amount of energy released by one event of $\log_{10}(M) = 10$ and $10^{3.4}$ events of $\log_{10}(M) = 8$. Replacing a large event with an equivalent number of smaller events will result in a violation of the G-R relationship. Stated differently, given the G-R relationship, the log-linear relationship of the third seismic law, and a

minimum amount of energy to dissipate, large seismic events are unavoidable (Gere and Shan, 1984).

Other seismic laws

There are two other seismic laws reported in literature that are not considered in this study. These laws focus on the fractal nature of event locations and the relationship between the size of an oncoming large event, given the area of abnormal seismic activity. Laws I-III above are concerned more with a characterization of the seismic activity than giving an in-depth analysis of the seismic nature of the seismic data set. The other two seismic laws are not expressed in the components used for surrogate analysis.

3.2.4 Linear statistics

We consider hypothesis testing of the T4 under the null that the data are sampled from an autoregressive moving average (ARMA) process, possibly transformed by a fixed nonlinear invertible function. The null hypothesis specifies structure exhibited by a red colored noise. The red colored noise is constrained to have the same probability distribution function (PDF) and autoregressive coefficients as the data set.

The surrogate sets are sampled without explicitly estimating the PDF or autoregressive coefficients of the red colored noise process. The surrogate data and the observed data share the same linear statistics, i.e. red colored noise parameters. The relationships between the linear statistics for the data set and the surrogate sets are checked to ensure that the correct red colored noise was sampled for the surrogate set.

The PDF of the data set is compared to the surrogate sets by pairing and plotting estimated percentiles for each surrogate set with the PDF of the data set. If the percentiles show a perfect linear relationship, then the two PDFs will correspond when plotted on a quantile-quantile (q-q) plot.

More formally, suppose x is distributed according to $x \sim f_X(x)$, with PDF $f_X(x)$ and y is distributed as $y \sim f_Y(y)$, then the quantile-quantile plot is the graphical

representation of the set of pairs of values:

$$\{(x_\alpha, y_\alpha) \bullet \alpha \in (0, 1)\}$$

such that α , $\alpha = Pr(x \leq x_\alpha) = Pr(y \leq y_\alpha)$, is the probability that x and y are less than or equal to x_α and y_α , respectively. If the pairs of values are proportional and share a similar domain, $x_\alpha \propto y_\alpha$, the two PDFs are at least as proportional and share their domains extensively. Since the α values are not available for a sampled data set, they are estimated by pairing the rank ordering of a sample of x and a sample of y .

The quantiles of a distribution are the three points corresponding to $\alpha \in \{0.25; 0.5; 0.75\}$. On the quantile-quantile plots the quantiles are used to compare $f_X(x)$ and $f_Y(y)$ from their samples. The quantiles are connected using straight line segments. The line segments serve as a reference to the proportionality of the (x_α, y_α) -pairs. The line segments will demonstrate an $x = y$ functional relationship if the distributions are the same. For example, the q-q plots of 2 different distributions are shown in Figure 3.2.4.

The estimated autocorrelation coefficients of the data set are compared to those of the surrogate sets by estimating the lag of the coefficients for each surrogate set and plotting each as a function of the corresponding coefficients. If the two sets of autoregressive coefficients are the same, the plot will demonstrate an $x = y$ functional relationship for the estimated autocorrelation coefficients. An autocorrelation coefficient is the standardized covariance between a component of a data set and the same component separated by a fixed lag.

More formally, suppose $\{AC_n^i\}$ and $\{AC_n^{data}\}$ are the autocorrelation coefficients for lags $n = 1 \dots N$ of a component of a multivariate surrogate set, $i \in \{1 \dots N_{sur}\}$, and the corresponding component of the multivariate data set, respectively. The graphical comparisons between the coefficients of the surrogates and the data set is a plot of the set of pairs, $\{(AC_n^{data}, AC_n^i) \bullet i = 0 \dots N \text{ and } n = 1 \dots N_{sur}\}$. That is, suppose $\{x_t\}$ is a data component with $t = 1, 2, \dots T$ data points, and $var(x_{t+n})$ the variance of $\{x_{t+n}\}$, $t = 0, 1, 2, \dots T - n$, then

$$AC_n^{data} = \frac{1/(T-1) \sum_{r=1}^T x_r x_{r+n}}{\sqrt{var(x_t) var(x_{t+n})}} \quad (3.6)$$

Since the $\{AC_n^{data}\}$ coefficients are fixed, the comparison to the $\{AC_n^i\}$ coefficients can be formalized in a confidence bound around the $x = y$ functional relationship

Comparing a Normal(0,1) to an Uniform(0,1) distribution using a q–q plot.

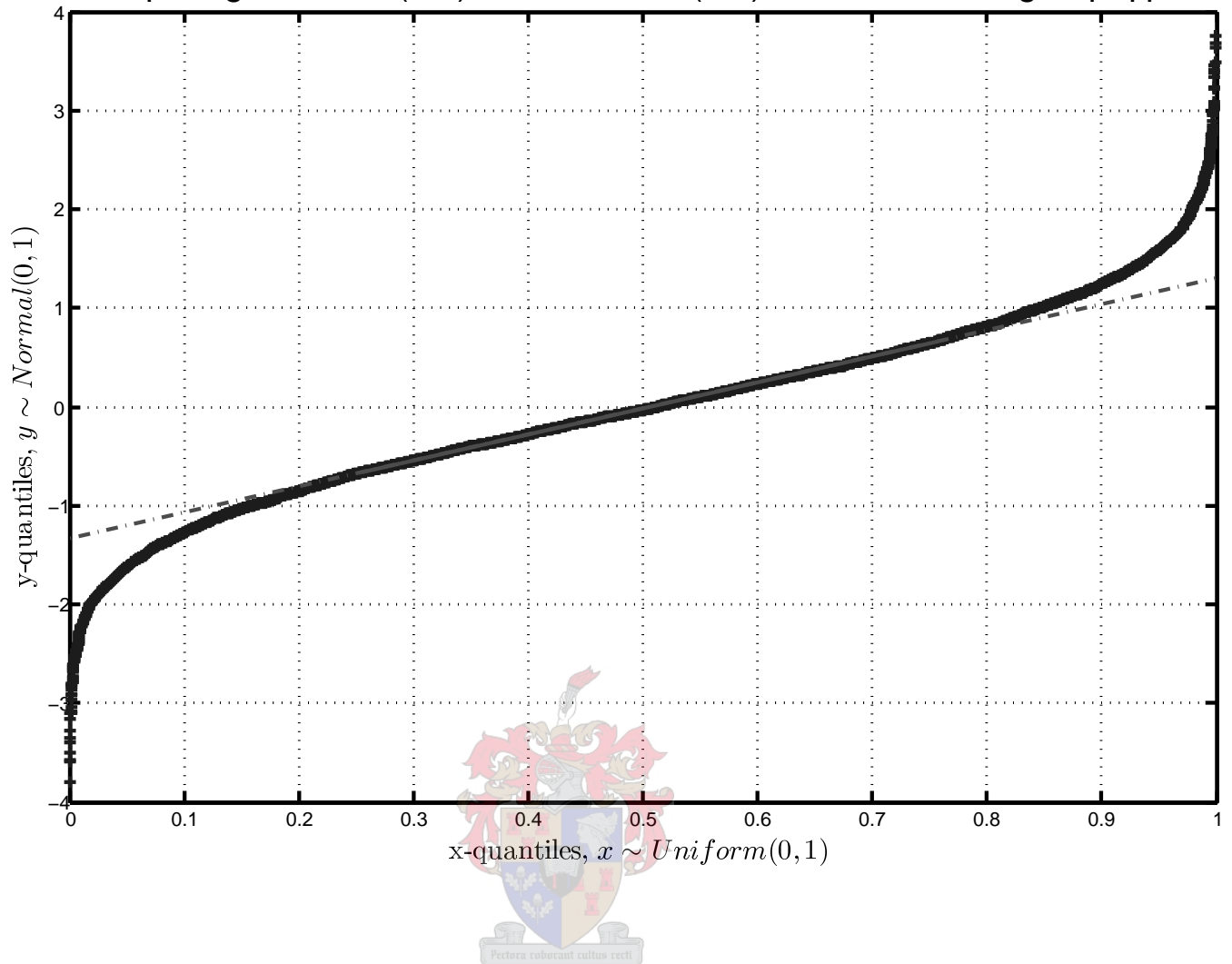


Figure 3.7: A quantile-quantile plot comparison between a Normal(0,1) and an Uniform(0,1) distribution. The dashed red line indicates the location of the quantiles, the blue pluses connects the (x_α, y_α) -pairs. The domains of the distributions differ as well as the proportionality of the α -pairs.

using the Fisher transform for correlation coefficients. That is, suppose $Z_\alpha > 0$ is the critical value for the standardized normal distribution larger than zero at probability $1 - \alpha$, then

$$AC_n^i(\alpha) \in \left[\frac{1/ke^{-2Z_\alpha\sigma} - 1}{1 + 1/ke^{-2Z_\alpha\sigma}}, \frac{1/ke^{2Z_\alpha\sigma} - 1}{1 + 1/ke^{2Z_\alpha\sigma}} \right] \quad (3.7)$$

The confidence bounds for rejecting that 2 sets of autocorrelation coefficients are the same.

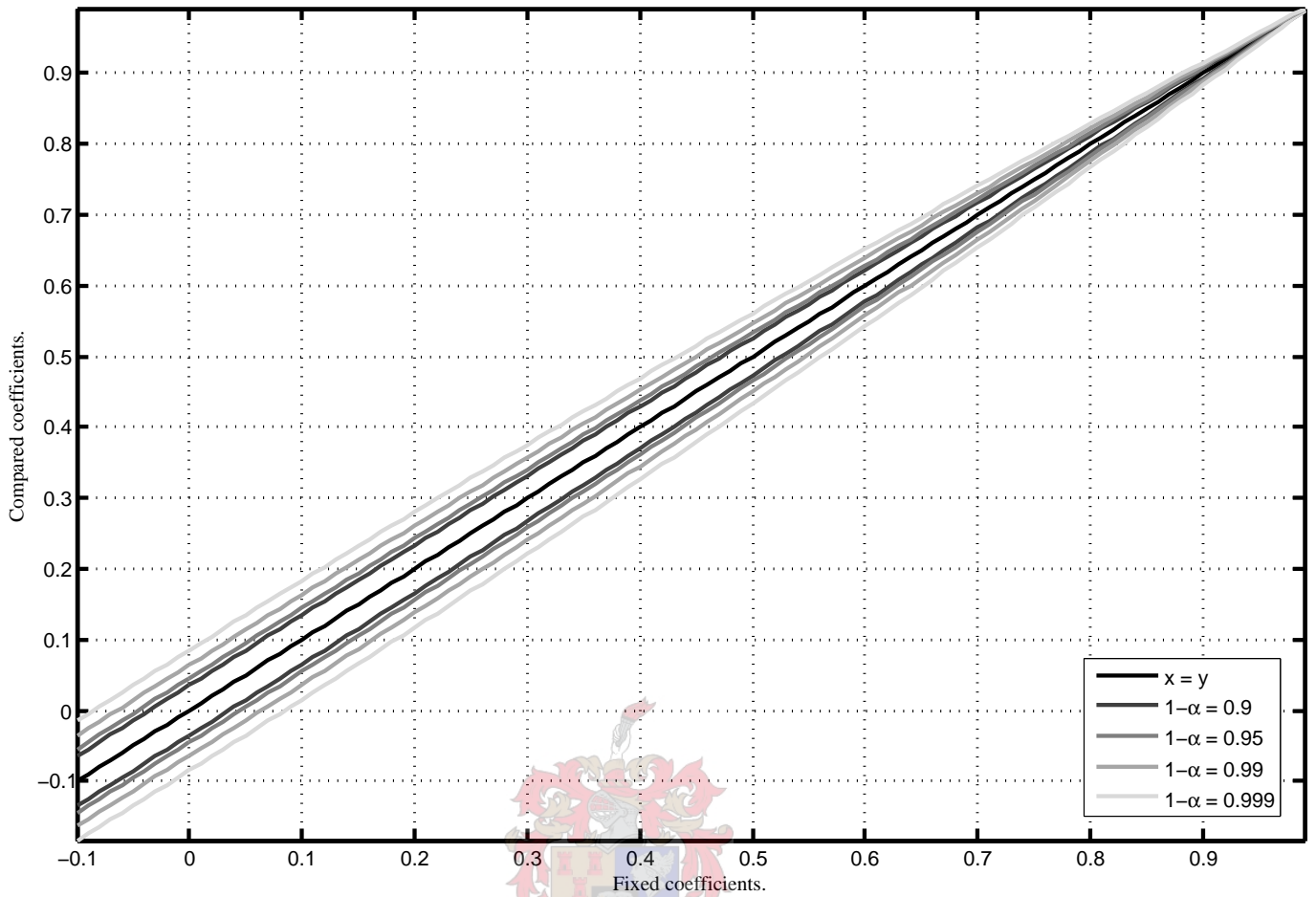


Figure 3.8: The contours of the probability density drop off around the null hypothesis that the correlation coefficients are the same for the number of data points used in the surrogate data analysis.

where

$$k = \frac{1 + AC_n^{data}}{1 - AC_n^{data}} \quad (3.8)$$

$$\sigma = \sqrt{2/(T - n - 3)} \quad (3.9)$$

provides the confidence boundary for rejecting the null that autocorrelation coefficients for the surrogates are the same as those of the data set.

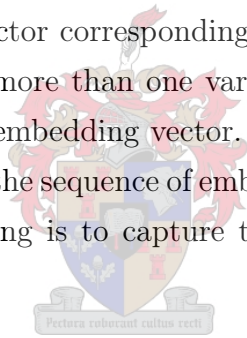
Figure 3.8 shows confidence intervals for different α 's for the null that the autocorrelation coefficients of the surrogate data do not differ from that of the data. If the autocorrelation coefficients of surrogate data lie outside the $x = y$ relationship demarcated by the α boundary, it can be concluded that the autocorrelation coef-

ficients are different. The variance term in the confidence boundary is set at the number of points used in computing the autocorrelation coefficients for the T4 data set and its surrogates. Note that the graph is only drawn for coefficients in the range of -0.1 to about 1. The actual range of the function is from -1.0 to 1.0 but the functional relationship for the lower range of coefficients is symmetrical to the upper range of coefficients and only the upper range was drawn here. More information on the PDF and autocorrelation coefficients can be found in Law and Kelton (1991).

3.2.5 Embedding

As mentioned previously, hypothesis testing involves comparing invariant properties of surrogate data and the original data in respective reconstructed phase spaces. The reconstructed phase space is obtained using delay embedding methods.

A delay embedding of observed data is a set of vectors of fixed dimension, with adjacent components in a vector corresponding to consecutive observations. If the observed system consists of more than one variable, the different variables can be combined to obtain a single embedding vector. The components of a multivariable embedding vector consists of the sequence of embedding components of each variable. The purpose of the embedding is to capture the time invariant properties of the observed system.



Suppose a data set of $i \in \{1, 2, 3, \dots, N\}$ observations consists of three variables, Dt_i , $\log(M_i)$ and $\log(E_i)$. Denoting the lag length for each vector by lag_{dt} , lag_M and lag_E and the observation indices of the first component as $k \in \{i, i + 1, i + 2, \dots, (N - \max(lag_{dt}, lag_M, lag_E) + 1)\}$, then the three consecutive multivariable embedding vectors for the data set are given by

$$\begin{aligned}
 \overline{Seis}_k &= (Dt_k, Dt_{k+1}, \dots, Dt_{k+lag_{dt}-1}, M_k, \dots, M_{k+lag_M-2}, M_{k+lag_M-1}, E_k, \dots, E_{k+lag_E-1}) \\
 \overline{Seis}_{k+1} &= (Dt_{k+1}, Dt_{k+2}, \dots, Dt_{k+lag_{dt}}, M_{k+1}, \dots, M_{k+lag_M-1}, M_{k+lag_M}, E_{k+1}, \dots, E_{k+lag_E}) \\
 \overline{Seis}_{k+2} &= (Dt_{k+2}, Dt_{k+3}, \dots, Dt_{k+lag_{dt}+1}, M_{k+2}, \dots, M_{k+lag_M}, M_{k+lag_M+1}, E_{k+2}, \dots, E_{k+lag_E+1})
 \end{aligned} \tag{3.10}$$

The dimensions of the set of vectors are $D_{embed} = \sum_{i \in \{dt; M; E; \}} lag_i$. The set of vectors

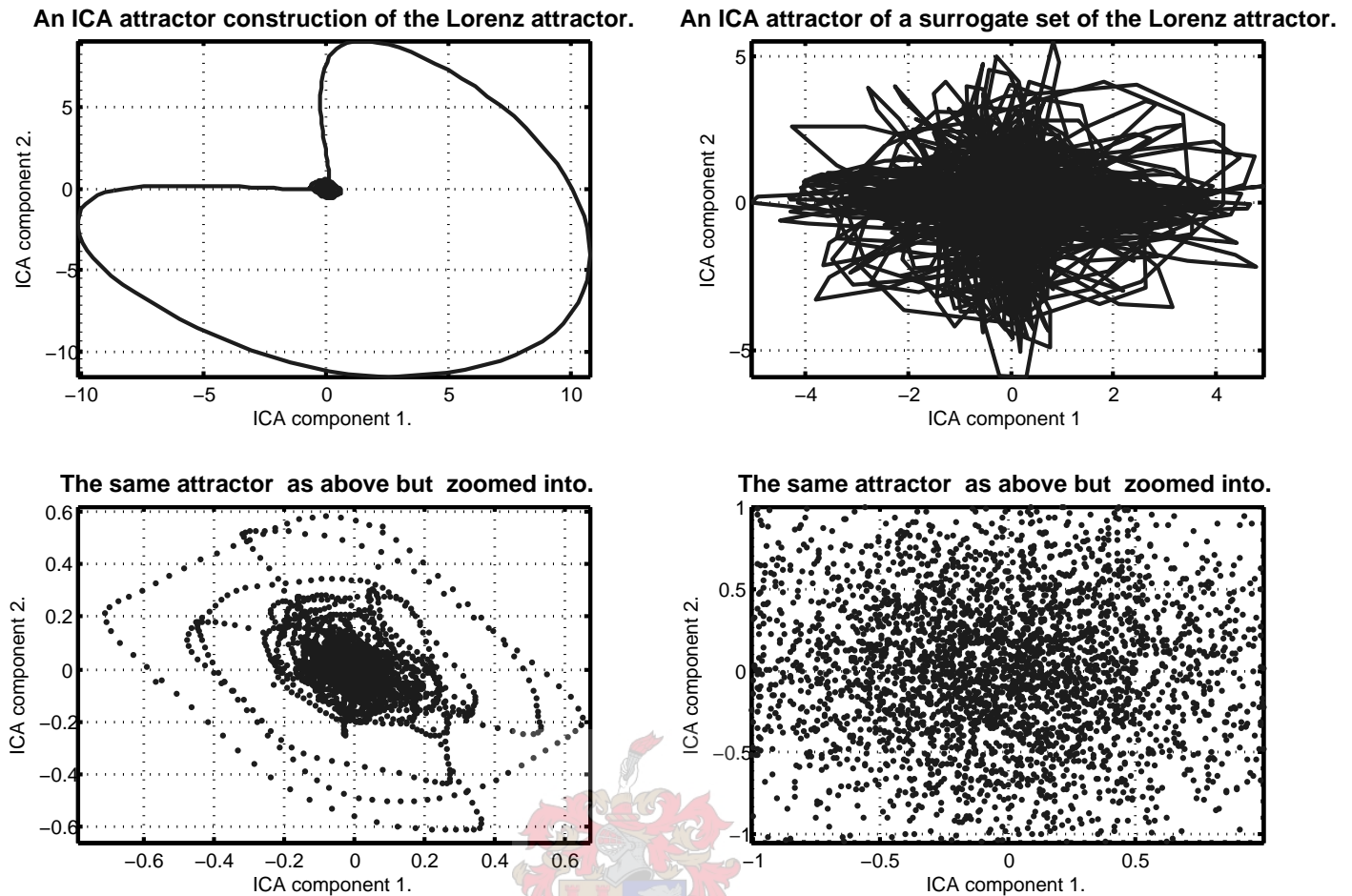


Figure 3.9: A picture of two ICA attractors. The two on the left are the Lorenz attractor the, two on the right a surrogate data set of variables. The bottom graphs are the same attractors as the top graphs, but are enlarged.

represents a sample of a phase space diffeomorphically equivalent¹ to the state space of the original system. The dimension of the reconstructed phase space can be reduced using a linear map of full rank, since the equivalence is maintained under such a transformation. This is important in situations when there is insufficient data to work in a high dimensional space.

The reconstructed attractor has a number of invariant properties not affected by linear transformation such as the average mutual information and the correlation dimension. These properties serve as a signature for the attractor. However, these are affected by the choice of embedding delay lag and the sampling rate of the data.

¹Diffeomorphical equivalence exists between a vector field and its transform if the vector field is subject to a continuous map with a continuous differentiable inverse.

3.2.6 Independent component analysis

The test statistics for hypothesis testing are sampled from the reconstructed phase space of the two populations of data sets. The phase space is obtained by lag embedding of the observed variables. The dimension of the embedded vector space for the surrogate analysis of T4 is too large to accurately sample the test statistics. Fortunately, the power of the test statistics to reject the hypothesis is unaffected by reducing the dimension of the embedding with a linear map of full rank (Sauer, 1991). The test statistics are more readily sampled from the lower dimensional phase space than the original embedding space. One such dimensionality reduction method is independent component analysis.

Independent component analysis (ICA) (Haykin, 1999; Hyvärinen, 1999) seeks the inverse of a mixing, linear transformation W of full rank for a number of statistically independent signals s_i of different non-Gaussian distributions. The mixing matrix and the signals are unknown, and only a finite sample of the mixed signals, $y_i = A \times s_i$, are available to infer W . The number of mixed signals are set *a priori*, providing a linear mapping W of full rank to reduce the dimension of the original mixed signals.

The inference mechanism is based on the fact that a linear combination of two independent non-Gaussian random variables is closer or equal to a Gaussian random variable than the original variables taken separately. Each rank of the inverse linear transformation is then picked as a linear combination of the mixed signals that maximizes the non-Gaussianity of the source signal, and is orthogonal to the other inverse linear transformations to form W of full rank. The measure of non-Gaussianity is an approximation of negentropy. The Gaussian distribution has the maximum entropy score for any distribution with a given mean and variance. The negentropy is the difference between the entropy score for a normal distribution of similar mean and variance as the signal and the actual score of the signal itself. Maximizing the negentropy results in a random variable of least Gaussianity.

Computing the entropy for a given signal is computationally intensive and, therefore, a robust approximation of the negentropy is usually used. The negentropy measure is a generalized measure of the kurtosis or sharpness of the distribution. Maximizing the negentropy approximation consists of iteratively adjusting the inverse mixing matrix until convergence is reached (Hyvärinen, 1999, 2003). Each independent

component estimate is initialized from a random number generator, which introduces a variance into the value of the final linear transform for the same set of parameters and mixed signals. The specific implementation of the ICA algorithm used in the thesis can be found in Appendix B.1.

ICA requires that the variables be uncorrelated or else the assumptions of the optimization technique fails. Hence, before applying ICA a data set needs to be pre-processed to remove cross-correlations in the data. This can be achieved using principal component analysis (PCA). PCA defines a linear map of full rank mapping a correlated set of random variables on to corresponding scores with least possible correlation (Haykin, 1999). In essence, PCA is coordinate rotation of basis vectors such that each of the new basis vectors represents an uncorrelated portion of the variance of the vector set. Rotation being a linear transformation, and as a result of the central limit theorem, if a large initial dimension can be reduced through the rotation, the resulting observations would be closer to normal than the original observations. PCA can also be used as a dimension reduction technique. However, in the case of the T4 data the size of the variance in the rotated directions was found to be too small to effectively reduce the dimension.

The ICA linear map is used to reduce the dimension of the reconstructed phase space of T4 without affecting the power of the test statistics. Significant dimension reduction using only the PCA map on the reconstructed phase space of T4 and its surrogates was not observed. ICA successfully converged to a transform into a lower dimension without significant loss of the variance observed in the reconstructed phase space of T4.

3.2.7 Discriminating statistics

The average mutual information and $D_c(\epsilon/\epsilon_o)$ scores can be used as test statistics on the population of reconstructed phase spaces (Judd, 1994; Kantz and Schreiber, 1997). These test statistics have the power to discriminate between different classes of reconstructed phase spaces. Analytical distributions for the test statistics under assumption of a specified system are not generally available. Numerical realizations of the test statistics for some classes of parameters are available. If the population of a given test statistic has been sampled n times, the variance of the population has been sufficiently observed to provide a confidence interval on the statistic of $1 - \frac{1}{n-1}$ and a hypothesis test based on the statistic at an $\alpha = \frac{1}{n-1}$ level of confidence

(Schreiber and Schmitz, 1996, 2000).

Average mutual information

The first test statistic we will consider is the average mutual information (AMI). AMI is a measure of the amount of information the value of one variable provides on the value of another variable. Formally, the average mutual information is a measure between two dependent, discrete measurements, X_i and Y_i , $i = 1 \dots N$, with $X_i, Y_i \in \mathcal{N}$, of the average amount of information X_i provides on Y_i (Fraser and Swinney, 1986). If (X_i, Y_i) is sampled sufficiently from the trajectory of an orbit which is long enough, the AMI provides a quantitative characterization of the trajectory's invariant qualities. AMI is derived from Shannon's information theory and defined here as:

$$AMI(X, Y) = \sum_{X_i \in M} \sum_{Y_i \in S} P_D(X_i, Y_i) \log_2 \left(\frac{P_D(X_i, Y_i)}{P_M(X_i)P_S(Y_i)} \right) \quad (3.11)$$

where P_D is the joint probability distribution of the discrete random variables X_i and Y_i , P_M and P_S are the distributions of X_i sampled from the population of $M \subset \mathcal{N}$ and Y_i sampled from the population of $S \subset \mathcal{N}$. All the probability distributions are estimated empirically using a histogram. Terms of the summation for which the denominator will result in zero are assigned zero values. AMI is measured in bits when base two logarithm is used as above.

The AMI of the reconstructed high dimensional phase space is measured by taking the norm of consecutive vectors in the domain of the ICA map. The time ordered sequence of norms provides a measure of the bumpiness of the phase space. A predictable system typically has a smoother phase space compared to red colored noise. Thus, AMI scores on the variables populating respective reconstructed phase spaces can be used to distinguish a red colored noise process and a predictable data set. The surrogate approach is particularly useful in this respect since no analytical model exists relating the distribution of the average mutual information to the red colored noise specified by the null hypothesis.

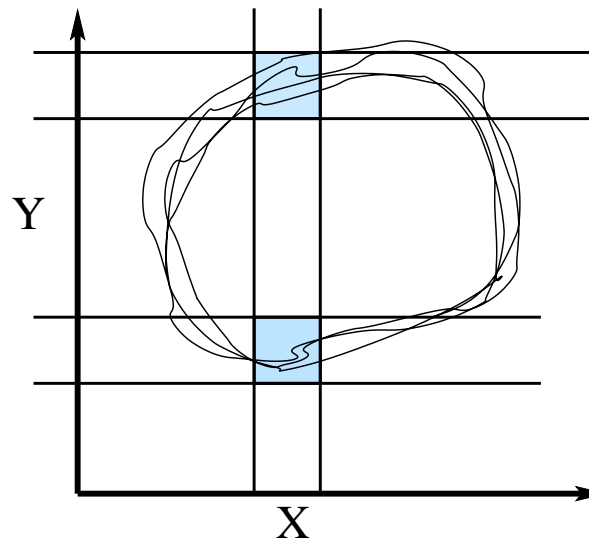


Figure 3.10: The AMI between X and Y is a measure of how much the value of X is pinned down given the value of Y , standardized to the case where there is no dependence. The AMI is computed by sampling the shared distribution of X and Y .

Correlation Dimension Estimate

The second of the test statistics considered is the correlation dimension estimate. The correlation dimension separates predictable data from red colored noise based on the self-similar behavior of the reconstructed phase spaces.

The correlation dimension characterizes the extent of the self-similarity of a reconstructed vector field. It is used to track the influence of the parameters of the reconstruction process on the scaling behavior of the reconstructed field. Possible scaling properties give an indication that the reconstructed phase space does not fill the observation space. A lower dimension measure in a higher dimensional observation space indicates that the observed vector field has been constrained to the lower dimension. A phase space reconstruction of red colored noise will eventually fill the space as the dimension increases.

The correlation dimension measures the geometrical structure of the sample of points over a range of small scales, $d = D_c(\epsilon)$, $\epsilon \in (\epsilon_{min}, \epsilon_o)$ (Judd, 1992, 1994). The dimension estimate is the exponent of the probability drop off in the tail of the inter-point distribution for small scale lengths ϵ for a general family of distributions. Any distribution will suffice as long as the tail of the distribution over small inter-point lengths is an asymptotic approximation of a probability measure of a cross-

product of a separable self-similar measure and an absolutely continuous measure. The functional form of the tail of the probability measure is given by

$$P(\epsilon) = \epsilon^d \left(\sum_{i=0}^t a_i \epsilon^i \right), \quad (3.12)$$

where $P(\epsilon) = Pr(|X - Y| < \epsilon)$ is the probability measure on the distribution of inter-point distances. A binning procedure is used to approximate the distribution of inter-point distances. Only the small scale characteristic of the distribution is of interest, hence all the inter-point distances greater than ϵ_o are grouped in the largest bin. Bin edges follow an even geometrical sequence starting from ϵ_o and ending in ϵ_{min} such that $\log(\epsilon_{min}) = \log(\epsilon_o) - n * b$, for n bins.

The dimension estimator $D_c(\epsilon/\epsilon_o)$ characterizes a self-similar scaling range if an ϵ -region (on a logarithmic scale) with an approximately constant value can be identified. The estimate has the power to characterize self-similar scaling behavior, or lack thereof, for a range of non-fractal probability measures. The power to discriminate is useful because the use of a test for self-similarity with a test statistic that is only valid under the assumption of self-similarity might lead to invalid conclusions. The estimate is accurate over a dimension range between one and four. The correlation dimension over the scaling range is quoted as the y-axis position of the flat portion.

The D_c estimates provide an indication of the scattering of the reduced vectors populating the phase space. If the data set scatters differently from red colored noise of the surrogates, the data will exhibit some structure on small scales that is not explained by a red colored noise process.

3.2.8 Surrogate data

Surrogate data analysis of T4 involves performing a hypothesis test that the data exhibits no more structure than a pre-specified null hypothesis. The hypothesis is conducted by sampling surrogate data sets from the same distribution as T4, assuming T4 was sampled from a red colored noise, i.e. if the null hypothesis were true. The T4 data set consists of a 3-dimensional data set exhibiting auto- and cross-correlations, with each component sampled from a unique non-normal distribution.

To generate surrogate data, the observed data are first shuffled and the amplitudes iteratively adjusted in the Fourier spectrum transform domain to maintain the rank

ordering and the energy spectrum of the original data (Schreiber and Schmitz, 1996). The adjustment is made by re-ordering the rank order of the surrogate variable after fixing the energy spectrum of each of the observed variables, and mapped by the inverse Fourier transform in sequential iterations. The cross-correlations are maintained by adjusting the angle of the surrogate variable in the Fourier domain of the iteration. In short, a surrogate data set is a shuffle of the original data set in such a way that the auto- and cross-correlations are maintained. If the data set is large enough and sampled from a specified red colored noise, the resulting surrogate data set will be sampled from the same source.

Suppose $\{\overline{X}_n^m\}$ is an ordered sequence of m -dimensional vectors of length $n = 1 \dots N$, $fftm(\cdot)$ is the fast Fourier transform for m components of each frequency, and $ifftm(\cdot)$ is the corresponding fast inverse Fourier transform. Then, $\{\overline{z}_k^m\}$ is the m -dimensional sequence of complex numbers of length, $k = 1 \dots K$ defined as

$$\{\overline{z}_k^m\} = \{|\overline{z}_k^m| \exp^{i \arg(\overline{z}_k^m)}\} = fftm(\{\overline{X}_k^m\}) \quad (3.13)$$

with $|\cdot|$ denoting the norm and $\arg(\cdot)$ denoting the angle of the complex numbers (Weisstein, 2001).

The sequence of complex values $\overline{a} + i \overline{b} = ifftm(\overline{z}_k^m)$ is the result of the inverse fast Fourier transform. For the multivariate surrogate generation \overline{b} is nonzero since $\{\overline{z}_k^m\}$ is not symmetrical (MATLAB[®], 2001).

Furthermore, let $\{rank(\overline{X}_n^i)\}$ denote a vector of the rank ordering of one of the components of \overline{X}_n^m such that

$$\{\overline{C}_n^i\} = \{\overline{X}_{rank(\overline{X}_n^i)}^i\} \quad (3.14)$$

is the non-descending ordered sequence of numbers in $\{\overline{X}_n^i\}$ for each of the components of $\{\overline{X}_n^m\}$, i.e. $i = 1 \dots m$, and

$$\{\overline{C}_n^m\} = \{\overline{X}_{rank(\overline{X}_n^m)}^m\} \quad (3.15)$$

denotes the sequence of m dimensional vectors with each entry a component of $\{\overline{X}_n^m\}$ ordered in ascending order.

Let $\{\overline{rand}_n^m\}$ be a sequence of random numbers of dimension m . Then, the surrogate variable \overline{rV}_n^m for \overline{X}_n^m is initially sampled as

$$\{\overline{rV}_o\} = \{\overline{X}_{rank(\overline{rand}_n^m)}^m\} \quad (3.16)$$

and iterated to the correct auto-correlation structure as follows:

1. Map the random variable, \overline{rV}_i into the frequency domain:

$$\{\overline{rZ}_i^m\} = fftm(\{\overline{rV}_k\}) \quad (3.17)$$

2. Adjust the angles, $\overline{\rho}_k^m = arg(\overline{rZ}_k^m) + \alpha_k$ with:

$$\alpha_k = \left| \arctan \left(\frac{\sum_{m=1}^M \sin(arg(\overline{z}_k^m) - \overline{\rho}_k^m)}{\sum_{m=1}^M \cos(arg(\overline{z}_k^m) - \overline{\rho}_k^m)} \right) \right| + \pi * q/2 \quad (3.18)$$

$$q \in \{0; 1; 2; 3\} \text{ such that} \quad (3.19)$$

$$\sum_{m=1}^M (\alpha_k - \pi * q/2 - arg(\overline{z}_k^m) + \overline{\rho}_k^m) \text{ is a maximum.} \quad (3.20)$$

3. Adjust the amplitudes and the angles of the Fourier surrogates and map the adjusted surrogates back using $ifftm(\cdot)$

$$a_n^m + i * b_n^m = ifftm(|\overline{z}_k^m| \exp^{i\overline{\rho}_k^m}) \quad (3.21)$$

$$\overline{s}_{i_n}^m = a_n^m \quad (3.22)$$

4. Scale the values of the amplitude adjusted random variable, \overline{s}_n^m to the correct distribution:

$$\overline{rV}_{k+1} = \{C_{rank(\overline{s}_{i_n}^m)}^m\} \quad (3.23)$$

5. Repeat the iteration until convergence is reached:

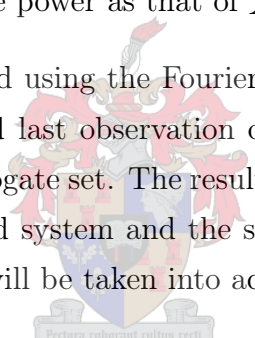
$$rank(\overline{s}_{i_n}^m) \text{ is the same as } rank(\overline{s}_{i+1_n}^m) \quad (3.24)$$

The iteration is repeated until the rank ordering adjustment converges. An implementation of the algorithm is provided in Appendix B.2.

Note that the surrogate generating mechanism optimizes the surrogate multivariable data set to a given cross-correlation for a fixed distribution of each of the components. It does not ensure that the surrogate data set has the same multivariate distribution. The correspondence in the cross-correlations between the surrogate data and the original data set will ensure the same multivariate distribution if the components of the original data are distributed normally (Law and Kelton, 1991).

As an example of the iteration sequence consider Table 3.1. The first column of the table is a 10-dimensional vector of observations, each consisting of one component. The 10-dimensional vector rV is the randomly shuffled sequence of \bar{X} observations. In the first iteration, fixing the power spectrum of rV_i results in a re-shuffling of \bar{X} to form a new rV_{i+1} . Once the order of observations in rV_i are fixed, the phases of rV_{i+1} are fixed at the same power as that of \bar{X} .

Surrogate data sets computed using the Fourier spectrum tend to distribute a difference between the first and last observation of a data set as energy in the high frequency domain of the surrogate set. The resulting difference in the high frequency domain between the observed system and the surrogate system influences the test statistics. This observation will be taken into account in the subsequent analysis.



\bar{X}	First iteration			Second iteration		Third iteration	
	$\bar{\Sigma}$	rV	arg(fft(rV))	rV	arg(fft(rV))	rV	arg(fft(rV))
0	10	1	0	0	0	0	0
1	0	1	1.7407	2	0	2	0
1	3.8042	0	-2.1991	1	-2.1991	1	-2.1991
2	0	1	-0.0766	1	0	1	0
1	2.3511	1	-2.8274	1	-2.8274	1	-2.8274
0	0	0	0	0	0	0	0
1	2.3511	2	2.8274	2	2.8274	2	2.8274
1	0	2	0.0766	1	0	1	0
2	3.8042	1	2.1991	1	2.1991	1	2.1991
1	0	1	-1.7407	1	0	1	0

Table 3.1: Three iterations of an IAAFT surrogate generated for the 10-dimensional vector \bar{X} , depicted in the first row. The $\bar{\Sigma}$ row provides the energy spectrum of \bar{X} . rV_i is the shuffled sequence of \bar{X} serving as the surrogate. The iterations continue until the phases of rV_{i+1} are fixed at the same power as that of \bar{X} .

3.2.9 Hypothesis testing

A **test statistic** is a sampled value from an unknown distribution. A **parameterized** test statistic belongs to a parameterized distribution of unknown value. A **parameterized distribution** defines a family of distributions, each specific distribution in the family defined by a different value of the parameter. A **hypothesis** about a test statistic is a pair of propositions, each specifying a set of mutual exclusive parameters, restricting a parameterized test statistic to a mutual excluding subset of its family of distributions under assumption of each of the propositions. The pair of propositions is known as the **null hypothesis** and the **alternative hypothesis**, each hypothesis corresponding to one of the propositions. A **hypothesis test** rejects that the null hypothesis is true in favor of the alternative hypothesis if the observed test statistic of unknown distribution falls within the α -probable tail portion of the distribution defined by the assumed value for the parameter.

The probability of rejecting a null hypothesis as false if the distribution of the test statistic holds under the null hypothesis is specified by the probability α . The probability that the hypothesis is not rejected given that it is false, is the probability that the test statistic was sampled from the distribution specified by a false null hypothesis and a true alternative hypothesis. Since the distribution of the test statistic under assumption of the alternative hypothesis is not always known, the probability that a false null hypothesis is not rejected is not always known. The difficulty in setting up a hypothesis test is in finding the uniquely specified distribution of the test statistic under assumption of the null hypothesis. On establishing the distribution of the test statistics under assumption of the zero hypothesis, the critical value at which the hypothesis is rejected is fixed (Dudewicz and Mishra, 1988).

The hypothesis testing procedure can be summarized as follows.

1. Generate sequences of red colored noise similar to the data set;
2. Measure the test statistics for each sequence;
3. Evaluate sufficient test statistics for each sequence to give the correct probability of false rejection;
4. Compute the test statistic for the data set;
5. Compare the surrogate test statistics to the data test statistic;

The realization of two test statistics with a difference in distribution parameter value.

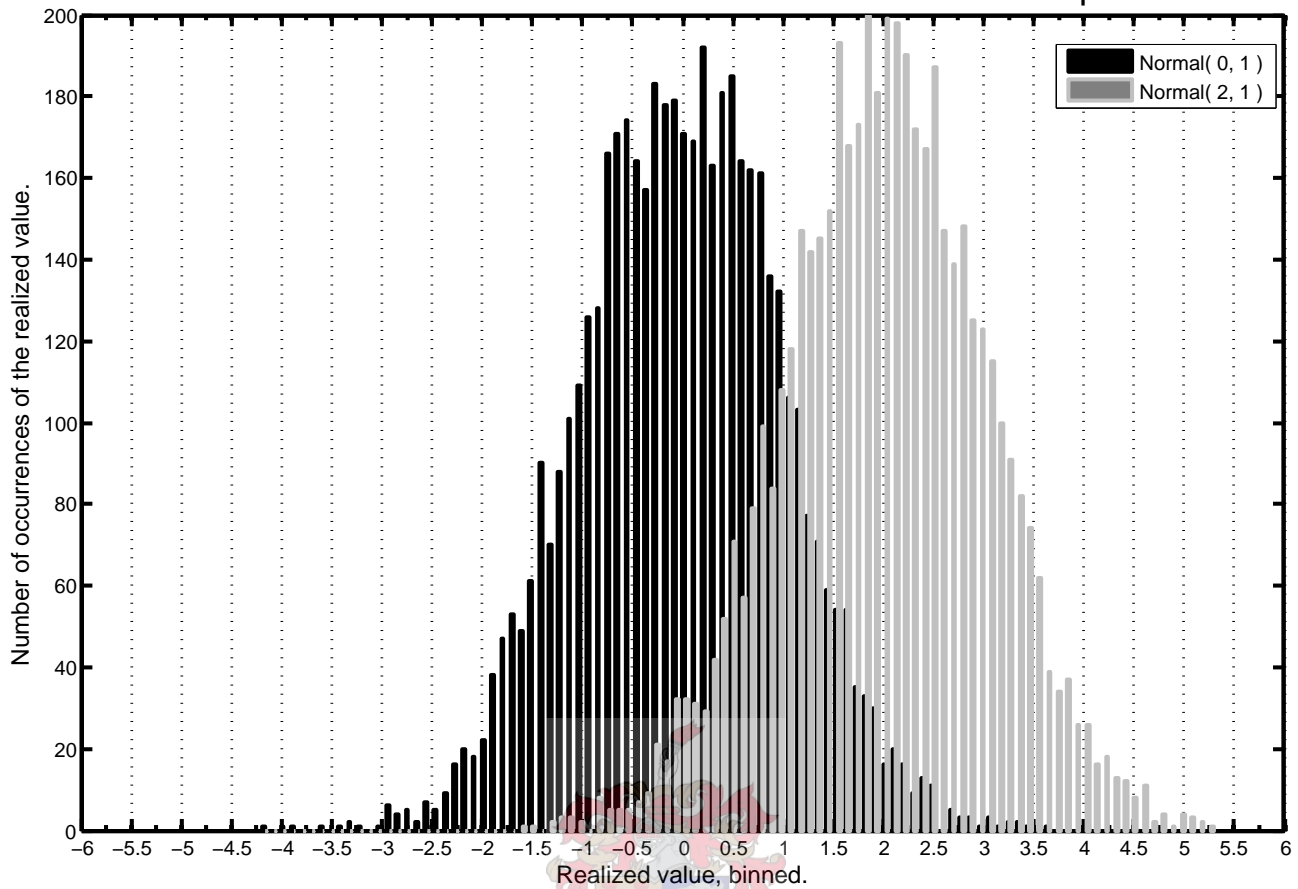


Figure 3.11: Two histograms of the realized values of two test statistics belonging to the same family of distributions but with different parameter values, changing the distribution of each.

6. Test the null hypothesis that the data belongs to the population of red colored noise if the computed test statistics do not differ.

Figure 3.11 depicts the realization of two test statistics with normal distributions of unit variance, but different means of 0 and 2, $normal(0, 1)$ and $normal(2, 1)$. Each test statistic is sampled from a parameterized distribution belonging to the family of normal distributions with unit variance. The parameter separating the family of distributions into different groups is the mean of each distribution. The hypothesis that the means of the two distributions are the same can be tested by sampling a number of realizations from $normal(0, 1)$ and sampling one realization of $normal(2, 1)$. If 11 values of $normal(0, 1)$ are sampled and a value of $normal(2, 1)$ is larger than the largest value of the 11, the hypothesis that $normal(0, 1)$ and $normal(2, 1)$ are the same distribution can be rejected at an $\alpha = 1/10$ level of false rejection. Since the two distributions are only separated by their mean, if the

distributions are not the same, the hypothesis that their means are the same can be rejected.

3.2.10 Generating synthetic seismic data

The mechanisms for generating synthetic seismic activity can be divided roughly into two categories. The first category consists of the point process arising from the interaction between small units, each one governed by the same set of laws and driven by a noise signal. It is commonly classified as a form of cellular automaton models. The second category consists of the direct transformation of a noise signal into the point process.

A cellular automaton was the first model used in the synthesis of self-organized criticality behavior (Barriere and Turcotte, 1994). The relationships between cellular automaton to seismicity of a single fault (Ben-Zion, 1996; Kaiser and Tang, 1998) and seismic active region (Gabrielov et al., 2000) have been discussed. Other cellular automaton models have been used to simulate the general failure process (Newman, 1995). In each of the cellular automata the system is driven in some random way until a unit reaches a threshold value and fails. The failed unit interacts with its associated neighboring units according to some set of functions. This might result in more units failing. A seismic event is associated with a sequence of failures, or the failure of a single unit with a large threshold value. The resulting kinematic description of the failure process (Pollard, 2000) is a general description of some of the seismic modeling algorithms (Ben-Zion, 1996), corresponding to discrete event simulations (Law and Kelton, 1991).

The direct transformation of a noise signal into a point process forms the basis of constructing a probability density field for the occurrence of seismic events. The construction of a probability density for the occurrence of seismic events implies the transformation of a noise signal into a point process. The probability field implies the construction of synthetic seismic activity. Examples of this style of synthetic seismic activity include Akkaya and Yüçemen (2000); Akkaya and Yüçemen (2002); Helmstetter and Sornette (2002); Hooge et al. (1994); Kagan and Vere-Jones (1996); Sornette et al. (1992).

The problem of the existence of repeatable behavior exhibited by a specified seismically active area observed in the mines has been highlighted in Chapter 1. Repeata-

bility is usually initially tested in terms of the hypothesis that the system is separable from stationary red-colored²(Schreiber and Schmitz, 2000). A deterministic system can be distinguished from a stochastic system in a hypothesis test using appropriate pivotal test statistics, even if the stochastic system exhibits similar observational properties to the seismic data. Distinguishing a stochastic and another system depends on the test statistic's power to differentiate between the two (Dudewicz and Mishra, 1988).

3.2.11 ISSI's attractor analysis

ISSI used seismic event measurements from a given volume of rock and analyzed reconstructed attractors using each variable separately. They used the estimated autocorrelation function to determine the delay lag. For the given lag, the Grassberger-Procaccia (Lai and David, 1998) method was used to estimate a correlation dimension. Based on this analysis, it was concluded that an attractor did exist since in three of the data set since the plot of the dimension versus the correlation dimension showed a plateau region between the dimensions of 4 to 8 for all the variables.

ISSI's analysis of the attractor can be critiqued from a number of angles. The descriptive statistics used made it very difficult to analyze other data sets using the same method. Also, the Grassberger-Procaccia (Lai and David, 1998) method was not originally developed for determining a point estimator, especially for systems exhibiting high-dimensional chaos (Kantz and Schreiber, 1997). It gives an indication of low-dimensional chaos with the distinction made on a series of curves. The range of dimensions they used to plot the D_c estimate curve ranged from 1 to 10. Not all point processes convey attractor information in this range of values. It has been reported that integrate-and-fire point processes convey attractor information in the time intervals between events(Sauer, 1994). In all data sets analyzed, only one data set consisted of inter-event time intervals, and the time intervals was measured only between consecutive events not taking the effect that different event sizes could have on the overall seismic rate(Helmstetter and Sornette, 2002).

Figure 3.13 shows correlation dimension estimates for estimated attractors of the

²**Red-coloured noise** (Schreiber and Schmitz, 1996) is known as an autoregressive moving average sampled from a normal distribution and scaled by an invertible nonlinear function, $s(\cdot)$:

$$s_n = s(y_n), \quad y_n = \sum_{i=1}^M a_i y_{n-i} + \sum_{i=0}^N b_i \eta_{n-i} \quad (3.25)$$

where the sequence of η_{ns} is sampled from Gaussian uncorrelated random variables of zero mean.

The inter-event time intervals of T4, D_t , in order of occurrence.

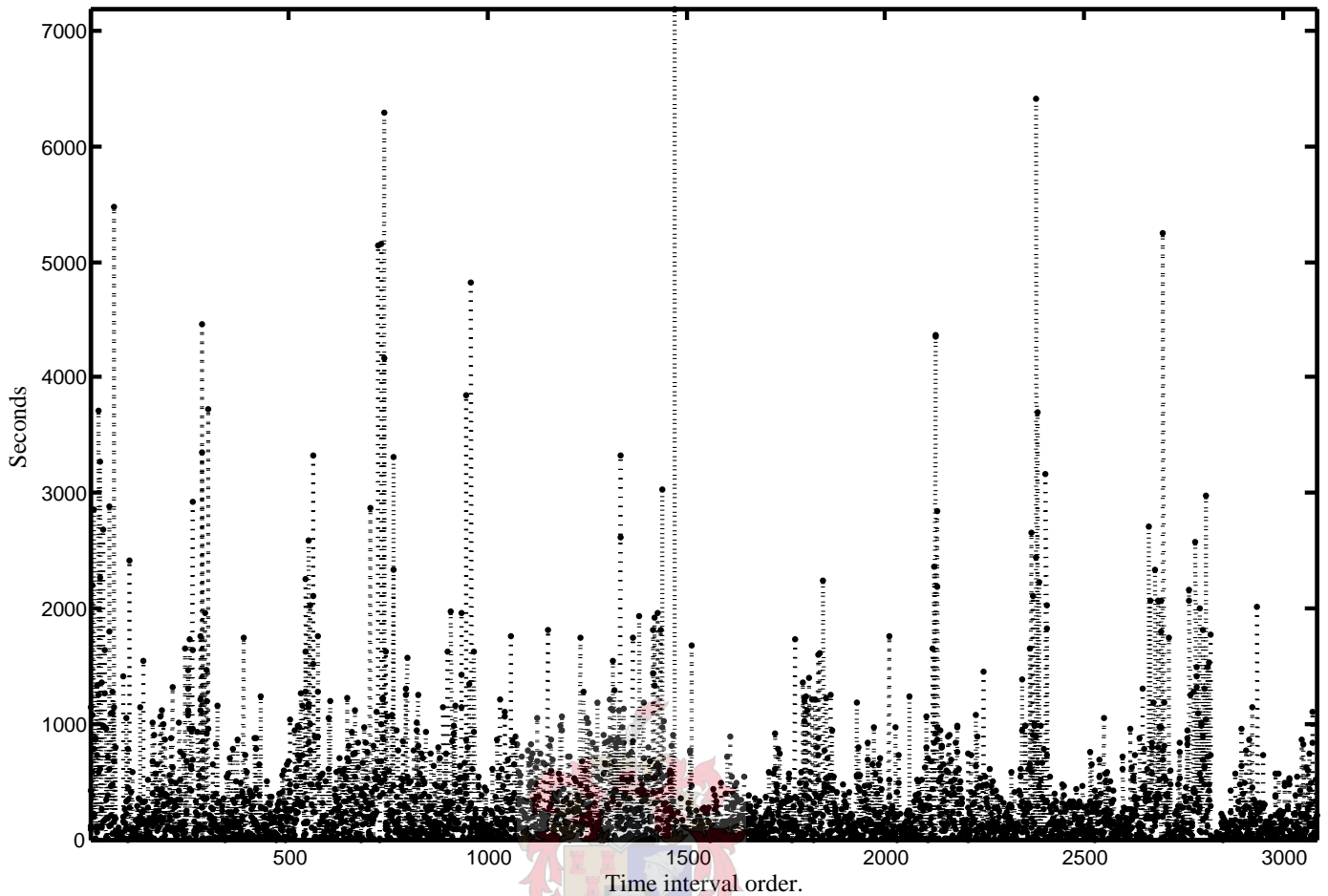


Figure 3.12: The time interval between two consecutive events, for all the events in the T4 Data set.

indicated data using Judd's algorithm. The collection of correlation dimension measures, $D_c(\epsilon/\epsilon_0)$, in Figure 3.13 is a series of estimates of the dimension of the reconstructed attractors of the T4's time intervals, depicted in Figure 3.12, and that of a set of surrogate data sets. In Judd's algorithm, correlation dimension estimates $D_c(\epsilon/\epsilon_0) > 0$ are evaluated over a range of scales $\log(\epsilon/\epsilon_0) \in (-\infty, 0]$, proportional to a specified largest scale ϵ_0 . In this context scale refers to a length in the inter-point distances, and the log over the largest scale results in the 0 on the log-scale x-axis of Figure 3.13. If the sampled attractor exhibits self-similarity then Judd's correlation dimension estimate would give nearly the same $D_c(\epsilon/\epsilon_0)$ estimate over a range of ϵ/ϵ_0 's.

Each surrogate set sur_i is a sample from a random variable, $sur_i \in Sur$ of data sets. Each data set in Sur is a sequence of samples from a stochastic system with

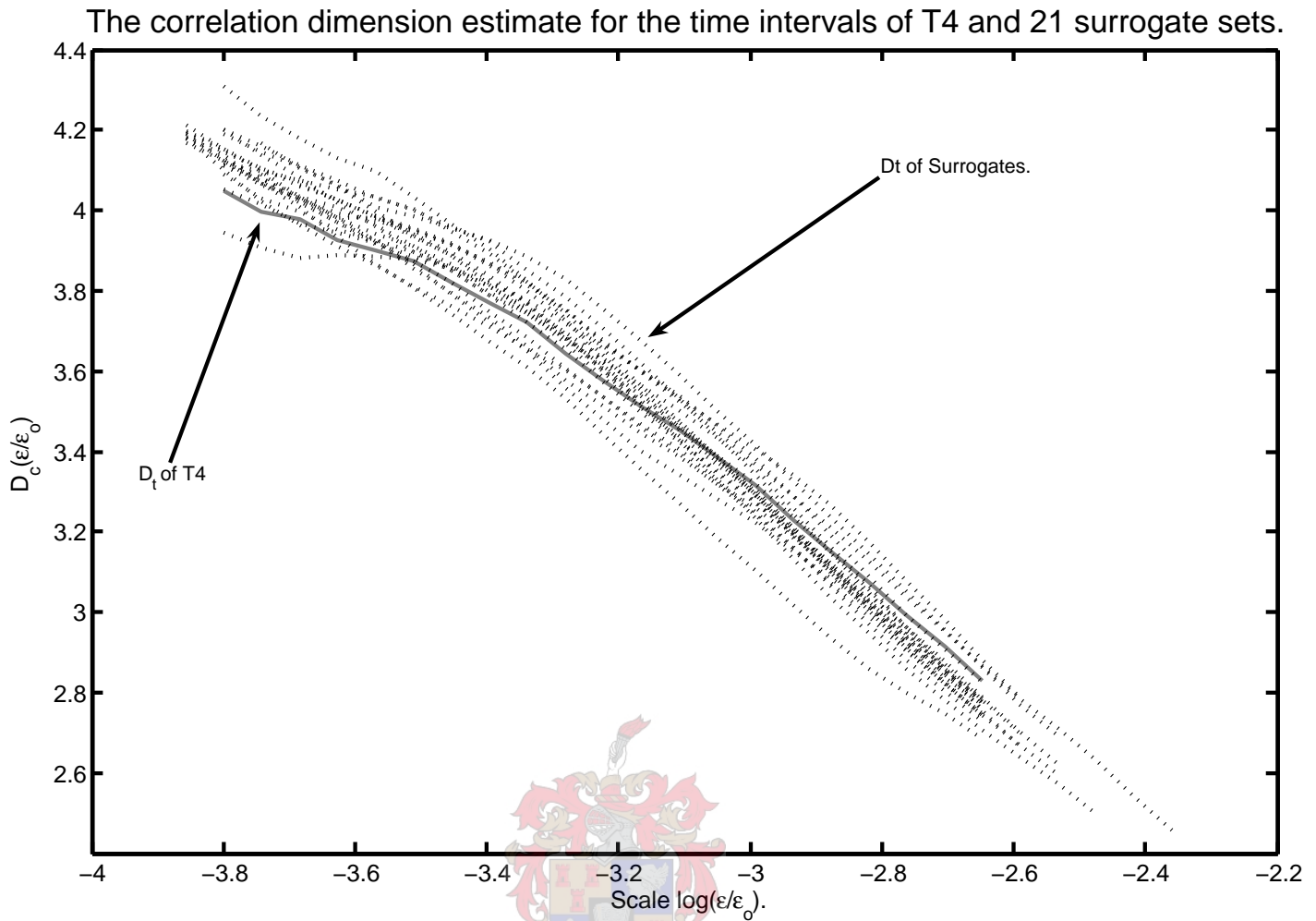


Figure 3.13: The estimate of the correlation dimension of the time intervals of the T4 data set, and a set of surrogates, according to the Judd method. The arrow marked with, D_t of T4, indicates the solid line of the data set. The other arrow the time intervals of the surrogate data sets.

a specified distribution and autocorrelation structure. T4's time interval sequence belongs to *Sur*, as T4's statistics specifies the distribution and coefficients. Each data set in the collection of reconstructed attractors was constructed according to the method previously used by ISSI for attractor reconstruction from a sequence of time intervals. The idea of generating surrogate time intervals and surrogate testing of time intervals can be found in, for example, Sauer (1994).

From Figure 3.13 it can be concluded that T4 does not separate well from the random variables. Also, T4's correlation dimension is not clearly established since no scaling range giving a constant estimate can be identified. The inter-event time intervals of T4 are not separable from the population of surrogates *Sur*. The hypothesis that the sequence of time intervals of T4 was drawn from a population of stochastic

systems with a specified linear structure, therefore, cannot be rejected based on the D_c estimate.

Since the publication of ISSI's latest reported work (1997), new and improved techniques on establishing and testing attractors have been proposed and successfully applied in other domains of nonlinear dynamical systems theory. It would be useful to extend the previous ISSI analysis with a dependence test, Judd's correlation dimension estimate, determinism test and surrogate analysis.

3.3 The seismic behavior of T4

The T4 data set follows a number of seismic laws or relationships known to exist in seismicity in general. These are the Gutenberg-Richter relationship, occurrence of large seismic event clustering in time, and the log-linear relationship between the $\log(E_i)$ and $\log(M_i)$. The hypothesis test is structured in such a way that these relationships can be explained by the structure of the red colored noise. However, large event clustering might violate the assumption of stationary red colored noise.

When conducting the hypothesis test, selected discriminating statistics are sampled from the reconstructed phase space of components in T4. These are compared to similar measures for the observed data. Each combination of compared components highlights the structure of the reconstructed attractor that is sampled by that combination.

The first step in the hypothesis test is to determine parameters of the seismic laws when fitted to T4. In the second step, the T4 data is characterized as mining-induced seismicity by determining and interpreting test statistics used in the hypothesis tests.

3.3.1 Seismic laws and stationarity

Seismic laws are used to indicate which relationships should be factored into the structure exhibited by the red colored noise. The laws demonstrate which portions of T4 are usable in the hypothesis test. The red colored noise sampled under assumption of the null hypothesis is assumed stationary, that is, the PDF or autocorrelation coefficients do not vary with time. Due to the nature of sampling, small events in T4 are not considered part of a stationary process. Seismic laws are used to identify

the stationary parts of T4.

Law I: Gutenberg-Richter relationship

The Gutenberg-Richter relationship states that an inverse log-linear relationship exists between the number of events and the size of the event. The relationship for the components of T4 are sampled as,

$$Y_{X^L} = \log_{10}\left(\sum_i I(X_i \geq X^L)\right)$$

with $X \in \{Dt, \log(M), \log(E)\}$ such that the y-axis depicts the number of events in T4 larger than the corresponding variable value on the x-axis. Figure 3.14 depicts the sampled scores. The top graph depicts the relationship for Dt_i . The middle graph depicts the relationship for $\log(M_i)$ and the bottom graph for $\log(E_i)$. Each graph shows the measured relationship with the number of events on the y-axis and the corresponding event size on the x-axis. On each graph a model is fitted to the measurement, as well as the coefficients of the model. The models in the graph are fitted on the stationary measurements only.

Three observations can be made from the graphs and the fitted models in Figure 3.14:

- The quadratic models fitted in Figure 3.14 to Y_{M^L} and Y_{E^L} demonstrate deviation from the norm in the T4 data set. The functional relationship of Y_{M^L} and Y_{E^L} does not appear to be a single linear relationship, but rather consists of two linear relationships of different slope, interchanging at $M^L \approx 9.4$ and $E^L \approx 1.6$.

A possible explanation for the change of slope is that a large amount of energy is utilized by the system for fault formation. Therefore, the system producing the smaller events behaves like a seismic system dissipating more energy than the system producing the larger events (McGarr, 2000). If this is the case, the GR relationship is validated for T4 over smaller events.

- The quadratic models show a deviation from the law for small scales. The deviation indicates the non-stationarity in the data set introduced by the sampling process. In the hypothesis test the events below the stationary threshold are omitted from the data set since the hypothesis is sensitive to changes in stationarity. Note, however, that omitting the non-stationary events in $\log(M_i)$ does not necessarily result in the removal of all the non-stationary events, since $\log(E_i)$ still demonstrates the non-stationary signature Darpahi-Noubary (2002).

- The model fit of the time intervals shows that omitting the non-stationary sampled events does not affect the shape of the frequency size relationship of the time intervals Dt_i significantly. A possible reason for this is that the inter-event time intervals and the size of the events are not significantly related.

The frequency-size relationships in the seismic data do not affect the hypothesis of red colored noise because the noise has the same distribution as the components of the data set.

Law II: Large event clustering

The second seismic law concerns the clustering in time of larger seismic events, especially around the large events in the data set. The clustering in T4 is observed by relating the moving average of event size to the corresponding moving average of inter-event time intervals for large events, Figure 3.15. The same variables are related for the whole data set instead of just the large events, Figure 3.16. The relationship is depicted for all three variables.

The depiction of the relationship for the time intervals Dt_i of all the events is an indication of the size of the autocorrelation coefficient for consecutive time intervals. The relationship was added for completeness. A negative linear trend exists between the moving average of event size and the corresponding inter-event time interval average for the $\log(M_i)$ and $\log(E_i)$ components, Figure 3.15. The trend is not evident in the Dt_i component or in the component containing all the events. The trends in Figure 3.15 show that larger events tend to cluster around the large events in the data set.

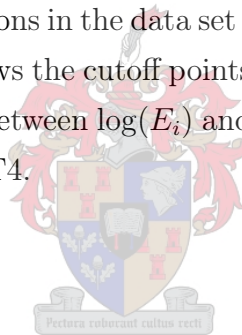
However, the clustering around the large events in the data set did not generally hold for all events as shown in Figure 3.16. The occurrence of aftershocks reported in literature is normally more visible than the results reported here. This can be attributed to at least two deviations observed in T4 and not in seismic data sets in literature. Firstly, the size of the events considered for the generation of aftershocks is significantly smaller. A smaller event size represents a larger example set and demonstrates that the law is actually visible under the smaller events as well. Secondly, the dual system discussed under the G-R relationship might have a dampening effect on the aftershock mechanism compared to the one observed in crust-scale earthquakes. In the crust-scale mechanism a number of readily avail-

able fault lines exist to produce after-shocks. In the mining scenario, as discussed here, opportunities to produce secondary, larger earthquakes might not be so readily available.

The clustering of larger events around large events might cause the rejection of the hypothesis that the structure in T4 is different from red colored noise. The clustering will cause a rejection if the clustering cannot be explained by the auto- and cross-correlation coefficients of the red colored noise. If more structure exists than found in red colored noise modelling, the data should be investigated beyond its linear properties, despite the size of the deviation.

Law III: Linear relationship between $\log(E_i)$ and $\log(M_i)$

The third seismic law relationship is depicted in Figure 3.17 in the scatter plot between corresponding pairs of $\log(E_i)$ and $\log(M_i)$ with a trend line added. Similarly, the other cross-correlations in the data set are also demonstrated in Figure 3.17. The $\log(E_i)$ - $\log(M_i)$ plot shows the cutoff points for the small non-stationary events. Other than the linear trend between $\log(E_i)$ and $\log(M_i)$ no other trends are evident between the components in T4.



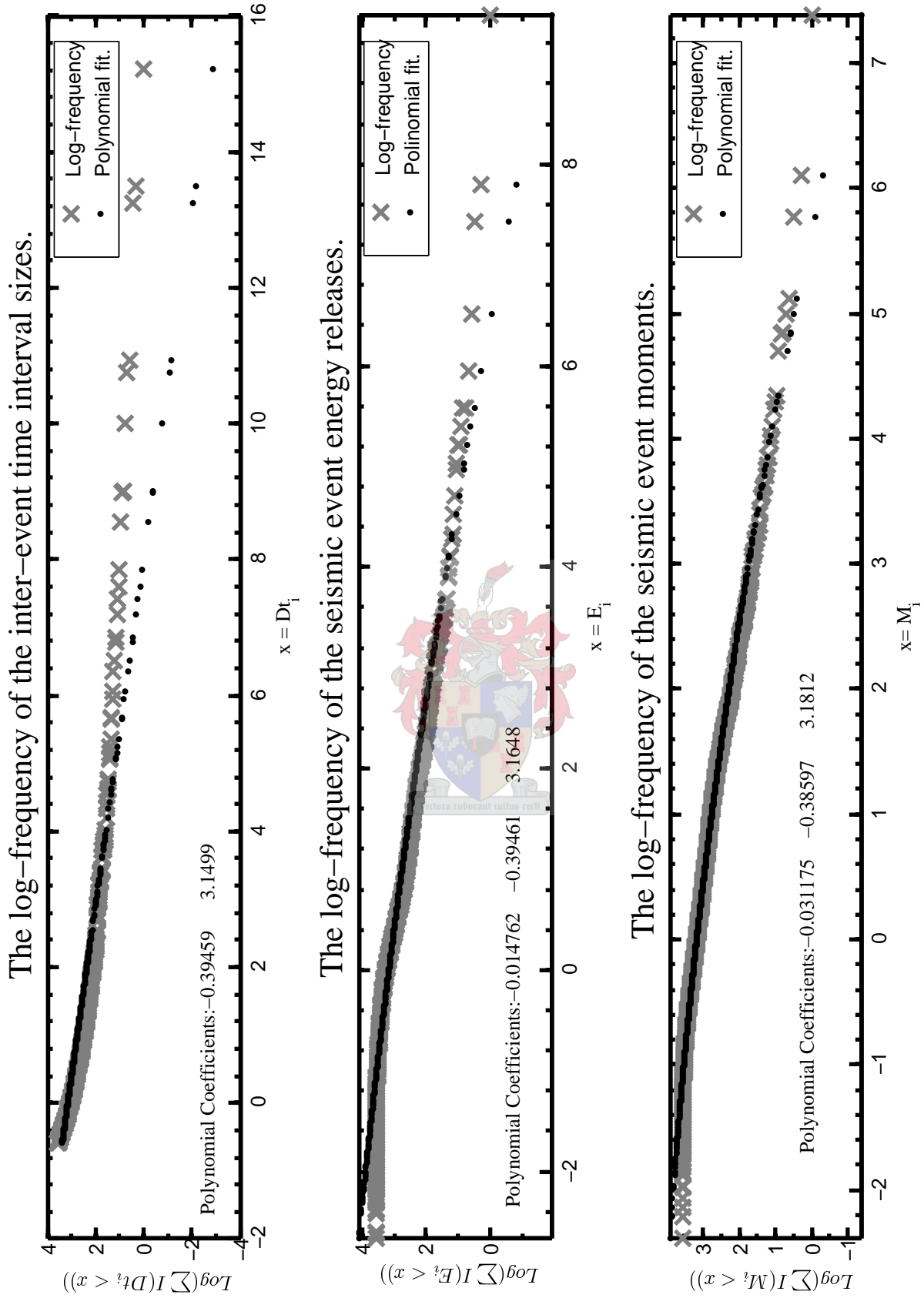
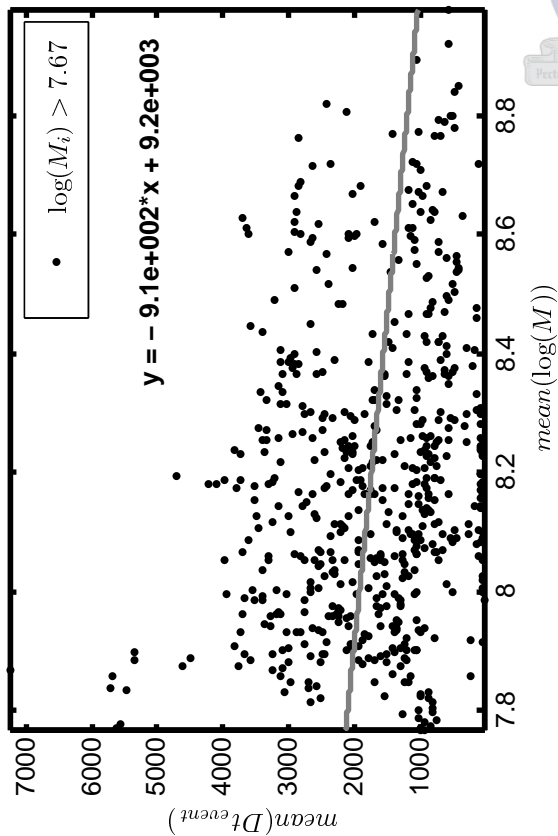
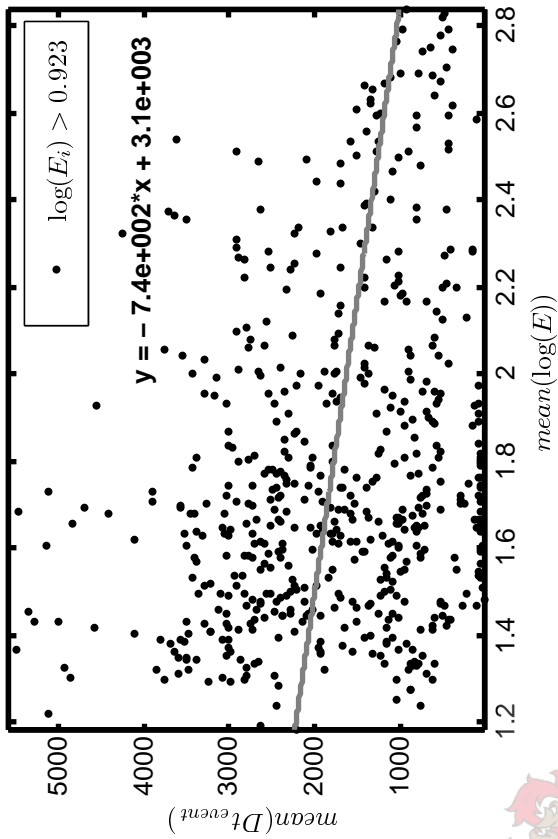


Figure 3.14: The GR relationships in the components of T4 Dt_i , $\log(M_i)$ and $\log(E_i)$.

The arrival rate vs. the average event size for 10 consecutive large events.



The arrival rate vs. the average event energy release for 10 consecutive large events.



The arrival rate vs. the average inter-event time interval for 10 consecutive long time intervals.

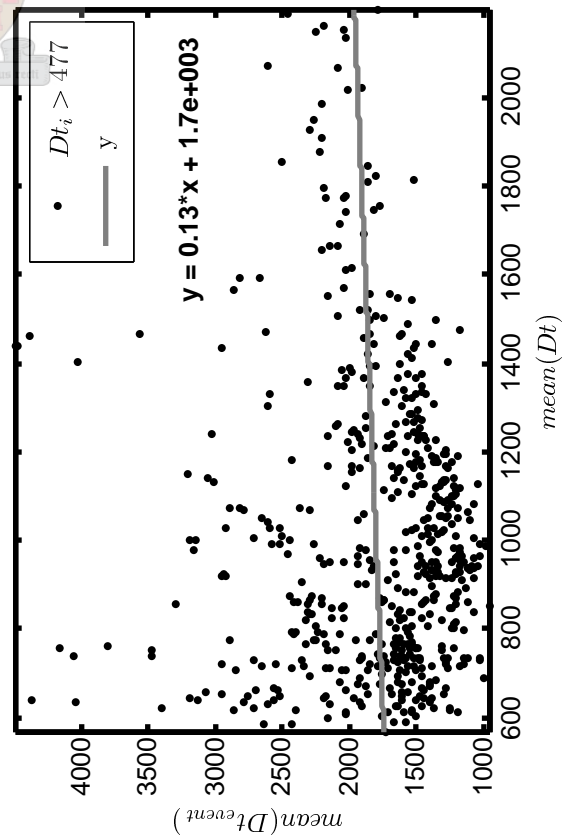
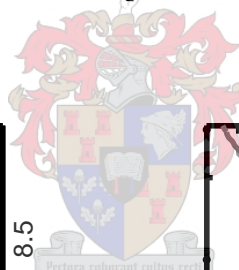
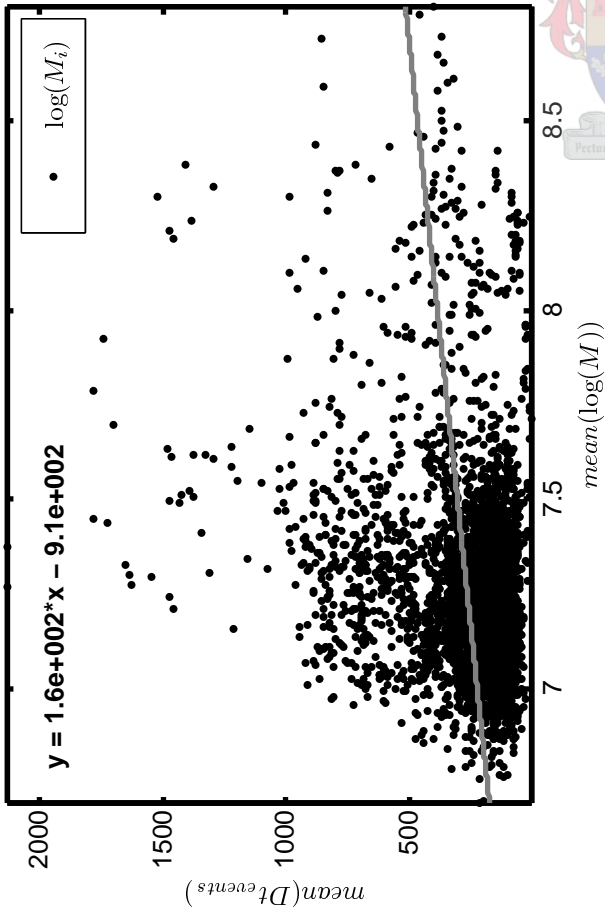
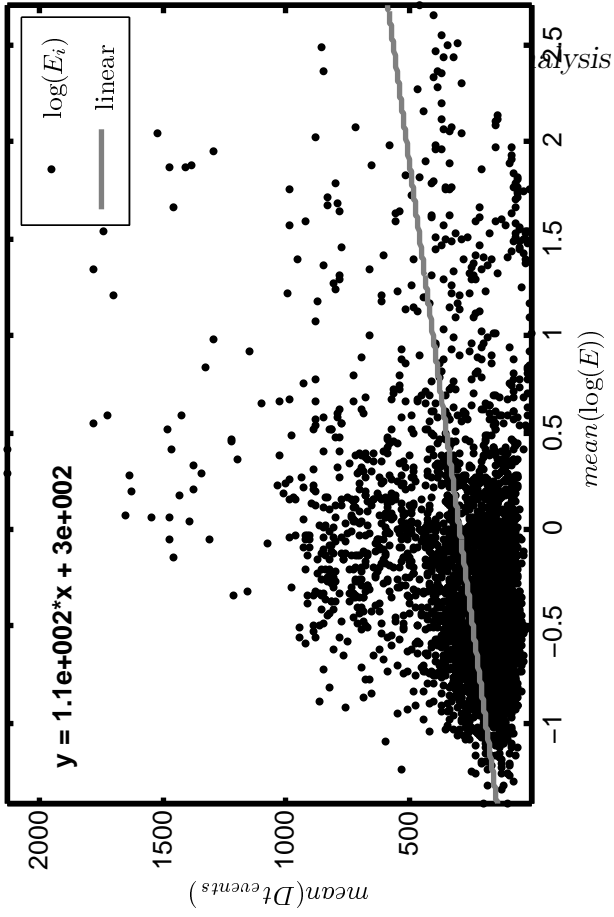


Figure 3.15: T4, clustering of large events for event sequences of length 10.

The arrival rate vs. the average event size for 10 consecutive events.



The arrival rate vs. the average event energy release for 10 consecutive events.



The arrival rate vs. the average inter-event time interval for 10 consecutive events.

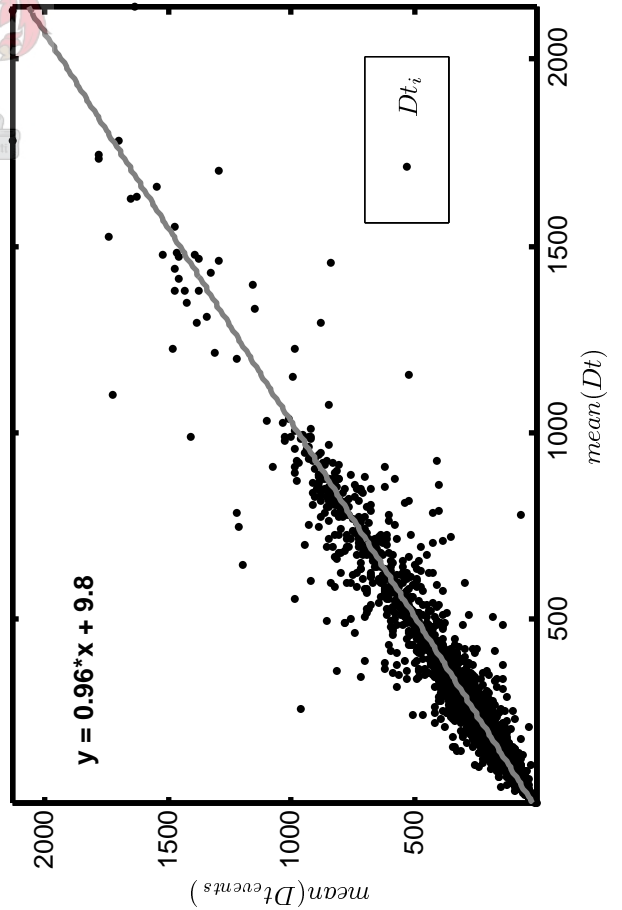


Figure 3.16: The lack of event clustering in the event sequences of length 10.

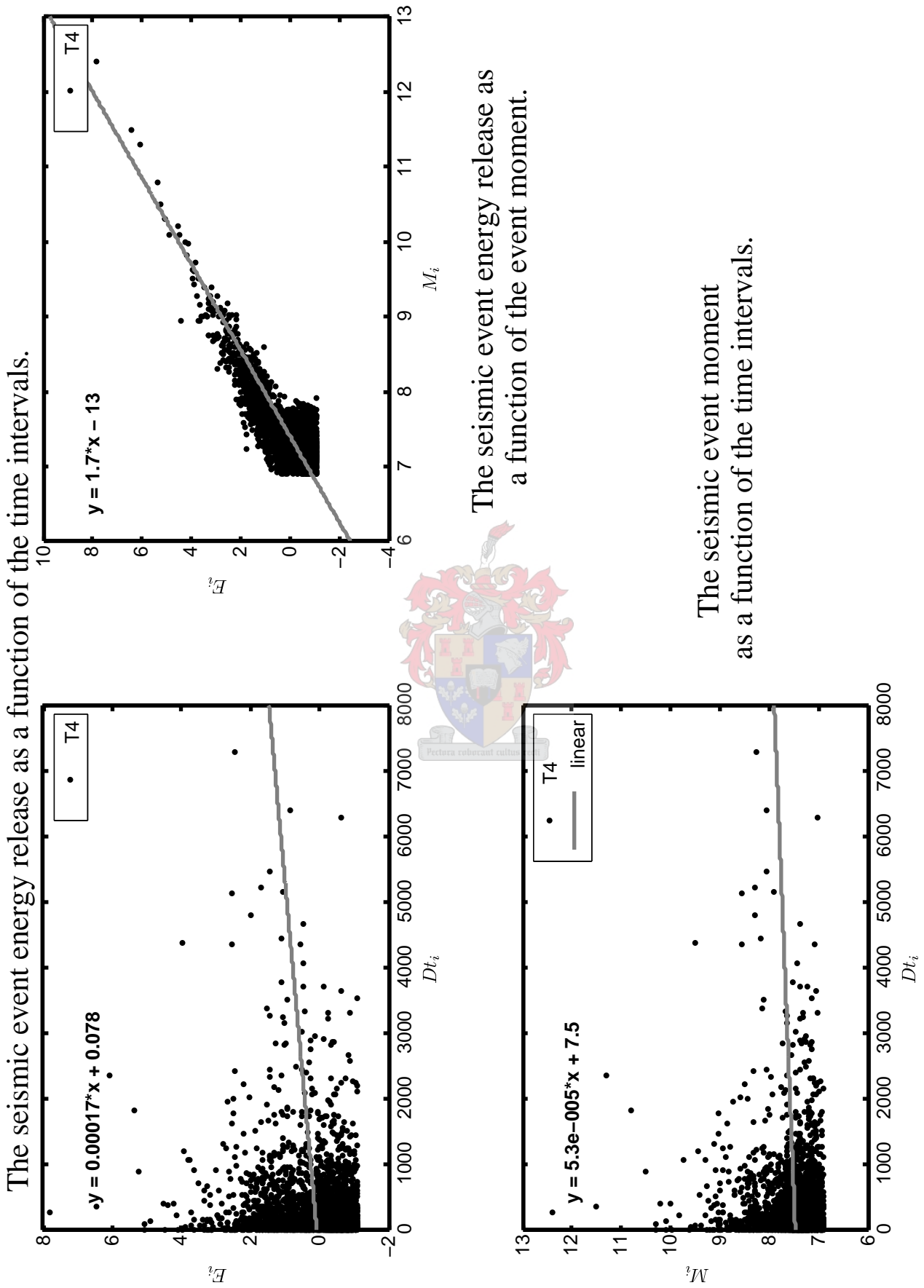


Figure 3.17: The cross correlations between the components of the T4 dataset.

The surrogate hypothesis test can be rejected for surrogate data test involving combinations of $(\log(M_i), \log(E_i))$ if the dependence between $\log(M_i)$ and $\log(E_i)$ is not sufficiently captured by the linear relationship between the two components. Figure 3.17 demonstrates the observed dependence between the two components. The drop-off from the trend between $\log(E_i)$ and $\log(M_i)$ on the small scales is addressed in the surrogate case by maintaining the distribution of each variable. Hence the surrogates should not separate from the data set purely based on the drop-off from the trend in Figure 3.17.

3.3.2 Nonlinear measures on attractors

The average mutual information and correlation dimension statistics are measured for different, but comparable phase space reconstructions of T4. Each reconstruction is done in the same way and mapped onto the same 10-dimensional space thus ensuring similar properties in each reconstruction. Each of the components are independent of each other based on the measure of independence employed by the ICA algorithm. The two statistics are characteristic of a specific attractor and invariant to linear transformations on the attractor (Kantz and Schreiber, 1997).

Suppose the data set was generated by the same underlying phase space and the phase space participates in the generation of each of the parameters. Reconstructing the phase space from the different variables into a 10-dimensional space with independent components should result in similar measures of the AMI and $D_c(\epsilon/\epsilon_o)$ statistics as for the original attractor; if the dimensionality of the dimensional space guarantees unfolding of the same underlying attractor. If the original attractor requires more than 10-dimensions to unfold properly, the attractor can not be reconstructed properly with the tools and techniques followed in this study and the surrogate hypothesis tests can not be rejected.

The $D_c(\epsilon/\epsilon_o)$ and AMI measures are sampled from a set of 10 dimensional vectors, $\{s_n^{10}\}; n = 1 \dots N$. The set of vectors is the resultant ICA transform of the multi-dimensional embedding of the components of T4 obtained according to

$$s_i^{10} = W_{10, D_{embed}} \times \overline{Seis_i}, \quad i = 1 \dots K. \quad (3.26)$$

where $W_{10, D_{embed}}$ is the estimated separation matrix. The lags for $\overline{Seis_i}$ (that is lag_{dt} ,

lag_M and lag_E) are taken as the first point of decorrelation of the autocorrelation function.

In the estimation of W using ICA, pre-processing was achieved with PCA, with 99.9% of the variance retained in the transform to obtain decorrelated $\{\overline{Seis_n}\}$. Measurement and sampling noise as well as a lack of variable mixing obscures reconstructed phase space.

The ICA estimation starts off with a randomly initialized map and iteratively adjusts the map to find an optimal fit according to the selected measure of independence. The measure of independence for the independent components are invariant to mirror and sign changes. ICA does not give exactly the same transform in repeated estimation of W . Therefore, to keep track of the variance two maps were estimated for each phase space reconstruction. The total number of reconstructed phase spaces, each supposedly a transform from the phase space producing T4, is 14, that is two (2) times seven (7). Computing W twice while keeping everything else constant in the phase space reconstruction is a simple mechanism for keeping track of the variance introduced in the computation. If the computation of W introduced no variance then the 14 reconstructions would consist of two groups of 7 identical reconstructions. As it is, the only difference between the two groups was introduced by the computation of each of W 's. The variance introduced in the phase space reconstruction should not affect the hypothesis tests and repeating the result allows for a test on that.

The phase space reconstructions were performed twice for all possible embeddings of the three variables: $\{(1 : [Dt_i]); (2 : [\log(M_i)]); (3 : [\log(E_i)]); (4 : [\log(M_i), \log(E_i)]); (5 : [Dt_i, \log(E_i)]); (6 : [Dt_i, \log(M_i)]); (7 : [Dt, \log(M_i), \log(E_i)])\}$, for both the AMI and $D_c(\epsilon/\epsilon_o)$ test statistics.

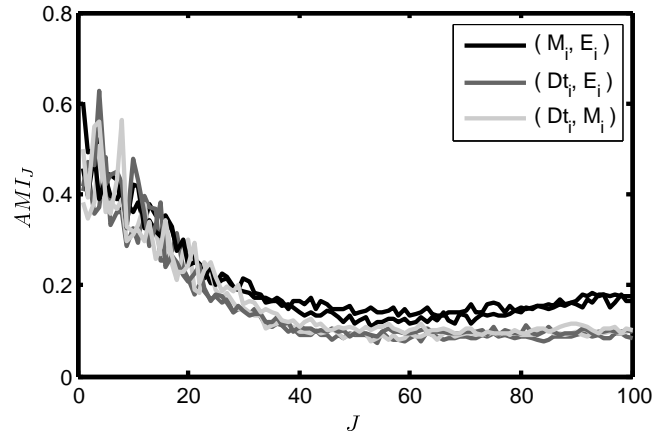
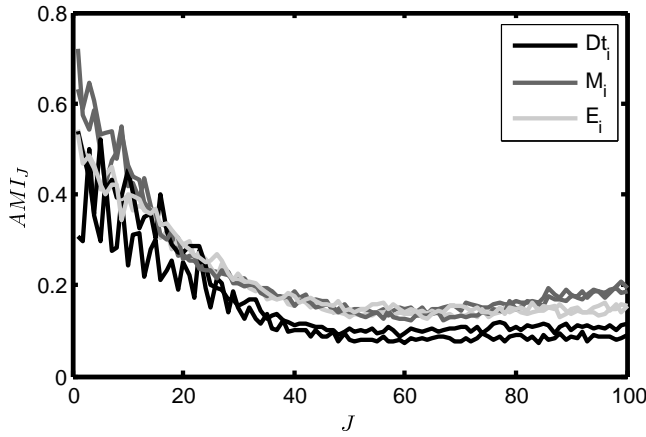
Average Mutual Information

$AMI(X, Y)$ is a measure of two one-dimensional variables that characterizes the global bumpiness of the phase space. $AMI(X, Y)$ is computed for a delay lag of the Euclidean norm, $|x^{10}| = \sqrt{\sum_{m=1}^{10} ((x_m^1)^2)}$ of s_n^{10} :

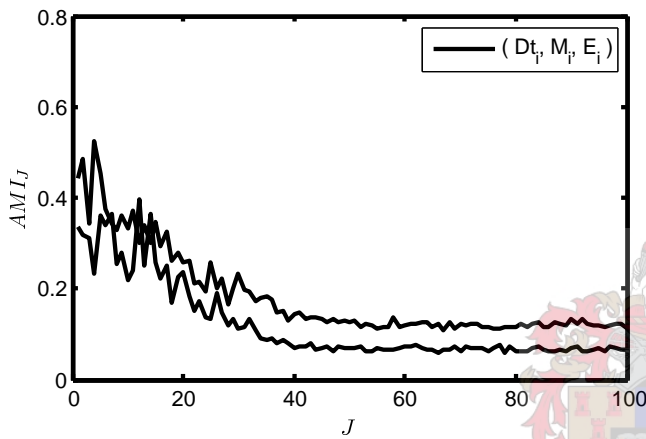
$$AMI_J = AMI(|s_n^{10}|, |s_{n+J}^{10}|), \quad J = 1, 2, 3, \dots, 100$$

where J is the shift in the lag between the components of the sequence of norms, $|s_n^{10}|$. Each computation of AMI_J represents an AMI score at a lag of J between

The AMI of the norm of two ICA scores of the seismic components using 25 symbols.



The AMI of the norm of two ICA scores of two seismic components using 25 symbols.



The AMI of the norm of two ICA scores of all the seismic components using 25 symbols.

Figure 3.18: The AMI_J scores for the norm of the 10 dimensional ICs of the T4 dataset as a delay lag in J .

two components of the sequence of $|s_n^{10}|$ norms.

The AMI of the delay lag of the norm for the reconstructed phase space is still a measure of the global bumpiness of the attractor because of the relative closeness two consecutive embedding vectors will have in the reconstructed phase space.

The AMI_J scores for the 14 reconstructed phase spaces of T4 are displayed in Figure 3.18. The Figure contains three pairs of axes, with each pair having the same units. The x -axis tracks the lag J while the y -axis tracks the value of AMI_J (in bits).

Each variable was binned into 25 bins for the computation of the AMI_J scores. A value of 1 to 25 requires about 4.5 bits of information to be specified uniquely. If the AMI score resulted in values of around 4.5 bits, the lag of variables can uniquely identify the resulting trajectory in the attractor up to the bin with. The AMI_J scores vary from about 0.6 to 0.1 from lag 1 to lag 50. Interestingly, $0.1 * 50$ equals 5, but since the lag of scores are based on correlated variables the lag of AMI scores is not an indication that $|s_n^{10}|$ can be uniquely identified from $|s_{n+J}^{10}|$, $J = 1 : 100$.

The jumps in the AMI_J scores from one J to the next are an indication that the original phase space is obscured by some kind of noise and the reconstructed phase space is rather bumpy. The drop from a high value to a low value in AMI_J indicates that the sequence of s_n^{10} are dependent on the time ordering. If the sequence of s_n^{10} were independent the size of the AMI_J scores would have been smaller and no degradation of AMI_J scores would have occurred with an increase in lag J .

The phase spaces constructed from the time intervals appear to be more bumpy in the variance from one AMI_J score to the next than is the case in other reconstructed phase spaces. This is further supported by the fact that the phase spaces with component Dt_i included score lower than the rest of the computed AMI_J scores. The exception is the combination $(Dt_i, \log(M_i))$ which seems to give the optimal reconstructed phase space. The variance introduced by the computation of the W 's did not affect the result significantly.

The slightly better reconstruction using the combination $(Dt_i, \log(M_i))$ compared to any of the other variable combinations can be explained as follows. Firstly, Any one of the seismic source variables does not provide a sufficient mix of variables from the original phase space (Mendecki et al. (1997) as discussed in Chapter 2). Secondly, the $\log(E_i)$ and Dt_i components have more noise associated with them than the $\log(M_i)$ variable, as observed from their AMI_J scores. The noise argument is ironic since the time stamp is the only variables that have no error term associated with it. The dynamics driving the timing of the events and the occurrence of them might be independent of each other.

Correlation Dimension

A total 14 phase spaces were reconstructed from the components of T4: all the different combinations of variables and two ICA maps for each. The small scale scattering of a reconstructed phase space is controlled by the amount of measurement noise, sampling noise and mixing of the variables from the original phase space.

The $D_c(\epsilon/\epsilon_o)$ can be considered a measure of the dimension of the reconstructed attractor in the phase space if:

- a flat portion of $D_c(\epsilon/\epsilon)$ exists over a range of $\log(\epsilon/\epsilon_o)$'s;
- the pairs of vectors used for the distance measures s_n^{10} are uncorrelated.
- the dimension estimate is less than about 3.1.

None of these conditions hold for the $D_c(\epsilon/\epsilon_o)$ estimates of the 14 attractors shown in Figure 3.19. As a measure of the scatter of the reconstructed phase space, it can still be used as a test statistic in the surrogate data analysis. The Dt_i variable as opposed to the AMI_J scores seems to reduce the small-scale scattering. Any phase space reconstruction with Dt_i included in the reconstruction seems to scatter less than without it. The variance introduced by the computation of the IC's did not seem to affect the results.

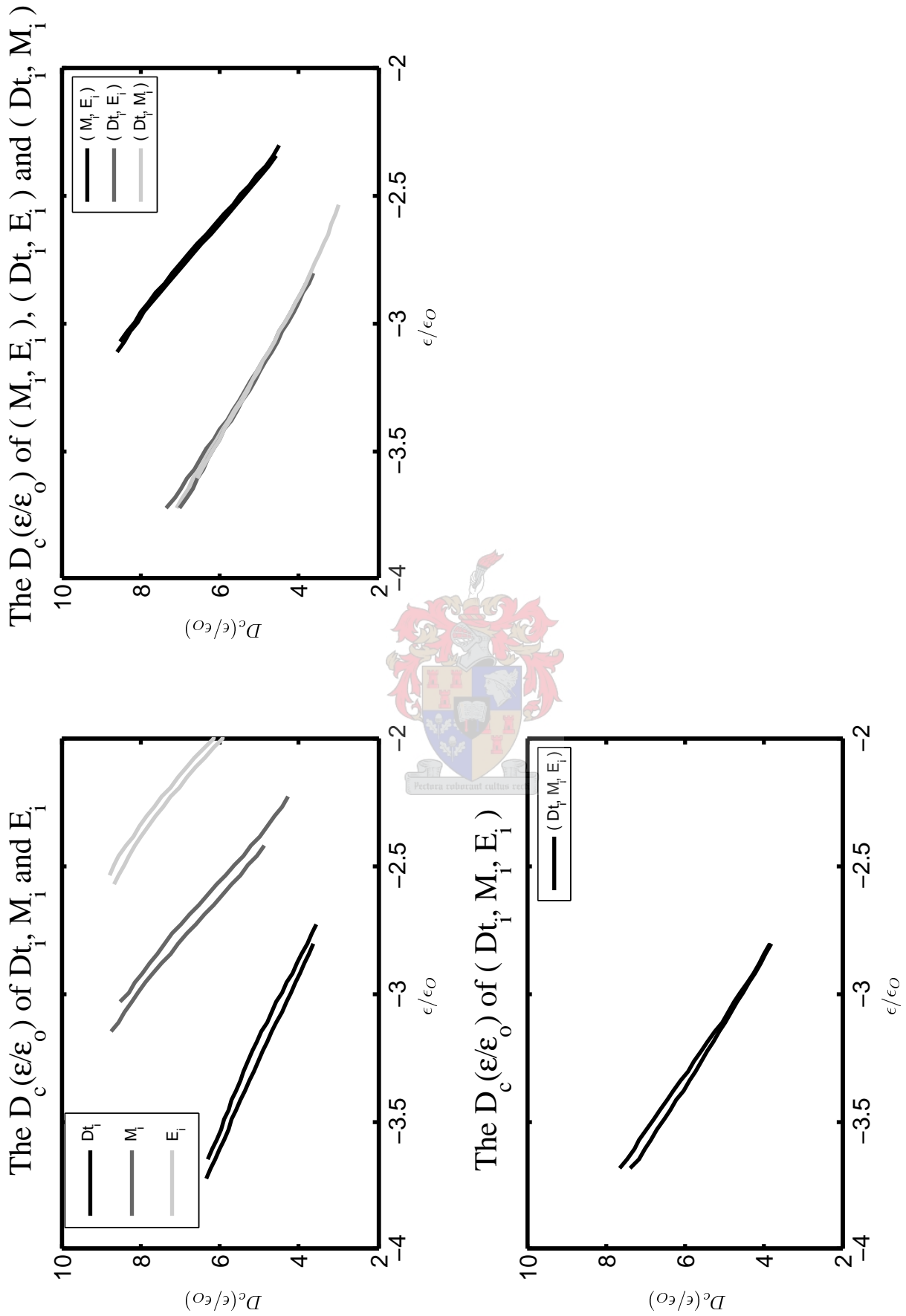


Figure 3.19: The 7 phase space reconstructions are grouped into 3 groups. The top left graph is for single component phase spaces, the top right for pairs of components and the bottom graph for all three components

3.4 Surrogate data and Hypothesis tests

Results from the previous section showed that none of the reconstructions were similar nor was an ideal reconstructed phase space obtained. This could be because the data were either not uniformly sampled from the same red colored noise, or the system is simply not deterministic. The next test for determinism is to fit a model on the data to see if the future evolution of the system can be derived using measurements collected in the past. Since the reconstructions did not show small-scale or large-scale structure different from a red colored noise source, constructing a model more complex than an autoregressive moving average must be approached cautiously.

In the surrogate data analysis, a null hypothesis test of stationary red colored noise generating source is assumed. The $D_c(\epsilon/\epsilon_o)$ and AMI_J scores are the test statistics with unknown distributions under the assumption of the null. As previously discussed, surrogate analysis provides a mechanism for sampling portions of the red colored noise from which T4 could have been sampled if the null hypothesis is true. The ratio of surrogate sample to the observed data is critical, and guidelines on choosing the size of the surrogate set to establish the confidence bound exist (Schreiber and Schmitz, 1996, 2000).

Constructing the hypothesis tests involves a number of steps. First, a set of surrogate data satisfying the red colored noise null hypothesis is generated. Then the test statistics are evaluated for the surrogate data. Finally, the sets of test statistics can be compared to see if each reconstructed phase space was sampled from the same underlying red colored noise system.

If the test statistics for the different surrogate reconstructed phase spaces differ, then the difference in the autocorrelation coefficient and noise distribution has a significant influence on the final reconstructed phase space. If the initial autocorrelation structure and distribution of the sampled system influences the final reconstructed phase space, it would be difficult to establish if the different components are sampled from the same underlying phase space. If such an underlying system can be established, even if it represents only some type of noise, it could serve as a basis model to generate synthetic seismic activity similar to that of the data set.

Figure 3.20 shows a comparison between the 3 components of $(Dt_i, \log(M_i), \log(E_i))$ and a realization of the surrogate data. The two clusters in the figure appear to

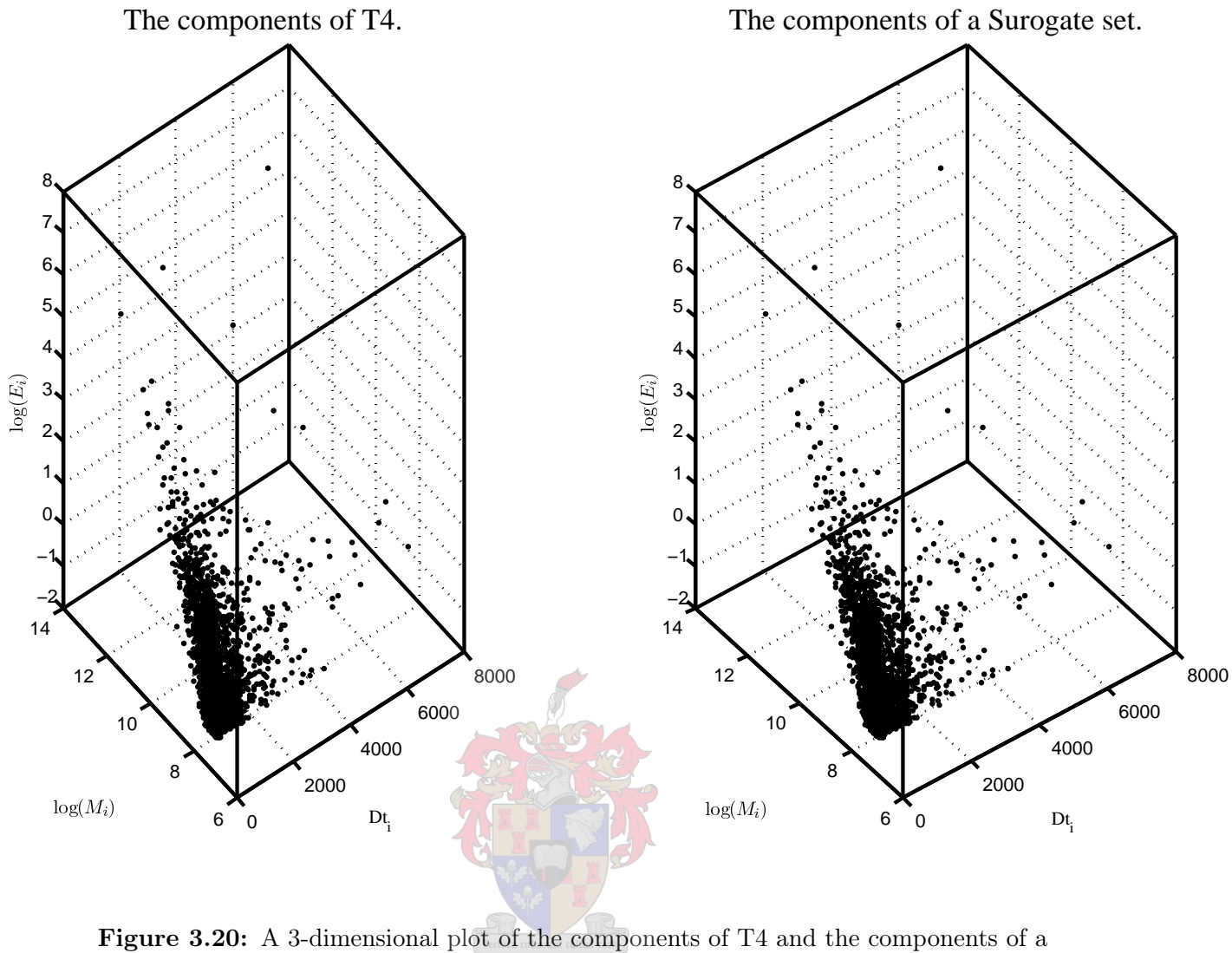


Figure 3.20: A 3-dimensional plot of the components of T4 and the components of a surrogate set.

have superficially similar distribution, orientation and location in the 3-dimensional space.

In Figure 3.21 are plots of the scores in order sequence in which the AMI_J statistic is sampled for a surrogate set as well as the T4 data set. The score is computed as the norm of the 10-dimensional independent components $|s_n^{10}|$ representing the reconstructed phase space of T4 and the surrogate set. The scores from the surrogate set appear to be slightly more scattered than the scores from T4. Other than the scattering the two sets of scores appear to be sampled from the same stationary autoregressive moving average (ARMA) process, scaled with the same static non-linear function. Two comparisons help to illustrate that the surrogate generating mechanism gives noise similar to the original data.

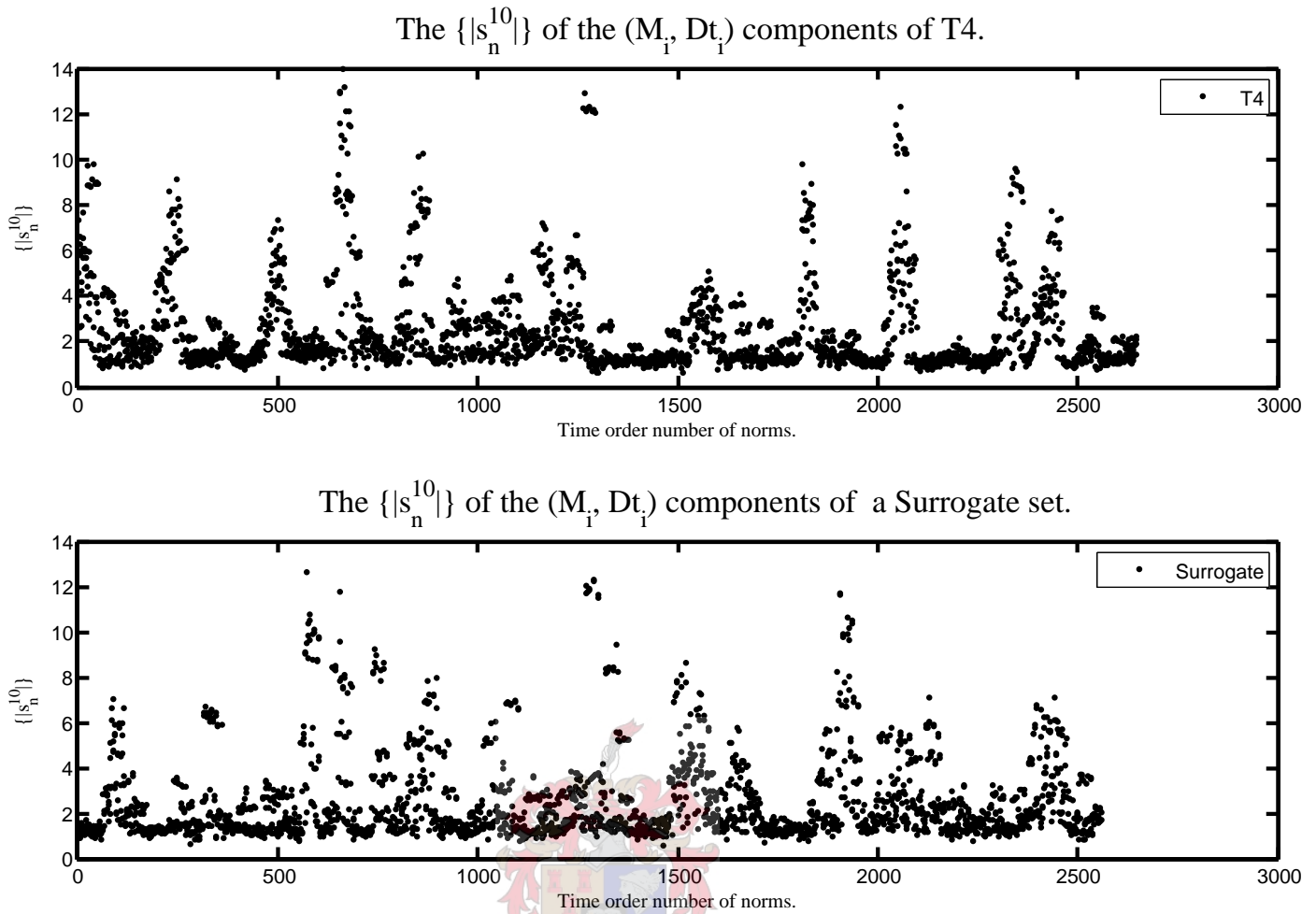


Figure 3.21: The norm of the 10-Dimensional phase space reconstruction for $(\log(M), Dt_i)$ of T4, top, and a surrogate set, bottom. Note the scattering in the surrogate set compared to the T4 set.

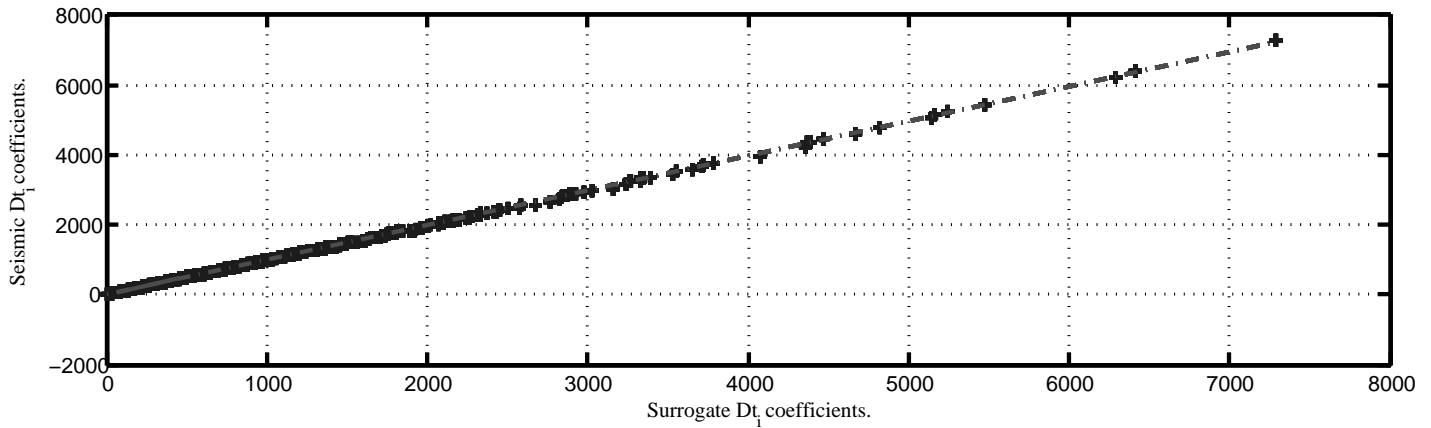
3.4.1 Comparison of linear statistics

T4 is identical to the surrogate population of red colored noise in that:

1. The components of T4 have a similar distribution to the corresponding components of the surrogates as visualized in a quantile-quantile plot.
2. The components of T4 have a similar set of autoregressive coefficients to those of the surrogates, demonstrated by plotting the corresponding coefficients on the same axes.
3. The components of T4 have a similar set of cross correlation coefficients to those of the surrogates, demonstrated by plotting the corresponding coefficients from a surrogate set and T4 on the same axes.

In all the above cases, the domain of each graph should correspond and the functional relationship should be a straight line. If any of these properties do not hold for any of the surrogates then the hypothesis test is invalid. Figure 3.22 shows the

The quantile–quantile plot of the Dt_i s for surrogate and seismic data.



The corresponding autocorrelation coefficients of the Dt_i s, surrogate and seismic data.

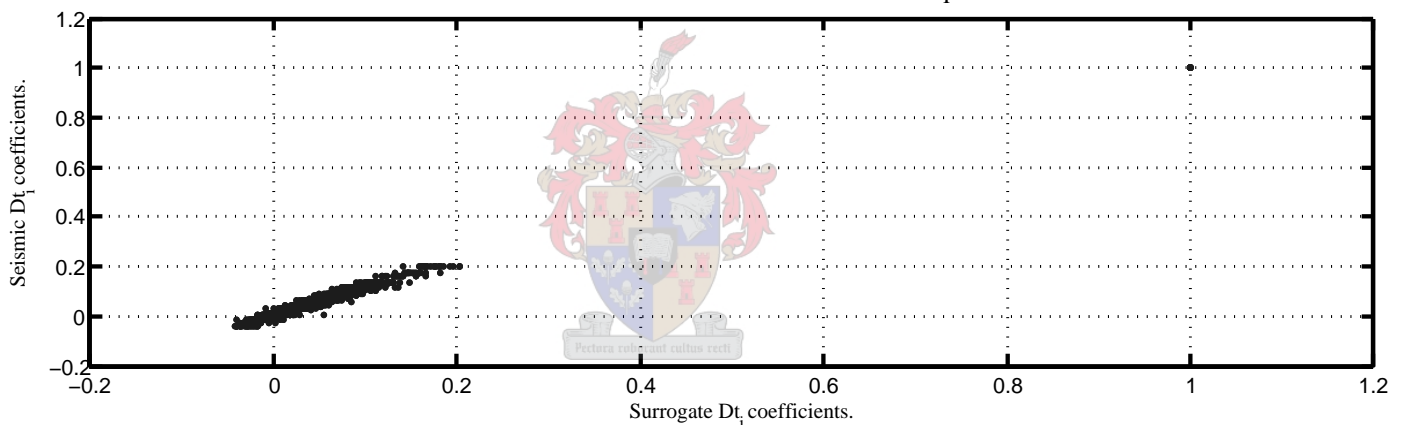
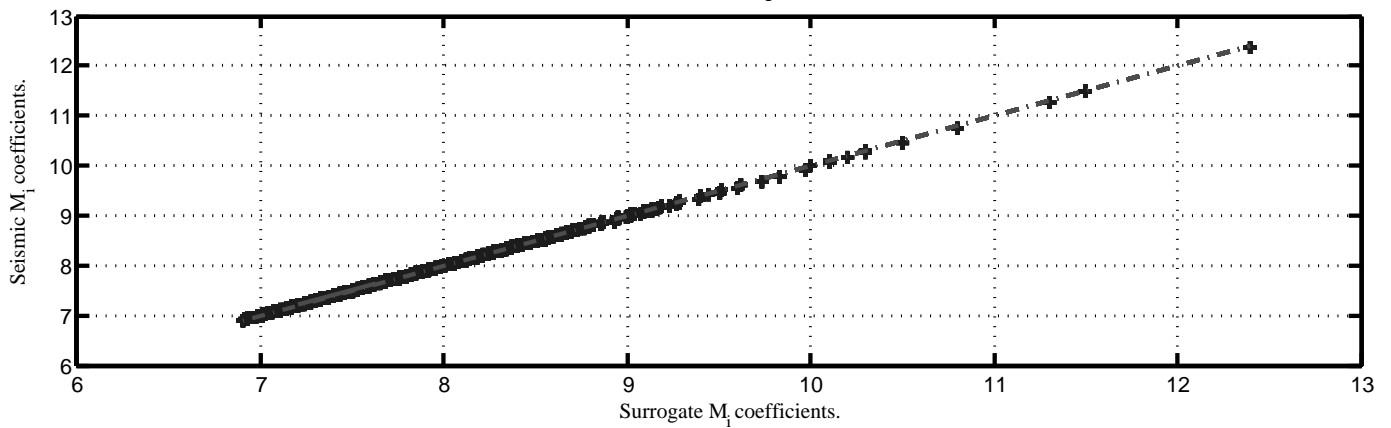
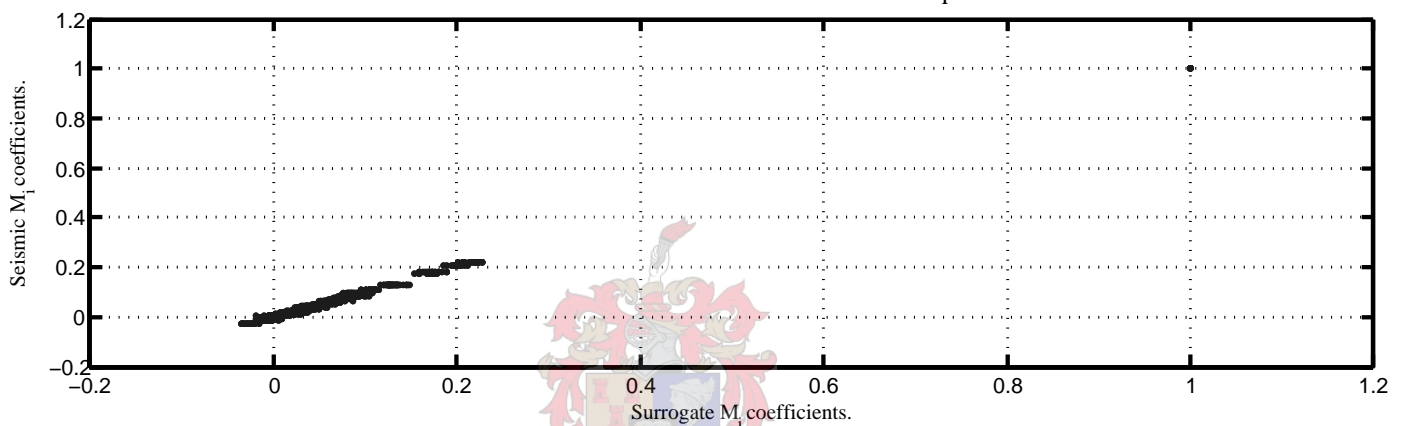


Figure 3.22: q-q plot and Acc plots of T4 and the surrogates for Dt_i

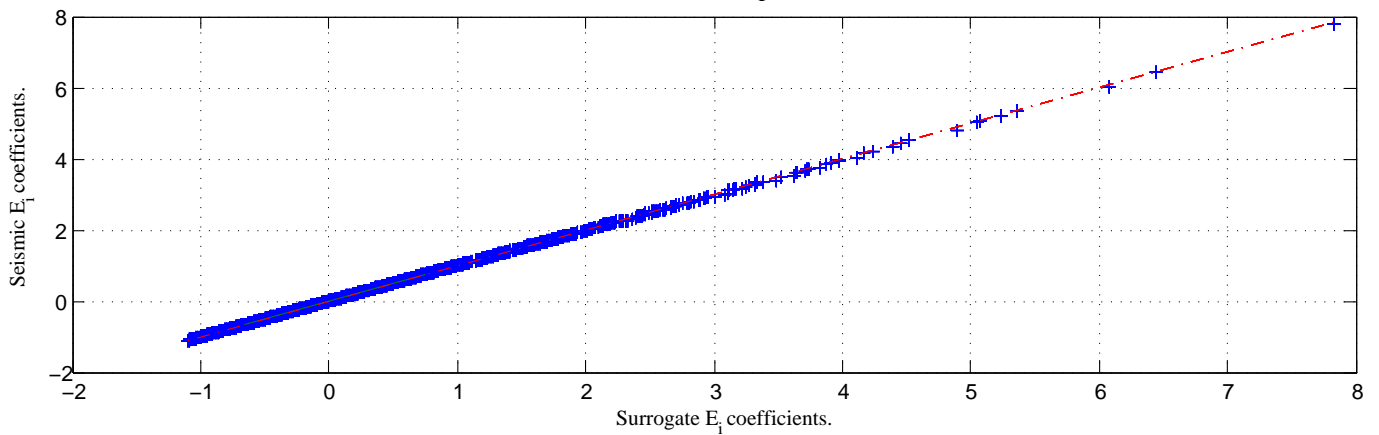
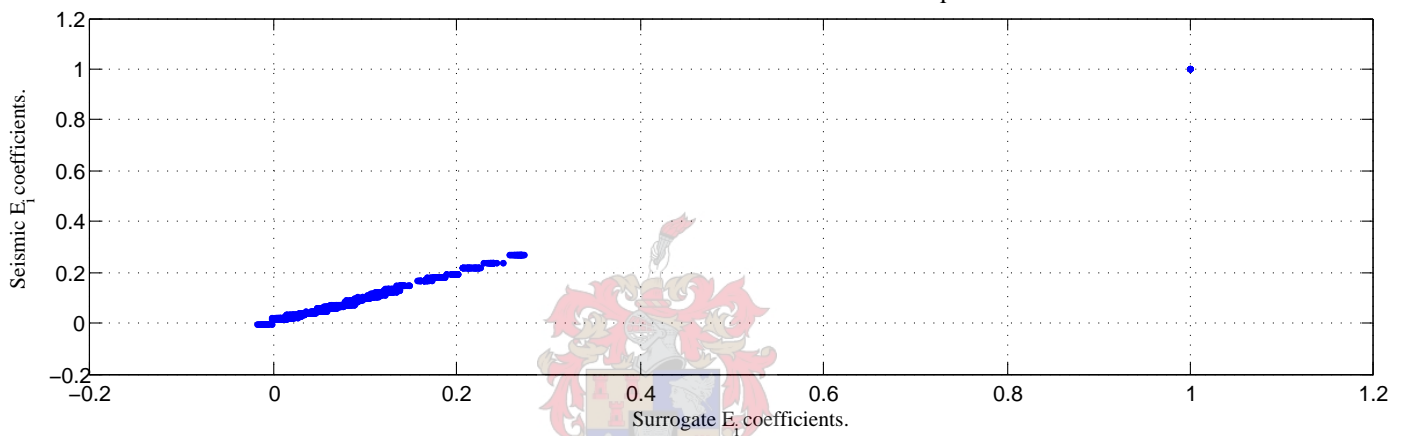
quantile-quantile plot for Dt_i and the comparative autoregressive coefficient plot in the bottom plot. The graphs show the desired variable domain as well as straight line functional relationship. Figure 3.23 is the quantile-quantile plot for the $\log(M_i)$ component followed by the comparative autoregressive coefficient plot. The two figures show the desired variable domain as well as straight line functional relationship. Figure 3.24 is the quantile-quantile plot for the $\log(E_i)$ followed by the comparative autoregressive coefficient plot. The two figures show the desired variable domain as well as straight line functional relationship. Figure 3.25 plots the cross-correlation coefficients for the pairs of components for surrogates as well as those of T4 on the same axes.

The quantile–quantile plot of the M_1 s for surrogate and seismic data.The corresponding autocorrelation coefficients of the M_1 s, surrogate and seismic data.**Figure 3.23:** q-q plot and Acc plots of T4 and the surrogates for the $\log(M_i)$

Based on the comparisons it can be concluded that the coefficients from the $(Dt_i, \log(E_i))$ and $(Dt_i, \log(M_i))$ were probably sampled from the same system. A small bias towards larger coefficients is demonstrated in the $(\log(M_i), \log(E_i))$ pair for the surrogates. The slightly larger cross-correlation coefficients have an expected influence on the $D_c(\epsilon/\epsilon_o)$ and AMI_J scores, although the difference does not affect the conclusion that it is not possible to discriminate between T4 and an autoregressive process on the basis of linear statistics only.

3.4.2 Comparative non-linear statistics and hypothesis tests

Using a similar procedure as in the previous section, the hypothesis tests were performed for the 14 phase space reconstructions using $D_c(\epsilon/\epsilon_o)$ and AMI_J scores as the non-linear test statistics. Each of the 14 comparisons is a comparison between

The quantile–quantile plot of the E_1 s for surrogate and seismic data.The corresponding autocorrelation coefficients of the E_1 s, surrogate and seismic data.**Figure 3.24:** q-q plot and the Acc plot for T4 and its surrogates for $\log(E_i)$.

components from the original data set with the same set of surrogates. A total of 21 surrogate sets were initially sampled but due to the problems with the convergence of W in the ICA a few test statistic measures were not realized for some surrogates. Twenty one surrogates translate to a confidence level of $\alpha = 0.05$, 16 surrogates translate to a confidence level of $\alpha = 0.06$, representing the bounds on the confidence level in each of the hypothesis tests.

Since each of the phase space reconstructions originates from the same system of red colored noise or surrogate set, the statistics should converge to the same set of values. The variance in the test statistics across the phase space reconstructions cannot be attributed to the variance introduced by the ICA. Therefore, such variance needs to be taken into careful consideration.

The set of figures below reconstruct phase spaces from different groupings in the

The cross correlations of the surrogate sets compared to the seismic data set.

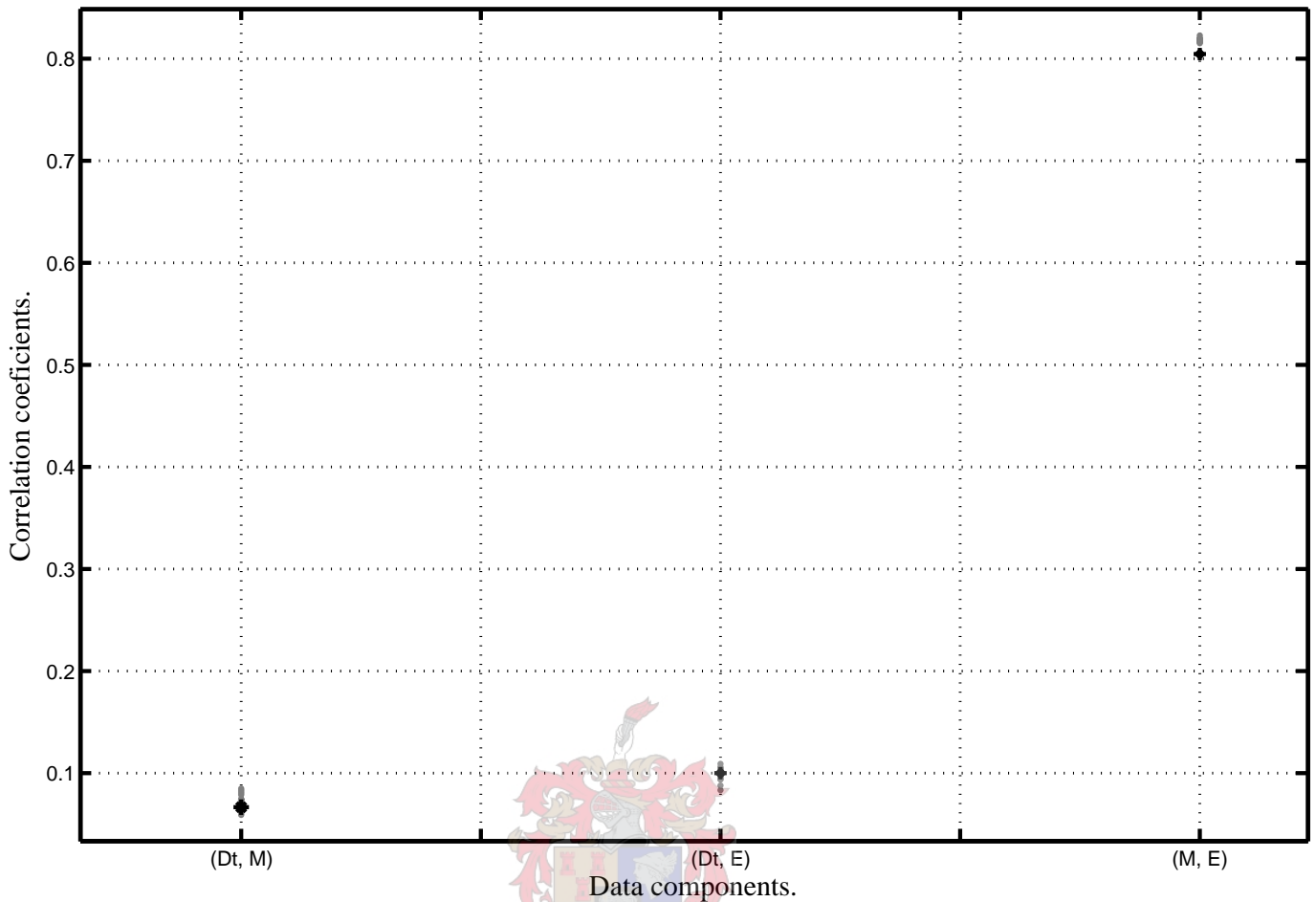


Figure 3.25: The black dots are the cross correlations of T4 and the gray dots are the cross-correlations of the surrogate data sets. The two sets of points do not separate significantly.

following order:

$\{(1 : (Dt)); (2 : (\log(M_i))); (3 : (\log(E_i))); (4 : (\log(M_i), \log(E_i))); (5 : (Dt, \log(E_i))); (6 : (Dt, \log(M_i))); (7 : (Dt, \log(M_i), \log(E_i)))\}$.

An $\alpha = 1/15$ level of confidence applies in all cases.

Figure 3.26 is a comparison of the phase space reconstruction of the inter-event time intervals Dt_i for the data and surrogate sets which shows a separation between the test statistics of the data and surrogates. The hypothesis that the Dt_i 's of the data is red colored noise can be rejected in both the small scale scattering as well as the large scale bumpiness of the reconstructed phase space.

Figure 3.27 shows a corresponding comparison for the seismic moments $\log(M_i)$ of

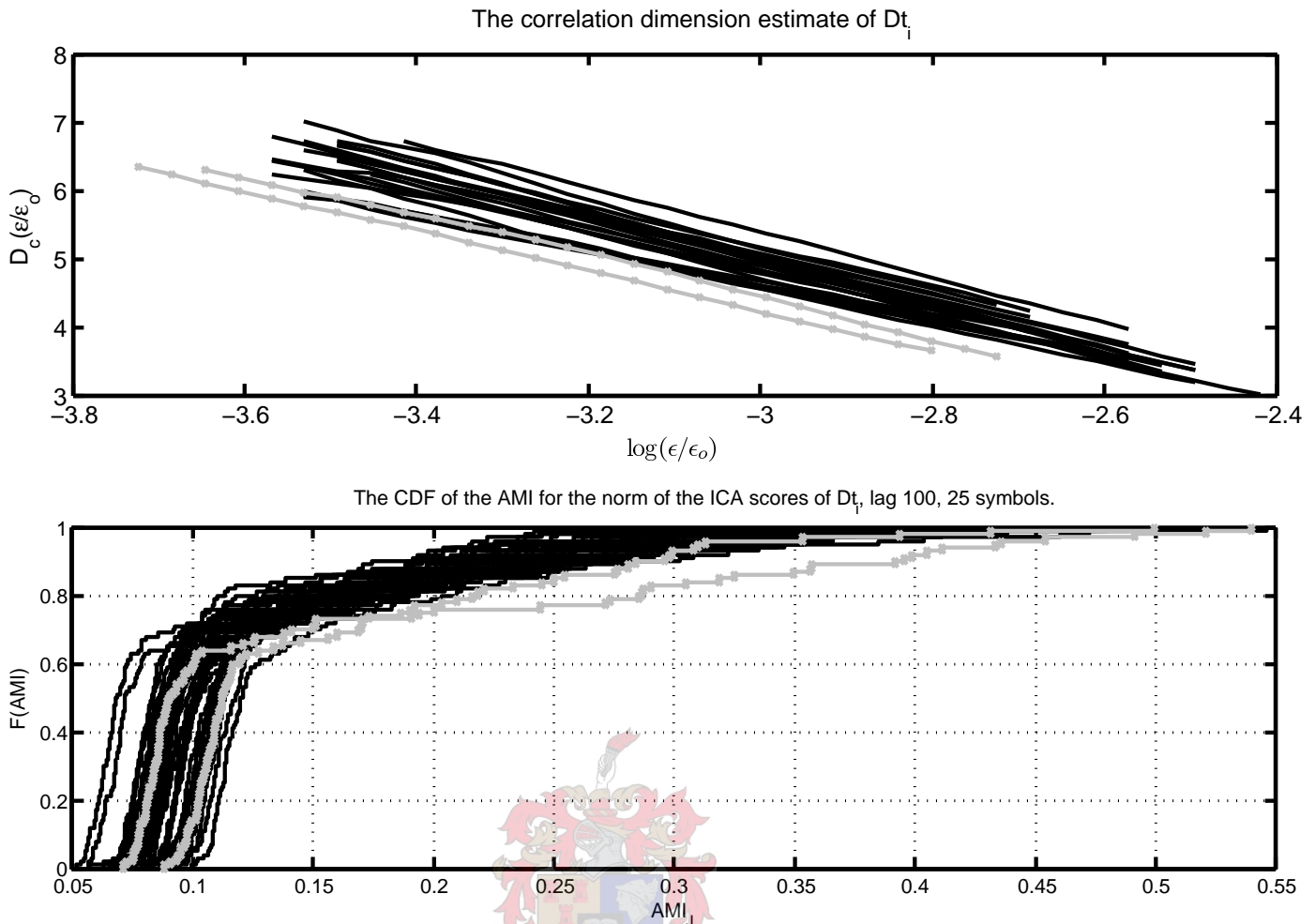


Figure 3.26: Surrogate analysis of Dt_i . The top graph is of the correlation dimension and the bottom of the distribution of the AMI_J scores. The test statistic realizations of the surrogates are in black and those of the data set in gray. Two grays are present, since each test was repeated with a different separation matrix from the ICA.

the data and surrogates. Similarly, the test statistics of the data set separate from those of the surrogates. Therefore, the null hypothesis that the $\log(M_i)$ s of the data set is red colored noise is rejected in both the small scale scattering as well as the large scale bumpiness of the reconstructed phase space.

In the case of seismic energy $\log(E_i)$, there was no separation between the test statistics of the data and surrogates, Figure 3.28. Hence, the hypothesis that the $\log(E_i)$'s of the data are consistent with red colored noise cannot be rejected in both the small scale scattering and large scale bumpiness of the reconstructed phase space.

The red colored noise process is rejected in the case of the inter-event time interval

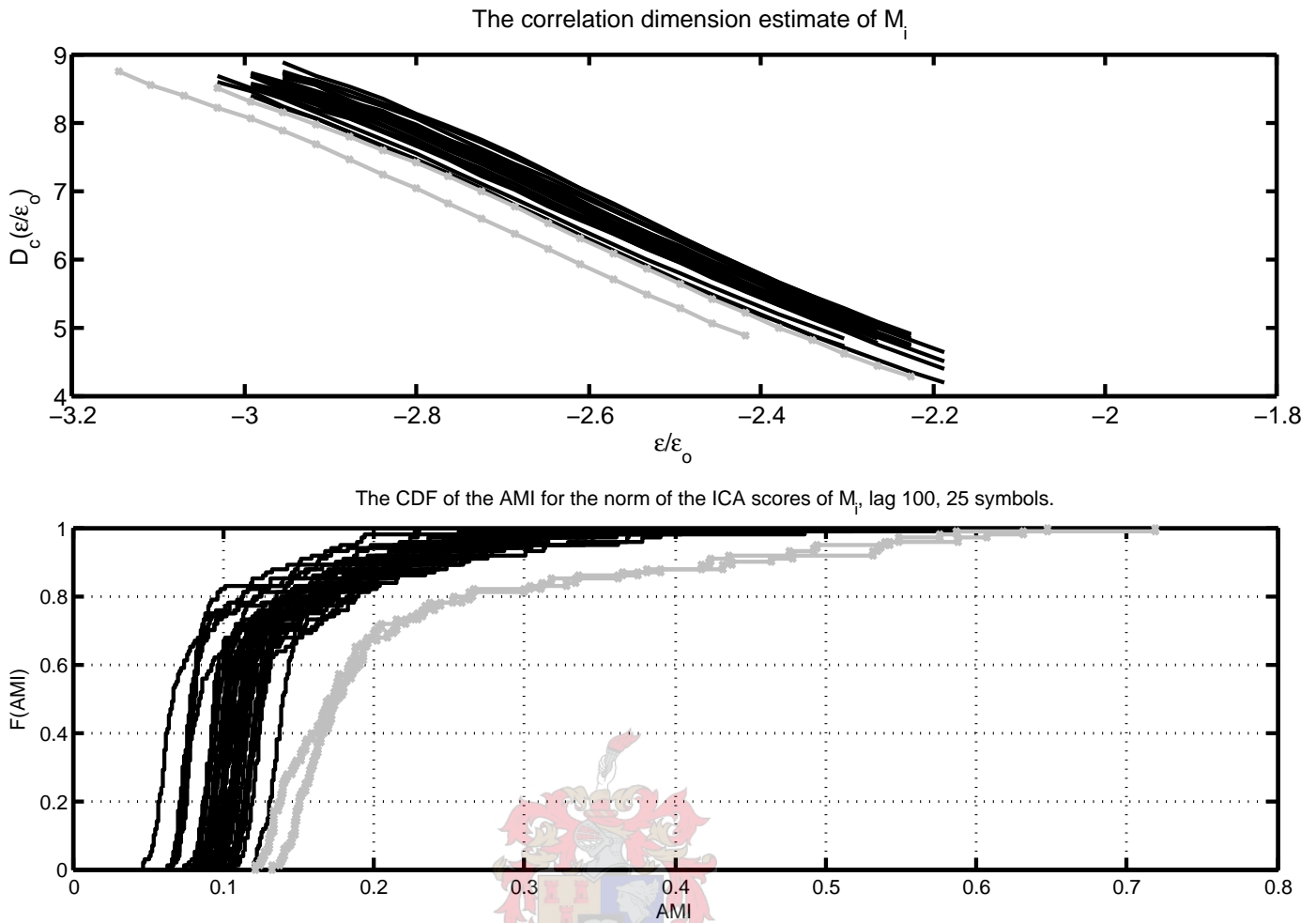


Figure 3.27: Surrogate analysis of $\log(M_i)$. The top graph is of the correlation dimension and the bottom of the distribution of the AMI_J scores. The test statistic realizations of the surrogates are in black and those of the data set in gray. Two grays are present, since each test was repeated with a different separation matrix from ICA.

and seismic moment $(Dt_i, \log(M_i))$ as indicated in Figure 3.29. The separation between the test statistics is not as marked as in the reconstruction from Dt_i or M_i individually.

The test statistics computed for the phase space reconstruction using the inter-event interval and seismic energy $(Dt_i, \log(E_i))$ separate significantly better than when only $\log(E_i)$ is used, Figure 3.29. However, the measure of the large scale bumpiness does not separate as well as the small scale behavior of the reconstructed phase space. In any event, the hypothesis that the $(Dt_i, \log(E_i))$'s of the data set are consistent with a red colored noise process can be rejected in both the small scale scattering as well as the large scale bumpiness of the reconstructed phase space.

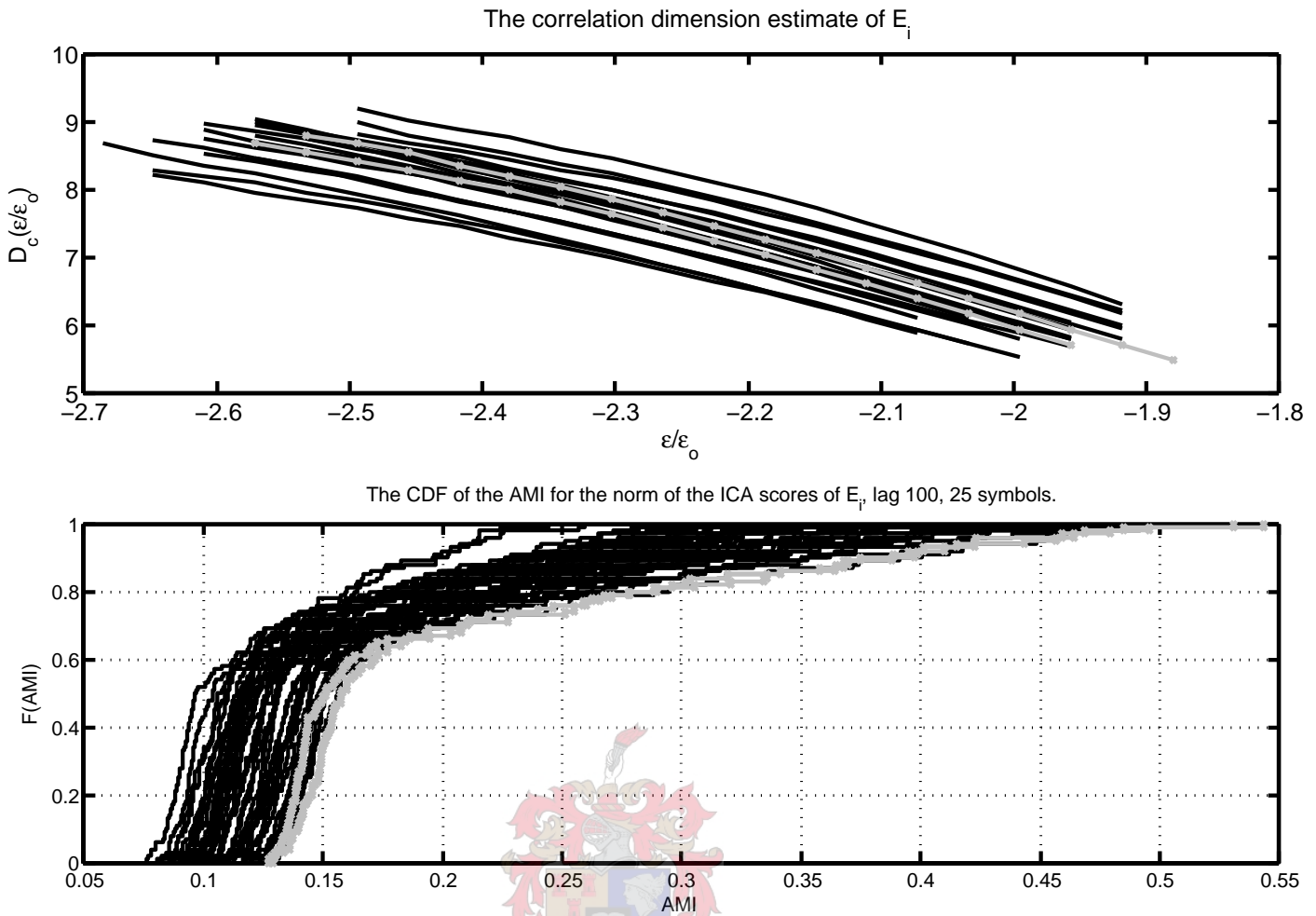


Figure 3.28: Surrogate analysis of $\log(E_i)$. The top graph is of the correlation dimension and the bottom of the distribution of the AMI_J scores. The test statistic realizations of the surrogates are in black and those of the data set in gray. Two grays are present, since each test was repeated with a different separation matrix from the ICA.

Different conclusions are obtained for the two test statistics in the comparison of the phase space reconstruction of the seismic moment and seismic energy ($\log(M_i), \log(E_i)$) of the data and surrogate sets, Figure 3.31. While the AMI_J test statistic shows a separation between the data and surrogates, no separation is observed for the $D_c(\epsilon/\epsilon_0)$ test statistic. The measure of the small scale behaviour does not separate at all compared to the large scale bumpiness of the reconstructed phase space. Thus, the null that the $(\log(M_i), \log(E_i))$'s of the data is red colored noise cannot be rejected for the small scale scattering. However, it can be rejected for the measure on the large scale bumpiness of the reconstructed phase space.

Finally, with respect to the inter-event time intervals Dt_i , the red colored noise null is marginally rejected in both the small scale scattering as well as the large scale

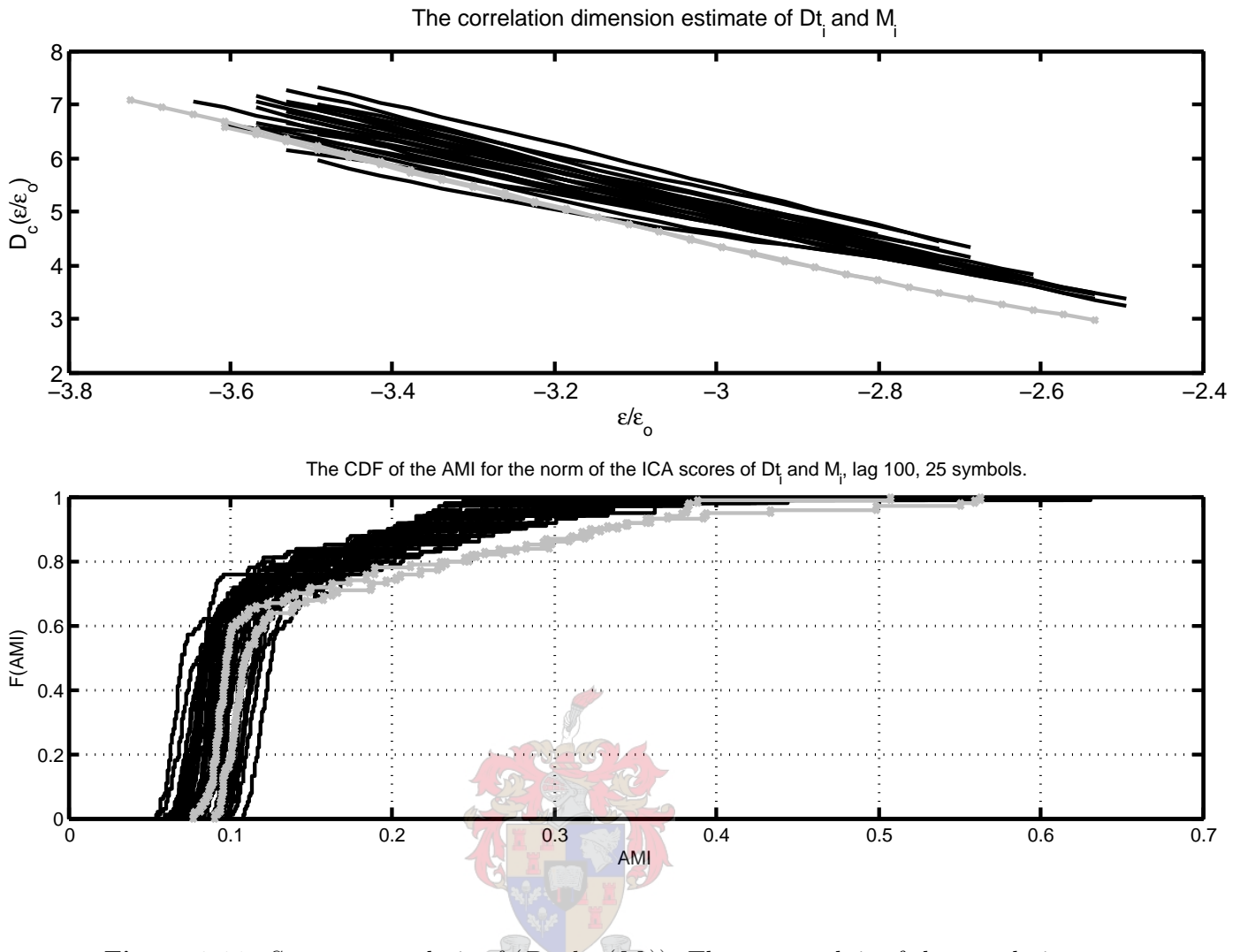


Figure 3.29: Surrogate analysis of $(D_{t_i}, \log(M_i))$. The top graph is of the correlation dimension and the bottom of the distribution of the AMI_J scores. The test statistic realizations of the surrogates are in black and those of the data set in gray. Two grays are present, since each test was repeated with a different separation matrix from the ICA.

bumpiness of the reconstructed phase space, Figure 3.32.

From the foregoing, the only data component that did not separate for the set of test statistics is the $\log(E_i)$ variable. The D_{t_i} and $\log(M_i)$ reconstructions separated best. The difference in the separation behavior $\log(E_i)$ and $\log(M_i)$ is interesting due to the correlation between the 2 variables. However, it was not possible to explain the cause of this. In general, the T4 data set cannot be labeled as merely due to the sampling of stationary red colored noise. The data set was sampled over a period of 12 days and the non-stationary events due to sampling were removed from the set. The event removal did not affect the frequency of time interval size behavior of the inter-event time intervals and the time intervals separated the best among all the variables.

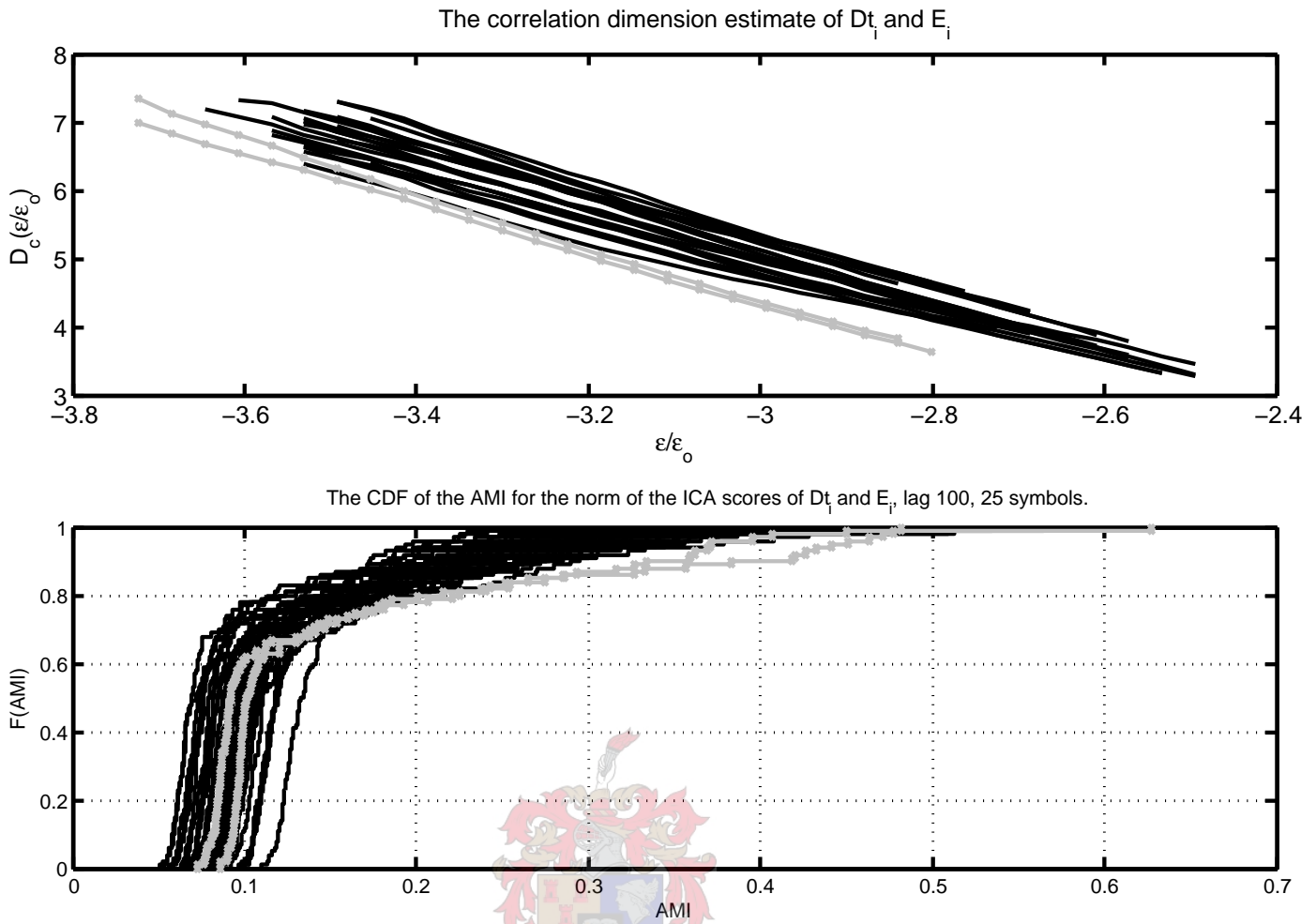


Figure 3.30: Surrogate analysis of $(Dt_i, \log(E_i))$. The top graph is of the correlation dimension and the bottom of the distribution of the AMI_J scores. The test statistic realizations of the surrogates are in black and those of the data set in gray. Two grays are present, since each test was repeated for with a different separation matrix from the ICA.

The AMI scores for the surrogates showed similar behaviour, implying they belong to the same underlying phase space system. On the other hand, the scores for the data set showed variation and did not appear to be sampled from the same underlying phase space system.

The $D_c(\epsilon/\epsilon_0)$ scores for the surrogates did not display the same behaviour as the AMI scores. The phase space reconstruction containing the Dt_i variable behaved the same. The two other phase space reconstructions containing the M_i variable behaved similarly. Finally, the $D_c(\epsilon/\epsilon_0)$ scores for E_i behaved differently from the others. It is not clear why the small scale scattering in the red colored noise phase spaces differed in these groups. The separation is an indication that the ICA dimension reduction map allowed even changes in the noise structure to reflect onto the lower

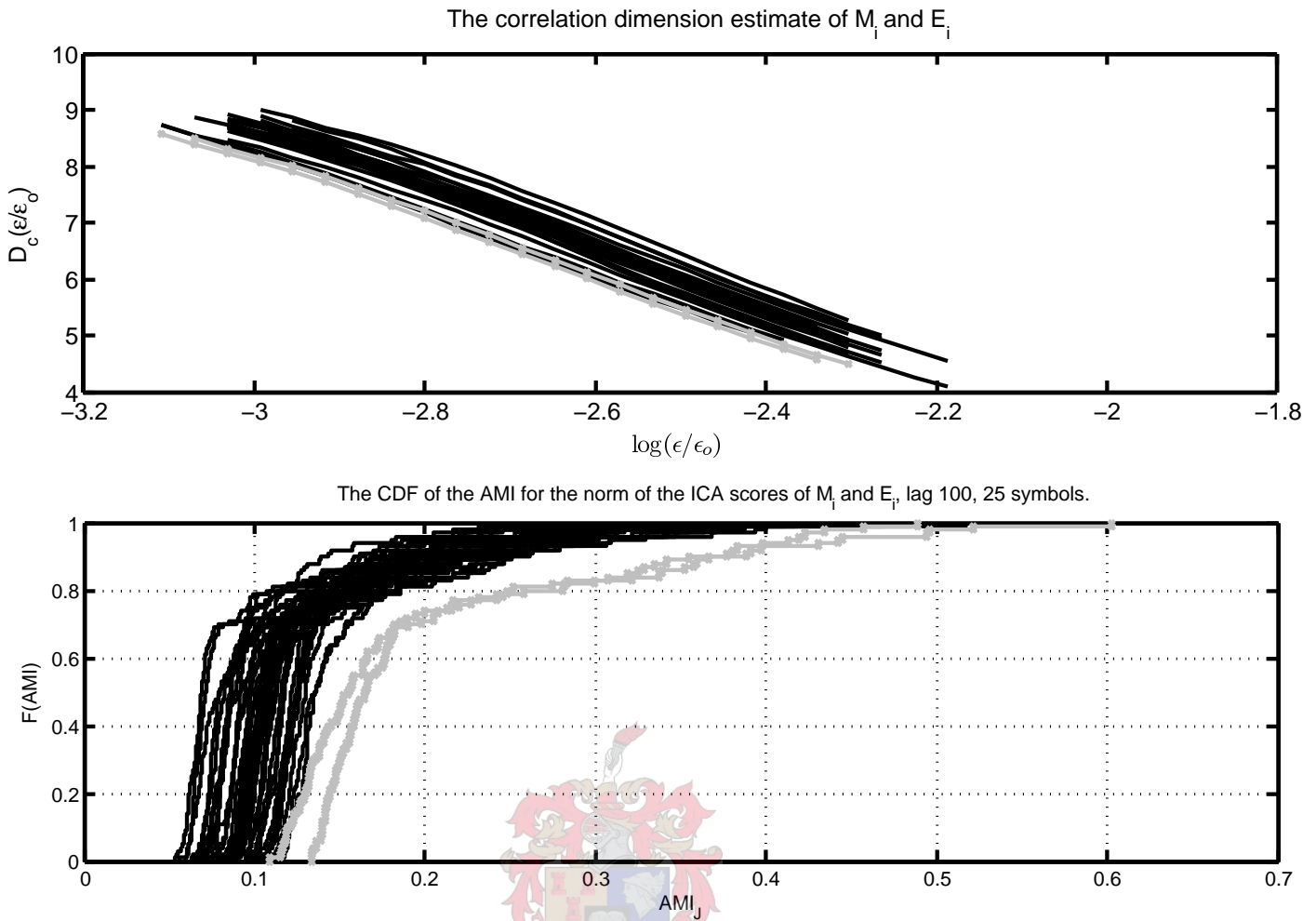


Figure 3.31: Surrogate analysis of $(\log(M_i), \log(E_i))$. The top graph is of the correlation dimension and the bottom of the distribution of the AMI_J scores. The test statistic realizations of the surrogates are in black and those of the data set in gray. Two grays are present since each test was repeated for with a different separation matrix from the ICA. Note that the parameter, ϵ/ϵ_0 to the correlation dimension is on a log scale.

dimensional projection but did not introduce the behavior itself. If the ICA map construction introduced the behaviour, then the correlation between the repeated maps for the ICA map construction would not have reflected the same results.

3.5 Concluding Remarks

Surrogate data analysis is a method that allows to assess whether the structure in a data set is explained by a specified null hypothesis (Schreiber and Schmitz, 1996, 2000). Surrogate data analysis was performed on three of the components of the T4 data set, viz. Dt_i , $\log(M_i)$ and $\log(E_i)$. The null hypothesis was that

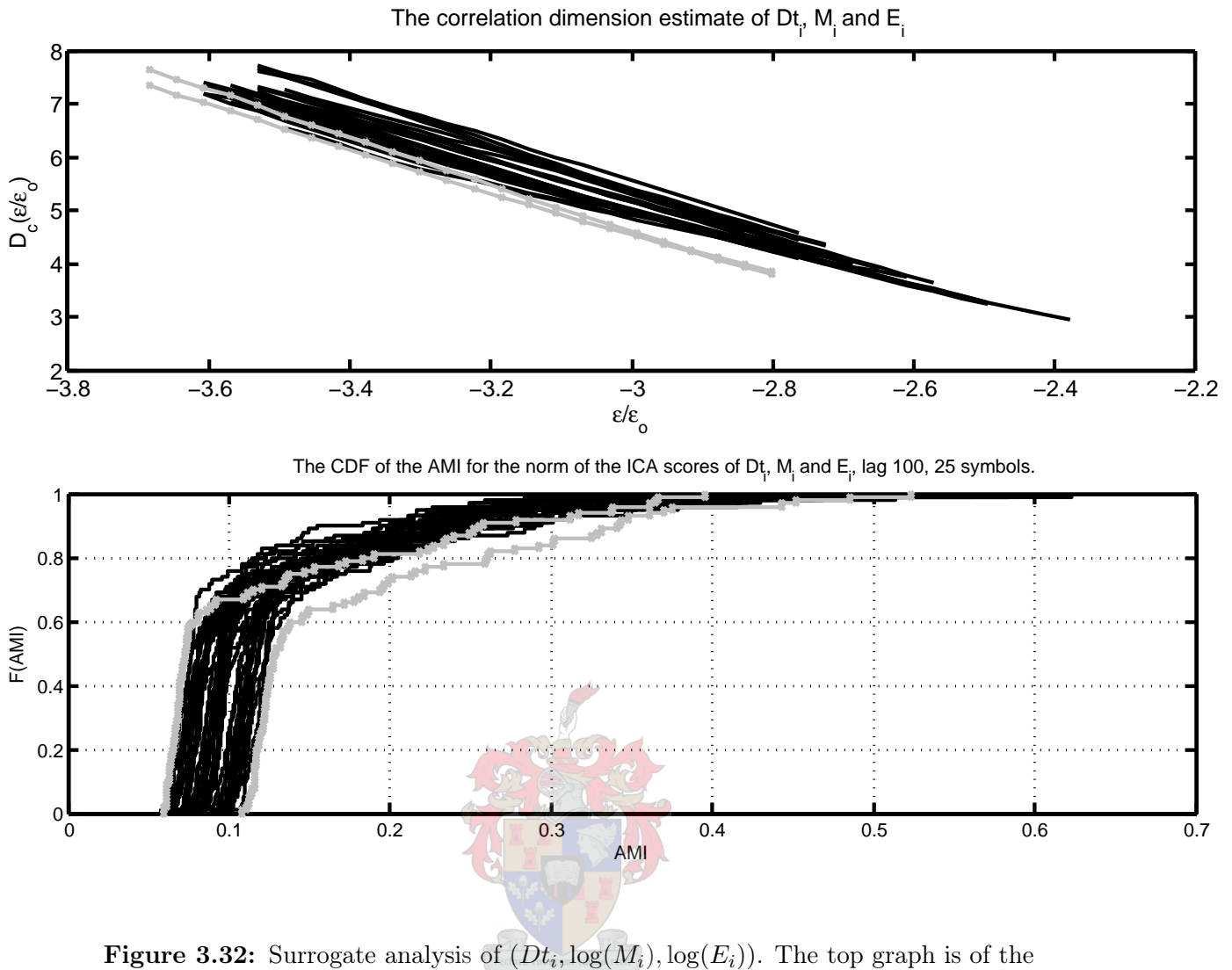


Figure 3.32: Surrogate analysis of $(Dt_i, \log(M_i), \log(E_i))$. The top graph is of the correlation dimension and the bottom of the distribution of the AMI_J scores. The test statistic realizations of the surrogates are in black and those of the data set in gray. Two grays are present since each test was repeated for with a different de-mixing matrix from the ICA.

the three components of T4 were sampled from a three-dimensional red colored noise process. The test statistics used in the hypothesis test were the correlation dimension $D_c(\epsilon/\epsilon_0)$, and the average mutual information lag AMI_J on the sequence of vectors representing the reconstructed phase space from the three variables. The major findings from the analysis can be summarized as follows:

- The hypothesis can be rejected at a confidence level of at least $\alpha = 0.06$.
- T4 did not demonstrate a linear scaling range in the 10-dimensional space of the reconstructed attractor, and hence did not show an actual correlation dimension during the construction of the hypothesis test.

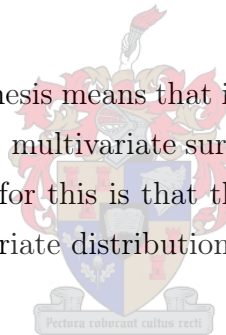
- The AMI_J scores did show that the some structure existed in the phase space beyond red colored noise, but the trajectory was far from smooth, as would be required for high dimensional nonlinear systems to be significantly predictable.

The phase space was reconstructed by means of multivariate delay embedding using a delay lag of 1 and a dimension per component equal to the first point of decorrelation in the autocorrelation plots.

- The dimension of the embedding was successfully reduced to 10 independent components using independent component analysis. It was not possible to achieve the same levels of dimension using a PCA map while maintaining 99.9 % of the variance in the reconstruction.

- An immediate consequence of the rejection of the red colored noise process for the T4 data is that better predictability of the system can be expected using nonlinear models compared to autoregressive moving average models. This issue is investigated in the next chapter.

- Rejection of the null hypothesis means that iteratively amplitude adjusted Fourier spectrum transform (IAAFT) multivariate surrogates cannot serve as synthetic seismic data. A possible reason for this is that the IAAFT multivariate surrogates do not enforce the same multivariate distribution of the data set on the surrogates.



Chapter 4

Modelling $T4$ using Long Short-Term Memory

4.1 Introduction

A Long Short-Term Memory (LSTM) is a flexible, general map and inference procedure capable of inferring context sensitive grammars from examples. LSTM belongs to the class of recurrent neural networks with time varying inputs. The network iterates over consecutive input vectors to compute successive internal states. Subsequently, output and error signals are computed from the internal states on request. The LSTM architecture was first proposed in Hochreiter and Schmidhuber (1991). In a series of subsequent publications, the architecture was extended and applied to different benchmark problems.

Traditional time delay neural network architectures, which are trained using the first derivative of the error signal, have trouble learning sequences of values with long noisy time lags. An analysis of the first derivative error signal back propagation in networks with time-varying inputs show that the size of the error gradient decreases exponentially as it is back propagated into the network. On the basis of this analysis, the LSTM architecture was constructed to provide a solution to the long time lag, vanishing error gradient problem. A shortcoming of the original architecture is that it can only process input sequences of a finite length in a batch process style. The internal state of the network becomes unstable after it has been presented with an excessive number of input vectors.

Gers and Schmidhuber (2001); Gers et al. (1999) proposed extensions of the original

architecture which possess the same functionality as the original with the added advantage of continual prediction. This form of LSTM can be collapsed to an equivalent of Real Time Recurrent Learning by fixing one of the weights in the network to a dominating value (Haykin, 1999). The architecture investigated in Gers and Schmidhuber (2001) was used in modelling the transition table for small context-free and context-sensitive grammars. The transition table was generalized from strings taken from the grammar.

Another LSTM architecture was applied in modelling the Mackey-Glass time series and the Class A laser data set of the Santa Fé time series competition (Gers et al., 2001, 1999). It was illustrated that the LSTM architecture performed worse than other time window approaches in modelling and predicting the time series. It must be mentioned that the values of the time series were presented to the network one at a time, and to model the time series all the relevant information had to be stored in the context values of the network. A time window approach was not used in learning the future unfolding of the time series from past information. Gers et al. (2001) concluded that the LSTM can “track the strongest eigen-frequency in the task but was unable to account for high-frequency variance”. Bakker et al. (2000) obtained improved results on the same data set using a time delay input to an feed forward neural network, having reduced the dimensionality of the input series by a PCA linear transformation.

In this chapter an LSTM-based predictive model is used on the T4 data. The LSTM-based model is subsequently compared with the autocorrelation function of the T4 data set. If the structure in the T4 system is fully described by its autocorrelation function, the LSTM network should not be able to predict future events any better than the autocorrelation function. This would be equivalent to testing the null hypothesis that the T4 system is identical to an autoregressive system. Note that Long Short-Term Memory is applied to a point process (viz. T4) only for the purpose of demonstrating that the T4 data set is more predictable than red colored noise without attempting to construct an optimal model to predict seismic activity. The presented methodology does not include a mechanism to optimize the model of T4.

The functional form of the LSTM networks implemented for this thesis can be found in Appendix A.

4.2 Modelling Using Mean LSTM Networks

Long Short-Term Memory is a parameterized map, $\overline{X}_{sk}(t) \xrightarrow{LSTM_k} \overline{Y}_{sk}(t)$ that maps a sequence of input vectors, $\overline{X}_{sk}(t)$ to a sequence of corresponding output values, $\overline{Y}_{sk}(t)$. The network input and output is a sequence of time ordered vectors denoted by t . The input and output pair are divided into a training set and a validation set denoted by s . Each LSTM map, i.e. $\xrightarrow{LSTM_k}$, maps a given sequence of time ordered training and validation inputs to a corresponding set of outputs. The mean estimator $\overline{Y}_s(t)$ is then an average of the LSTM estimates for the desired output vector at time step t which belongs to either a training set or a validation set. The mean estimator is computed as

$$\overline{Y}_s(t) = 1/N \sum_{k=1}^N \overline{Y}_{sk}(t) \quad (4.1)$$

$$\overline{X}_{sk}(t) \xrightarrow{LSTM_k} \overline{Y}_{sk}(t) \quad (4.2)$$

A number of maps, $k = 1 \dots N$, of the same architecture (see Appendix A.1, 102) and learning parameters are fit to different selections of training and validation sets. The parameters of network are estimated only on the training set, and the actual performance of the network is measured on an independent validation set. Constructing and evaluating the mean estimators $\overline{Y}_s(t)$ from the LSTM estimators $\overline{Y}_{sk}(t)$ is done according the following sequence of steps:

1. Dataset setup

The maps from network inputs to network outputs need to be associated with each other ($\overline{X}_{sk}(t), \overline{Y}_{sk}(t)$). The sampled association is the basis for fitting the map represented by the LSTM networks.

2. Training and validation set selection

The data set is divided into two mutually exclusive training and validation sets. The training set is used in the network's parameter fit. The validation set determines the performance of the fit. The two types of sets are denoted by s and each selection denoted by k .

3. LSTM training

To fit the network a number of network and training parameters need to be

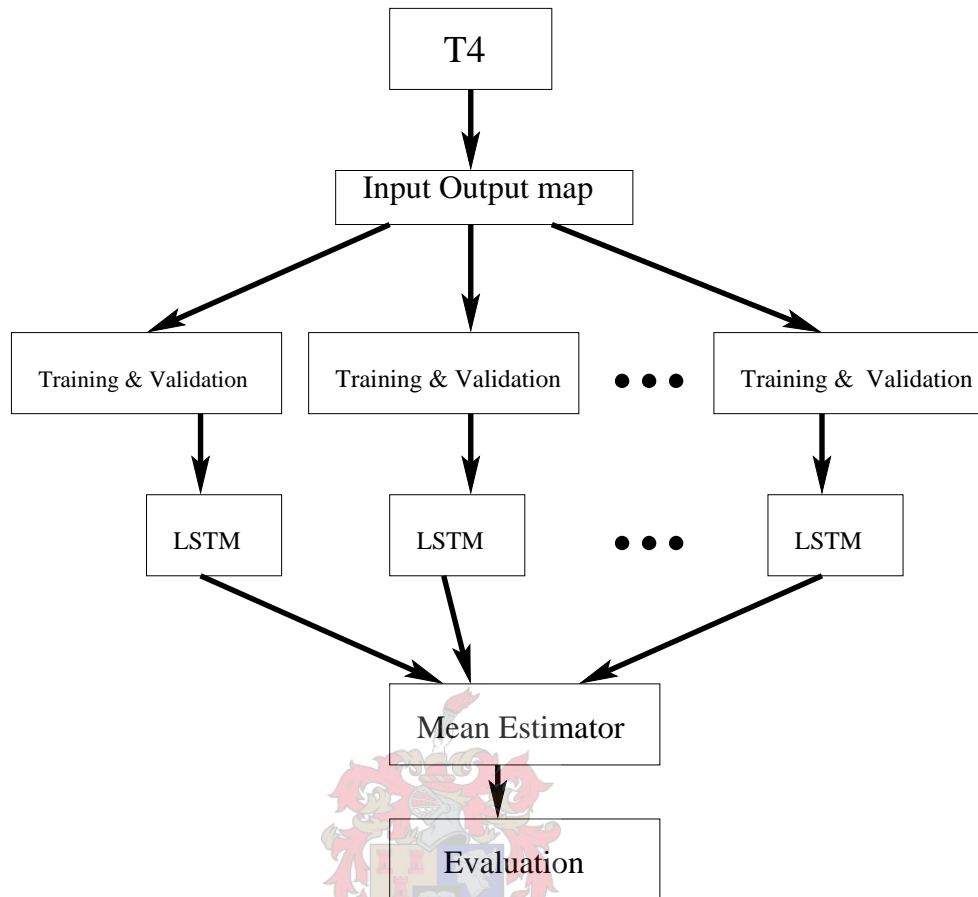


Figure 4.1: A schematic representation of the 7 steps for modeling and evaluating the components of T4. The schematic starts with the T4 data set and ends with the comparison between the mean model and the desired components.

defined. These parameters are the size and shape of the network, the learning rate, bias term, weight initialization, weight update and error generation strategy.

4. LSTM estimates

During the iterative fit of the network to the data set an optimal weight set is reached for that specific training sequence. A set of $\overline{Y}_{sk}(t)$ estimates is associated with the optimal weight set.

5. Repetition

Steps 1 to 4 are repeated to generate a number of optimal weight sets and associated $\overline{Y}_{sk}(t)$ estimates. Each repetition is indexed by k .

6. Mean estimates

A mean validation estimator is obtained by averaging the validation set estima-

tors for the different fits from step 5. A mean training estimator is constructed from the individual training set estimators for the different fits sampled in step 5.

7. Prediction test

The objective of fitting the network is to establish if the future evolution of T4 can be derived from the past behavior with improved performance than is possible using autoregressive models. The future evolution modelled by the mean estimators is evaluated by conducting two hypotheses tests. The first hypothesis is that the correlation between the estimators and the desired values are greater than zero. The second hypothesis is that correlation coefficients are larger than the autoregressive coefficients.

The 7 steps are schematically shown in Figure 4.1. Results from LSTM modelling are reported later in the chapter.

4.3 Data set training and validation

The first three stages in constructing and evaluating the mean estimator consists of fitting an LSTM model to a set of input-output pairs previously partitioned into a training and validation set. This section of the chapter deals with these three closely related steps.

The input data $\overline{X}_{sk}(t)$ is obtained from the two 10-dimensional ICA scores used in the surrogate data analysis of Figure 3.32, i.e. the two sets of ICA scores for $\{(Dt_i, \log(M_i), \log(E_i))\}$. Each of the ICA score vectors was defined in equation (3.26) and is denoted by s_{ai}^{10} with index i iterating over the sequence and a denoting a set of ICA scores, i.e.

$$\overline{X}_{sk}(t) = (s_{1t}^{10}, s_{2t}^{10}). \quad (4.3)$$

The set of targets for the network, \overline{Y}_i is constructed from a 5-dimensional lag of the components of T4. The vector of components is shifted one step into the future from each input vector. Thus, $(s_{1i}^{10}, s_{2i}^{10})$ is associated with time stamps $i+1$ to $i+6$ of T4's components. The outputs are scaled to be within the range $[-0.5, 0.5]$ using

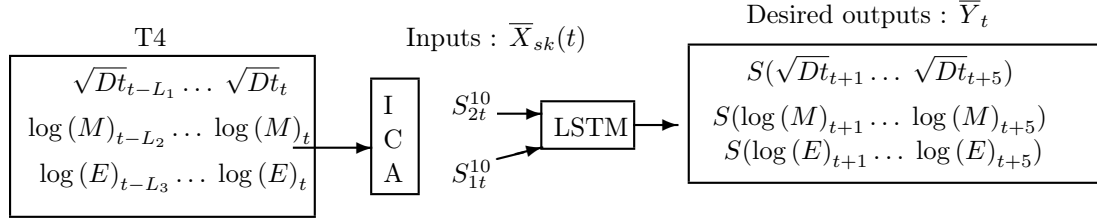


Figure 4.2: The input-output map for modeling T4 with LSTM.

a function $S(\cdot)$, that is,

$$\bar{Y}_t = S(\sqrt{D}_{t+1}, \dots, \sqrt{D}_{t+5}, \log(M)_{t+1}, \dots, \log(M)_{t+5}, \log(E)_{t+1}, \dots, \log(E)_{t+5}) \quad (4.4)$$

The desired output for the data set is referred to as \bar{Y}_t since each lag of T4 components is uniquely identified through its time stamp t . The time interval components of T4 are scaled to their square root prior to transforming using $S(\cdot)$. Note that all the correlation tests are done on the desired values and not on the original components of T4. Figure 4.2 is a schematic representation of input-output map for modelling the components of T4 using LSTM. The schematic starts with the inputs as a delay lag of T4 components, starting at time stamp t . The schematic ends with the outputs as a delay lag of 5 for the components of T4 starting at time stamp $t+1$. The LSTM network inputs are explicitly defined in equation 4.3 and the outputs defined in equation 4.4.

The data set is partitioned using a non-replacement random selection such that 90% are assigned to a training set and the rest to a validation set. Each training set is fitted with a network of 4 blocks consisting of 1 cell each. The bias term was fixed at 0.95 and the learning rate alternated between 1×10^5 and 1×10^6 . Each fit was iterated over 25000 epochs. An error signal was generated for every vector of the training set. Weight updating occurred each time an error signal was generated. The fitting procedure was repeated over 149 training and validation set selections.

The training sequence for each epoch was monitored using a Normalized Prediction Error (NPE) for higher dimensions given by

$$NPE_k = \frac{\text{mean}(|\bar{Y}_{sk}(t) - \bar{Y}_i|)}{\text{mean}(|\bar{Y}_i - \text{mean}(\bar{Y}_i)|)} \quad (4.5)$$

The $\text{mean}(\cdot)$ denotes the mean of a sequence of values and the $|\cdot|$ denotes the euclidean norm. The NPE_k score is the ratio of the average distance of the error of the

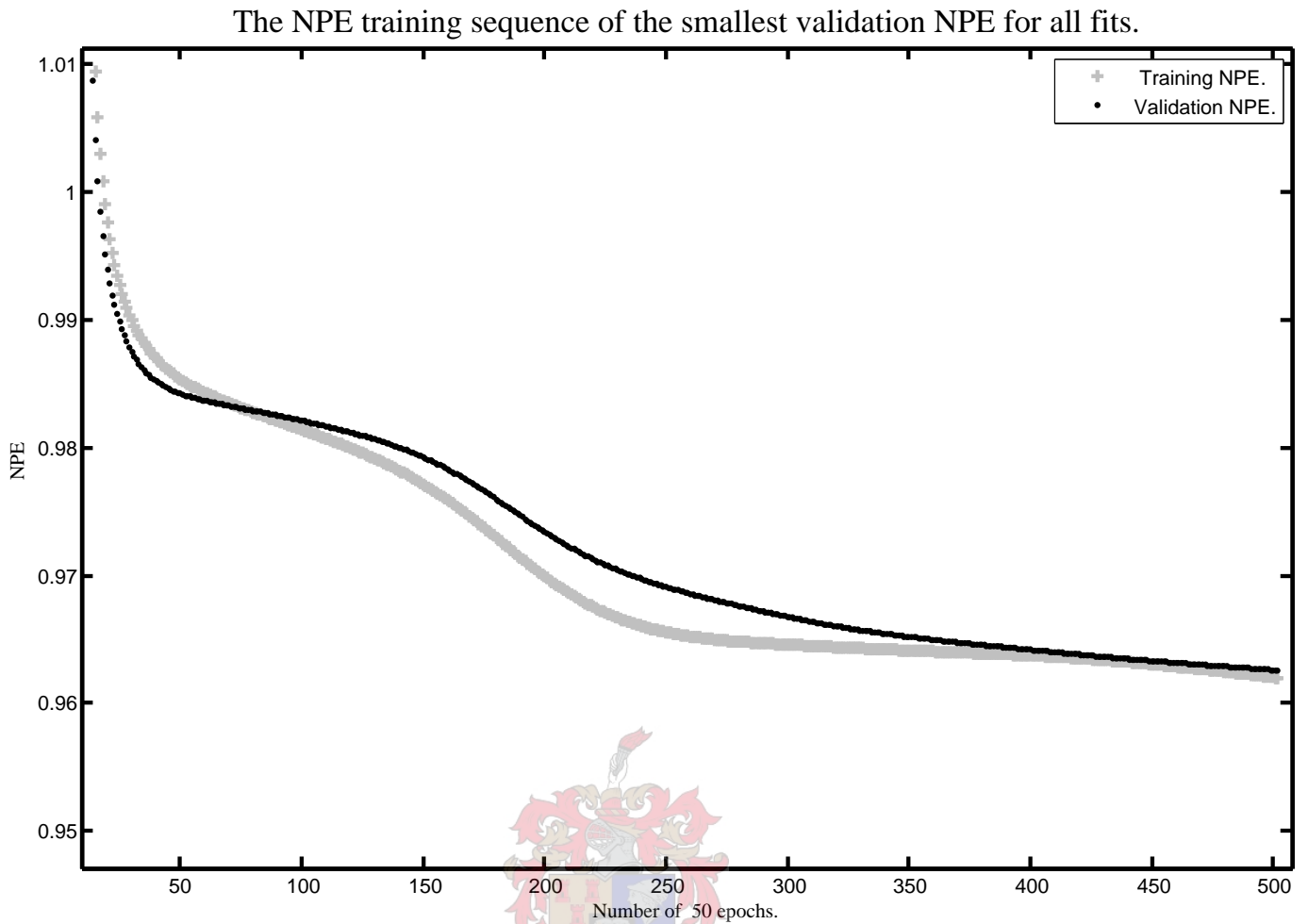


Figure 4.3: An error per epoch plot for an LSTM training sequence.

estimator compared to the average distance of the error of the mean estimator $mean(\bar{Y}_i)$. Figure 4.3 shows a sequence of NPE errors for the training and validation sets during training. The training was stopped after 25000 epochs. Clearly, data training and fitting does occur with the validation set tracking the behaviour of the training set. Figure 4.4 gives the NPE errors for each of the 149 LSTM maps. An LSTM map with minimum training error was sampled from the 25000 epoch sequence. The jump in NPE validation error is the difference in the learning rate. The initial network was trained with the smaller learning rate. A few of the errors on the right of the graph indicate over-fitting of the training set as validation error greater than unity occurred for relatively small training errors. A larger number of the other fitted networks shows that more epochs of training maybe required for an optimal fit. Figure 4.5 gives the number of times a 15-dimensional vector of T4 components was selected as either a training or a validation vector. The mean

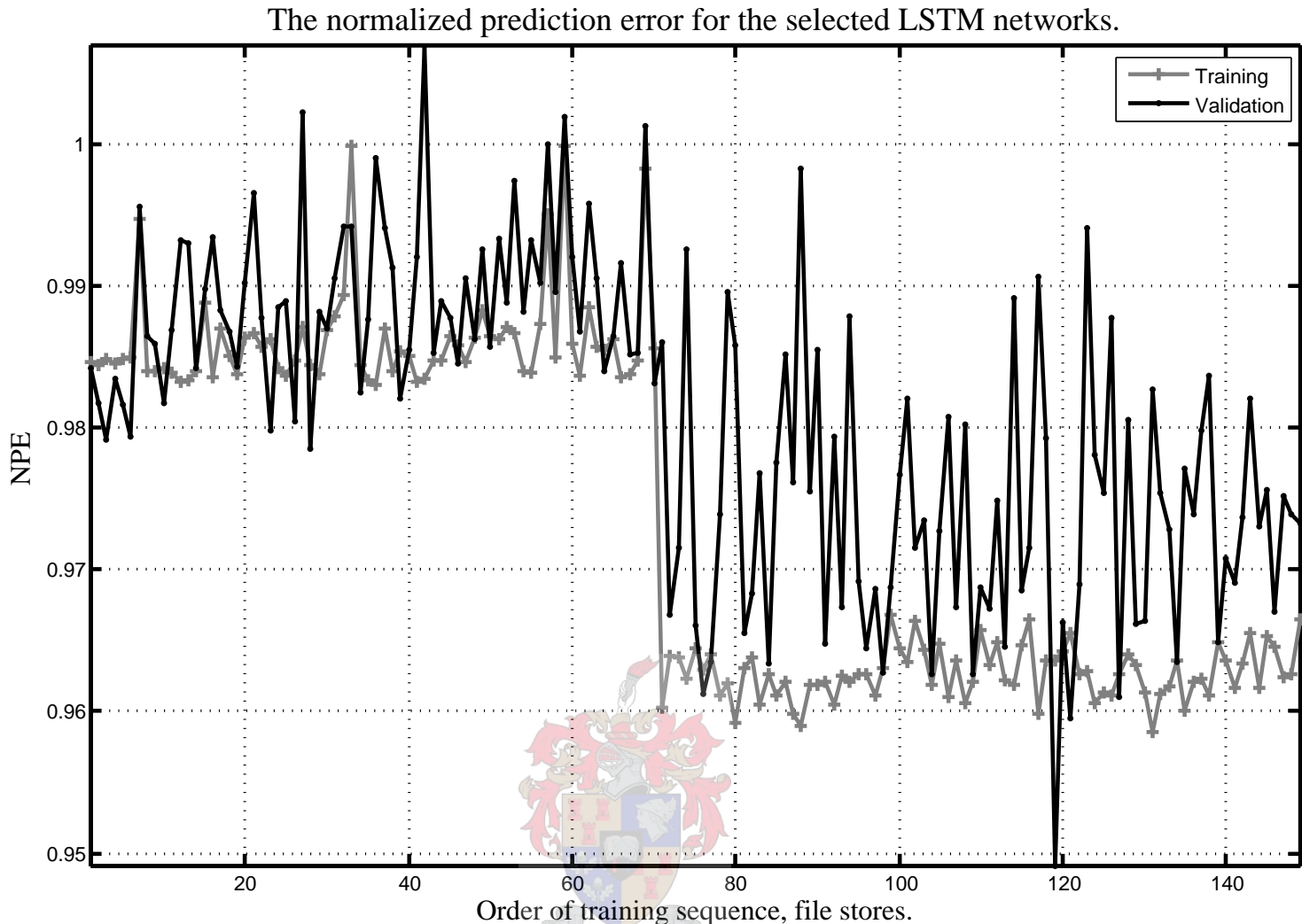


Figure 4.4: Training and validation errors for the selected LSTM networks.

validation estimator is constructed for values with at least 3 validation estimates. Also shown in the plot are the number of points used in computing the correlation statistics (2525 - almost the whole data set!). The large number of data points in the mean estimators helps to provide tight confidence bounds on the test statistics used in hypothesis testing. The figure shows that no portion of the domain of \bar{Y}_i was left out in either the training or validation mean estimators.

No effort was made to find an optimal training strategy or network architecture for the particular problem of the T4 map. The focus was on establishing the expected behavior of any map when relating the past behavior of T4 to its future behavior. Any valid map from the past to the future unfolding of the system performing better than an ARMA model would have been sufficient. Finding the training and set selection procedure that best solves the T4 problem is a problem that is beyond the scope of this study.

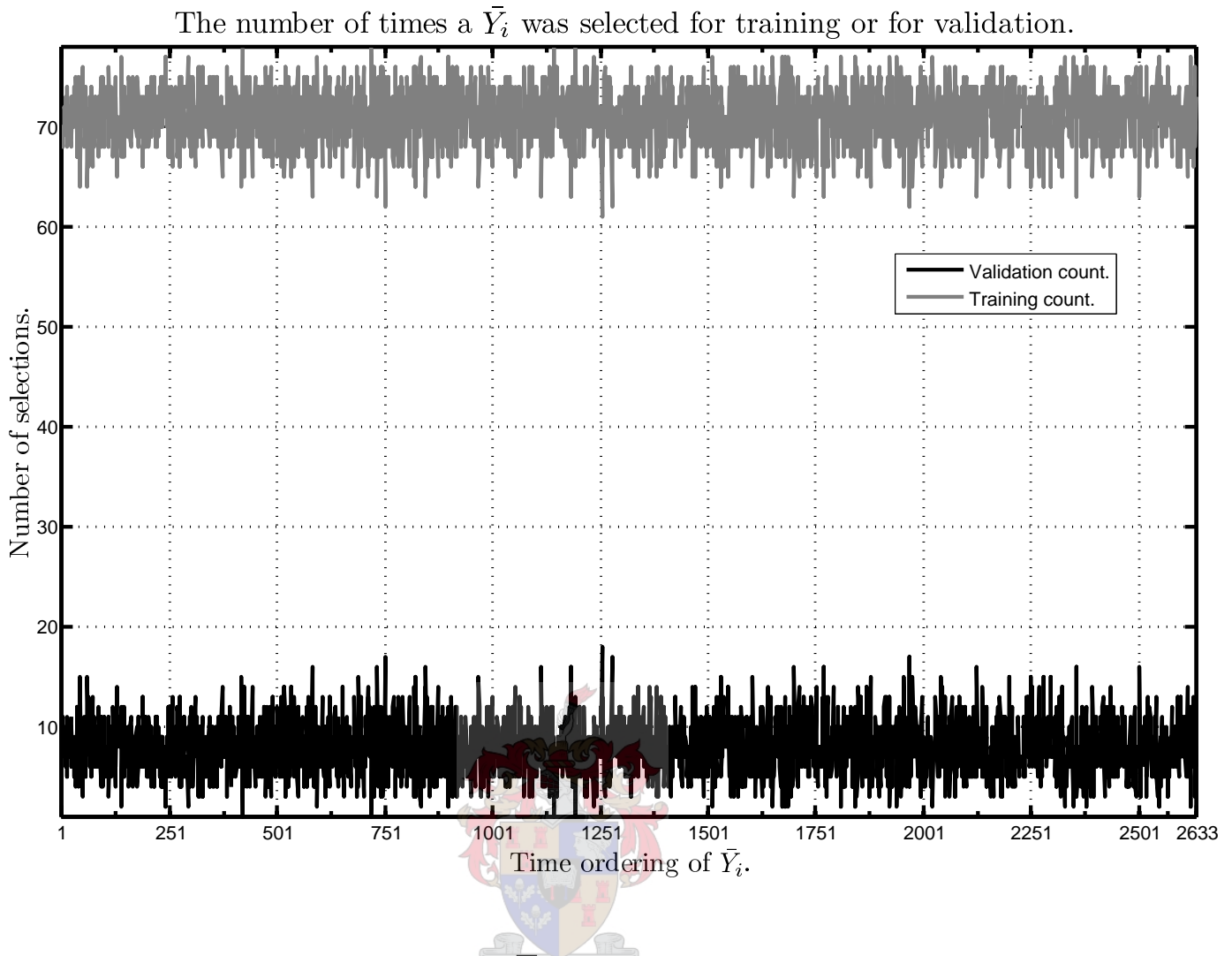
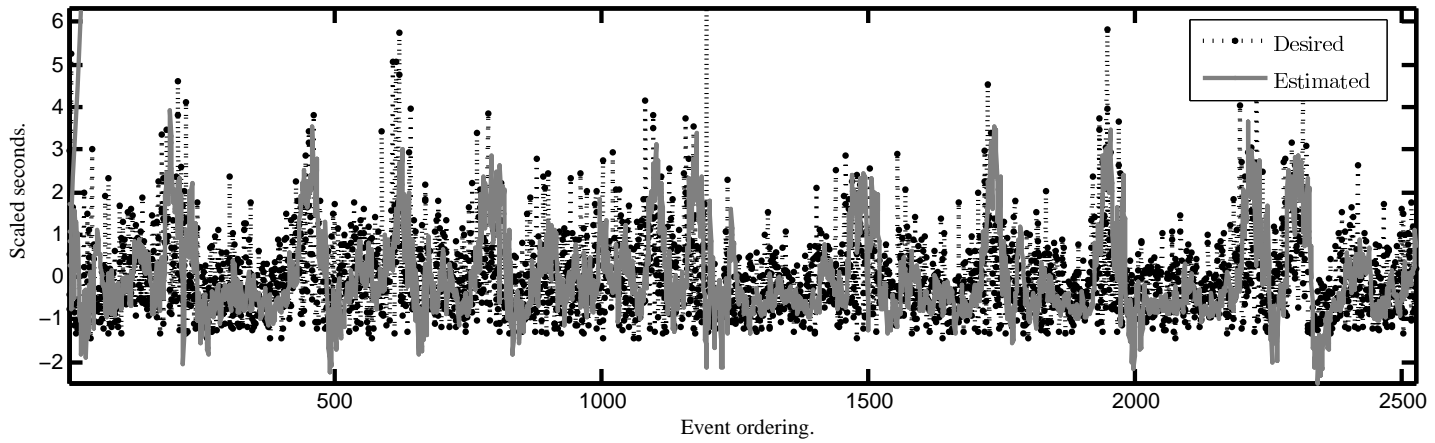


Figure 4.5: The number of times a \bar{Y}_i vector was selected for training or for validation.

4.4 Estimating the components of T4

A total of 149 LSTM fits to a 90% training set and a 10% validation set were used to construct a mean validation estimator and a mean training estimator for the 15-dimensional desired output vectors. Each desired output vector represents a lag of 5 observations from the components of T4 beyond the time stamp of the input vectors. A lag of 5 observations into the future represents a prediction of 5 consecutive components for each network output. The following hypothesis tests include correlation coefficients of estimators for predictions up to 5 step ahead. Figure 4.6 displays the mean validation estimator for one-step prediction in comparison to expected values. The top graph displays the relationship of the two variables to event ordering, and shows that the estimated values constructs an event series that is similar to the original event. The mean estimators have the ability to model T4

The scaled time intervals $S(Dt_{i+1})$ and their mean estimator in event order.



The scaled time intervals $S(Dt_{i+1})$ and their mean estimator in event order, standardized.

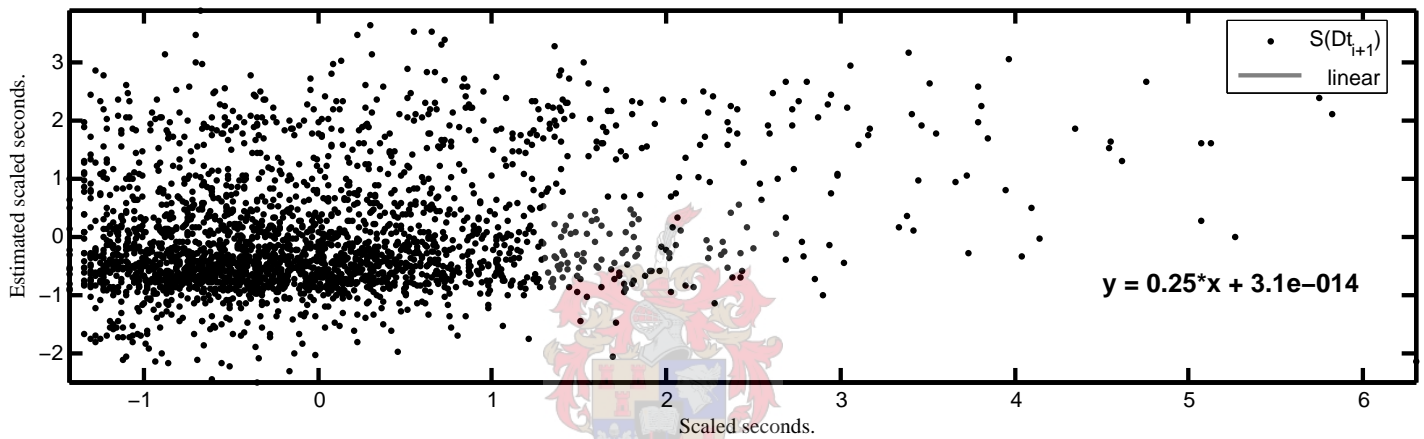


Figure 4.6: A plot of the estimate of $S(Dt_{i+1})$ in comparison to the actual values.

otherwise the mean validation estimator would have been some function completely different to the original data, for example, a flat line. In the bottom graph which shows the relationship between the original component and its estimator, a linear trend can be discerned. Although the trend is not large, it is significant due to the large number of points that are used in the estimate. The linear trend does not show that the mean validation estimator results in an event sequence similar to the original data set. A more extensive relationship between desired estimators and the validation estimators might be established using a different, time dependent measure between the two sequences. A possible measure would be the AMI between the two variables. Figure 4.7 shows the autocorrelation coefficients for the 15 mean validation estimates. The estimator results in a significantly more autocorrelated event sequence than the original data set. The result is not surprising since the LSTM maps used in the estimate have only cell to output connections. The auto-

The autocorrelation coefficients of the components of the LSTM based mean estimator.

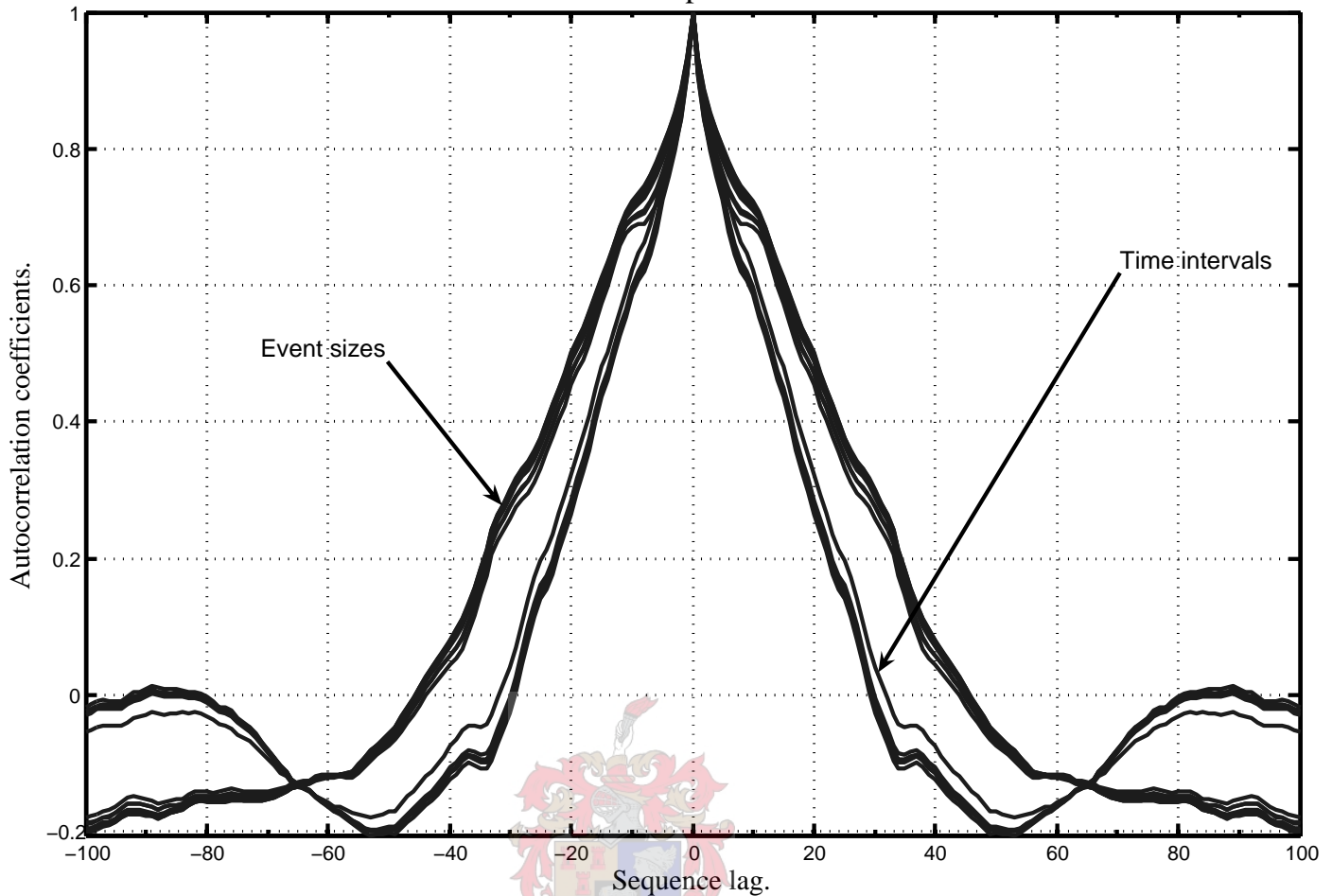


Figure 4.7: The autocorrelation coefficient estimates for the modeled components of the T4 dataset.

correlation sequences separate into two groups: the estimators for the time intervals on the inside, and the estimators for the event sizes on the outside. Figure 4.8 shows the cross-correlations in the mean validation estimator. Note that the $E(+1)$ -label refers to the $S(\log(E_{i+1}))$ component of T4. The estimators for the time intervals are almost the same values, as are the estimators for the event sizes. However, the estimators for the time intervals differ from the estimators for the event sizes. The large cross-correlations and autocorrelation values in the estimators are undesirable because they are not present in the original data set. Thus, LSTM models a completely different system generating similar event sequences but with different auto- and cross-correlation structures. It appears as if LSTM is estimating two variables instead of three.

Maintaining the auto- and cross-correlation structure in a training set might be

The cross correlation structure of the LSTM based mean estimator.

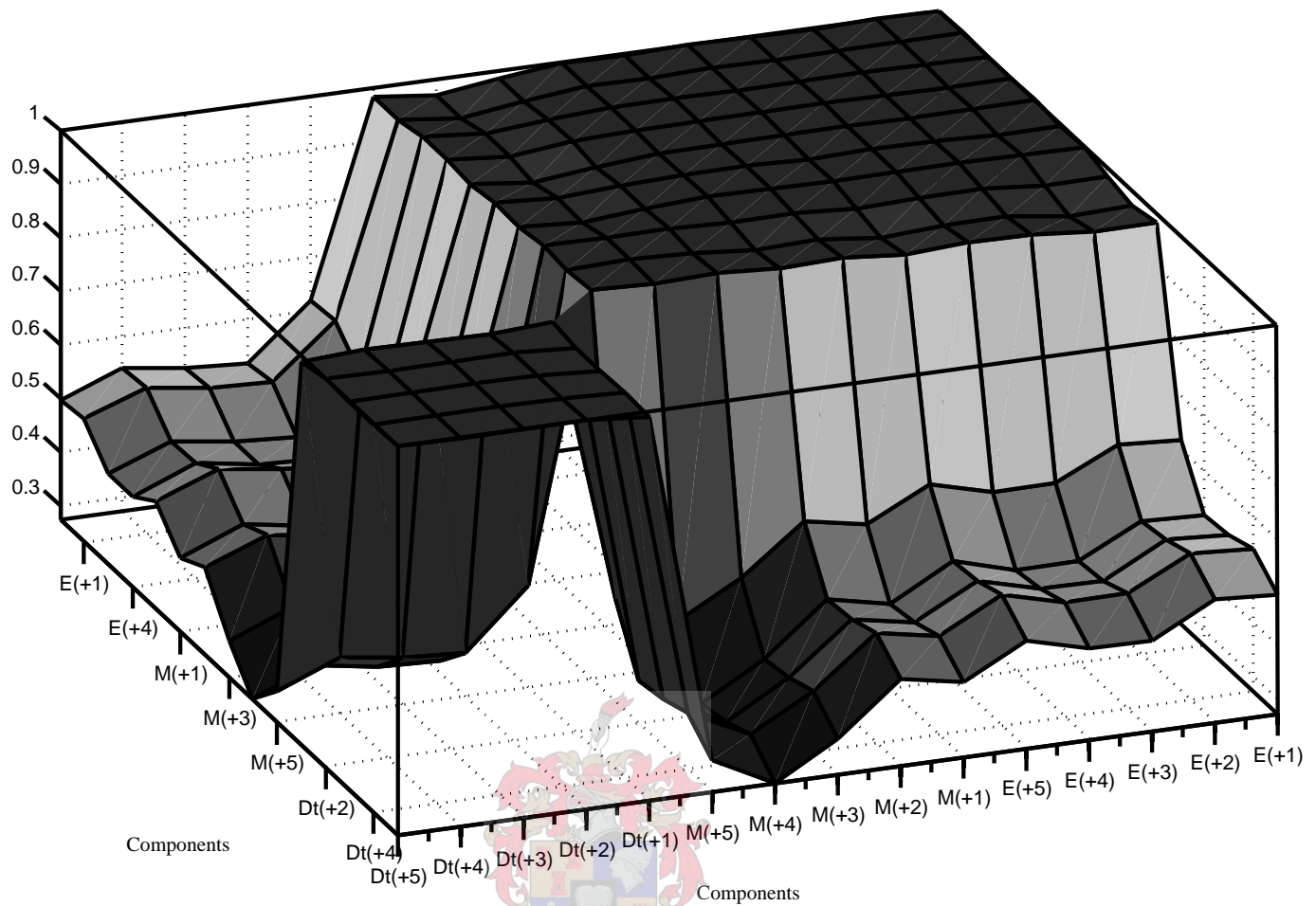


Figure 4.8: The cross correlation coefficients for the modeled components. The $E(+1)$ -label refers to the $S(\log(E_{i+1}))$ component of T4.

another source of error that can help in the search direction of the weight updating. That is, the emergent correlation structure in the estimators is an inherent part of the LSTM network after 25000 epochs of training.

4.5 Hypothesis tests on correlation coefficients

Hypothesis tests were conducted on the mean validation and training LSTM estimators to establish whether they predict the future evolution of the system better than an autoregressive moving average according to the following 3 steps:

1. The correlation coefficients of the mean estimators,

R_{si} , $i \in Comp = \{Dt(+k); E(+k); M(+k) \bullet k = 1 \dots 5\}$ are compared to the

autocorrelation coefficients Acc_i , $i \in Comp$, of the component at the same lag of estimation;

2. The R_{si} values are subjected to the hypothesis test that they are equal to zero;
3. The R_{si} values and the Acc_i values are subjected to the hypothesis test that they are of the same size.

The indices in $Comp$ corresponds to the labels on the figures demonstrating the component and offset corresponding to each estimator.

The hypothesis test in step 2 that the correlation coefficient ρ between a component of T4 and a mean estimator is zero, i.e.

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

uses the test statistic

$$t = \frac{R\sqrt{N-2}}{\sqrt{1-R^2}} \sim \text{Student's-T}(N-2) \quad (4.6)$$

with $T(N-2)$ distributed as Student's t -distribution with $N-2$ degrees of freedom under assumption of the H_0 hypothesis. The value of $N = 2525$ is the number of pairs of values on which the R value is computed.

The comparison in step (1) above is depicted in Figure 4.9. The R_{si} values are larger than the Acc_i values for both the training and the validation sets. The comparative $R_{H_0} = 0.0038$ value for $N = 2525$ is the cutoff value at which the hypothesis will be rejected at an $\alpha = 0.999$ level of confidence. The alpha value is rather large but demonstrates the importance of the large number of points resulting from the mean estimators

$$R_{H_0} = 0.0038 = \sqrt{T(N-2)_\alpha^2 / (N-2 + T(N-2)_\alpha^2)}. \quad (4.7)$$

Since the cutoff value is smaller than the estimates, the hypothesis that the correlation between the desired values and the estimated values is zero can be rejected at almost any level of confidence.

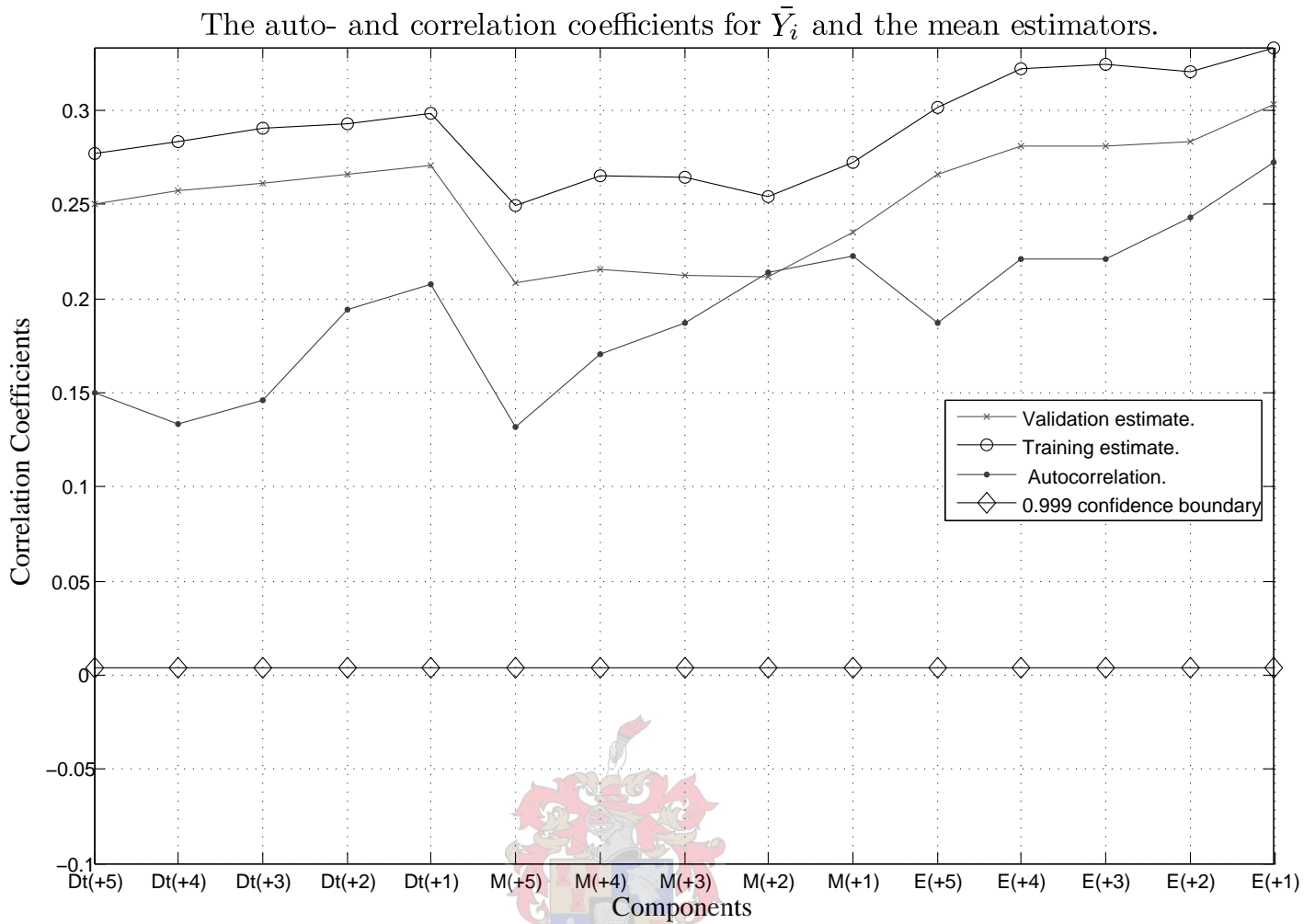


Figure 4.9: The correlation coefficients between the \bar{Y}_i s and the estimator; the autocorrelation coefficients of T4; the critical value for rejecting that the correlation coefficients (not the autocorrelation coefficients) are equal to zero. The $E(+1)$ -label refers to the $S(\log(E_{i+1}))$ component of T4.

Figure 4.9 depicts four sets of correlation coefficients: (i) correlation coefficients of the mean training estimator; (ii) correlation coefficients of the mean validation estimator; (iii) the autocorrelation coefficients for a lag of data components; and (iv) the cutoff values for rejecting the hypothesis that the correlation coefficients for the mean validation estimates are not zero. Comparing the four sets leads to the following conclusions:

- The training and validation correlation coefficients are larger than zero;
- The training and validation correlation coefficients appear to be larger than the autocorrelation coefficients;
- The training and validation correlation coefficients degrade with time: The

correlation coefficients can only degrade if some kind of structure was present;

- The training and validation correlation coefficients degrade in the same way: The LSTM networks could train more.

The hypothesis test in step 3 that the correlation coefficients from the mean estimator is essentially the same as the autocorrelation coefficients, i.e.,

$$H_0: \rho_{Acc_i} = \rho_{si} = \rho$$

$$H_1: \rho_{Acc_i} \neq \rho_{si}$$

is based on the test statistic:

$$z(R_{si}, Acc_i) = \frac{F(R_{si}) - F(Acc_i) - F(\rho) - F(\rho)}{\sqrt{1/(N_{Acc_i} - 3) + 1/(N_{si} - 3)}} \sim \text{Normal}(0, 1) \quad (4.8)$$

$$F(R) = 1/2 \log_e \left(\frac{1+R}{1-R} \right) \quad (4.9)$$

with $\text{Normal}(0, 1)$ distributed as Gaussian with zero mean and unit variance under assumption of the H_0 hypothesis. The values of N_{Acc_i} and N_{si} are the number of pairs of values on which the R values are computed. $F(R)$ is the Fisher transform of correlation coefficients and the test statistic z is derived from the distribution of $F(R)$. In the hypothesis test $z(R_{si}, Acc_i)$ and the probability of realizing $z(R_{si}, Acc_i)$ under the null hypothesis are evaluated.

Figure 4.10 depicts the realized test statistic for the hypothesis that the correlation coefficients for the estimators are the same as the autocorrelation coefficients between two components separated by the same lag. The test statistics for $\log(M_{i+3})$, $\log(M_{i+2})$, $\log(M_{i+1})$ could probably have been sampled from a normal distribution of zero mean and unit variance. The other correlation coefficients of the other components were probably not sampled from the same distribution. The hypothesis that the correlation coefficients for the other estimator are the same as the autocorrelation coefficients for the components can be rejected, possibly with a small error. A critical value of 1.96 on the Y-axes corresponds to an error of $\alpha = 0.975$ and most of the realized values are above a critical value of 2 standard deviations.

The mean estimators demonstrate more structure in T4 than could be expected of red colored noise because:

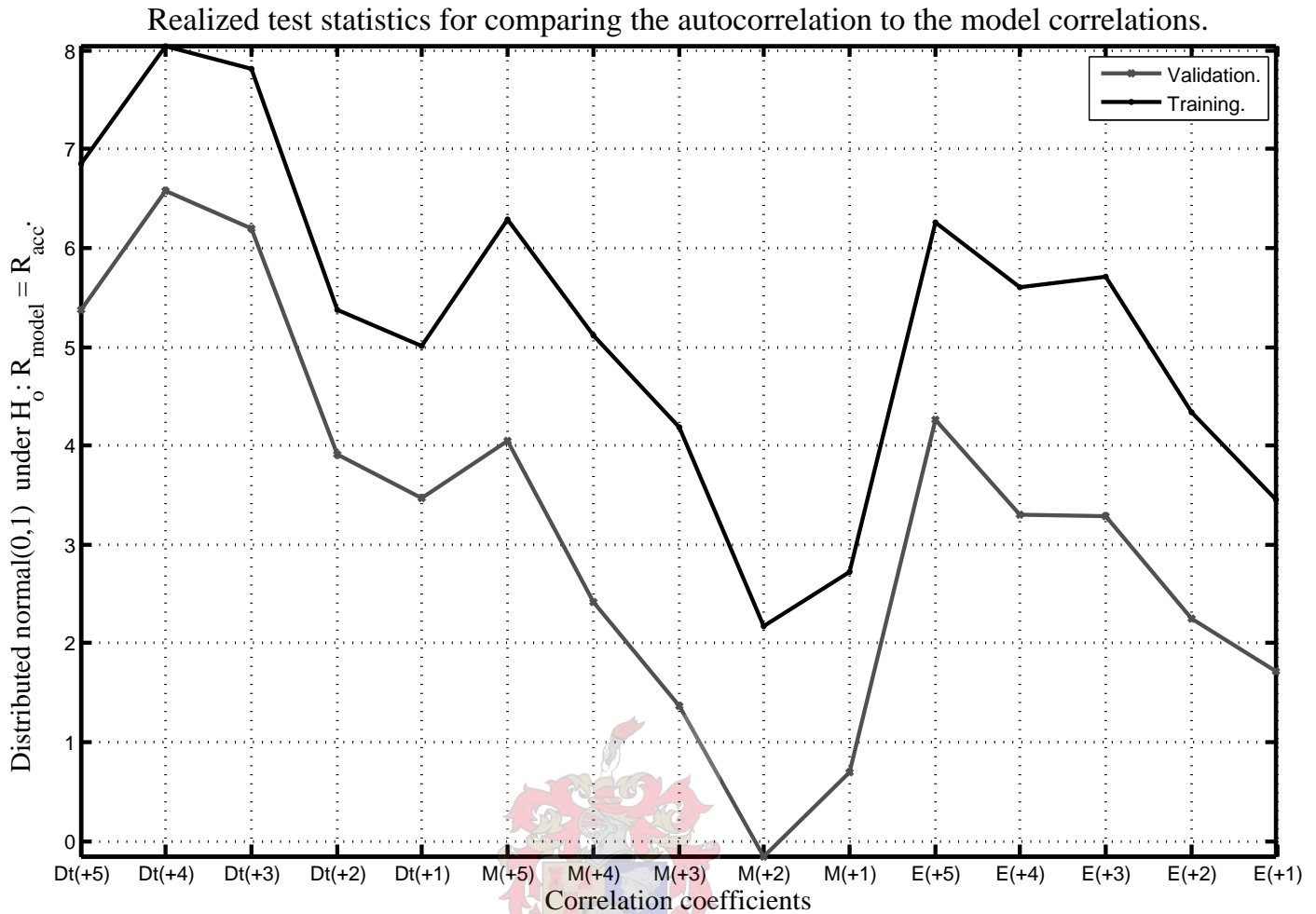


Figure 4.10: The hypothesis test that the correlation of the LSTM estimators are significantly different from the autocorrelation coefficients. The y-axis is the value of the test statistic. Under assumption of the null hypothesis the test statistic should be distributed $normal(0, 1)$. The $E(+1)$ -label refers to the $S(\log(E_{i+1}))$ component of T4.

- All the correlation coefficients for the estimators and the components are significantly different from zero;
- The correlation coefficients for the Dt_i s and $\log(E_i)$ s components are significantly different from the autocorrelation coefficients;
- The correlation coefficients for the Dt_i s and $\log(E_i)$ s components are larger than the corresponding autocorrelation coefficients;
- The correlation coefficients between the estimated components and the actual components decrease persistently with an increase in time.

4.6 Concluding Remarks

The chapter demonstrated that the components of T4 ($\log(E_i)$ and Dt_i) can be predicted beyond the structure exhibited by red colored noise (viz. an autoregressive moving average, scaled with a fixed invertible nonlinear function) The prediction was conducted on a lag of values into the future and the predictor outperformed the autocorrelation coefficients on all the lags.

The estimator was constructed by taking the mean estimator for the validation sets of a number of LSTM fits on a training set. The mean estimator was successful even though neither the LSTM network architecture nor the learning parameters were optimized for the fitting problem. The LSTM model showed that it could generate output similar to the desired output but with a different autocorrelation structure. The number of valid estimates resulting from the mean estimator scheme helps to improve on the level of significance for the hypothesis tests relying on the estimators. The optimal network architecture and learning parameters for fitting LSTM on the T4 problem is an open research issue.



Chapter 5

Conclusions

In this study, seismic data (referred to as T4) were analyzed by comparing it with artificially generated data (surrogate data) that had the same autocorrelation and probability density functions as the T4 data set. The surrogate data were generated as realizations of the estimated autocorrelation and probability density functions of the T4 data set. A total of 16 such data sets were generated in order to get reliable estimates of the statistics on which the comparisons of the T4 and surrogate data sets were based.

Three statistics derived from each of the data sets, viz. the average mutual information, the correlation dimension and the prediction error, were used in the comparisons. The study demonstrated that the T4 seismic system can be distinguished from an autoregressive system, possibly sampled through a nonlinear invertible function, due to its time invariant (viz. persistent) properties. The hypothesis that T4 was sampled from the specified autoregressive system was rejected on all three measures.

A possible explanation for the distinction between T4 and the noise system is the seismic law regarding the clustering of large events (Section 3.3.1, page 53). The other 2 seismic laws relevant to the hypothesis test on T4 were sufficiently explained by the null hypothesis. Hence, the noise generating mechanism cannot be used for generating synthetic seismic data nor for determining the expected future unfolding of a seismic generating system.

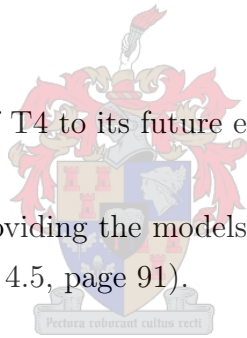
The measures of the small- and large-scale scattering on the flow of the T4 state space suggested that T4 does not represent a deterministic system (Section 3.3.1, page 51). This conclusion was supported by the test for a one-to-one map from past T4 states to the future unfolding of T4 using Long Sort-Term Memory, (Chapter 4). Even

though the LSTM model did predict the future behavior of T4 more successfully than expected for systems consistent with null hypothesis, the model only predicted a small portion of the variance in the T4 data set (Section 4.5, page 91).

In the phase space reconstruction of T4 the study demonstrated that:

- The ICA map was an effective dimension reduction tool that maintained more of the original variance of T4 than PCA could for the same number dimensions.
- The probability distribution of the AMI score serves as a means to discriminate the surrogate data from the original data set (for example Figure 3.26, page 72).
- The IAAFT surrogate data generator cannot be used as a synthetic seismic generator (Section 3.4.2, page 69) because of the power of the average mutual information and the correlation dimension to discriminate between T4 and its surrogate sets.

In the nonlinear modelling of T4 to its future events the following were observed:

- 
- LSTM succeeded in providing the models that out performed the autoregressive modelling (Section 4.5, page 91).
 - A large bias term helped in the modeling of T4. Initial tests on fitting an LSTM model to T4 struggled to converge to a decreasing error term.
 - The optimal fit for the T4 estimator from LSTM was not reached, an issue that can be addressed in future studies.
 - Constructing a mean estimator from a number of LSTM fits gave a model for T4 over the whole domain of its variables with sufficient power to model the future unfolding of T4 better than the estimated autocorrelation function.

The hypothesis test can be developed and applied to the study of seismicity in general since T4 behaves in a similar manner to seismicity reported in the literature (Section 3.3.1, page 51).

This study suggests that seismic data like T4 can be used to improve current predictions on seismic hazards in mines.

Appendix A

The rule derivations for the LSTM network

A.1 LSTM forward pass

Let $DS = (\bar{X}(t), \bar{D}(t)), t \in \{i : \mathcal{N} \mid i = 1 \dots n\}$ be a data set consisting of a sequence of paired values. Long Short-Term Memory both an iterative optimization algorithm and a parameterized map,

$$\bar{X}(t) \xrightarrow{LSTM} \bar{Y}(t)$$

The parameters of the map are updated at each iteration of the optimization algorithm in the general direction that minimizes the individual error terms $\bar{e}(t) = \frac{1}{2} \sum_{i=1}^O (d_i(t) - y_i(t))^2$. The iterated terms $d_i(t)$ and $y_i(t)$ are the components of $\bar{D}(t)$ and $\bar{Y}(t)$, respectively, O being the size of the output dimension. The iterator of the error terms,

$$t \in \{i : 0 \dots n \mid \bar{e}(i) \text{ is an error term} \bullet i\} = subDS$$

can iterate over any subset of indices to the data set.

The LSTM map consists of two maps. The first assimilates input vectors into an internal storage vector

$$(\bar{H}(t-1), \bar{X}(t)) \xrightarrow{LSTM_H} (\bar{H}(t))$$

and the second maps components of the interval storage vector onto the estimated vector outputs

$$((\bar{H}(t)) \xrightarrow{LSTM_O} (\bar{Y}(t))$$

Input vectors are assimilated into the hidden vector because the first map uses the current input vector in combination with components of the previous hidden vector to compute the current hidden vector. The second map allows LSTM to map output vectors disjunct from the input vectors.

The two LSTM maps are described in terms of the artificial neural networks (ANN) literature. A map consists of a weighed, directional, connected network of smaller units. Each unit is stimulated with an activation value to which the unit responds, as it sends or propagates a signal to the units it is connected to, in the proper direction of connection. The input to the network is viewed as signals propagated into the network from input units and output from the network of units, viewed as signals propagated to output units. The two LSTM networks can then be combined to form a single LSTM network of units and weighted connections propagating signals. In this implementation of LSTM the activation value of a unit is a linear combination of the signals and the weighted connections. The activation function mapping the activation value to the response signal of the unit is mostly a logistic sigmoidal function. One deviation in LSTM from the network of connections of other ANNs is multiplicative connections between some of the response signals of a normal unit and the activation value of the multiplicative unit. In the LSTM architecture these are referred to as multiplicative gates since they scale the size of the activation value flowing into a unit.

The minimum unit of computation in the $LSTM_H$ map is known as an LSTM memory block. A memory block is a conglomerate of units, each performing a specialized function and connected to each other in an unorthodox manner, compared to normal ANNs. The LSTM architecture allows for a variable amount of memory blocks. A memory block consists of an array of memory cells and three multiplicative gates. Each of the cells and the multiplicative gates has an activation value and an associated response value at that time step. Each memory cell of a memory block consist of a self connected linear unit known as the “Constant Error Carousel” (CEC), an input squashing unit, and an output squashing unit. A cell’s input unit maps its activation value with function $g(\cdot)$. The response to the cell’s activation value is scaled with the first of the *block’s* multiplicative gates. The activation value of the CEC is computed by adding the scaled response of the input squashing unit to the previous response of the CEC, scaled with the second of the *block’s* multiplicative units. The CEC responds with the same value as its activation value. The CEC’s response is the activation value of the output squashing function. The response of

the output squashing function $h(\cdot)$ is scaled with the third of the *block's* multiplicative gates, resulting in the cells response value. The response from a memory block is a vector consisting of the response values of the *block's* cells. The values used in the hidden vector to combine with the input vector to assimilate into the internal memory of the network are the response values of the cells and the units responsible for the multiplicative gates. The internal memory of the network consists of the hidden vector and the response values of the CECs. All the memory blocks in this implementation of LSTM have the same amount of cells.

Let

- $In = \{k, K : \mathcal{N} \mid k = 1 \dots K \text{ and } k \text{ is an index to an input component } \bullet k\}$ be the set of indices of the elements in the input vector;
- $Gates = \{in, scale, out\}$ be the set of indices to the multiplicative gates in the LSTM architecture;
- Let $Blocks = \{b, B : \mathcal{N} \mid b = 1 \dots B \text{ and } b \text{ is an index to a memory block } \bullet b\}$ be the set of indices to the memory blocks in die LSTM architecture;
- Let $Cells = \{c, C : \mathcal{N} \mid c = 1 \dots C \text{ and } c \text{ is an index to a cell in a memory block } \bullet c\}$ be the set of indices to the cells of each of the memory blocks in the LSTM architecture. In cases where two of the same elements of the natural numbers \mathcal{N} refer to different components in the same set and lead to an ambiguity, a distinction will be made.

Then $c \in BC = Cells \times Blocks$ is the set of indices to each cell of each memory block in the LSTM architecture and $G = Gates \times Blocks$ is the set of indices to each multiplicative gate for all blocks in the LSTM architecture. This defines the index $h \in Hide = \{Cells \times Blocks \cup Gates \times Blocks\}$ as the set of indices for all the hidden units of the LSTM architecture. It follows that $m \in LSTM = Out \cup Cells \times Blocks \cup Gates \times Blocks \cup In$ refers to an index to any of the components in the LSTM architecture. A weighted connection from unit $l \in LSTM$ to unit $m \in LSTM$ is indicated by W_{ml} , the activation value of unit m at input value $t \in DS$, is indicated by $net_m(t)$ and the response as $Y_m(t) = Y_m(net_m(t))$. A correction has to be added to the usage of the indices in *Hide*. The cells of the memory blocks are denoted by BC . The usage of this index depends on the unit it refers to. As source it refers to the response of the cells of the memory blocks.

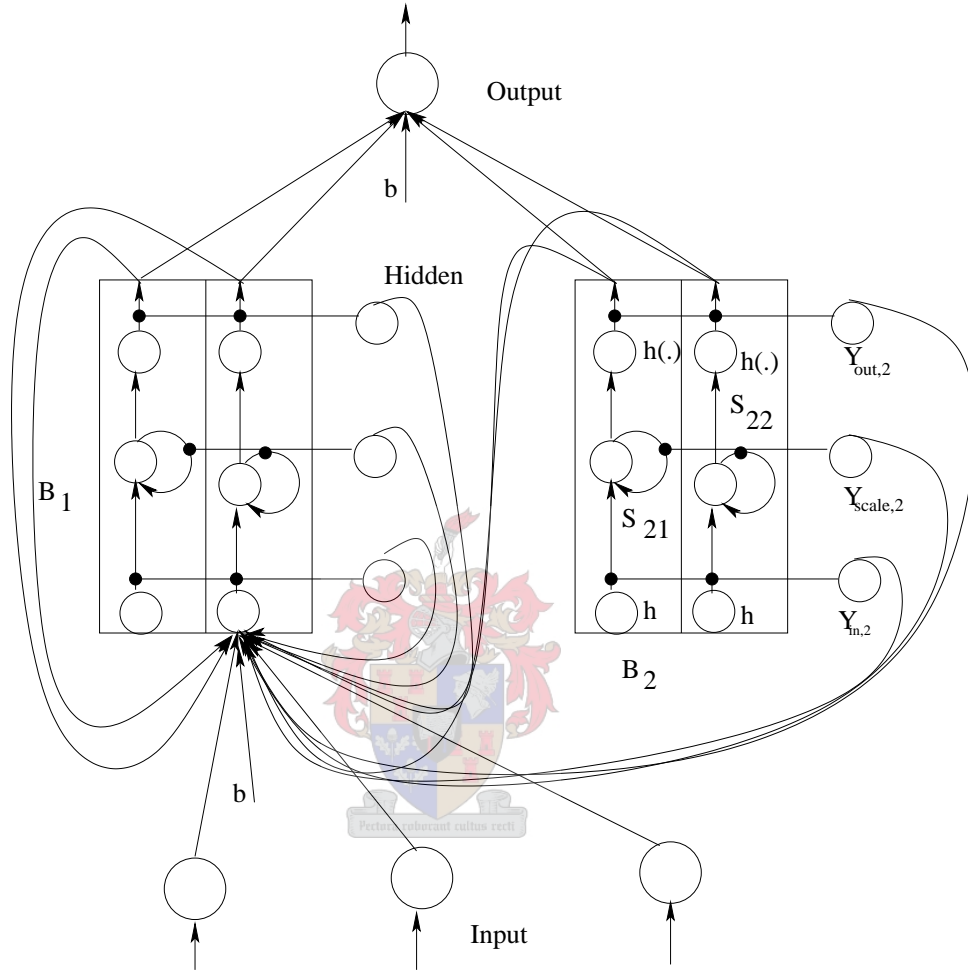


Figure A.1: A graphical representation of an LSTM network with 3 input nodes, two memory blocks, B_1 and B_2 , with two CECs each and one output node. The output layer is connected like the architectures used in our study: no gate node to output node connections. The hidden layer demonstrates only the connections to one node, including the bias connection, b . For B_2 the input and output squashing functions, $g(\cdot)$ and $h(\cdot)$, are indicated as well as the multiplicative scaling gates: $Y_{(input,2)}$, cell input gate; $Y_{(scale,2)}$, cell forget gate; $Y_{(output,2)}$, cell output gate. The CECs are indicated with S_{21} and S_{22} .

As a destination it refers to the input squashing unit for each of the cells. The combination does not result in an ambiguity in this implementation of the LSTM map since no unit requires both indices as a source and a destination at the same

time.

The $LSTM_H$ map is then initiated by the activation values for the hidden units:

$$net_m(t) = \sum_{i \in I} W_{m,i} * X_i(t) + \sum_{h \in Hide} W_{m,h} * Y_h(t-1) \quad (A.1)$$

$m \in Hide.$

The $LSTM_H$ map is then completed by the update of the CEC's and,

$$S_c(t) = S_c(t-1) * Y_{scale \times b}(net_{scale \times b}(t)) + g(net_c(t)) * Y_{in \times b}(t) \quad (A.2)$$

$c \in BC, b \in Blocks$

the computation of the response for each cell for each blocks.

$$Y_c(t) = Y_c(net_c(t)) \text{ with} \quad (A.3)$$

$$net_c(t) = h(S_c(t)) * Y_{out \times b}(net_{out \times b}(t)), c \in BC, b \in Blocks$$

The $LSTM_O$ map is then computed from the response signals from the array of cell and gate units of the blocks:

$$Y_k(t) = Y_k(net_k(t)) \text{ with} \quad (A.4)$$

$$net_k(t) = \sum_{h \in Hidd} W_{k,h} * Y_h(t), k \in Out$$

The LSTM architecture discussed in the thesis and the map discussed above is not a completely connected network of units. The connection of W_{mn} , $m \in Out$, $n \in In$, is not a part of the LSTM map as defined above, although the LSTM architecture allows for such connections in principle.

A.2 LSTM backward pass

The LSTM map from input to output vectors is fitted to the desired output vectors by iteratively changing the weights of the connections between the units of the map.

The optimization algorithm for the parameters in this implementation uses a gradient descent adaption of the back propagation algorithm for the special network architecture. The direction of the gradient is provided by the partial first derivative of the error distributed over the linear map of the activation value for a unit. The first derivative error information is distributed throughout the LSTM map across all units and the associated weight change computed.

A fully connected LSTM architecture includes two recurrent connects. The first is in the hidden vector used as input in combination with the input vectors. The second is in the self connected units of the CECs. The hidden units of memory blocks follow the same approach of first derivative gradient descent as the truncated Back Propagation Through Time used in an Elman ANN (Zurada, 1992). Error signals flowing out of the memory block are truncated after they have been used in computing the change in weights immediately connected to the memory block. No error signal flows back into a memory block once it has left any of the other memory blocks. The same is not true for the CECs, as the error signal in the self connected unit is allowed to propagate back in time as many steps as it has been iterated itself. The partial derivative information relating a change in response value in the CEC to the weights responsible for its inputs is maintained in a manner similar to Real Time Recurrent Learning (RTRL) while the forward pass of signaling is conducted (Haykin, 1999).

The implementation of LSTM used in this study is not a completely connected network. Input signals are connected only to the indices of the hidden units. The response of the hidden units are connected to the full range of hidden units. None of the multiplicative gate units are connected to the output units, only the array of cells of the memory blocks are connected to the outputs units.

In the following $\Delta W_{mn}(t)$ $m, n \in LSTM$ refers to the change in the weighted connection connecting unit n with unit m as a result of the error signal generated at the input-output pair of t . The derivative of a function $f(\cdot)$ is denoted $\dot{f}(\cdot)$. It is important to note which activation of the hidden vector is responsible for a weight change at the presentation of error signal t . Signals indexed with a $t - 1$ are either

the activation value of a previous layer or from a previously presented vector.

The weight updates of this implementation of the LSTM architecture differs slightly from the weight updates used in the literature in that a momentum term is added to the weight updates. The momentum term is a parameter added to the training process with a value of 0 for the LSTM maps in the technical reports,

$$\begin{aligned} \Delta W_{mn}(t) &= \beta \Delta W_{mn}(t-1) + \alpha \delta_m(t) Y_m(t), \\ m, n &\in LSTM, \alpha; \beta \in \Re \end{aligned} \quad (\text{A.5})$$

In the above general weight update rule for the back propagation error gradient descent type of optimization $\delta_m(t)$ provides the direction for that specifically weighted connection of the local gradient to minimize the loss function for output vector number t . The value of $\delta_m(t)$ is computed with the back propagation of the error signal according to the partial derivative chain rule. The rules computing the value of $\Delta W_{mn}(t)$ $m, n \in LSTM$ in terms of the indices for the units of the LSTM map are presented below. For a full analysis of the partial derivatives upon which these rules are based see Hochreiter and Schmidhuber (1991) or Gers et al. (1999).

During successive iterations of the LSTM map, the partial rate of change for a given CEC in terms of the weights responsible for its updated values are maintained, as with RTRL. Let the rate of change of every CEC in terms of a change in weight affecting the value of the CEC be denoted by:

$$DS_{bm}^c(t) \equiv \frac{\partial S_c(t)}{\partial W_{bm}} \text{ with} \\ m \in In \cup Hide, b \in Hide \text{ and } c \in BC \quad (\text{A.6})$$

This results in a table of $DS_{**}^*(t)$ indexed by the source and location indices defined above. The table is initialized as $DS_{**}^*(0) = 0$ for all elements. All the entries in the table not updated are fixed at zero. The rules to maintain the definition of the table during successive iterations of the LSTM map, using the above defined indices are:

for $b \in \{in\} \times Blocks \subset G$ (A.7)

$$DS_{bm}^c(t) = DS_{bm}^c(t-1)Y_b(t) + g(net_c(t))\dot{Y}_b(net_b(t))Y_m(t-1), m \in Hidd$$

$$DS_{bm}^c(t) = DS_{bm}^c(t-1)Y_b(t) + g(net_c(t))\dot{Y}_b(net_b(t))Y_m(t), m \in In$$

for $b \in \{scale\} \times Blocks \subset G$ (A.8)

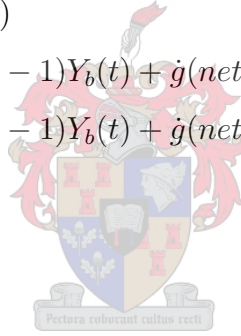
$$DS_{bm}^c(t) = DS_{bm}^c(t-1)Y_b(t) + s_c(t-1)\dot{Y}_b(net_b(t))Y_m(t-1), m \in Hidd$$

$$DS_{bm}^c(t) = DS_{bm}^c(t-1)Y_b(t) + s_c(t-1)\dot{Y}_b(net_b(t))Y_m(t), m \in In$$

for $b, c \in BC, b = (i, j)$ (A.9)

$$DS_{bm}^c(t) = DS_{bm}^c(t-1)Y_b(t) + \dot{g}(net_c(t))Y_{in \times j}(t)Y_m(t-1), m \in Hidd$$

$$DS_{bm}^c(t) = DS_{bm}^c(t-1)Y_b(t) + \dot{g}(net_c(t))Y_{in \times j}(t)Y_m(t), m \in In$$



For an error signal,

$$e_k = d_k(t) - y_k(t), k \in In$$

propagating into the network, the first step is to compute the local gradient δ_L , $L \in LSTM$ for the reduction of the error signal. The second step is to compute the change in weights for that local gradient in terms of the signal propagated over the connection or representative value of past signals.

The local gradients, in terms of e_k are computed as:

$$\text{for } k \in In \quad (A.10)$$

$$\delta_k(t) = \dot{Y}_k(net_k(t))e_k(t)$$

$$\text{for } outG = (out, b) \in \{out \times Blocks\} \subset G \quad (A.11)$$

$$\delta_{outG}(t) = \dot{Y}_{outG}(net_{outG}(t)) \left[\sum_{c \in Cells} h(S_{(c,b)}(t)) \sum_{k \in Out} w_{k(c,b)} \delta_k(t) \right]$$

$$\text{for } c = (i, b) \in BC \quad (A.12)$$

$$e_c(t) = Y_{(out,b)}(t) \dot{h}(S_c(t)) \left[\sum_{k \in Out} w_{kc} \delta_k(t) \right]$$

The associated weight changes are then, for $m \in LSTM$:

$$\text{for } k \in In \cup out \times Blocks \quad (A.13)$$

$$\Delta w_{km}(t) = \alpha \delta_k(t) Y_m(t)$$

$$\text{for } g = (i, b) \in \{out; scale\} \times Blocks \subset G \quad (A.14)$$

$$\Delta W_{bm}(t) = \alpha \sum_{c \in Cells} e_{(c,b)}(t) DS_{gm}^{(b,c)}$$

$$\text{for } c = (j, b) \in BC \quad (A.15)$$

$$\Delta W_{cm}(t) = \alpha e_c(t) DS_{cm}^c(t)$$

A.3 Functional LSTM correctness

“There is always one more bug...” is commonly known as the programmers’ rule of anthropology. The implementation of LSTM is no exception. The code was debugged in various ways, but unfortunately no guarantees can be provided that every bug in the implementation was found. In the literature LSTM’s functional correctness is demonstrated empirically by benchmark problems. A bug in an empirical model poses a problem only if it stops the network from doing what it is supposed to do. Benchmark problem solving does not provide an absolute reference for correctness.

The first step in forming a correct implementation was in picking an easy and accessible design for the implementation. LSTM was implemented as an abstract data type in a procedural subset of C++ (Stroustrup, 1997) using **for** loops and arrays. The C++ implementation is sufficient for research purposes, while not over specifying the structure of the implementation and creating unnecessary complexity. Weights and internal states were grouped together in such a way that the mapping in the network can be performed with as little iterations over indices as possible, forcing errors to occur in groups. During implementation, text output of computed values was used to detect obvious errors. After the C++ implementation was finished, a specific LSTM architecture was implemented on a spreadsheet and the results compared to the C++ implementation. Benchmark problems were used from the continual LSTM prediction paper (Gers et al., 1999) as well as the application to the Santa Fé laser data set (Gers et al., 2001). Implementing the LSTM algorithm would have been quicker if a constructed solution for a weight update iteration was available, similar to the one used in the spreadsheet.

The spreadsheet comparison would not have been viable without the implementation adhering to some variance reduction techniques, as proposed by Law and Kelton (1991). The idea is to fix the sequence of random numbers used in a specific initialization of LSTM as a seed to the random number generator. A fixed random seed will remove any stochastic component for a specific instance of a network. This type of random seed implementation and subsequent instance-variance control was found lacking in other implementations of numerical algorithms used in the thesis.

Appendix B

Algorithm listing

B.1 Code for ICA transforms

The Matlab code for transforming the embedded vectors, \mathbf{X} , into a vector space of the d number of independent components is given. The `ICAm` function is an interface to Hugo Gävert's implementation of the fixed point ICA approximation algorithm. His initial implementation combines a number of IC estimation scenarios. Only the code representing the method of analysis used in this thesis is listed. Even though this code was not specifically implemented in this thesis, it is included as a specification of the IC approximation mechanism. Only portions of the ICA implementation actually used for the thesis are presented.

```
function [vectors , W] = ...
ICAm(X, d, ica_count, parms, conv_step, pca_pers)
    %X      - The embedded vectors as rows.
    %d      - The number of ICs to compute.
    %ica_count- If the convergence takes more steps than this, stop
    %parms   - variance parameter to the Gaussian kernel
    %conv_step - the step size during convergence
    %pca_pers - total percentage of variance of in the first PCs.

    [ ap, mp, stdp ] = auto(X);
    [ E, Di ]        = pcamat(ap');

    %calculate the number of PCs

    pca_pers        = 0.99;
    dim              = 1;
    while sum(diag(Di((end-(dim-1)):end, (end-(dim-1)):end)))...
    ./sum(diag(Di))) < 0.99,
        dim = dim+1;
```

```

end

%cut down the dimension and handle a bit of noise
E = E(:,end-(dim-1):end);
Di = Di((end-(dim-1)):end, (end-(dim-1)):end);

%balance the input variables to remove
%spurious correlations, ie scale not rotate
[ v, wm, dwm ] = whitenv(ap', E, Di);
%infer an inverse to a proposed mixing matrix
[ A, W ] = fpica(v, wm, dwm, 'symm', d, 'gaus', ...
'gaus', 0, parms, conv_step, ica_count);
%deMix the proposedly mixed signals
vectors = (W*ap')';

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [newVectors, whiteningMatrix, dewhiteningMatrix]...
= whitenv(vectors, E, D);
%Whitens the data (row vectors).
%Returns the whitened vectors (row vectors),
%whitening and dewhitening matrices.
%
% vectors Data in row vectors.
% E Eigenvector matrix from function 'pcamat'
% D Diagonal eigenvalue matrix from function 'pcamat'
% 24.8.1998
% Hugo äGvert
%
% =====
% Calculate the whitening and dewhitening matrices
whiteningMatrix = inv(sqrt(D)) * E';
dewhiteningMatrix = E * sqrt(D);

% Project to the eigenvectors of the covariance matrix.
% Whiten the samples and reduce dimension simultaneously.
newVectors = whiteningMatrix * vectors;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function [A, W] = fpica(X, whiteningMatrix, dewhiteningMatrix,
approach, numOfIC, g, finetune, a1, a2, myy, maxNumIterations);
%FPICA - Fixed point ICA. Main algorithm of FASTICA.
%
% Perform independent component analysis using
% Hyvarinen's fixed point algorithm. Outputs
% an estimate of the mixing matrix A and its inverse W.
% Follow the symmetrical method, converge to all the ICs
% at once.
%
% X

```

```

%:the whitened data as row vectors
% whiteningMatrix
%:can be obtained with function whitenv
% dewhiteningMatrix
%:can be obtained with function whitenv
% numOfIC [ 0 - Dim of whitesig ]
%:number of independent components estimated
% g      [ 'gaus' ]
%:the nonlinearity used
% finetune[ 'gaus' ]
%:the nonlinearity used in finetuning.
% a1     [0       ]      :unused
% a2
%:parameter for tuning 'gaus'
% mu
%:step size in stabilized algorithm
% maxNumIterations
%:maximum number of iterations
%
% 23.8.1999
% Hugo äGvert

% Default values
[vectorSize , numSamples] = size(X);
%stopping criterion
epsilon      = 0.0001;
myyOrig      = myy;
% When we start fine-tuning we'll set myy = myyK * myy
myyK         = 0.01;
%the nonlinearity used
usedNlinearity = 30;
finetuningEnabled = 1;
%the nonlinearity used in fine tuning
gFine        = 30 + 1;
%Watch and guide the convergence process
stabilizationEnabled = 1;
stroke       = 0;
notFine     = 1;
long        = 0;

%Dewhitened basis vectors.
A = zeros(vectorSize , numOfIC);
%Take random orthonormal initial vectors.
B = orth(rand(vectorSize , numOfIC) - .5);

Bold = zeros(size(B));
Bold2 = zeros(size(B));

%This is the actual fixed-point iteration loop.
for round = 1:maxNumIterations + 1,

```

```

if round == maxNumIterations + 1,
    A=[];
    W=[];
    return;
end
%Symmetric orthogonalization.
B = B * real(inv(B' * B)^(1/2));

%Test for termination condition. Note that we consider opposite
%directions here as well.
minAbsCos = min(abs(diag(B' * BOld)));
minAbsCos2 = min(abs(diag(B' * BOld2)));

if (1 - minAbsCos < epsilon)
    %Has convergence been reached?
    if finetuningEnabled & notFine
        fprintf('Initial_convergence ,_fine-tuning:_\n');
        notFine = 0;
        usedNlinearity = gFine;
        myy = myyK * myyOrig;
        BOld = zeros(size(B));
        BOld2 = zeros(size(B));
    else
        fprintf('Convergence_after_%d_steps\n', round);
        %Calculate the de-whitened vectors.
        A = dewhiteningMatrix * B;
        break
    end
elseif %stabilization Enabled
    %If convergence has not been reached, maybe some adjustments?
    if (~stroke) & (1 - minAbsCos2 < epsilon)
        %convergence in one direction and not in the other.
        fprintf('Stroke!\n');
        stroke = myy; => ~stroke == false
        myy = .5*myy;
        if mod(usedNlinearity,2) == 0
            usedNlinearity = usedNlinearity + 1;
        end
    elseif stroke
        myy = stroke;
        stroke = 0;
        if (myy == 1) & (mod(usedNlinearity,2) ~= 0)
            usedNlinearity = usedNlinearity - 1;
        end
    elseif (~long) & (round>maxNumIterations/2)
        fprintf('Taking_long_(reducing_step_size)\n');
        long = 1;
        myy = .5*myy;
        if mod(usedNlinearity,2) == 0
            usedNlinearity = usedNlinearity + 1;

```

```

        end
    end
end

BOld2 = BOld;
BOld = B;

switch usedNlinearity
    %gaussian kernel, entropy score.
    case 30
        U = X' * B;
        Usquared=U .^ 2;
        ex = exp(-a2 * Usquared / 2);
        gauss = U .* ex;
        dGauss = (1 - a2 * Usquared) .* ex;
        B = X * gauss / numSamples - ...
            ones(size(B,1),1) * sum(dGauss).* B / numSamples;
    case 31
        %gaussian kernel, refinement step.
        Y = X' * B;
        ex = exp(-a2 * (Y .^ 2) / 2);
        gauss = Y .* ex;
        Beta = sum(Y .* gauss);
        D = diag(1 ./ (Beta - sum((1 - a2 * (Y .^ 2)) .* ex)));
        B = B + myy * B * (Y' * gauss - diag(Beta)) * D;
    end
end
%Calculate ICA filters.
W = B' * whiteningMatrix;
return

```

B.2 Code for the surrogate computation

The Matlab code for generating an iaaft multivariate surrogate, through function `iaaftn` is given. See the code listings for details about the inputs and outputs.

```

function [r, s_bar, shift, cnt] = iaaftn( s, no_adjust, maxi, r )
%[r, s_bar, shift, cnt] = iaaft( s, no_adjust, maxi, r )
%Generate a IAAFT surrogate of s.
%Default :
% [r, s_bar, shift, cnt] = iaaftn( s, 0, 1000, [] )
%Inputs :
% s          nx1 - sampled data
% no_adjust  1x1 - [0|1] to shorten s to a small enough jump.
% maxi      1x1 - The maximum number of iterations
% r         nx1 - The surrogate at time step 0
%Outputs:

```



```

% r          nx1 - The surrogate with an exact distribution
% s_bar      nx1 - The surrogate with exact amplitudes
% shift.start 1x1 - The startoff point for the shortend s
% shift.slip  1x1 - The portion of the variance in s due
%                  to a jump between s1 and sn
% shift.jump  1x1 - The same as slip but for the first derivative.
%
%Reference:
% Surrogate time series
% T. Schreiber and A. Schmitz
% Physics Department, University of Wuppertal, D-42097, Germany

%Set up the reference amplitudes, rank order and inverse rank map
s_fft          = fftn(s);
s_amplitudes   = abs(s_fft);
s_rho          = angle(s_fft);
I              = 1:length(s);
s_rank_order   = cell(size(s,2),1);

sI             = cell(size(s,2),1);

for i = 1:size(s,2)
    [s_rank_order{i}, sI{i}] = sort(s(:,i));
end

if nargin < 2, maxi = 1000; end %Maximum iterations
if nargin < 3 | isempty(r), %start off surrogate
    r = zeros(size(s));
    for i = 1:size(r,2)
        r(:,i) = s(randperm(length(s)),i);
    end
end

% The first iteration, While loop repeats on the step failure
%Step one: "A Crude Fourier Filter"
cnt = 1;
%extract the angles for the surrogate
R    = angle(fftn(r));
%adjust the angles for the surrogate cross correlations
R    = adjust_phases( R, s_rho );
%a surrogate with the correct angles
s_bar = (ifftn(s_amplitudes.* exp(R .* j )));

%Step two: Adjust the distribution
done = 1;
for i = 1:size(s,2)
    s_bar(:,i) = abs(s_bar(:,i)).* sign(real(s_bar(:,i)));
%The rank ordering of the surrogate
    [temp, s_barI] = sort(s_bar(:,i));

```

```

%The back map from the rank to the timeseries ordering
    s_barI(s_barI) = I;
% Adjust the distribution of the surrogat with the correct angles
    r(:,i)         = s_rank_order{i}(s_barI);
    done           = done & all(sI{i}==s_barI);
    sI{i}          = s_barI;
end

while ~done & (cnt < maxi)
    % loop iterator
    cnt = cnt+1;

    %Step one
    %extract the angles for the surrogate
    R    = angle(fftn(r));
    R    = adjust_phases( R, s_rho );
    %a surrogate with the correct angles
    s_bar = (ifftn( s_amplitudes.*exp(R .* j )));

    %Step two: Adjust the distribution
    done = 1;
    for i = 1:size(s,2)
        s_bar(:,i) = abs(s_bar(:,i))...
            .* sign(real(s_bar(:,i)));
%The rank ordering of the surrogate
        [temp, s_barI] = sort(s_bar(:,i));
%The back map from the rank to the timeseries ordering
        s_barI(s_barI) = I;
    %Adjust the distribution of the surrogat with the correct angles
        r(:,i)         = s_rank_order{i}(s_barI);
        done           = done & all(sI{i}==s_barI);
        sI{i}          = s_barI;
    end
end

end

return

function [ phi ] = adjust_phases(gamma, rho)

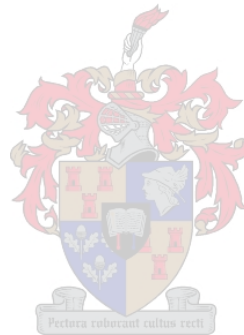
    angles = [0 pi/2 pi 3*pi/2];
    M      = size( gamma, 2 );
    alpha  = ...
        abs(atan(sum(sin(gamma-rho),2)./sum(cos(gamma-rho),2)));

    %compute the optimal alpha_k 's.
    for k = 1 : length( alpha )
        best_a = 0;
        max_score = 0;
        a_shift = 0;

```

```
    for a = angles
        a_shift = (a + alpha( k ));
        score = sum( cos( a_shift( ones(1,M) - gamma(k,:) + rho(k,:) ), 2);
        if score > max_score
            best_a = a_shift;
            max_score = score;
        end
    end
    alpha(k) = best_a;
end

phi = rho + repmat(alpha, 1, M);
return
```



References

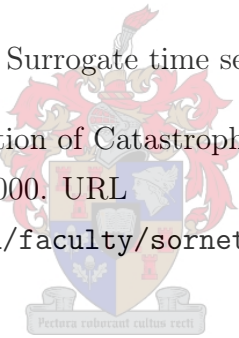
- R. Abercrombie, D. Agnew, and F. Wyatt. Testing a model of earthquake nucleation. *Bulletin of the Seismological Society of America*, 85:1873–1878, 1995.
- R. Abercrombie and J. Mori. Local observations of the onset of a large earthquake: 28 June 1992 Landers, California. *Bulletin of the Seismological Society of America*, 84:725–734, 1994.
- M. Abinante and L. Knopoff. Quasidynamic model for earthquake simulations. *Physical Review E*, 52:5676–5678, 1995.
- K. Aki. A probabilistic synthesis of precursory phenomena. Washington, D.C., 1981. ISBN 0-87590-403-3.
- A. Akkaya and M. Yüçemen. Estimation of earthquake hazard based on extremes of local integral random functions. *Engineering Geology*, 58:53–66, 2000.
- A. Akkaya and M. Yüçemen. Stochastic modeling of earthquake occurrences and estimation of seismic hazard: a random field approach. *Probabilistic Engineering Mechanics*, 17:1–13, 2002.
- R. Bakker, J. Schouten, C. Giles, F. Takens, and C. Van den Bleek. Learning chaotic attractors by neural networks. *Neural Computation*, 12:2355–2383, 2000.
- B. Barriere and D. Turcotte. Seismicity and self-organized criticality. *Physical Review E*, 49:1151–1160, 1994.
- D. Beck and B. Brady. Evaluation and application of controlling parameters for seismic events in hard-rock mines. *International Journal of Rock Mechanics and Mining Sciences*, 39:633–642, 2002.
- Y. Ben-Zion. Stress, slip, and earthquakes in models of complex single-fault systems incorporating brittle and creep deformations. *Journal of Geophysical Research*, 101:5677–5706, 1996.

- E. Blanter and M. Shnirman. Self-organized criticality in a hierarchical model of defects development. *Physical Review E*, 53:3408–3413, 1996.
- B. Bodri. A neural-network model for earthquake occurrence. *Journal of Geodynamics*, 32:289–310, 2001.
- D. Brillinger. Earthquakes, statistics of. New York, 1997. ISBN 0-471-11836-2.
- J. N. Brune. Tectonic stress and the spectra of seismic shear waves. *Journal of Geophysical Research*, 75:4997–5009, 1990. Correction, *J. Geophys. Res.*, 76, 5002, 1972.
- M. Cai, P. Kaiser, and C. Martin. Quantification of rock mass damage in underground excavations from microseismic event monitoring. *International Journal of Rock Mechanics and Mining Sciences*, 38:1135–1145, 2001.
- R. Castro and T. Sauer. Reconstructing chaotic dynamics through spike filters. *Physical Review E*, 59:2911–2917, 1999.
- D. Colombo, V. Gitis, and R. de Franco. Application of pattern recognition techniques to long-term earthquake prediction in central Costa Rica. *Engineering Geology*, 48:7–18, 1997.
- G. Darpahi-Noubary. The use of modern statistical theories in the assessment of earthquake hazard, with application to quiet regions of eastern North America. *Soil Dynamics and Earthquake Engineering*, 22:361–369, 2002.
- E. Dudewicz and S. Mishra. *Modern Mathematical Statistics*. John Wiley and Sons, Inc., Singapore, first edition, 1988. ISBN 0-471-60716-9.
- M. Eneva. Monofractal or multifractal: a case study of spatial distribution of mining induced seismic activity. *Nonlinear Processes in Geophysics*, 1:182–190, 1994.
- M. Eneva. In search for a relationship between induced microseismicity and larger events in mines. *Tectonophysics*, 289:91–104, 1998.
- M. Eneva and Y. Ben-Zion. Application of pattern recognition techniques to earthquake catalogs generated by model segmented fault systems in three-dimensional elastic solids. *Journal of Geophysical Research*, 102: 24513–24528, 1997a.

- M. Eneva and Y. Ben-Zion. Techniques and parameters to analyse seismicity patterns associated with large earthquakes. *Journal of Geophysical Research*, 102:17785–17795, 1997b.
- J. Fan and I. Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London, first edition, 1996. ISBN 0-412-98321-4.
- R. Feynman, R. Leighton, and M. Sands. *The Feynman Lectures on Physics: Commemorative Issue*, volume II. Addison-Wesley Publishing Company, Reading, Massachusetts New York, unknown edition, 1989. ISBN 0-201-51004.
- A. Fraser and H. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33:1134–1140, 1986.
- Y. Fujii, Y. Ishijima, and G. Deguchi. Prediction of coal face rockburst and microseismicity in deep longwall coal mining. *International Journal of Rock Mechanics and Mining Sciences*, 34:85–96, 1997.
- A. Gabriellov, V. Keilis-Borok, I. Zaliapin, and W. Newman. Critical transitions in colliding cascades. *Physical Review E*, 62:237–249, 2000.
- R. Geller, D. Jackson, Y. Kagan, and F. Mulargia. Earthquakes cannot be predicted. *Science*, 275:1616–1617, 1997.
- J. Gere and H. Shan. *Terra non Firma: Understanding and Preparing for Earthquakes*. W.H. Freeman and Company, New York, first edition, 1984. ISBN 0-7167-1497-3.
- F. Gers, D. Eck, and J. Schmidhuber. Applying LSTM to time series predictable through time-window approaches. Technical report, IDSIA, Switzerland, 2001.
- F. Gers and J. Schmidhuber. Long short-term memory learns context free and context sensitive languages. Technical report, IDSIA, Switzerland, 2001.
- F. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: continual prediction with LSTM. Technical report, IDSIA, Lugano, January 1999.
- R. Habermann. Precursory seismicity patterns: stalking the mature seismic gap. Washington, D.C., 1981. ISBN 0-87590-403-3.
- D. Hanson, T. Vandergrift, M. DeMarco, and K. Hanna. Advanced techniques in site characterization and mining hazard detection for the underground coal industry. *International Journal Coal Geology*, 50:275–305, 2002.

- S. Haykin. *Neural Networks a Comprehensive Foundation*. Prentice Hall, Upper Saddle River, second edition, 1999. ISBN 0-13-273350-1.
- A. Helmstetter and D. Sornette. Diffusion of epicenters of earthquake aftershocks, Omori's law, and generalized continuous-time random walk models. *Physical Review E*, 66:061104-1-061104-24, 2002.
- S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735-1765, 1991.
- C. Hooge, S. Lovejoy, D. Schertzer, S. Pecknold, J. Malouin, and F. Schmitt. Multifractal phase transitions: the origin of self-organized criticality in earthquakes. *Nonlinear Processes in Geophysics*, 1:191-197, 1994.
- A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *To appear in IEEE Trans. on Neural Networks.*, unknown:unknown, 1999.
- A. Hyvärinen. The fastica package for matlab, September 2003. URL <http://www.cis.hut.fi/projects/ica/fastica/index.shtml>. [cited 2004, 31 January].
- ISSI©. ISS International Limited, 2003. URL <http://www.issi.co.za>. [cited 2004, September 21].
- K. Judd. An improved estimator of dimension and more comments on providing confidence intervals. *Physica D*, 56:216-228, 1992.
- K. Judd. Estimating dimension from small samples. *Physica D*, 71:421-429, 1994.
- Y. Kagan and L. Knopoff. Statistical short-term earthquake prediction. *Science*, 236:1563-1567, 1987.
- Y. Kagan and D. Vere-Jones. Problems in the modelling and statistical analysis of earthquakes. New York, 1996. ISBN 0-387-94788-4.
- P. Kaiser and C. Tang. Numerical simulation of damage accumulation and seismic energy release during brittle rock failure- Part ii: Rib pillar collapse. *International Journal of Rock Mechanics and Mining Sciences*, 35:123-134, 1998.
- H. Kanamori. The nature of seismicity patterns before large earthquakes. Washington, D.C., 1981. ISBN 0-87590-403-3.

- H. Kantz and T. Schreiber. *Non-linear Time Series Analysis*. Cambridge University Press, Cambridge, first edition, 1997. ISBN 0-5217-55144-7.
- Y. Lai and L. David. Effective scaling regime for computing the correlation dimension from chaotic time series. *Physica D*, 115:1–18, 1998.
- A. Law and W. Kelton. *Simulation, Modelling and Analysis*. McGraw-Hill International Editions, Inc., Singapore, second edition, 1991. ISBN 0-07-100803-9.
- J. Li, Y. Chen, and H. Mi. $\frac{1}{f^\beta}$ temporal fluctuation: detecting scale-invariance properties of seismic activity in North China. *Chaos, Solitons and Fractals*, 14: 1487–1494, 2002.
- V. Mansurov. Prediction of rockbursts by analysis of induced seismicity data. *International Journal of Rock Mechanics and Mining Sciences*, 38:893–901, 2001.
- MATLAB[®]. Matlab function reference: fft. *Release 12.1*, Version 6.1.0.450, May 2001.
- A. McGarr. Energy budgets of mining-induced earthquakes and their interactions with nearby stopes. *International Journal of Rock Mechanics and Mining Sciences*, 37:437–443, 2000.
- A. Mendecki, J. Niewiadomski, S. Radu, M. Sciocatti, G. van Aswegen, and C. Funk. *Seismic Monitoring in Mines*. Chapman and Hall, London, first edition, 1997. ISBN 0-412-75300-6.
- W. Newman. Log-periodic behaviour of a hierarchical failure model with applications to precursory seismic activation. *Physical Review E*, 52:4827–4835, 1995.
- A. Pavlov, O. Sosnovtseva, E. Mosekilde, and V. Anitshchenko. Chaotic dynamics from interspike intervals. *Physical Review E*, 63:036205–1–036205–5, 2001.
- D. Pollard. Strain and stress: discussion. *Journal of Structural Geology*, 22: 1359–1367, 2000.
- R. Poplawski. Seismic parameters and rockburst hazard at Mt Charlotte mine. *International Journal of Rock Mechanics and Mining Sciences*, 34:1213–1228, 1997.

- W. Pytel. Rock mass–mine workings interaction model for Polish copper mine conditions. *International Journal of Rock Mechanics and Mining Sciences*, 40: 497–526, 2003.
- T. Rikitake. *Earthquake Prediction*. Elsevier Scientific Publishing Company, Amsterdam, first edition, 1976. ISBN 0-444-41373.
- S. Russell and P. Norvig. *Artificial Intelligence*. Prentice-Hall, Inc., Englewood Cliffs, first edition, 1995. ISBN 0-13-360124-2.
- T. Sauer. Embedology. *Journal of Statistical Physics*, 65:579–616, 1991.
- T. Sauer. Reconstruction of dynamical systems from interspike intervals. *Physical Review Letters*, 72:3811–3814, 1994.
- T. Schreiber and A. Schmitz. Improved surrogate data for nonlinearity tests. *Physical Review Letters*, 77:635–638, 1996.
- T. Schreiber and A. Schmitz. Surrogate time series. *Physica D*, 3-4:346–382, 2000.
- D. Sornette. Scientific Prediction of Catastrophies: A New Approach, March 2000. URL <http://www.ess.ucla.edu/faculty/sornette/catastrophies.asp#catastrophies>. [cited 2003, August 11].
- 
- The image is a faint watermark of a university crest, likely from the University of California, Los Angeles (UCLA), given the URL in the reference above. The crest features a shield with various symbols, topped with a crown and surrounded by a banner with the motto "Pectora cubant celsus recti".
- D. Sornette, C. Vanneste, and L. Knopoff. Statistical model of earthquake foreshocks. *Physical Review A*, 45:8351–8357, 1992.
- R. Stein. Earthquake conversations. *Scientific American*, 1:60–67, 2003.
- B. Stroustrup. *The C++ Programming Language*. Addison-Wesley, Massachusetts, third edition, 1997. ISBN 0-20-188954-4.
- F. Vieira, D. Diering, and R. Durrheim. Methods to mine the ultra-deep tabular gold-bearing reefs of the Witwatersrand basin, South Africa. Colorado, 2001. ISBN 0-87335-193-2.
- I. Vorobieva. Prediction of a subsequent large earthquake. *Physics of the Earth and Planetary Interiors*, 111:197–206, 1999.
- Eric Weisstein. Mathworld—a wolfram web resource, March 2001. URL <http://mathworld.wolfram.com/>. [cited 2004, August 24].

- Eric Weisstein. Scienceworld—a wolfram web resource, September 2004. URL <http://scienceworld.wolfram.com/physics/P-Wave.html>. [cited 2004, September 16].
- I. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, first edition, 1999. ISBN 1-55860-552-5.
- K. Yamashina. *Some empirical rules on foreshocks and earthquake prediction*. Washington, D.C., 1981. ISBN 0-87590-403-3.
- J. Zurada. *Artificial Neural Systems*. West Publishing Company, St. Paul, NW, first edition, 1992. ISBN 0-314-93391-3.

