Helsinki University of Technology Laboratory of Computational Engineering Publications
Teknillisen korkeakoulun Laskennallisen tekniikan laboratorion julkaisuja
Espoo 2005 REPORT B44

Statistical Inference and Random Network Simulation

Jani Lahtinen

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

Helsinki University of Technology Laboratory of Computational Engineering Publications
Teknillisen korkeakoulun Laskennallisen tekniikan laboratorion julkaisuja
Espoo 2005 REPORT B44

Statistical Inference and Random Network Simulation

Jani Lahtinen

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Department of Electrical and Communications Engineering, Helsinki University of Technology, for public examination and debate in Auditorium S4 at Helsinki University of Technology (Espoo, Finland) on the 28th of January, 2005, at 13.

Helsinki University of Technology
Department of Electrical and Communications Engineering
Laboratory of Computational Engineering

Teknillinen korkeakoulu
Sähkö- ja tietoliikennetekniikan osasto
Laskennallisen tekniikan laboratorio

Distribution:
Helsinki University of Technology
Laboratory of Computational Engineering
P. O. Box 9203
FIN-02015 HUT
FINLAND
Tel. +358-9-451 5324
Fax. +358-9-451 4830
http://www.lce.hut.fi

Online in PDF format: http://lib.hut.fi/Diss/2005/isbn9512274906/

E-mail: Jani.Lahtinen@hut.fi

ISBN 951-22-7489-2 (printed)
ISBN 951-22-7490-6 (PDF)
ISSN 1455-0474
paino
Espoo 2005

Abstract

The scope of this dissertation is twofold, in the sense that it deals on one hand with statistical inference and on the other hand with random graphs. Due to inherent randomness in both areas the scope can also be seen as onefold, which is further united methodologically by the attempt to build models of random processes involved and by simulating their behaviour.

The statistical part of the thesis follows the Bayesian theory of probability, and applies it to a fault diagnostic setting. This part also contains an exploration of metrics on probability distributions, in which the introduction of a new metric is one of the main contributions. This new metric is constructed from utilities of the samples instead of the more conventional entropy–based metrics. In Bayesian methods the simulation of samples from distributions is an integral part of the analysis. It also becomes the leading principle in the evaluation of the proposed metrics. This metric is shown to be useful in statistical inference in some cases where the probabilities are difficult to compute. The problem of uncomputable likelihoods is analysed also from the Bayesian perspective and two branches emerge: the kernel estimate and the indirect inference.

In the analysis of random graphs the attention is on the small–world property, requiring that any two sites in the network are joined by only a short path with a relatively small average number of connections per site. Again one of the main tools in analysing complex graphs is by simulation of random dynamics on the graphs. The first dynamic property that is analysed is the spreading phenomenon. Spreading means the number of unique sites a random walker on the graphs goes through. This number is shown to have transition points relative to the small–world control parameter. Apart from the spreading phenomenon the thesis also studies the self–organised criticality properties through the so called sandpile model on the one dimensional small–world networks. In this setting of self–organised criticality there are interesting behaviours that are absent in the standard 1–dimensional sandpile model. Both the spreading and the sandpile model are analysed with two forms of disorder: quenched and annealed. The quenched case corresponds to a simulation setting on an ensemble of random graphs, whereas in the case of annealed disorder the simulation is performed on a regular graph but the dynamics also allow random moves to other sites. The annealed form allows

simpler analytic tools to be used, but the quenched form corresponds more closely to natural systems. Even though these forms of disorder are different it is shown that the annealed systems can be made to behave in a qualitatively similar fashion as the quenched case.

Preface

This thesis for the degree of Doctor of Technology has been prepared in the Laboratory of Computational Engineering at the Helsinki University of Technology during 2000-2004.

The author wishes to thank the staff of the Laboratory of Computational Engineering for the support and atmosphere. The author is specifically grateful to Dr. Jorma Rissanen for his advice on information theory, and to Dr. Jukka Heikkonen, Dr. Aki Vehtari and prof. Kimmo Kaski for general guidance. Also thanks to the co–authors of some of the publications dealing with the random networks goes to professors Janos Kertész and Kimmo Kaski.

The work was partially funded by the Academy of Finland, project No. 1169043 (Finnish Centre of Excellence Programme 2000-2005) and the Finnish KAUTE foundation.

Jani Lahtinen

List of symbols

$\#(x, A)$	number of times x appears in the set A
θ	parameter of distributions
$\kappa(x)$	scaling function
μ_X	distribution of the random variable X
$C(x\|\|y)$	pairwise cost–function of x and y
$\mathbb{D}$	sample–space of random variables
$\mathsf{E}\{f(X)\}$	mean value of $f(X)$
$\mathcal{F}$	parameter space
$K(\theta\|\|x)$	pairwise cost–function of θ and x
$lcm(m, n)$	least common multiple of m and n
$M_{i,:}$	ith row of the matrix M
$M_{:,i}$	ith column of the matrix M
$N(s)$	distribution of the number of the avalanche of size s
$\mathsf{P}\{X = x\}$	probability of the event $X = x$
$P_{ij}(t)$	transition probability from i to j of a random walker
P_{tr}	traversal probability of an avalanche
$Q(t)$	average number of unique sites visited by a random walker in time t
$\mathbb{S}$	superset of infinite subsets of $\mathbb{D}$
$S(x_{1:m}\|\|y_{1:n})$	transformation discrepancy of $x_{1:m}$ and $y_{1:n}$
$\mathbf{W}$	transition matrix of a random walker
$X_{1:m}$	multi–set of IID random variables $\{X_i\}_{i=1}^m$
X, Y, Z	random variables on $\mathbb{D}$ are written with capital letters
$x_{1:m}$	multi–set of elements $\{x_i\}_{i=1}^m$
$\bar{x}_{1:m}$	sample mean of $x_{1:m}$
$\tilde{x}_{1:m}$	sample standard deviation of $x_{1:m}$
$x_{1:m}^{\alpha}$	set $x_{1:m}$ repated α times

Contents

Chapter 1

Introduction

This dissertation is composed of two parts, statistical methods and random graphs, which have factors in common, namely randomness and statistical models. The first half of this dissertation deals with statistical inference, which is essentially an inversion process: there is a set of random observations, and a unifying pattern is sought that fits these observations. This pattern is a statistical model that describes the probabilities of the events, already observed and the ones yet to be observed [28, 56]. The statistical models are of great utility in practise where no phenomenon is truly free from randomness. They become handy in signal processing [33], pattern recognition [77], and finance [16]. Within this framework Lahtinen and Lampinen have analysed a fault diagnostic system for identifying the status of devices based on counted events [48]. There it was shown how latent, unobservable, states of the system can be identified based on the observations that do not contain explicit information about them. This is discussed further in Chapter 2.

Choosing the model, or model selection, can be done in various ways [56]. One is the Bayesian inference, in which the calculus of probability is utilised to obtain a posterior probability for the models. Posterior meaning the probability of the models after the Bayes' rule has been applied using the likelihood of the observations given the model and the probability of the models prior to using the observations. The Bayesian statistics is essentially an update process, which can be thought to begin with *null* information and by utilising the observations attains a more accurate model. This means that all the models are considered equally likely [8]. There are many more ways to choose the model based on observations [20].

In Bayesian statistical analysis the calculation of the posterior most often produces analytically unsurmountable problems. This obstacle can however be overcome by using numerical methods, namely simulating random samples from the posterior distribution [73]. The principle of random simulation and inference

based on them can be utilised in many ways, such as in approximating the posterior integral, and estimating the size of sets of objects meeting a specified criterion [12]. In Bayesian methodology the use of Markov Chains provides the theoretical justification and mathematical means for this [27].

When comparing models and their performance, there is a need for a metric. The examination of one novel metric is a major contribution of this dissertation, and is dealt in detail in Chapter 3. Perhaps the most common metric of probability distributions is called the Kullback–Leibler divergence [41], which is based on the information theory of Shannon [81], which has a connection to the theory of Kolmogorov complexity [50] and minimum description length principle by Rissanen [72]. The algorithmic minimum description length principle states that the predicted optimal model is the one which generates the observed data with the shortest description in terms of computer programs.

In a great many cases the model intergrals cannot be calculated analytically and simulational methods must be used [21]. The metrics can also be devised by simulating random samples from the models, to represent it, and comparing these with the ones obtained from other models. This alternative metric extends the metric on samples to a metric on the models, and thus provides perhaps an intuitive yardstick for statisticians. The metrics may also come to use in Bayesian analysis when the likelihood of the observations cannot be handily computed, but when generation of random samples from the model is still feasible. This metric is the sum of distances between pairs of elements in two sets; with a minimisation over the possible ways to choose the pairs. When the number of samples goes to infinity we can consider the resulting limiting value as the distance between two models [43]. This metric is entirely new in this field, and may be useful when comparing models in sample spaces which have a natural metric.

There is also a possible application for this: In cases where the likelihood function of the model is not easily computable, one can perform the model selection by generating samples from the models, and choosing the model for which the total distance to the observed data is the smallest. This was originally the topic of the study of Diggle and Gratton [21]. There has also been a similar proposal to this effect called *indirect inference* proposed by Gourieroux [31] in which samples are generated from the model and a statistic is computed from these, then this statistic is matched to the one obtained from the data. Lahtinen and Heikkonen have shown that the use of the metric on sample space will also provide means to perform the inference [44], which is the main topic of Chapter 4.

—

The second half of the dissertation focuses on the analysis and simulation of random graphs. A graph is a set of sites and a set of connections between these sites. The sites of the graph could be people, computers or even power plants, and

connections between them social acquaintances or electric cables. Graph theory is a very abundand branch of discrete mathematics, and many famous problems are associated with graphs, or are reducible to ones involving them. In the modern world where networks are vital to the functioning of the society and business, questions of efficiency and vulnerability of networks become important.

During the last few years an overwhelming amount of evidence has been accumulated about diverse networks showing *small–world* properties. This means that on average an arbitrarily selected site can be reached from another site in very few steps despite the fact that only relatively small number of connections are present in the graph [87]. It was noted by Stanley Milgram that it seems to take 6 handshakes to connect between any two people in the world [55]. The internet has similar characteristics, as most computers there are connected to some very central server [24]. The documents of WWW are very often connected to some relating important document, search engine, or collection [2]. The scientist tend to cooperate with famous scientists [69]. Watts and Strogatz were the first to suggested a simple mathematical model, which reflects the small world phenomenon: They proposed a regular lattice and then rewired some of the connections to form long range connections [88]. This model of small–world networks interpolates between a lattice and the so called Erdős-Rényi random graph [10, 87].

In addition to the interesting static structural properties in these networks, there is ever growing interest in dynamical processes operating on them. As a matter of fact it is expected that the underlying network topology should have a major impact on practically any phenomenon taking place in it. This view is supported by the recent results on the spectral density of the adjacency matrix of small world models, which show that these graphs produce a dramatic deviation from the semi-circle law of random graphs [25]. One can indeed infer a great deal about a network by performing a random walk on it [84]. For example the number of sites visited in a given time is a significant indicator of the structure of the network. This is also called *spreading*: over how much area does diffusion relocate a particle. The spreading phenomenon is the main topic of Chapter 5. The distribution of the number of visited sites has a transition which was first analysed by Jasch and Blumen [35]. The inaccuracy in the result of their analysis was corrected by Lahtinen et al. in [45]. The final result is that the distribution has the natural exponent of 2, i.e. the distribution has a transition that is proportional to the size of the area covered squared.

Another dynamic model is called the *sandpile model* [3]. This describes a process of loading the site of a given system with a burden, e.g. computational work for a computer. Eventully there comes a limit to how much load a single site can carry. When the limit is reached the load is transferred to the neighbours of that site. The neighbours themselves can also be excessively employed and the excess will need to travel on to their neighbours. Although this is not quite a physical model of sandpiles as they present themselves in sandy beaches, but a

simpler discrete system, it still gives insight to the mechanism of failures of an electrical power grid for instance [13]. In Chapter 6 the sandpile model on one dimensional random topology is investigated.

It is also interesting to ask to what extent the dynamical properties of small–world networks depend on the *quenched* character of the disorder, as opposed to the *annealed* disorder, where the connections are not frozen but are rewired during the time evolution of the system. Such a model with spreading was studied by Pandit and Amitkar [65], with the focus on the average access time, and Lahtinen and al. [46] investigated the scaling laws establishing that the annealed model can be extended into an equivalent model in the quenched setting. This kind of random walk system seems to bear some resemblance to the idea of the random walker making Lévy flight type jumps [37, 78, 83]. With the sanpile model a similar approach is the stochastic sandpiles considered previously by Manna [52]. Both annealed and quenched disorder are present also in the analysis of the one dimensional sandpiles by Lahtinen et al. [47], in Chapter 6.

—

This thesis is organised so that the next three chapters of the first part concentrate on statistical inference, and the second half of the thesis with two chapters on random graphs and their dynamic properties. Chapter 2 gives a general review of Bayesian statistical inference and its basic formulation. In Chapter 3 there is a treatise on the metrics on probability distributions. Then Chapter 4 focuses on the problem of uncomputable likelihoods. In Chapter 5 the spreading phenomenon on small–world networks is taken under scrutiny. After defining the basic results the attention is focused on the dynamics on graphs, i.e. the spreading. Then the Chapter 6 deals with the self–organised criticality in 1–dimensional small–world networks.

Chapter 2

Bayesian Statistical inference with a fault diagnostic application

Statistical inference is essentially an inversion process: there is a set of random observations, and a unifying pattern is sought that fits these observations. A statistical model is the pattern, describing the probabilities of the events, already observed and the ones yet to be observed. However, the problem is mostly much more complex than connecting the dots. There are too many models and too few observations so that no one model alone would explain the observed phenomena. This dissertation focuses on the Bayesian approach which was applied to a fault diagnostic system described later in this chapter. It was shown that the latent states of a device can be identified based on observations which do not have direct information about the inner states [48]

Here the focus is mainly on Bayesian statistical methods. In Bayesian statistics the conditional probability of different models given the observations is calculated using the rules of probability calculus. This probability of the models is called the posterior probability. Although Bayes' formula was discovered early, its use was scarce and for a long time statistical problems were mainly solved with other methods [8]. The difficulty lay in the integral that would be needed to utilise the posterior probability in analytical calculation. Once it was realised that in practical statistical applications stochastic integrals can be efficiently approximated by simulation of random samples, all the required integrals could now be handled by computers [73]. These simulational methods in this context are called Markov Chain Monte Carlo (MCMC) methods [27]. With the aid of MCMC one can generate simulated samples from the distributions in the modelling situation and using the law of large numbers to approximate relevant integrals.

There are many other approches to statistical inference, such as the traditional

statistical methods [75], neural networks [15] or even game theoretic settings [20], the model selection is treated as a game between the modeler and an opponent. These however are more specific cases. Nonetheless, the fundamental nature of statistical inference is a *no free lunch theorem* [51]: whatever the base assumptions are no method can truly outperform any other in general comparison.

In this chapter there is a short primer on Bayesian theory of probability, with an introduction to the basic Markov Chain Monte Carlo (MCMC) methods and their convergence tests. These concepts are applied in section 2.2 to a fault diagnostic system.

2.1 Bayesian inference

In the Bayesian approach each observation is conjectured to have a probability of occurrence, determined by the model under inspection. When in addition each model is assigned a probability, one arrives at the foundation of Bayesian statistical inference: based on this information the probability of a model given the observations is computed using the probability calculus.

A random variable X is a measurable function from a sample space $\mathbb{D}$ with a σ–algebra on $\mathbb{D}$, a super set of the sets of $\mathbb{D}$, and a measure P on that σ–algebra. A random variable X also has an associated distribution which is here always denoted as μ_X, and thus measure P of events of the random variable X is the integral over the distribution μ_X. An indexed sequence of independent and indentically distributed, IID, random variables is denoted as $X_{1:m} = \{X_i\}_{i=1}^m$ on a sample space $\mathbb{D}$, for which there is a multi-set of sample points $x_{1:m}$, where $x_i \in \mathbb{D}$. A multi–set is a set where duplicates of the same element are possible. The notation matches the capital lettered random variables with the lowercase sample–sets, and it is assumed that the multi-set $x_{1:m}$ is exchangable (infinitely so if the set is), i.e., $\mathsf{P}(X_1 = x_1, \ldots, X_m = x_m) = \mathsf{P}(X_1 = x_{\pi(1)}, \ldots, X_m = x_{\pi(m)}) \equiv \prod_i \mathsf{P}(X_i = x_i)$ for any permutation π of $\{1, \ldots, m\}$.

For a model, parametrised with θ, given the observations $x_{1:m}$ the *posterior distribution* is:

$$\mu_{\Theta'}(\theta|x_{1:m}) \propto \mu_{X|\Theta}(x_{1:m}|\theta)\mu_\Theta(\theta), \tag{2.1}$$

where $\mu_{X|\Theta}(x_{1:m}|\theta)$ is the *likelihood* and $\mu_\Theta(\theta)$ is the prior. The expression $\propto$ means that the parts are proportional to a multiplicative normalising constant, in this case $\mathsf{E}\{\mu_{X|\Theta}(x_{1:m}|\Theta)\}$. This term is usually omitted as in practice it is not needed in comparing models. The common use of the posterior distribution is to calculate the *predictive distribution*, which is the distribution of future events independent of the parameters:

$$\mu_{X'}(x'|x_{1:m}) = \mathsf{E}\{\mu_{X|\Theta}(x'|\Theta')\}. \tag{2.2}$$

This is an expectation over the posterior Θ', thus utilising the posterior distribution in the predictive distribution of future observations. This particular expectation cannot most often in practise be computed in an analytical form. In usual situations this expectation is complex and multi–dimensional that the usual numerical approximations also fail. Therefore the standard approach is to simulate samples from the posterior distribution of equation (2.1).

2.1.1 Random sampling of the posterior distribution

The most important tool in Bayesian inference is posterior sampling. This method allows one to generate random samples distributed according to the posterior disribution. These samples can then be used to estimate the predictive distribution of equation (2.2) as a sum over n samples $\{\theta_i'\}_{i=1}^n$:

$$\mu_{X'}(x'|x_{1:m}) \approx \frac{1}{n}\sum_{i}^{n} \mu_X(x'|\theta_i'). \tag{2.3}$$

This is called Monte Carlo integration. The integral can thus be replaced by a sum over a discrete set of properly distributed points [27].

In the rest of this section the random sampling techniques from a distribution are discussed. The practical methods of random simulation rely on Markov chains. A Markov chain is a sequence of random variables $X_1 \to X_2 \to \ldots \to X_m$ with the property that the element X_i is not dependent on the previous elements except X_{i-1}. This means that $\mu_{X_i}(x_i|x_1, x_2, \ldots, x_{i-1}) = \mu_{X_i}(x_i|x_{i-1})$. Thus the Markov property ensures that in order to generate a sample i all that is needed is the previous sample, which is very useful for efficiently generating a large number of samples.

2.1.2 Metropolis–Hastings algorithm

The basic method for sampling the parameters θ_i' distributed according to $\mu(\Theta')$ proceeds in steps. When the current state is θ_i' then the proposal ζ for the next sample in the sequence is drawn from a transition distribution $\mu_Z(\zeta|\theta_i')$, which is called *the Markov kernel*. The proposal is accepted with a probability $a(\theta_i', \zeta)$:

$$a(\theta_i', \zeta) = \min\left(1, \frac{\mu_{\Theta'}(\zeta|x)}{\mu_{\Theta'}(\theta_i'|x)}\frac{\mu_Z(\theta_i'|\zeta)}{\mu_Z(\zeta|\theta_i')}\right). \tag{2.4}$$

If accepted then $\theta_{i+1}' = \zeta$. This method is called the Metropolis–Hastings algorithm [28].

The choice of the kernel is decisive when applying the MCMC methods. A kernel with a too little variance will converge too slowly, as it may take many steps for the chain to extend over all significantly probable parameters. In turn a too wide kernel may not accept many proposals.

2.1.3 Gibbs sampling

An important practical method that is usually tried first is Gibbs sampling [30], where amultidimensional parameter is sampled one at the time keeping the others fixed. This can be viewed as a special case of the Metropolis–Hastings algorithm [28]. For a set of parameters $\theta_1, \theta_2, \ldots, \theta_d$ each θ_i is sampled from the conditional distribution given the other parameters $\theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d$ and the observations $x_{1:m}$.

Usually the parameter θ_i given the others can be presented analytically, and samples from it can be drawn without the need for the proposal–acceptance procedure. However, often it can be that this conditional distribution is difficult to define in closed form and then the Metropolis–Hastings step is needed also for the θ_i conditional to the other parameters.

The advantage that Gibbs sampling has when compared to the Metropolis–Hastings method is its faster sampling when the number of parameters is very large. Whereas the Metrpolis–Hastings method may not easily find an acceptable proposal, Gibbs sampling searches for a single new parameter at a time, which speeds the acceptance, but this has the limitation that in some multimodal cases the Gibbs sampling may not find all the modes. Also when there is need to make a proposal that has more, or fewer, dimensions than the previously accepted sample, these methods alone do not quite work. For this purpose there is an extension, which is studied next.

2.1.4 Reversible jump algorithm

A notable extension to the basic Metropolis–Hastings algorithm is the added ability to jump between spaces with different dimensions [32]. The method also carries the name Metropolis–Hastings–Green algorithm, or sometimes the *reversible jump MCMC*. This algorithm requires an additional latent variable ξ drawn from a distribution $\mu_\Xi(\xi|\theta_i')$, which is chosen to balance the scales between the numerator and the denominator of equation (2.4). Assume that θ_i' is in a subspace of dimension d, ξ is in an e–dimensional space, and the next proposal ζ is in a $d+e$–dimensional subspace. The algorithm needs a function f that maps a pair (θ', ξ) to ζ. The acceptance probability becomes $a_{RJ}(\theta_i', \zeta, \xi)$ reading as follows:

$$a_{RJ}(\theta_i', \zeta, \xi) = \min\Big(1, \frac{\mu_{\Theta'}(\zeta)}{\mu_{\Theta'}(\theta_i')} \frac{\mu_Z(\zeta|\theta_i')}{\mu_Z(\theta_i'|\zeta)} \frac{|J(\theta_i', \xi)|}{\mu_\Xi(\xi|\theta_i')}\Big), \tag{2.5}$$

where $J(\theta_i', \xi)$ is the Jacobian determinant of the function f. If the dimension decreases instead of the equation (2.5) one should use the resiprocal form:

$$a_{JR}(\theta_i', \zeta, \xi) = \min\Big(1, \frac{\mu_{\Theta'}(\zeta)}{\mu_{\Theta'}(\theta_i')} \frac{\mu_Z(\zeta|\theta_i')}{\mu_Z(\theta_i'|\zeta)} \frac{\mu_\Xi(\xi|\theta_i')}{|J(\theta_i', \xi)|}\Big). \tag{2.6}$$

With these amendments one can apply MCMC simulation for example in approximating the number of kernels in kernel–estimation [54], or selecting the input variables for a model [85].

2.1.5 Simulation convergence

The problem with the MCMC method is the determination of when there are enough samples. The simulation usually begins with an initial value which may be very located in a remote area of the sample space from the are wich contains the most probable events. The dynamics of the simulation, determined by equation (2.4), then move the focus into a more probable region of the parameter space. The identification of the initial period from the stable sequence is almost a similar unresolvable problem to the original decision–making process for the model itself, but many good methods still exist. The first is visual inspection of the statistics, which turns out to be usually very good for separating any initial burn–in period, in which the chain moves from the initial value to the main region. Also the consecutive samples in the chain are not quite independent. For this one can use the autocorrelations to extract a subset of more independent samples from the the simulated sequence [60]. Finally, in order to ascertain the convergence of the chain one general method is the Kolmogorov–Smirnov goodness–of–fit test [73], which is dealt with at end of this section. There are also many other method to evaluate the convergence of a chain [73], but the convergence testing in general is also a difficult problem that cannot be absolutely solved.

Use of the autocorrelation in MCMC

Samples in a sequence that is generated with MCMC methods are usually correlated. The immediately consecutive samples are always by the definition of Markov chains correlated, but usually this correlation extends much further depending on the kernel. The correlation can be radically reduced by choosing a subset of the actual samples. The autocorrelation time is the average number of steps in the sequence such that the samples that far apart are almost uncorrelated, and thus if one were to omit the samples in between the resulting sequence is uncorrelated. The autocorrelation time is obtained through the normalised autocorrelation sequence of a sequence of random variables $\Theta_{1:n}$ with lag k, which is defined as:

$$a_k = \frac{\sum_{i=1}^{k}(\theta_i - \bar{\theta}_{1:n})(\theta_{i+k} - \bar{\theta}_{1:n})}{\sum_{i=1}^{n}(\theta_i - \bar{\theta}_{1:n})^2}, \tag{2.7}$$

where $\bar{\theta}_{1:n}$ is the sample mean of $\theta_{1:n}$. This is an estimate of the covariance of the ith and the $(i+k)$th sample divided by an estimate of the variance of the ith sample.

From the set $\theta_{1:n}$ one can form a subset of samples by taking every t_{aut}th step, with

$$t_{\text{aut}} = 1 + 2\sum_{i=1}^{n-1} a_k. \tag{2.8}$$

Decimating the samples then guarantees that these reduced samples are correlated as little as possible [73].

Kolmogorov–Smirnov test

Next there is a discussion on the use of the Kolmogorov–Smirnov goodness–of–fit test to evaluate the convergence of the chain, i.e., when to stop the simulation. This test is a general non–parametric method of deciding whether a set comes from a given distribution. For a given continuous random variable Θ the KS–statistic is the maximum empirical deviation of the sample estimate from the true value of the cumulative probability:

$$\text{KS} = \max_i \left\{ \left| \mathsf{P}\{\Theta \leq \theta_i\} - \frac{\#(\theta_i, \theta_{1:n})}{n} \right| \right\}, \tag{2.9}$$

where $\#(x, A)$ is the number of elements in A smaller or equal to x. The null hypothesis is that the $\theta_{1:n}$ are distributed according to μ_Θ. The null hypothesis is rejected if the KS–statistic is greater than a given tolerance α_{KS}.

For testing of the convergence of an MCMC chain the distribution μ_Θ is replaced with another sequence, preferably obtained from a second independent chain:

$$\text{KS}_{\text{MCMC}} = \max_i \left\{ \left| \frac{\#(\theta'_i, \theta'_{1:m})}{m} - \frac{\#(\theta_i, \theta_{1:n})}{n} \right| \right\}. \tag{2.10}$$

In practice one can assume that the chain has converged if the value KS_{MCMC} is not too small or too large. As a rough rule of thumb one could use the relation: $0.1 \geq \text{KS}_{\text{MCMC}} \geq 0.9$.

2.2 Application to Fault Diagnostic

In this section Bayesian inference is applied to a fault diagnostic system. It is assumed that a device under inspection records the total number of some events in its lifetime. For example, the device can count the number times it has been turned on etc. Eventually the device will break down and the user brings it for repair. The goal then is to decide in what way the device is malfunctioning based on the information gained from these counter values.

The framework in this case is a set of counters, the final value of which is observed at the end of some period of time. During this time period the process has changed from the initial state to the final state at an unknown point of time.

For both the initial and final periods there are an unknown numbers of possible substates, i.e., event occurrence rates. For instance, a device used by a travelling businessman can record a different behaviour than that of an office worker.

The estimation task is complicated due to the fact that there is no prior knowledge about the event rates for either intact or faulty devices, hence the event rates for the states and the state transitions must be modelled simultaneously. Collecting the data during actual operation from a large number of devices causes additional complication in the model as the devices may not be exactly similar. For example, in a paper machine both the sensors and the production line hardware are continuously updated. Similarily in mass production devices, like in portable computers, the same model may contain various different hardware configurations and operating system versions, possibly affecting the rates of the monitored events. To account for this variation all the states are modelled as mixtures of processes, with an unknown number of substates. The substates are assumed to be constant during the operation, so that each device has zero or one unknown state transitions to be estimated.

Counter generation model

The process can be sampled in two ways, so that some of the devices have only gone through a single state, an intact device, and some have two states, an initially intact device which has then broken down. There may well be several inner states in which the device may be as broken or intact, see for example the Figure2.1.

The vector of values of m counters are denoted by $x_i \in \mathbb{Z}$. The latent, unobservable, variables determining the states of the process are denoted by $z_1 \in \{1, \ldots, k_1\}$ for the initial state and $z_2 \in \{k_1 + 1, \ldots, k\}$ for the broken state, with k_1 and k_2 the number of initial and broken states, respectively. The unobserved value of the counter i during the initial state of length υ is denoted by y_i, and the final observed value during time t is denoted by x_i. Each counter x is modelled as a Poisson process with parameters λ [56]. This means that in a given time the probability of observing one event measured by the counter is exponentially distributed, and thus is assumed not dependend on the previous events.

Assuming that there are n counters. There are $k = k_1 + k_2$ latent states in the model, where k_1 is the number of initial states and k_2 is the number of broken states. The matrix of Poisson rates in each of these states is $\lambda \in \mathbb{R}^{k_1 \times k_2}$. The probabilities of the k_1 initial states are denoted by $\omega \in \mathbb{R}^{k_1}$ and the matrix of the transition probabilities from the initial to broken states by $r \in \mathbb{R}^{k_1 \times k_2}$.

The device is initially in one of the k_1 intact states, z_1, with probabilities ω_{z_1}. At time υ the counter i will have the value y_i drawn from Poisson distribution with mean $\Lambda_{z_1 i}$. Then at time υ it makes a transition to a broken state z_2, of which there are k_2 possibilities, with a probability r_{z_1, z_2}. In this state the counter i is again generated at a different rate $\Lambda_{z_2, i}$. The total value of the counter i is then x_i.

Estimation of the posterior

In choosing the priors it is assumed that no useful knowledge is attainable about their form, and thus one should resort to non–informative forms. The prior distributions of the variables are

$$\begin{aligned}
K_1 &\sim \text{Uniform}\{1, \ldots, k_{max}\} \\
K_2 &\sim \text{Uniform}\{1, \ldots, k_{max}\} \\
\Omega &\sim \text{Dirichlet}(\underbrace{1, \ldots, 1}_{k_1 \text{ times}}) \\
R_{i,1:k_2} &\sim \text{Dirichlet}(\underbrace{1, \ldots, 1}_{k_2 \text{ times}}) \\
\Lambda_{i,j} &\sim \text{Gamma}(\alpha, \beta) \\
Z_1 &\sim \text{Bernoulli}(\omega) \\
Z_2 &\sim \text{Bernoulli}(T_{z_1,1:k_2}) \\
\Upsilon &\sim \text{Uniform}[0, t] \\
Y_i &\sim \text{Poisson}(\upsilon \Lambda_{z_1,i}).
\end{aligned}$$

Each state is thus considered equally probable. The weights of the states are Dirichlet distributed, which again means that all possible combinations of $\omega_1, \ldots, \omega_{k_i}$ are equally probable with the restriction that $\sum_{j=1}^{k_i} \omega_j = 1$. The Bernoulli distribution here is a discrete distribution where each of the k values have the corresponding probabilities in the parameter vector. The Γ–distribution is chosen as a prior for the Poisson rates because it is a conjugate prior of the Poisson distribution [28].

The likelihood of the observed counter values $x_{1:m}$ when the latent variables and parameters are given is then:

$$\mu_X(x|t, \upsilon, \lambda, z) = \prod_{i=1}^{m} \text{Poisson}(x_i|\lambda_{z_1,i}\upsilon + \lambda_{z_2,i}(t - \upsilon)). \qquad (2.11)$$

The Poisson rates can sampled using the common Gibbs sampling for Poisson distributions. These are first sampled for the initial states and then kept fixed for the sampling of the second state rates. The number of latent states can be chosen in both cases according to the most likely values based on the MCMC sampling with reversible jump steps, RJMCMC [32]. The posterior distribution of the parameters can be estimated with the Metropolis–Hastings–Green algorithm. Similar approaches for mixture distributions have been studied by Viallefont & al in [86] for Poisson mixtures and by Marrs in [54] for Gaussian mixtures. A difference here is that the device has a possible change of state from intact state to a defective state at an unknown point of time, whereas the methods in the

references assume that the system has always been in one state of which there are many choises. The RJMCMC jumps between dimensions are done with *Split–Merge* type reversible jump moves. Here the upper index is used to enumerate through the data samples, and their latent variables. This is done by repeating the following steps cyclically where in each step is described the parameters to be sampled while the others are given:

1. Draw each $\lambda'_{i,j}$, $k_1 < i \leq k$, from $\Gamma(\alpha + \sum_{l:\{z_2^l=i\}}(x_i^l - y_i^l), \beta + \sum_{l:\{z_2^l=i\}}(t^l - \upsilon^l)$.

2. Draw each $r_{i,:}$, ith row of r, from Dirichlet(A), where $A \in \mathbb{R}^{k_2}$, $A_j = 1 + \sum_l I\{z_1^l = i \wedge z_2^l = j\}$, where $I(a) = 1$ if a is true and 0 otherwise.

3. Draw each z_1^i from Bernoulli(B), where $B \in \mathbb{R}^{k_1}$, $B_j = \omega_j \prod_l \text{Poisson}(y_l^i | \upsilon^i \lambda_{z_1^i,l})$.

4. Draw each z_2^i from Bernoulli(C), where $C \in \mathbb{R}^{k_2}$, $C_j = \omega_{z_1^i} T_{z_1^i,j} \prod_l \text{Poisson}(x_l^i - y_l^i | (t^i - \upsilon^i)\lambda_{z_2^i,l})$.

5. Draw each υ^i and y^i from their posterior by Metropolis–Hastings procedure.

6. In the Reversible Jump step either decide to try a split or merge a random kernel (the Poisson rate parameters of some latent state) with probability $1/2$.

7. Use the split, or merge, map (see below) to a kernel κ chosen at random.

8. Reallocate the latent states $z_2^i = \kappa$, (or while merging $z_2^i = \kappa \vee z_2^i = \kappa + 1$) by drawing from Bernoulli(D), where $D \in \mathbb{R}^{k_2}$, $D_j = \omega_{z_1^i} T_{z_1^i,j} \prod_l \text{Poisson}(x_l^i - y_l^i | \upsilon^i \lambda_{z_2^i,l})$.

9. Accept the split proposal with probability

$$\min\{1, \frac{\mu_X(x|\zeta)}{\mu_X(x|\theta)} \frac{|J|}{\mu_Z(z)}\}, \tag{2.12}$$

where θ represents the distribution of all parameters, and $\mu_Z(z)$ is the reallocation probability of the latent z and $|J|$ is the Jacobian determinant of the split map (see below). In case of merge the acceptance probability is

$$\min\{1, \frac{\mu_X(x|\theta')}{\mu_X(x|\theta)} \frac{\mu_Z(z)}{|J|}\}, \tag{2.13}$$

where $\mu_Z(z)$ is the reallocation probability of the latent z if splitting from the new state back to the original with the map whose Jacobian is $|J|$.

The steps 1 to 4 follow the Gibbs sampling and the steps 5 is a standard Metropolis–Hastings jump, whereas the steps 6 to 9 describe the reversible jump.

In full Bayesian analysis no fixed values for any intermediate variables are estimated, but instead the posterior distribution of the variables is propagated throughout the analysis. The sequential estimation of the parameters can be justified by practical reasons: to simplify the analysis and to make the sampling faster.

Reversible jump step

The jump between dimensions here is a *Split–Merge* mapping [71]. There are many other forms this could be done. This has been found functional, along with a similar *birth–death* mapping [71]. The process is done by randomly choosing with equal probability either splitting or merging a state. In splitting a state the average rate of the two new states is preserved:

$$\omega_1'\lambda_1' + \omega_2'\lambda_2' = \omega\lambda. \tag{2.14}$$

The other parameter values are copied from the original one. The new values for ω_1, ω_2, λ_1 and λ_2 are then mapped so that all possible positive values of λ_1' and λ_2' satisfying equation(2.14) are equally probable. This is the following map, $(\lambda_i, \omega_i, u, v) \mapsto (\lambda_i', \lambda_{i+1}', \omega_i', \omega_{i+1}')$, in which the latent state i is split, and $u, v \in [0, 1]$ are drawn from the uniform distribution:

$$\begin{array}{lcl} \omega_i' & = & u\omega_i \\ \omega_{i+1}' & = & (1-u)\omega_i \\ \lambda_i' & = & v\lambda_i \\ \lambda_{i+1}' & = & \frac{\omega_i\lambda_i - \omega_i'\lambda_i'}{\omega_{i+1}'}. \end{array} \tag{2.15}$$

The Jacobian determinant is then

$$J = \frac{\omega_i\lambda_i}{u-1}. \tag{2.16}$$

When merging two states, the rate of the new state is solved from the equation 2.14 and the other parameters are copied from one of the two components chosen at random.

Example

Initially the modelling done in this section was done in cooperation with an industrial corporation, which promised to provide labeled data to be used in the posterior inference and testing. However in the end no such data was provided, and the functionality of this model can only be shown with a simulations. Nonetheless it gives a hint that the Poisson parameters can be extracted given that there are

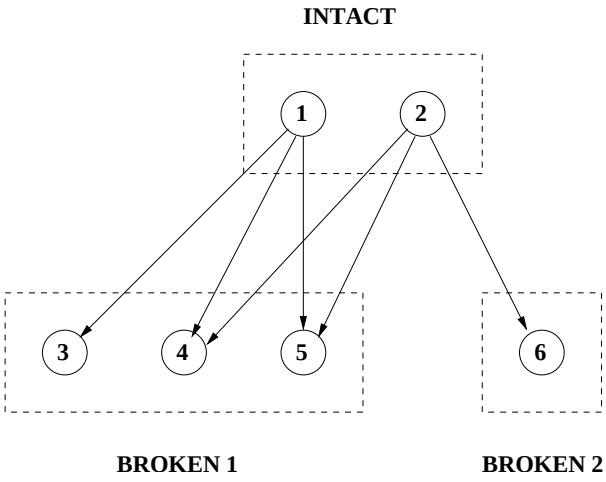

Figure 2.1: The state diagram of the example system.

enough counters and that the latent states are sufficiently separated in the parameter space.

Take for example the following: the initial, intact, states are labeled as $\{1, 2\}$ and the final, broken, states are $\{3, 4, 5, 6\}$, where the broken states are divided into two groups, this would present that the device has two different categories of malfunctions, with the set of states $\{3, 4, 5\}$ as one category (*class 1*) and the state $\{6\}$ alone (*class 1*), see Fig. 2.1. The inital states were equally probable with Poisson rates and transition matrix:

$$\Lambda = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ \hline 2 & 8 \\ 7 & 7 \\ 9 & 3 \\ 15 & 15 \end{pmatrix} \quad T = \begin{pmatrix} 0 & 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/4 & 0 \end{pmatrix} . \qquad (2.17)$$

The simulated data had 25 samples from the initial model, representing intact devices and 100 samples from the two state model, representing broken devices (see figure 2.2). The transition time was uniformly distributed.

The parameters of the intact devices were simulated for 1000 rounds and the parameters of broken devices were simulated for 3000 rounds, with prior Gamma(α, β) for $\Lambda_{i,j}$. The convergence of the MCMC simulation was tested using the Kolmogorov–Smirnov test [73] after a proper subset of the data samples was selected based on the autocorrelation time to avoid the dependence of consecutive samples [60]. The number of the latent states was identified very quickly

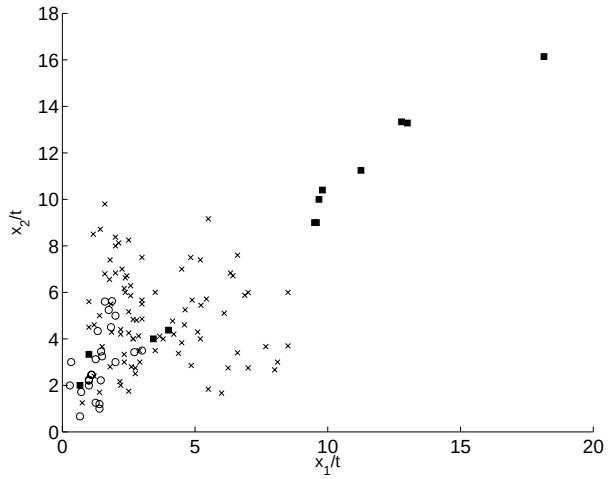

Figure 2.2: Samples of Example 1, intact devices marked with circles, broken class 1 with crosses and broken class 2 with boxes (the axes are the numbers of counters divided by time).

and the simulation remained very stable on the correct number of states. The estimated probabilities, take as the mean of the simulated samples, with the most likely number of initial states were:

$$\hat{\Lambda} = \left(\begin{array}{cc} 1.2 & 2.0 \\ 2.0 & 4.1 \\ \hline 2.0 & 8.6 \\ 7.3 & 7.2 \\ 9.1 & 3.1 \\ 15.6 & 15.4 \end{array} \right) \tag{2.18}$$

The transition probabilities to the first broken class were:

$$\hat{T} = \left(\begin{array}{ccc} 0.25 & 0.47 & 0.28 \\ 0.29 & 0.49 & 0.22 \end{array} \right). \tag{2.19}$$

In comparison to the true matrix 2.17 it can be seen that the matrices are not quite the same. This is because some of the observations could be explained as a transition from the intact state 1 into broken state 1, or as a transition from intact state 2 into broken state 4, both of which are not possible in the true matrix. This implies that the transition matrix is not quite identifiable.

The estimated distribution of the initial states was:

$$\hat{\omega} = \left(\begin{array}{c} 0.55 \\ 0.45 \end{array} \right), \tag{2.20}$$

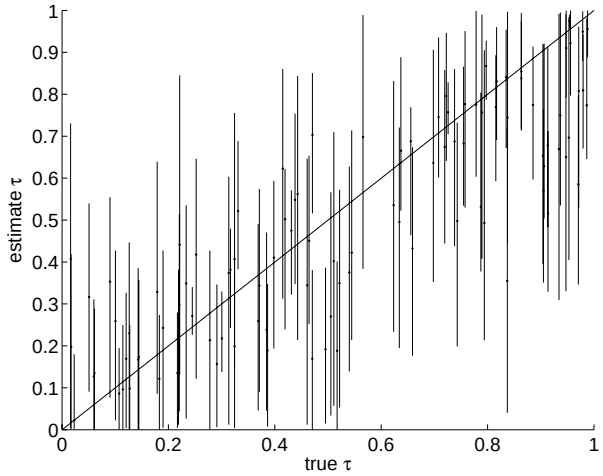

Figure 2.3: The true relative values of υ for the 2–D example compared against the median estimates with the 90% HPD intervals.

which corresponds sufficiently with the true equal probabilities of the initial states.

The estimated parameters were tested in a simulated classification task for 100 samples from the initial process, representing intact devices, and 500 samples from the two-state process, representing broken devices. The confusion matrix A, compared to the 3-Nearest Neighbour classifier is:

$$A = \begin{pmatrix} 0.92 & 0.080 & 0 \\ 0.22 & 0.72 & 0.064 \\ 0.13 & 0.14 & 0.75 \end{pmatrix} \quad A_{\text{3-NN}} = \begin{pmatrix} 0.57 & 0.43 & 0 \\ 0.13 & 0.85 & 0.011 \\ 0.085 & 0.64 & 0.27 \end{pmatrix} \tag{2.21}$$

From these matrices one can see that neither method mistakes an intact device with a broken one in class 2, the first row, but that the Bayesian classifier is much less likely to confuse an intact device with the broken one in class 1. One could also use the CART [11] but in cases such as this where the decision border is not parallel to the counter axes it performs rather poorly with too little data and the decision tree becomes very large.

The estimation of the υ parameters as the median of the samples for each data sample is plotted in figure 2.3. The lines are the 90% HPD intervals (*Highest Probability Density intervals*) [14]. The uncertainty of the estimates comes from the facts that the data does not contain direct information of the transition point, and that there is only one observation related to estimation of each transition point, and thus the estimates tend to come from the uniform prior. In estimation of the classification, the υ dependency was marginalised by summing over the estimated υ values of the simulations. The classification based on maximal probability bor-

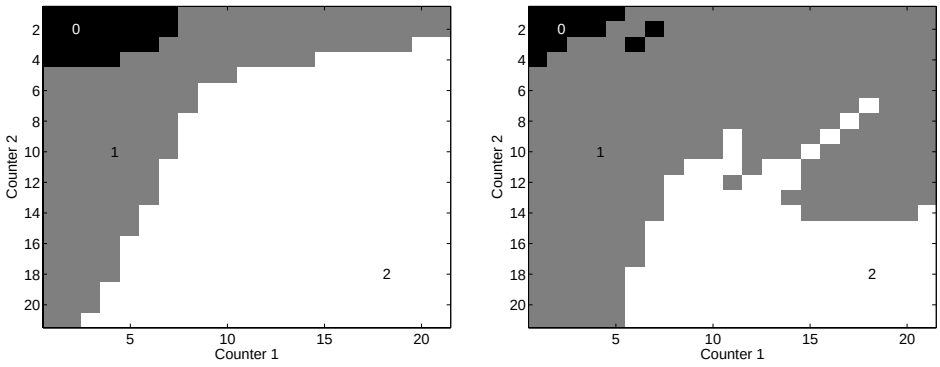

Figure 2.4: On the left is the maximal probability classification borders for the 2-dimensional example, the values of the counters in unit time, and on the right the 3–Nearest Neighbour classifier for the same data. The black is the area of intact devices (class 0) and the gray broken class 1 and white broken class 2.

ders compared with the 3–Nearest Neighbour classifier can be seen in figure2.4. It can be seen that that the Bayesian model gives a smoother transition between the classes, as it can be assumed.

2.2.1 Discussion

The two states of the process can be recognised by the final values of the counters when the dimension (the number of counters) is large. The one–dimensional case is not identifiable as the observed phenomena can be explained by varying the distribution of the transition point of which there is no direct information. Also if there are too few counters and too many latent states the counters may not contain enough information to separate all the inner states. The estimation becomes more difficult when the transition has occurred close to the end of the total time, in which case the counters only exhibit behavior of the initial states. In this estimation the availability of data for purely intact devices, and presence of more than one counter to record events, is critical.

The MCMC simulation results of the posterior distribution of the fault diagnostic example indicates that the parameters and the latent states can be identified from the 100 observations. It is also apparent that even when it is not possible to determine when the device was broken classification can still be done. The estimation is difficult for those observations for which the transition has occurred close to the end of the total time, as there has not been enough time for the events indicating the fault to accumulate. It was also assumed that estimation can be first done for the intact devices. It is also necessary to have one counter to record events. Even if the number of counters is very large it does not affect the func-

tionality of the model, however it may slow down the MCMC simulation.

Chapter 3

Metrics of probability distributions

Model selection is naturally also about model comparison, which needs a metric of some kind on the models, or a goodness–value. A metric is a value between two elements, whereas a goodness–value is assigned for a single element. The value of a model could be determined by its usefullness. This chapter introduces a new metric called *transformation discrepancy*, which thus can also be used in model selection schemes. Its advantages are that it is based on the natural metrics on the sample space or utilities of the models.

One value for comparison is the goodness-of-fit of the model to the observations [56]. Also in use are the entropy based metrics [42], which are based on Shannon's information theory[81]. However, the value could be defined from the usefulness of the model, such as the cost, or benefit gained by the decision based on it. This thesis presents a totally new kind of a metric called *transformation discrepancy* which essentially extends a metric on the samples to cover the distributions [43]. In applications such as clustering or image analysis one often needs a measure of similarity between sets, an image being also a set of sorts for which a similarity measures can be used [68]. Also Bennett et al. have a treatise on the framework of Kolmogorov complexity as such a measure [7]. In this chapter is presented a new metric based on a metric on the sample space.

When speaking about any objects a measure of difference gives the means of comparing them. This is equally true for locations in space, weights of items, and probability distributions. Formally such a measure is called a metric. A metric has three mathematical characteristics [74]:

Definition 1 *A metric is a binary function $\delta(\cdot,\cdot)$ such that*

1. *(Positivity)* $\forall x, y : \delta(x, y) \geq 0$ *and* $\delta(x, y) = 0$ *iff* $x = y$.
2. *(Symmetry)* $\forall x, y : \delta(x, y) = \delta(y, x)$.

3. (Triangle inequality) $\forall x, y, z : \delta(x, y) + \delta(y, z) \geq \delta(x, z)$.

3.1 Information–theoretic metric

A common form of a metric on probability distributions starts with the concept of information. Information contained in a random variable is a measure of disorder, or the average number of symbols that are needed to describe the observations [31, 18]. While the minimal code- engths required for the communication of an element x is undecidable, the logarithm of the probability $\log \mu_X(x)$, or the Shannon–Fano code length, is a useful upper bound [50]. The average of this code–length, and the measure of information, is called *entropy* [81]:

$$H(\mu_X) = \mathsf{E}\{\log \mu_X(X)\}. \tag{3.1}$$

When comparing a distribution of a random variable X to that of another random variable Y, their Kullback–Leibler discrepancy is the average difference of their Shannon–Fano code lengths:

$$KL\{\mu_X|\mu_Y\} = \mathsf{E}\{\log \mu_X(X) - \log \mu_Y(x)\} = H(\mu_X) - \mathsf{E}\{\log \mu_Y(X)\}. \tag{3.2}$$

This definition is not symmetric —one gets a different value if Y is compared to X, but it is zero if and only if X and Y are identically distributed. Of course there are natural extensions to a symmetric form and thus to a proper metric [42].

3.2 Utility based metrics

Metrics, like the prior, are mostly based on the views of the statistician, and the environmental constraints which essentially come from the purpose of the model and the decision based on it. In the previous section this purpose was data compression, and so also communication, but other applications of the statistical knowledge pertain. For example the expected return of a gambler, efficiency of a classifier, accuracy of measurements etc.

The concept of a *transformation metric* is slightly different from the stand points of the Kullback–Leiber type metrics. There the transformation refers to a process of changing the events for which the probability is calculated in one distribution such that the resulting events would match the probabilities of another distribution. The following treatise in this section on the transformation metrics can be illustrated by a physical methaphor: one could think of two clusters of particles. The energy required for moving the particles is prortional to the distance moved; naturally the very definition of work. For the two clouds of particles the metric between them is the minimal energy one would need to move the particles from one cloud into a configuration reminiscent of the formation in the other. In this kind of principle starts the building of the transformation metric.

3.2.1 From utilities to metric

When one can assign utility, or a cost, to each possible event, or there are readily available means for comparing two events, then this can be used to derive a metric on sets of events. A metric on distributions can then be obtained at the limit when the sizes of these sets grow to infinity.

The notation of $\mathcal{F}$ is used in this section for the space of parameters for a family of functions $\mathbb{D} \times \mathcal{F} \longrightarrow \mathbb{D}$. We call $\mathcal{F}$ *complete* if for all $x, y \in \mathbb{D}$ there exists $\theta \in \mathcal{F}$ such that, $f_\theta(x) = y$.

A utility function $U : \mathbb{D} \longrightarrow \mathbb{R}$ assigns a value to each element x in the sample space. Here the reference is to cost–like thinking: the lower the value the better. From this one can get a pairwise cost, or *discrepancy* $C : \mathbb{D} \times \mathbb{D} \longrightarrow \mathbb{R}$ of elements x and y by a path–integral; the total cost on a shortest path:

$$C(x\|y) = \min_{\gamma(x,y)} \int_\gamma dU, \tag{3.3}$$

where $\gamma(x, y)$ is a path connecting x and y. This cost is not quite a metric, as it is quite possible to have separate items x and y, $x \neq y$, such that $C(x\|y) = 0$, but this is not a problem per se, but only an indication that these elements are of equal value. Of course the cost–function $C(\cdot\|\cdot)$ may be naturally available directly –a metric on $\mathbb{D}$ for instance. Of course C need not be the result of a minimisation process but can be explicitly defined.

Another approach would be to use a cost function $K : \mathcal{F} \times \mathbb{D} \longrightarrow \mathbb{R}$ that measures the cost of different ways a can be transformed into something else by functions parametrised by $\theta \in \mathcal{F}$. Then a pairwise cost can be defined as $C(a\|b) := \min_\theta K(\theta\|a)$ such that $f_\theta(a) = b$, and ∞ if no such θ exists.

Once there is such a pairwise utility function, it forms a base for the discrepancy for sets of elements. This discrepancy is the sum of the pair–wise discrepancies of the elements in the two sets. Assuming that the two have equally many elements, then each element in one set can be matched with one in the other set. There are many such matching but the one that minimises the total is chosen. The discrepancy on sets becomes the total sum of the discrepancies of these pairs.

A *matching* of $x_{1:m}$ and $y_{1:n}$ is a multi-set R of k pairs of $x^\alpha_{1:m}$ and $y^\beta_{1:n}$, such that each element of x^α_i and y^β_j appears exactly once in some pair, and $m\alpha = n\beta = k$.

We then have a set of source points $x_{1:m}$ and targets $y_{1:n}$, but we still do not know which elements in $x_{1:m}$ are mapped to which ones in $y_{1:n}$. Define the *transformation discrepancy* S of sets $x_{1:m}$ and $y_{1:n}$:

Definition 2 *For multi-sets $x_{1:m}$ and $y_{1:n}$*

$$S(x_{1:m}\|y_{1:n}) = \frac{1}{k} \min_R \Big\{ \sum_{(x_i,y_j)\in R} C(x_i\|y_j) \Big\}, \tag{3.4}$$

where R is a k–matching of $x_{1:m}$ and $y_{1:n}$, with $k = lcm(m, n)$ is the least common multiple of m and n.

This is the average cost of mapping elements in the set $x_{1:m}$ to elements in the set $y_{1:n}$. The minimization problem in equation (3.4) is called the *minimal perfect bipartite matching* problem in computational complexity theory which can be solved in time roughly $\mathcal{O}(n^3)$ [17]. One algorithm for this problem is called the *auction algorithm*, which represents the situation as an auction: the other set of points are the bidders and the other are the items on auction. The algorithm proceeds in steps of bids until the bidder have received the items they want [9]. Another algorithm is called the *hungarian algorithm*, [9], which is much more abstarct. In this algorithm sets are marked and unmarked until the algorithm terminates.

Enlargening the sets $x_{1:m}$ and $y_{1:m}$ ad infinitum leads to a pairwise cost for probability functions. One would not venture far by assuming that probabilities are defined by infinite sets of samples: as all that can be reasoned about them is by statistics, and the plausibility of all such statistical inference demands that in the infinite limit the right conclusion can be reached.

Also there is a connection to communication, the theory of Kolmogorov complexity [50] and minimum description length principle [72]. The algorithmic minimum description length principle states that the optimal predictive model is the one which generates the observed data with the shortest description in terms of computer programs. The cost of transformation is analogous: the complexity of a string of symbols $y_{1:n}$ given another $x_{1:m}$, is the length of the shortest program that reads $x_{1:m}$ and outputs $y_{1:n}$. Consider, for instance, the following communication event: Alice and Bob both have access to a source producing a string x. Alice wants to transmit to Bob a string y, but instead of the string itself she transmits to Bob the description of the function, which Bob can apply on the string x to obtain y. Furthermore Alice might be able to send Bob for each symbol x_i individually a description of the function f_i, which, when applied on x_i, would produce the symbol y_i. The average amount of transmitted bits is the average $K(f_i)$ over the transmitted code lengths, $K(f_i)$, of the functions, f_i, during the transaction. If the sequence $y = x$, then no bits need be transmitted –Bob already knows y. Here $S(x_{1:m} \| y_{1:n})$ is the total amount of bits transmitted from Alice to Bob.

3.2.2 Extension to integral forms

Next it is shown how the above described discrete minimization process can be extended into continous models and integral forms. Such a formalism enables a much more powerful approach for the metric, and understanding of its behaviour.

The *dual* is a random transformation between two random variables on $\mathbb{D}$:

Definition 3 *The* dual *from X to Y is a random variable Θ on $\mathcal{F}$ conditional to X such that for all $x \in \mathbb{D}$:*

$$\mu_X(x) = \mathsf{E}\{\mathsf{J}_\Theta(x)^{-1}\mu_Y\big(f_\Theta(x)\big)|x\}, \tag{3.5}$$

where $\mathsf{J}_\Theta(x)$ is the Jacobian determinant of f_Θ at the point x.

The dual exists if for almost all x, $\mu_X(x) > 0$, there exists a parameter θ such that $f_\theta(x) = y$ and $\mu_Y(y) > 0$. The dual is not unique, which is easily demonstrated: let the sample spaces of both distributions be $\mathbb{R}$, and let the set of functions be the set of affine transformations on $\mathbb{R}$. If the distributions are $\nu(y) = \delta(y - y')$ and $\mu(x) = \delta(x - x')$, then any δ–function on $\mathcal{F}$ assigning positive probability to a transform of the form $\theta_1 y' + \theta_2 = x'$ is an admissible dual kernel.

In the case of a discrete sample space the dual is a transition matrix: given $p \in \mathbb{R}^d$ and $q \in \mathbb{R}^e$ with the property $\sum_i p_i = \sum_j q_j = 1$, a dual is a matrix $\Phi \in \mathbb{R}^{e\times d}$ also with $\forall i \sum_j \phi_{ij} = 1$ such that

$$\Phi p = q. \tag{3.6}$$

The solutions Φ to the system of equations 3.6, along with the constraints, can be parametrised by $de - (d + e) + 1$ parameters (de variables and $e + d - 1$ independent equations).

The transformation discrepancy S is in fact an average of C. It is the mean over a specific dual:

Theorem 1 $\lim_{m\to\infty} S(X_{1:m}\|Y_{1:m}) \xrightarrow{a.s.} s$ *and $s < \infty$ if and only if there exists a dual Θ such that $\mathsf{E}\{K(\Theta\|X)\} = s$ and for all duals Θ': $\mathsf{E}\{C(\Theta\|X)\} \leq \mathsf{E}\{K(\Theta'\|X)\}$.*

Proof: Assume there is a minimal dual Θ s.t. $\mathsf{E}\{K(\Theta|X)\} = s < \infty$ but $\lim_{m\to\infty} S(X_{1:m}\|Y_{1:m}) \xrightarrow{a.s.} s'$. If $|s'| = \infty$, this can only happen if there is $x_i \in \mathbb{D}$ such that $K(\theta|x) = \infty$ for all θ, and therefore there can be no Θ with $\mathsf{E}\{K(\Theta|X)\} < \infty$. If on the other hand $|s'| < \infty$ then there exist sequences $x_{1:\infty}$ and $y_{1:\infty}$ for which $S(x_{1:m}\|y_{1:m})$ converges to s. Define a sequence of matchings $R_1, R_2, \ldots$ such that R_m is the minimal matching for sets $x_{1:m}$ and $y_{1:m}$. For any subset $A \subseteq \mathbb{D}$ with $\mathsf{P}(X \in A) > 0$ define the subsets $A_n = \{a \in x_{1:n} \cap A\}$ and $B_n = \{b \in y_{1:n}|\exists a \in A_n : (a, b) \in R_n\}$. Take $T_n = \{\theta|x \in A_n, y \in B_n : f_\theta(x) = y\}$. Eventually then set T_n must be non-empty. Then by the law of large numbers T_n/n converges to the probability of a dual Θ' of X, and $S(X_{1:m}\|Y_{1:m})$ converges to $\mathsf{E}\{K(\Theta'\|X)\} = s'$, because each R_i is minimal $s = s'$. □

Thus at the limit when the number of samples goes to infinity, S becomes a metric on the probability distributions with some conditions on K, as will be shown later. The difference between the Kullback–Leibler distance and S is that

KL measures the difference between the code lengths of the elements in the sample space, while S is the cost that it takes to transform sample sets from one distribution to another.

3.2.3 Convergence of S

Next it can be shown that the transformation discrepancy converges with a proper selection of C, and also obtaining an average bound. Results of convergence bounds such as these are quite common in learning theory (see for example [20]), and here follow similar lines of reasoning. The most significant difference is that here the set of models, within which the bound is obtained, is the set of all models represented by finite (or infinite) sets of random samples. Parisi in [66] analysed the value of S as the sum of the matching problem with some simplifying assumptions on the distribution of the values of terms $C(x_i|y_j)$ with different values of x_i and y_j. The results of [66] however do not generalise well.

A restriction on the cost function C must first be imposed. This restriction should still cover as many forms of C as possible. It is here chosen as:

Definition 4 *The function C is* universal *if $|\mathsf{E}\{C(X|Y)\}| < \infty$ for all random variables X and Y with finite variance.*

Using such a cost function it can now be shown that the transformation discrepancy S converges:

Theorem 2 *If C is universal and $\mathcal{F}$ is complete, then for all IID random variables $X_{1:\infty}$ and $Y_{1:\infty}$ with finite variance,* $\lim_{m\to\infty} S(X_{1:m}\|Y_{1:m}) \xrightarrow{a.s.} s < \infty$.

Proof: For a complete $\mathcal{F}$ there exists a dual Θ. Since C is universal $\mathsf{E}\{K(\Theta|X)\} < \infty$ and therefore by Theorem 1 $\lim_{m\to\infty} S(X_{1:m}\|Y_{1:m}) \xrightarrow{a.s.} s < \infty$. □

Also the same can be now formulated with expectations instead of limiting values on the sizes of the sets involved:

Theorem 3 *If C is universal and $\mathcal{F}$ is complete, then for all IID random variables $X_{1:\infty}$ and $Y_{1:\infty}$ with finite variance there exists a constant c such that for all m* $\mathsf{E}\{|\mathsf{E}\{S(X_{1:m}\|Y_{1:m})\} - S(X_{1:m}\|Y_{1:m})|\} \leq \frac{c}{\sqrt{m}}$.

Proof: By Theorem 2 S converges and therefore by Theorem 1 S is an average over some dual. The value of $S(X_{1:m}\|Y_{1:m})$ is a random approximation of the integral $\mathsf{E}\{C(\Theta|X)\} < \infty$. By the Koksma-Hlawka inequality [64] we know that the error will be
$\mathsf{E}\{|\mathsf{E}\{S(X_{1:m}\|Y_{1:m})\} - S(X_{1:m}\|Y_{1:m})|\} \leq \frac{c}{\sqrt{m}}$, for some c proportional to the variation of $C(X|Y)$. □

Thus we know that the value of the transformation discrepancy entails a random error of the order $1/\sqrt{m}$ with regards to the number of samples. This is analogous to the error of numerical integration with random samples.

3.2.4 S as a Similarity Measure

The transformation discrepancy measure can as well be a useful tool in evaluating the similarity of objects, such as images or documents of text. For this the function $S(\cdot|\cdot)$ can be treated as a measure of similarity between discrete subsets of the sample space.

Denote the set of all infinite discrete subsets $x_{1:\infty}$ of $\mathbb{D}$ that satisfies $\lim_{m\to\infty} \sum_{i=1}^{m} \mu(x_i) \longrightarrow \infty$ for some probability distribution μ with finite variance as $\mathbb{S}$. When sets $x_{1:\infty}, y_{1:\infty} \in \mathbb{S}$ are such that for all subsets the limiting distributions of $x_{1:\infty}$ and $y_{1:\infty}$: $A \subset \mathbb{D}$: $\lim_{m\to\infty} |x_{1:m} \cap A|/m = \lim_{n\to\infty} |y_{1:n} \cap A|/n$ (the average number of elements in A are equal) they are considered equivalent, $x_{1:\infty} \sim y_{1:\infty}$.

Often such measures of similarity are required to be monotonic; i.e. similarity of a set does not decrease by taking the union with a third set. Here, however, there is not quite such a strong relation, and one has to settle for a weaker weighted form:

Theorem 4 *(Weighted monotonicity) If C is universal and $\mathcal{F}$ is complete then for all $x_{1:\infty}, y_{1:\infty}, z_{1:\infty} \in \mathbb{S}$: $(m+n)S(x_{1:m} \cup y_{1:n} \| z_{1:o}) \leq mS(x_{1:m}\|z_{1:o}) + nS(y_{1:n}\|z_{1:o})$.*

Proof: First note that $S(x^{\alpha}_{1:m} \| y^{\beta}_{1:n}) = S(x_{1:m}\|y_{1:n})$.

$$(m+n)S\big((x_{1:m} \cup y_{1:n})^o \| z^{m+n}_{1:o}\big) = \frac{1}{o} \sum_{(w_i,z_j)\in R_1} C(w_i\|z_j), \tag{3.7}$$

where R_1 is the minimal matching of $(x_{1:m} \cup y_{1:n})^o$ and $z^{m+n}_{1:o}$.

$$mS(x^o_{1:m}\|z^m_{1:o}) = \frac{1}{o} \sum_{(x_i,z_j)\in R_2} C(x_i\|z_j), \tag{3.8}$$

where R_2 is the minimal matching of $x^o_{1:m}$ and $z^m_{1:o}$.

$$nS(y^o_{1:n}\|z^n_{1:o}) = \frac{1}{o} \sum_{(y_i,z_j)\in R_3} C(y_i\|z_j), \tag{3.9}$$

where R_3 is the minimal matching of $y^o_{1:n}$ and $z^n_{1:o}$. Adding equations (3.8) and (3.9) there are the same terms as in (3.7) but summed over a different matching.

$$\sum_{(w_i,z_j)\in R_1} C(w_i\|z_j) \leq \sum_{(x_i,z_j)\in R_2} C(x_i\|z_j) + \sum_{(y_i,z_j)\in R_3} C(y_i\|z_j) \tag{3.10}$$

$$= \sum_{(w_i,z_j)\in R_2\cup R_3} C(w_i\|z_j) \tag{3.11}$$

as the pairing R_1 is by definition optimal. □

The relation in Theorem 4 is an equality when the samples are equivalent. Easily, if $x_{1:m}^{\alpha} = y_{1:n}^{\beta}$ for some α and β, in which case R_1, R_2 and R_3 in the proof of Theorem 4 are the same, but there is a more general result:

Theorem 5 *If C is universal and $\mathcal{F}$ is complete and $x_{1:\infty} \sim y_{1:\infty}$ then for all $z_{1:\infty} \in \mathbb{S}$: $\lim_{m\to\infty} S(x_{1:m}|z_{1:n}) = \lim_{m\to\infty} S(y_{1:m}|z_{1:n})$.*

Proof: Let $X_{1:\infty}$ be distributed according to the limiting distribution of $x_{1:\infty}$. $\lim_{m\to\infty} S(X_{1:m}\|z_{1:n}) = \lim_{m\to\infty} S(x_{1:m}\|z_{1:n})$. Then note that $y_{1:\infty}$ has the same limiting distribution. □

Finally it can be seen that S is in fact a metric on the space of probability distributions, at least on the set where models are represented by sample sequences.

Theorem 6 *If C is a metric on $\mathbb{D}$, then S is a metric on $\mathbb{S}$.*

Proof: First noting that $S(x_{1:\infty}\|y_{1:\infty})$ is the limiting sum of elements $C(x_i\|y_j)$. It can be seen that of the three parts following the Definition1 of a metric of which the the 3^{rd} property, the triangle inequality, requires the most attention: Assuming that $\forall a, b, c \in \mathbb{D} : C(a\|b) + C(b\|c) \geq C(a\|c)$ it is needed to be shown that $\forall x_{1:\infty}, y_{1:\infty}, z_{1:\infty} \in \mathbb{S} : S(x_{1:\infty}\|y_{1:\infty}) + S(y_{1:\infty}\|z_{1:\infty}) \geq S(x_{1:\infty}\|z_{1:\infty})$. For some $m \geq 1$ take finite subsets sets $x_{1:m}$, $y_{1:m}$ and $z_{1:m}$. Let the optimal matching of $x_{1:m}$ and $y_{1:m}$ be R_1, of $y_{1:m}$ and $z_{1:m}$ be R_2 and of $x_{1:m}$ and $z_{1:m}$ be R_3. Take other subsets $a_{1:n} \subseteq x_{1:m}$, $b_{1:n} \subseteq y_{1:m}$ and $c_{1:n} \subseteq z_{1:m}$ for some $1 \leq n \leq m$ such that for all $1 \leq i \leq n : (a_i, b_i) \in R_1$, $(a_i, c_i) \in R_3$ and for $1 \leq i \leq n-1 : (b_i, c_{i+1}) \in R_2$ and $(b_n, c_1) \in R_3$. The sets $a_{1:m}$, $b_{1:m}$ and $c_{1:m}$ can always be found, as the item a_1, for instance, can be chosen. This item will have some pairs b_1 and c_1, but the pair of a_1 in R_2 is not b_1 then the loop is continued by adding the item b_2 for which (a_1, b_2) is in R_2. This process is continued until for some b_k the pair (a_1, b_k) is in R_2, which will inevitably be found as the sets $x_{1:m}$, $y_{1:m}$ and $z_{1:m}$ were finite. The sums in $S(x_{1:m}\|y_{1:m})$, $S(x_{1:m}\|z_{1:m})$ and $S(y_{1:m}\|z_{1:m})$ are composed of sums over such cyclic subsets. Assume by contradiction that

$$\sum_{i=1}^{n} C(a_i\|b_i) + \sum_{i=1}^{n-1} C(b_i\|c_{i+1}) + C(b_n\|c_1) < \sum_{i=1}^{n} C(a_i\|c_i). \tag{3.12}$$

However on the left hand side of Equation 3.12 it is known that for all $i : C(a_i\|b_i) + C(b_i\|c_{i+1}) \geq C(a_i\|c_{i+1})$. Thus for the left-hand-side of Equation 3.12 it applies that

$$\sum_{i=1}^{n} C(a_i\|b_i) + \sum_{i=1}^{n-1} C(b_i\|c_{i+1}) + C(b_n\|c_1) \geq \sum_{i=1}^{n-1} C(a_i\|c_{i+1}) + C(a_n\|c_1), \tag{3.13}$$

which in turn by Equation 3.12 must be less than $\sum_{i=1}^{n} C(a_i \| c_i)$. However the matching R_3, part of which the pairs in the sum $\sum_{i=1}^{n} C(a_i \| c_i)$ are, was assumed optimal. Thus the triangle inequality applies for all m, and then also for the limit $m \to \infty$. □

3.2.5 Example

These ideas can be illustrated with a simple example. Here a metric form of S is chosen, such that for a very simple set of considered distributions this metric would allow a consistent model selection.

First take a metric in 1 dimensional space, and then apply it to a simple set of distributions. Let the sample space be the half-space $\mathbb{R}^+ = \{x : x \in \mathbb{R}, x > 0\}$. Take the set of functions as the set of multiplications by scalar: $f_\theta(x) = \theta x$. Define the cost of f_θ by

$$K(\theta \| x) = |\log \theta|, \tag{3.14}$$

and then a pairwise form C can be derived as explained in the beginning of Section 3.1 resulting to:

$$C(x \| y) = |\log \frac{y}{x}|. \tag{3.15}$$

The identity map f_1 will be assigned the minimal discrepancy. This metric is symmetric $C(x \| y) = C(y \| x)$, and the triangle inequality applies with equality. For $x, y, z \in \mathbb{R}^+ : x \leq y \leq z$:

$$\begin{aligned} C(x \| y) + C(y \| z) &= |\log \tfrac{y}{x}| + |\log \tfrac{z}{y}| \\ &= \log y - \log x + \log z - \log y \\ &= \log \tfrac{z}{x} = C(x \| z). \end{aligned} \tag{3.16}$$

It can then be shown that the minimal dual can be chosen as a single continuous function ψ: the dual distribution $\mu_\Theta(\theta|x) = \delta\big(\psi(x) - f_\theta(x)\big)$, for a spesific function ψ, which is explained below. First note that the matching in equation (3.4) preserves the order of the elements:

Proposition 1 *If $x_1 \leq x_2$ then $\psi(x_1) \leq \psi(x_2)$.*

Proof: In the following assume that $x_1 < x_2$, and writing $y_1 = \psi(x_1)$, $y_2 = \psi(x_2)$, for a contradiction assume that $y_2 < y_1$. Using the equation (3.16), one needs to check the following four cases:

i) $x_1 \leq y_2 < y_1 \leq x_2$. Looking at the pairs (x_1, y_1) and (x_2, y_2) that must occur in the sum of equation (3.4):

$$C(x_1 \| y_1) + C(x_2 \| y_2) \tag{3.17}$$

$$= C(x_1 \| y_2) + C(y_2 \| y_1) + C(x_2 \| y_1) + C(y_1 \| y_2) \tag{3.18}$$

$$= C(x_1 \| y_2) + C(x_2 \| y_1) + 2C(y_1 \| y_2) \tag{3.19}$$

$$\geq C(x_1 \| y_2) + C(x_2 \| y_1). \tag{3.20}$$

Therefore, the optimal matching in equation (3.4) cannot contain the pairs (x_1, y_1) and (x_2, y_2).

ii) $y_2 \leq x_1 < x_2 \leq y_1$. Then

$$C(x_1\|y_1) + C(x_2\|y_2) \tag{3.21}$$
$$= C(x_1\|x_2) + C(x_2\|y_2) + C(x_2\|x_1) + C(x_1\|y_1) \tag{3.22}$$
$$= C(y_1\|x_1) + C(y_2\|x_2) + 2C(x_1\|x_2) \tag{3.23}$$
$$\geq C(y_1\|x_1) + C(y_2\|x_2). \tag{3.24}$$

iii) $y_2 \leq x_1 \leq y_1 \leq x_2$. Then

$$C(x_1\|y_1) + C(x_2\|y_2) \tag{3.25}$$
$$= C(x_1\|y_1) + C(x_2\|x_1) + C(x_1\|y_2) \tag{3.26}$$
$$= C(x_1\|y_1) + C(x_2\|y_1) + C(y_1\|x_1) + C(x_1\|y_2) \tag{3.27}$$
$$= C(x_1\|y_2) + C(x_2\|y_1) + 2C(x_1\|y_1) \tag{3.28}$$
$$\geq C(x_1\|y_2) + C(x_2\|y_1). \tag{3.29}$$

iv) $y_2 < y_1 \leq x_1$ then

$$C(y_1\|x_1) + C(x_2\|y_2) \tag{3.30}$$
$$= C(x_1\|y_1) + C(x_2\|y_1) + C(y_1\|y_2) \tag{3.31}$$
$$= C(x_1\|y_1) + C(x_2\|x_1) + C(x_1\|y_1) + C(y_1\|y_2) \tag{3.32}$$
$$= C(x_1\|y_2) + C(x_2\|y_1) + 2C(x_1\|y_1) \tag{3.33}$$
$$\geq C(x_1\|y_2) + C(x_2\|y_1). \tag{3.34}$$

□

As the elements are in the same order, as long as neither distribution contains atoms, the dual mapping has to be a deterministic function.

Next consider uniform distributions with a parameter $\theta > 0$:

$$\mu(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{if } x \in (0, \theta] \\ 0 & \text{otherwise} \end{cases} . \tag{3.35}$$

Let $X_{1:\infty}$ have a distribution $\mu(\cdot|\theta)$ and let $Y_{1:\infty}$ have a distribution $\mu(\cdot|\rho)$. When the functions are scalar multiplications, and because the matching is order–preserving at the limit $m \longrightarrow \infty$ there is a matching defined by a continuous function $\psi(x) = \frac{\rho}{\theta}x$, and then

$$\lim_{m\to\infty} S\{X_{1:m}\|Y_{1:m}\} = \int_0^\theta \mu(u|\theta)K(\frac{\rho}{\theta}|u)\,du \tag{3.36}$$
$$= \int_0^\theta \frac{1}{\theta}|\log(\frac{\rho}{\theta})|\,du \tag{3.37}$$
$$= |\log(\frac{\rho}{\theta})|. \tag{3.38}$$

This is a proper metric for the considered distributions. Thus the minimal transformation discrepancy estimate ρ of θ is consistent. Note that this is the same as the Kullback–Leibler divergence $KL(X|Y)$, if $\theta \leq \rho$, but if $\theta > \rho$ then $KL(X|Y) = \infty$.

3.2.6 Discussion

In this section a metric on probability distributions based on a metric on the sample space was introduced. It was shown that this metric can be calculated as a minimised sum of the pairwise costs of the samples from the two distributions. Also it was established that with some restrictions on the underlying metric this sum converges into an expectation of the pairwise cost. This metric can serve in model selection or classification of random groups. A small example was used to show some analytic results that can be obtained with this metric.

Chapter 4

Uncomputable Likelihoods

There are cases where the likelihood term $\mu_X(x|\theta)$ in equation (2.1) cannot be practically computed. Then the inference can still be carried out with a simulational method by sampling elements $y_{1:n}$ according to the distribution μ_X and comparing them to the observations $x_{1:m}$. This problem was addressed by Diggle and Gratton [21], and the term *indirect inference* was advanced by Gourieroux et al. [31]. In this chapter a new approach is suggested which is based on the metric considerations of chapter 3 [44].

This problem arises for example when one attempts to infer the generating dynamics of a moving particle. The dymanics of the particle is defined by a possibly stochastic system of differential equations. In this system the particle begins its movement from a random initial point and then follows a distinct trajectory. Here the motion can be considered to be due to one of many possible models, of which correct one is sought. However, the dynamical process, which governs the action of the particle, may not be time–invertible and thus cannot be traced back to its origin when only the final resting place of the particle is seen. Also the distribution of the random initial condition may not be propagated analytically to the final state. The lack of invertibility and the randomness of the system prevents computing the probability of the observations.

This kind of estimation of the likelihood function also provides means for a goodness–of–fit test. Given two sets of samples for which the approximation of the likelihood can be calculated can be viewed as the probability that they have the same distribution. Gelman et al. referred to such an approach as *realised discrepancies* [29].

The rest of this chapter is organised as follows. First methods for estimating the posterior probability of observations, when the exact probability is not practical to compute, are discussed. At the end of the chapter there are simple artificial examples in which the methods are applied.

4.1 The estimation model

The estimation of the likelihood function can be constructed so that for each data sample from the model, called a *replica*, a noisy observation is made. The noise may change the replica a small amount, relative to the metric of the sample space. When the variance of this added noise is brought to zero, which in this approximation is required to happen as the number of replicas approaches infinity, naturally the the approximating likelihood should approach the noisless likelihood.

The model is the same as in Chapter 2 Equation 2.1:

$$\mu_{\Theta'}(\theta|x_{1:m}) \propto \mu_{X|\Theta}(x_{1:m}|\theta)\mu_{\Theta}(\theta), \tag{4.1}$$

where the value of the likelihood term $\mu_{X|\Theta}(x_{1:m}|\theta)$ was not possible to be computed, but from which samples can be generated.

The approximating model has additional variables: the latent random variable Y is distributed identically as X and are conditional to the model parameter θ, and the observable variables $\hat{X}_{1:m}$ are conditional to $Y_{1:n}$. With this notation the original model is discriminated from the approximation. The hierarchical posterior probability density of θ, and the replicas $Y'_{1:n}$, is then:

$$\mu_{\hat{\Theta}',Y'_{1:n}}(\theta, y_{1:n}|x_{1:m}, \rho) \propto \mu_{\hat{X}_{1:m}}(x_{1:m}|y_{1:n}, \rho)\, \mu_Y(y_{1:n}|\theta)\, \mu_{\Theta}(\theta). \tag{4.2}$$

Let us concentrate on the modelling of the right-hand side latent variable likelihood $\mu_{\hat{X}_{1:m}}(x_{1:m}|y_{1:n}, \rho)$. In the models to follow one should note that the nuissance parameter ρ is not usually identifiable by the data, and thus is mostly defined by its priors.

When modelling the predictive distribution of X, also the latent $Y_{1:n}$ needs to marginalised. This is done as:

$$\mu^n_{\hat{\Theta}'}(\theta|x_{1:m}, \rho, M) = \int \mu_{\hat{\Theta}',Y'_{1:n}}(\theta, y_{1:n}|x_{1:m}, \rho, M)\, dy_1 \ldots dy_n, \tag{4.3}$$

where n is superscripted to the left hand side because the dependency on the number of replicas remains. The integration of the equation can be done efficiently by MCMC simulation.

4.2 Kernel estimate

Here is presented the first of the two methods suggested by the hierarchy of equation (4.2). By using the standard method in Bayesian inference the latent variables are marginalised. In fact, it then becomes evident that this corresponds to the standard kernel estimate of the likelihood function.

For simplicity set $m = n$, and $\hat{X}_{1:m}$ are IID and for each i conditional to a latent variable y_i. In this case the realization of each Y_i as individually perturbed

to produce the observation of $\hat{X}_i$. From this scheme follows the formula for the latent variable likelihood:

$$\mu_{\hat{X}}(x_{1:m}|y_{1:n},\rho) = \prod_{i=1}^{n} \mu_{\hat{X}_i}(x_i|y_i,\rho). \tag{4.4}$$

When the sample space is d–dimensional, each factor of the product is defined by a *kernel density function* $\omega_\rho(x) = \omega(\frac{x}{\rho})/\rho^d$ with bandwidth ρ. This kernel function is maximised at zero, and is usually symmetric. One very often uses the Gaussian kernels of the form $\omega(x) \propto e^{-\|x\|^2}$. Then the latent variable likelihood gets the form

$$\mu_{\hat{X}}(x|y,\rho) = \omega_\rho(x-y). \tag{4.5}$$

When performing the marginalization of (4.3) over Y' with n MCMC samples $\{y^1,\ldots,y^n\}$ one gets

$$\mu_{\hat{X}}(x|\theta,\rho) = \int \mu_{\hat{X}}(x|y,\rho)\,\mu_Y(y|\theta)\,dy \approx \frac{1}{n}\sum_{j=1}^{n} \omega_\rho(x-y^j) \tag{4.6}$$

as an approximation of the likelihood with a kernel estimate with bandwith ρ.

The bandwith can be asymptotically chosen as a plug-in estimate:

$$\rho = \frac{c}{\sqrt[d+4]{n}}, \tag{4.7}$$

where c depends on $\|\nabla^2 p(x|\theta)\|_{L_2}$ and the choice of kernel [82]. When n goes to infinity the estimate on the right-hand side of (4.6) approaches the likelihood $\mu_X(x|\theta)$, recovering the original posterior of equation 2.1.

4.3 Indirect inference

If one calls as a direct approach the computation of $\mu_X(x|\theta)$, when possible, and using this knowledge to infer θ, then the alternative, when samples of X can be drawn, can be referred to as indirect. Construct a model where all $X_{1:m}$ are conditional to all $y_{1:n}$, by defining a binding probability density $\mu_{\hat{X}_{1:m}}(x_{1:m}|y_{1:n},\rho)$, where one assumest that $\hat{X}_i$ are exchangeable, but not independent.

4.3.1 Method of Gourieroux

Gourieroux et al. proposed the first method for this kind of a problem and dubbed it indirect inference [31]. To start they use the model to generate a set of samples. For these samples they compute a statistic, and then compare this statistic to the corresponding statistic of the observations. The best model is chosen by comparing the statistics of the simulations to the statistics to the observations, by a metric

on the statistics. In this framework the difficulties of the computation of the likelihood have always been about marginalization over some latent variables, which presently can well be solved by MCMC simulation, see for example [73, 60].

In this methodology, in absence of a way to evaluate the function $\mu_X(x|\theta)$, one uses a simpler model (called the *auxiliary model*), with a parameter $\zeta \in Z$. One assumes that this parameter can be easily estimated as $\hat{\zeta}(x_{1:m})$, which is a *statistic* of the observations $x_{1:m}$ —a measure computed from a set of data values. By generating a set of random samples $y_{1:n}$ of $Y_{1:n}$ one then tries to find a parameter θ that minimises distance of the parameters, relative to some metric on Z. The Gourieroux et al. defined a *binding function* as that closest parameter given the observations.

In the spirit of what follows we could define a probability in the model (4.2):

$$\mu_{\hat{X}_{1:m}}(x_{1:m}|y_{1:n}) = Ne^{-m\delta(\hat{\zeta}(x_{1:m})|\hat{\zeta}(y_{1:n}))}, \tag{4.8}$$

where N is a normalizing constant. The multiplier m is in the exponent to make the probability dependent on the number of samples, which is required for the posterior probability to converge correctly. One should note that this dependence is not otherwise present in the formula through the metric δ.

The critisism of this method is firstly: closeness in the metric used in the definition of the binding function does not imply that the functions behave in a similar manner. Rather, it might be wiser to use metrics on probability distributions, like total variance or the Kullback–Leibler divergence, but this would make the evaluation of the distance more complicated. Secondly: in order for this method to be a useful way of estimating θ in terms of an easier estimate $\hat{\zeta}$, it should be clear that the simpler model must then be a sufficient statistic for θ. An analysis of this may be a difficult task since the likelihood with the model, with parameter θ, is hard to compute. Also if the binding provides a consistent estimator for θ, it implies that the space of distributions defined by Θ has to be a subset of the models Z. However it does not assume a metric on the sample space which may be a benefit.

4.4 Inference with transformations

Another variant of the indirect inference technique, where the problems mentioned above are corrected is to utilise a metric on the sample space. In this method the metric binding function is replaced with a metric on probability distributions. A practical form, easily computable for sequences of random samples, is the *transformation discrepancy* measure S of equation 3.4.

Basically one would look at the probability of seeing two sets of samples a given discrepancy apart subject to a hypotesis that they have the same distribution. Naturally this probability depends on the distribution in question, and the metric used. Thus it would be hard to say anything properly general, but one can guess

that a form decreasing with distance is appropriate. In fact one can take what is given in physics, the exponential of the negative distance, for this distribution:

$$\mu(x_{1:m}|y_{1:m}) \propto e^{-mS(x_{1:m}|y_{1:m})}. \tag{4.9}$$

The exponent m is also required to get a dependency on the number of samples. Without this factor the inference based on this probability would be consistent but not efficient; the average is correct but it has too wide variance.

A similar approach to the kernel estimate is to marginalise over the latent $Y_{1:n}$ as in section 4.2. When adding an additional multiplier, a weight ρ, which now is proportional to m, on the exponent of the right hand side. of the equation (4.9) one has

$$\mu_{\hat{X}_{1:m}}(x_{1:m}|y_{1:n}, \rho) = Ne^{-\rho S(x_{1:m}|y_{1:n})}, \tag{4.10}$$

where N is a normalizing constant. With the theory of simulated annealing [1] we can again recover the original posterior of equation (2.1), as stated in Theorem 7.

Theorem 7

$$\lim_{\rho\to\infty} \mu_{\Theta'}(\theta|x_{1:m}, \rho) \propto \mu_X(x_{1:m}|\theta)\, \mu_\Theta(\theta). \tag{4.11}$$

Proof: When $\rho \longrightarrow \infty$, $\mu_{\hat{X}_{1:m}}(x_{1:m}|y_{1:m}, \rho)$ approaches the δ–distribution such that $\delta(x_{1:m}, y_{1:m}) = 0$ if $x_{1:m}$ is not the same sequence as $y_{1:m}$ upto the order of the symbols. Further $\int\int \mu_{\hat{X}_{1:m}}(x_{1:m}|y_{1:m}, \rho)\mu_Y(y_{1:m}|\theta)\, \mu_\Theta(\theta)\, dy_1 \ldots dy_m = \mu_X(x_{1:m}|\theta)\, \mu_\Theta(\theta)$. □

A different perspective is to marginalise the latent $Y_{1:n}$ to compare different values of θ. When the number of replicas is increased, the measure $S(x_{1:m}|y_{1:n})$ becomes no longer a random variable. This is the content of the next Theorem 8.

Theorem 8 *If C is universal and $\mathcal{F}$ is complete and Y has finite variance then*

$$\lim_{n\to\infty} \frac{\mu_{\hat{\Theta}', Y'_{1:n}}(\theta, Y_{1:n}|x_{1:m}, M)}{\mu_Y(Y_{1:n}|\theta)} \xrightarrow{a.s.} \lim_{n\to\infty} \mu^n_{\hat{\Theta}'}(\theta|x_{1:m}, M). \tag{4.12}$$

Proof: First, Theorem 3 implies that $S(x_{1:m}|Y_{1:n})$ converges almost surely to some constant $s < \infty$ when $n \longrightarrow \infty$ and Y has finite variance. Then the expression on the left hand side is a limit of:

$$\mu_{\hat{X}_{1:m}}(x_{1:m}|Y_{1:n}, \rho)\, \mu_\Theta(\theta), \tag{4.13}$$

which according to its definition in (4.9) depends on n through $S(x_{1:m}|Y_{1:n})$ and thus converges to some value q by Theorem 3. Likewise the right hand side is the average over the equation (4.13), which by theorem 3 also converges to the same value q. □

Theorem 8 implies that the term $\mu_Y(y_{1:n}|\theta)$ can be ignored provided that there are sufficiently many replicas when calculating the posterior marginal (4.3).

The transformation discrepancy can be used to evaluate the posterior goodness–of–fit posterior p–value, which here is the probability of observing a larger discrepancy of two sets, which are identically distributed, than the discrepancy of $x_{1:m}$ and replicas [29]. This as an average:

$$p_{\text{val}}(x_{1:m}) = \mathsf{E}\{\mathsf{P}(S(Z_{1:m}|Y_{1:n}) \geq S(x_{1:m}|Y_{1:n}))\}, \tag{4.14}$$

where $Z_{1:m}$ and $Y_{1:n}$ are IDD, and the average is also taken over the posterior distribution of the model parameters. This is the average probability of observing values of S larger that $S(x_{1:m}|Y_{1:n})$. A p–value close to 0 would imply a good fit: an unlikely thing to see larger discrepancies than that of the observations. This can be easily simulated with MCMC by generating samples θ_i from the posterior and for each generated θ_i generate two sets of replicas $z_{1:m}$ and $y_{1:n}$. The posterior p–value is estimated as the ratio of incidences where $S(z_{1:m}|y_{1:n})$ was larger than $S(x_{1:m}|y_{1:n})$.

4.4.1 Examples

Uniform distribution

Taking as the first example the simplest: the underlying true model is a uniform distribution over the real interval [0, 1] — a distribution hardly uncomputable but which is used for the sake of an exeample. Using the three methods discussed, the kernel model, the transformation model and the Gourieroux model in the succession is demonstrated that the distribution can be estimated using these indirect methods.

In the kernel model take the Gaussian kernels, and Scott's rule for the bandwidth: $\rho = \tilde{y}_{1:n} n^{-1/5}$ [80], where $\tilde{y}_{1:n}$ is the sample standard deviation of $y_{1:n}$.

In terms of Gourieroux et al. the Gaussian distribution as the auxiliary model with the estimated parameters

$$\zeta(x_{1:n}) = \begin{pmatrix} \bar{x}_{1:m} \\ \tilde{x}_{1:m} \end{pmatrix}, \tag{4.15}$$

where $\bar{x}$ is the sample mean of $x_{1:m}$. Under the assumption that the true distribution is uniform the above mapping does provide sufficient statistics to infer the bounds of the uniform distribution (the parameters of a uniform distribution given the mean and standard deviation are $\big(\min(A, B)\ \ \max(A, B)\big)^T$, where $A = \mu_x - \sqrt{3}\sigma_x$ and $B = \mu_x + \sqrt{3}\sigma_x$).

The measure of the magnitude of the error by the L_1-measure of probability distributions is:

$$\varepsilon = \int |\nu(u) - \hat{\nu}(u)|\, du, \tag{4.16}$$

where ν is the true probability distribution function, and $\hat{\nu}$ is the estimate obtained from the data.

Assume that the lower bound lies in the set $\{-1, -0.9, \ldots, 0.5\}$ and the upper bound in $\{0, 0.1, \ldots, 2.5\}$, and that a priori all these values are equally probable. When there are $m = 10$ observations, one needs to generate n replicas for each of these 220 different models. The values of the kernel estimate, the transformation discrepancy S estimate, and the Gourieroux estimate, for the exact same observations x and replicas y on this grid serve for comparison. To obtain statistics the process for 100 observation sets is repeated. The estimation was trialed with the L_1-measure ε in (4.16) of the true distribution against the Bayesian posterior predictive distribution of equation (2.2). The efficiency of the maximum likelihood point-estimates: choosing the model maximizing the likeilihood, or its estimate, is also of interest here.

In Figure 4.1 is plotted the average error as a function of the number of replicas for each three methods. One should pay attention to the efficiency of the different methods to use the information in the n replicas. It can be seen that the indirect methods reach the base level of the average error of the Bayesian posterior with the true likelihood with 20 replicas, and they are more efficient than the kernel method in utilizing the replicated data. The fact that the kernel goes below the base line at $n = 16$ can be accounted for by statistical fluctuations rather than that the kernel method would be capable of extracting more information than the true likelihood. This assumption is supported by Figure 4.2, where is plotted the standard deviation of the error. It can be seen that the kernel method is roughly 3 times more volatile than the others.

In Figure 4.3 is plotted the average error for the maximum likelihood estimates, with the true ML–estimate (a uniform distribution on $[\min(x_{1:m}), \max(x_{1:m})]$) as the base line. All the methods are in this sense about equivalent, reaching the base line after $n = m = 10$ replicas, and outperforming after that mainly because of the finite grid for the parameters: a quantization effect. Also the standard deviations of the maximum likelihood estimates are similar as can be seen from Figure 4.4. Thus these three approximation functions have their extreme values at about the same location.

The Lorenz system

As the second example is chosen a chaotic system having a strange attractor. A chaotic system is such that a small perturbation in the location of a particle induces a large deviation in the future position of the particle. Due to the nature of the system it becomes untraceable to know where a given particle was before, within finite accuracy. For this model analytic results are hard to come by and therefore only the MCMC simulation results are shown.

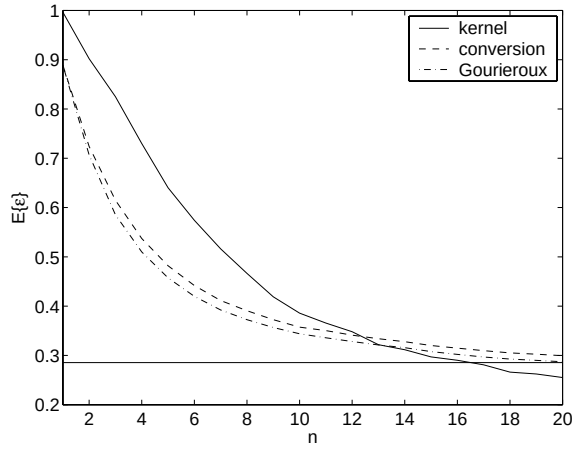

Figure 4.1: The average L_1-error ε of the posterior predictive densities as a function of the number of replicas n. The vertical line is the average error when the true likelihood is used.

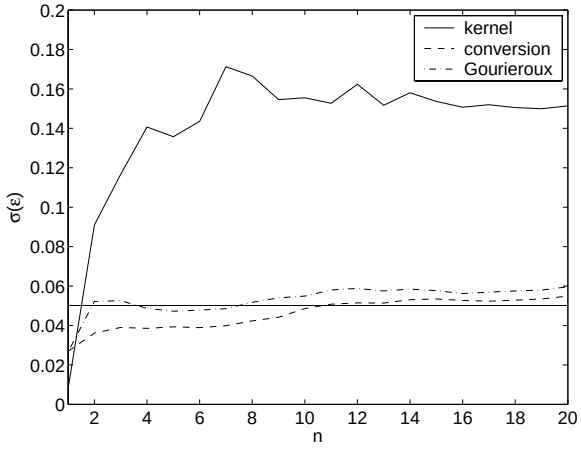

Figure 4.2: The standard deviation of the L_1-error ε of the posterior predictive densities as a function of the number of replicas n. The vertical line is the standard deviation of the error when the true likelihood is used.

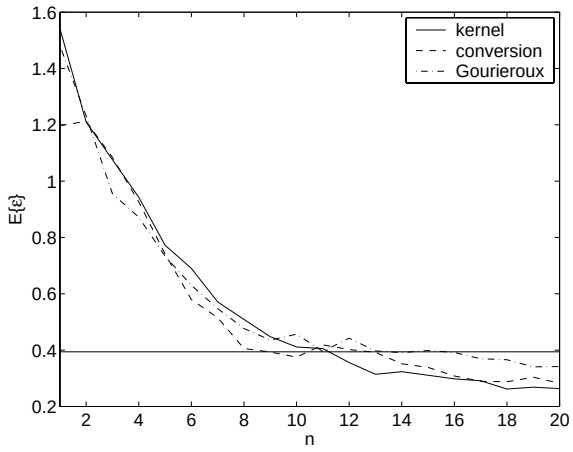

Figure 4.3: The average L_1-error ε of the maximum likelihood estimates as a function of the number of replicas n. The vertical line is the average error of the true ML–estimate.

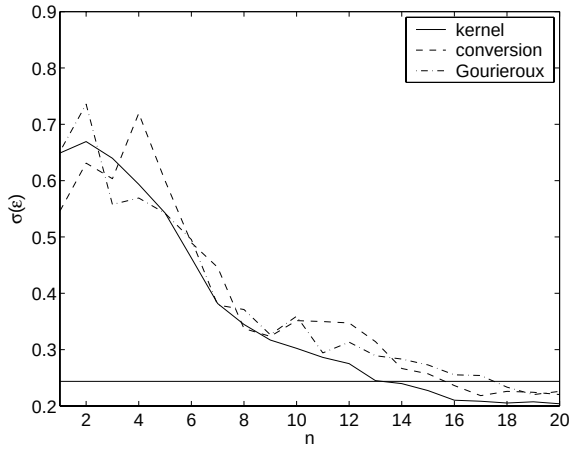

Figure 4.4: The standard deviation of the L_1-error ε of the maximum likelihood estimates as a function of the number of replicas n. The vertical line is the standard deviation of the error for the true ML–estimate.

The Lorenz attractor [79] is a chaotic system of differential equations:

$$\begin{aligned} \dot{z_1} &= q_1(z_2 - z_1) \\ \dot{z_2} &= q_2 z_1 - z_2 - z_1 z_3 \ , \\ \dot{z_3} &= z_1 z_2 - q_3 z_3 \end{aligned} \tag{4.17}$$

where q_1, q_2, and q_3 are the model parameters and z_1, z_2 and z_3 are spatial locations.

The standard choice for the parameters are $q_1 = 10, q_2 = 28$ and $q_3 = 8/3$. The initial value of z is $(0\ 0\ 0)^T + \eta$, where $\eta \sim N(0, 10^{-1})$ is a Gaussian noise term. For statistical observation we take m points from a numerical simulation of $T = 10^4$ time steps, sampled at intervals of $\Delta t = T/m$, but assuming here that the order of the observed samples is not known, or is not relevant.

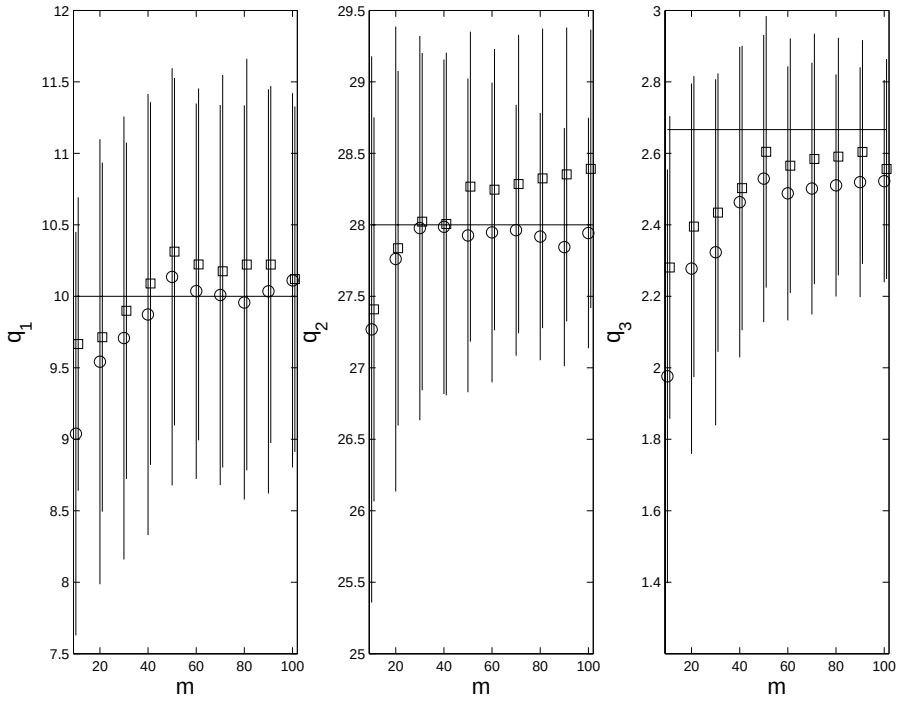

Figure 4.5: Plot of the average posterior estimated paramers for the Lorenz system with the transformation (∘) and the kernel method (□) as a function of m. The vertical lines are the standard deviations of the estimates. The horizontal lines are the true values.

The hypothesis is that the first parameter is in $q_1 = \{5, 6, \ldots, 13\}$, the second in $q_2 = \{23, 24, \ldots, 31\}$ and the third in $q_3 = \{1, 1\frac{1}{3}, 1\frac{2}{3}, \ldots, 3\frac{2}{3}\}$, each in equal prior probability. We generate $n = 100$ samples for each of these 729 cases of parameters, and evaluate the kernel estimate of the likelihood and the value of

S for these points. Then one can obtain the Bayesian predictive distribution in equation (2.2). In case of the kernel method Gaussian kernels can be used and choosing the bandwidth with Scott's rule [80]. The method of Gourieroux cannot be applied here because there is no simple sufficient statistic that is known that would enable its use.

As an estimate take the average parameters over the posterior probability. In Figure 4.5 are plotted the averages of the estimates over the data and their standard deviations for the transformation and kernel estimates as a function of the number of observations m, when the number of replicas is $n = 100$. The system parameters can be estimated with the observations within reasonable bounds. It can be seen that the first and the second parameters are unbiased after roughly 50 samples, but the third seems interestingly still biased after 100 samples, which can be warranted as a property of the system rather than a flaw in the methods. While the kernel gets on average closer with the third parameter it has a larger variance.

4.4.2 Discussion

In this chapter it was shown that even if the value of the likelihood function cannot be computed the inference can be carried out by adding a latent layer to the hierarchy. This lead to two estimates: the kernel estimate and the indirect inference. The examples demonstrated how this methodology can be applied, and that the correct model can be selected. The Gourieroux method has its distinct problems requiring a binding function, which must provide sufficient statistics which in turn may be difficult to prove as the original model had difficulties in the analysis. The kernel and the transformation methods are about equal in performance relative to the number of samples drawn from the likelihood.

Chapter 5

Spreading on random graphs

A graph is said to have the small–world propery if the average distance between the sites is small compared to the size of the graph and average number of connections in the sites. The distance between two sites is the length of a path that connects them. Another interesting property of such complex networks is the distribution of the degree, the number of connections to a given site. Graphs as associative constructions can serve as models for many kinds of natural systems, such as social relationships [55] or computer networks [24], and their analysis sheds light on phenomena like epidemic spreading, data network vulnerability and collapse of transportation routes. This dissertation provides analysis on these phenomena, specifically what is later called spreading dynamics [45, 46].

Real–world networks are commonly characterised by a large number of parameters, but in relation to small–world networks is the average distance between their sites [63, 70, 22]. It has turned out that there is a rich family of small–world networks which differ in many other respects. For example, the degree distribution of the sites is Poissonian for the Watts–Strogatz graphs while many real-world networks are often scale–free, i.e., they have a power law decay for the degree distribution. To explain this behaviour models of preferential growth have been introduced [5, 76]. Thus small–world networks are very interesting graphs not only because of these properties of distance and degree, but also because they are simple models that sometimes provide exact solutions [62, 46] and because they are directly appliable, e.g. in polymer physics [37].

Properties of random graphs can be largely investigated by looking at their response to dynamic randomness of some sort in simulations. Such simulations could correspond to performing a random walk in the graph, or passing messages between random sites. Once that is done simply looking at the statistics one can categorise the network. There are two different forms of disorder for dynamic simulation, i.e., *annealed* and *quenched* [67]. In the case of quenched disorder the graph is generated before the actual dynamic process that is studied, and is

kept fixed during it. In annealed disorder the connections of the graph are randomly updated during the process. The annealed dynamics can often be elegantly expressed with stochastic equations of updates, where the next state depends only on the previous state. Thus the annealed model provides more tools for explicit analysis than the models of quenched disorder [53, 46].

Also of great importance in the analysis of these graphs is the concept of *mean–field approximation*. Mean–field in statistical physics means that the internal interaction forces are replaced with an external field. Here it means specifically that the randomness of the different realizations of the graphs is replaced with their corresponding average. This sort of analysis was done on small–world networks by Newman et al. [61], and for the Barabási model by Fronczak et al. [26]. Both of these papers analysed the clustering phenomenon, i.e., the behaviour of the formation of large connected components.

The spreading phenomena in networks are perhaps one of the most direct examples of dynamical processes reflecting the small–world properties. In direct spreading of e.g. a disease, the sites of the graph get infected by the rule that infection propagates each time step to all uninfected neighbours of already infected sites [59]. Then the simplest example of non–trivial dynamics could be that of a diffusing particle in the network. This in turn is related to the intensively studied process of random walks in random environments as is evident from the two comprehensive volumes by Hughes [34], and ben-Avraham and Havlin [6]. Recently some related papers have been published on the issue of diffusion in small–world networks, see for example the study of spectral properties of the Laplacian in them [57]. In addition Pandit and Amitkar [65] have presented some numerical and analytic results for the spreading phenomenon being characterised by the average access time to the sites of the system. Furthermore, Jasch and Blumen [35] published simulation results for spreading in small–world networks using random walk dynamics with the main quantity of interest being the average number of distinct sites visited at a given time. This work was also done independently by the author, Kertész and Kaski [45], and obtained scaling more accurately than what was reported by Jasch and Blumen.

The rest of this thesis deals with the concept of random graphs. First in this chapter there is a cursory view of the models, followed by an analysis of the spreading phenomenon on small–world networks with quenched and annealed disorder.

5.1 Models of random graphs

Models of random graphs are usually described by a formation process, e.g. adding new connections between the sites by a rule which may depend on the previously added connections [23]. Below there are the three simplest basic models -the

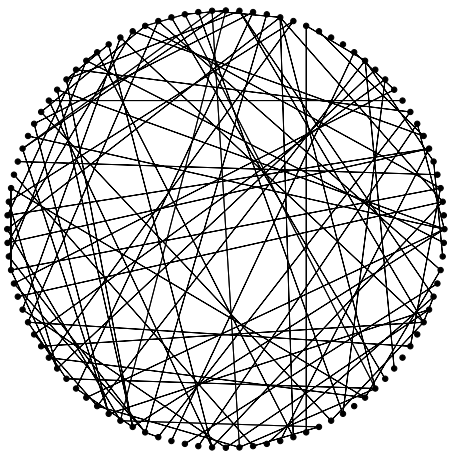

Figure 5.1: An Erdős–Rényi graph with 100 sites.

Erdős & Rényi, the Watts & Strogatz, and the Barabási networks.

5.1.1 Erdős and Rényi graphs

The Erdős and Rényi graph has a distribution on the connections with a fixed number of sites n. In this model there is a probability p, which is a function of n, for an connection to exist between any two vertices i and j [10]. Naturally this means that on average there are about pn^2 connections in such a graph.

Erdős and Rényi offered a simple proof for the remarkable phase transition relative to the value p in the limit when $n \to \infty$: when p is very small the graph is quite obviously very disconnected, but when p crosses over a threshold a large component emerges. Meaning that, if $p(n) \geq (\log n + c)/n$ then the probability for the graph to be connected is greater than $1 - 4e^{-c}$. The transition thus happens at a small probability p. In figure 5.1 an example of this kind of a graph with 100 sites and 100 connections is shown. In this case $p \approx 0.02$.

5.1.2 Small–world graphs

It was observed that, for instance in the case of social networks, the distance of any two people is remarkably short considering the size and the complexity of the network. The distance here is marked by the number of acquaintances such that a person would know someone who then knows someone else ultimately connecting any two persons in the network in such a relationships. The first observation and a proposal for a simple model archieving this small–world property in a random graph was reported by Watts and Strogatz [88]. Their model essentially lays

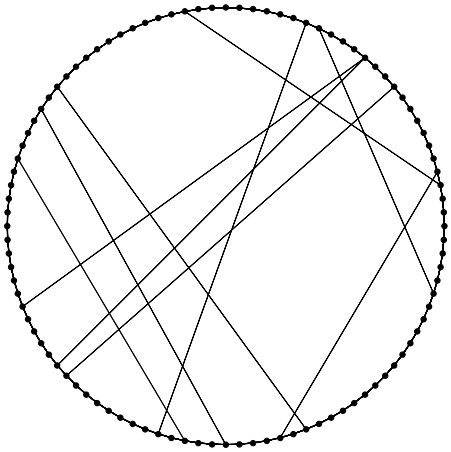

Figure 5.2: A Watts–Strogats small–world graph with 100 sites.

out a regular lattice, in which the end–point of any of the connections is rewired with probability p into a randomly chosen new site. The same effect can also be archieved by adding some connections between any random sites. The added long range connections provide a passage that significantly shortens the distances. In figure 5.2 there is an example of a Watts–Strogatz small–world network with 100 sites, 100 connections and the parameter is $p = 0.1$. An extension to this consept was introduced by Kleinberg [39] such that the underlying graphs is any lattice to which long range connections are added.

5.1.3 Scale–free graphs

In natural systems it is widely observed that there is a distinctive lack of a characteristic degree, i.e., no particular number of connections is dominant. The models that have such a propery are called *scale free* networks. This was first taken under scrutiny by Barabási and Albert [5]. Their proposal was a simple construction of a growing network, realised by beginning from a small initial graph, and adding each time step a new site, which is then connected to an older site with a probability proportional to the degree of the site to be connected to. This eventually gives rise to a degree distribution following a power-law behaviour, i.e., the degree has the relation $N(k) \sim k^{-\gamma}$ with some exponent γ, for the number of sites $N(k)$ having the degree k. Again in figure 5.3 is an illustration of a scale–free graph of Barabási–Albert type, where it can be seen that some connections are very highly connected.

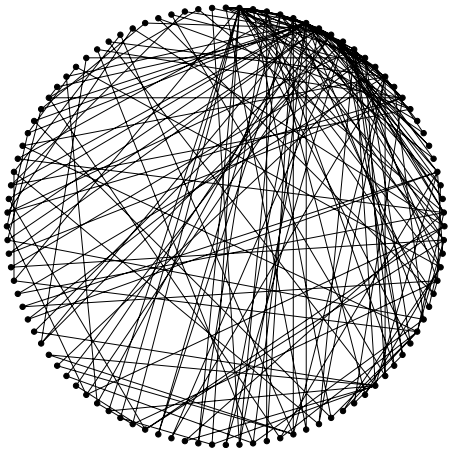

Figure 5.3: A scale–free graphs with 100 sites.

5.2 Spreading on small-world networks

In this section the dynamical spreading phenomenon is defined and then applied to small–world networks. It is shown that there are transitions relative to the small–world parameter p in the distributions of the number of distinct visited sites and the return probability, i.e., the probability of the walker to return the same site, in the Watts–Strogatz type graphs.

Spreading imitates in a way the diffusion process of a substrate in a medium. Here the graph is the medium and the substrate is a random walker. The average number of unique sites the walker visits in a given time, $Q(t)$, is an indicator to watch for as is the probability of the walker to return to the initial site in a given time, $P_{00}(t)$. Note that the choice of the origin is not relevant as the system is homogenous and any site could be chosen.

For comparison it is known that diffusion in a 1–dimensional lattice follows the power–law with an exponent of one half:

$$Q(t) \sim \sqrt{t}, \tag{5.1}$$

For higher dimension, however, the spreading turns out to be linear:

$$Q(t) \sim t. \tag{5.2}$$

In small–world graphs $Q(t)$ shows an interesting crossover from the initial $\sqrt{t}$ behaviour that is characteristic for the one-dimensional case to $Q(t) \propto t$ behaviour describing the high dimensional or random graph situation [35, 45]. As a function of p and t, $Q(t)$ has a scaling form:

$$Q(t) = t^{1/2}\kappa_Q(tp^\alpha) \tag{5.3}$$

where κ_Q is a universal scaling function with the following properties:

$$\kappa_Q(x) \propto \begin{cases} \text{const} & \text{for} \quad x \ll 1 \\ \sqrt{x} & \text{for} \quad x \gg 1 \end{cases} \tag{5.4}$$

It is expected that $\alpha = 2$ since in the system there exists a basic length scale $l \propto 1/p$, characteristic of the average distance between sites having long range connections for which the walker needs $t_l \propto l^2$ steps to sweep through. Thus the argument of the scaling function κ_Q in equation (5.3) should be t/t_l [35, 45].

Annealed spreading

Here the analysis is taken from the perspective of a system with annealed disorder. It is shown that although the system is different the dynamics can be transformed into a form that accurately corresponds to the dynamics of quenched disorder.

The movement of the random walker is governed by the simple master equation:

$$\partial_t P_i(t) = \sum_{j=1,N} T_{ij} P_i(t) \tag{5.5}$$

where the continuum time limit has been applied. Instead of discrete time steps, time here is now a continous variable. Here $P_i(t)$ is the probability that the walker is at site i at time t and T_{ij} is the transition rate from site i to site j written as follows

$$T_{ij} = W_{ij} - \delta_{ij} \tag{5.6}$$

where W_{ij} is the transition matrix of the following form:

$$\mathbf{W} = (1-p)\mathbf{W}^{(S)} + p\mathbf{W}^{(L)}. \tag{5.7}$$

Here the superscripts (S) and (L) refer to short and long range jumps, respectively.

The zeroth row of the short range transition matrix $\mathbf{W}^{(S)}$ reads as follows

$$W_{0\,\cdot}^{(S)} = \frac{1}{2k}(0, \underbrace{1, \ldots, 1}_{k \text{ times}}, \underbrace{0, \ldots, 0}_{N-2k-1 \text{ times}}, \underbrace{1, \ldots, 1}_{k \text{ times}}). \tag{5.8}$$

A similar equation can written to the long range transition matrix $\mathbf{W}^{(L)}$:

$$W_{0\,\cdot}^{(L)} = \frac{1}{N-2k-1}(\underbrace{0, \ldots, 0}_{k+1 \text{ times}}, \underbrace{1, \ldots, 1}_{N-2k-1 \text{ times}}, \underbrace{0, \ldots, 0}_{k \text{ times}}). \tag{5.9}$$

Then the i^{th} row of the transition matrices is obtained by cyclically shifting the 0^{th} row to the right. Matrices $\mathbf{W}$ and $\mathbf{T}$ have the Toeplitz form, i.e., T_{ij} depends only on the site difference $(i-j)$. Therefore, the right hand side of equation5.7

is a convolution which after spatial Fourier transform leads the master equation to the following form

$$\partial_t \hat{P}_q(t) = (\hat{W}_q(t) - 1)\hat{P}_q(t). \tag{5.10}$$

With the initial condition

$$P_i(0) = \delta_{0i} \tag{5.11}$$

the formal solution is as follows

$$\hat{P}_q(t) = \exp[\int_0^t (\hat{W}_q(u) - 1)du]. \tag{5.12}$$

This solution can be easily evaluated for the matrix $\mathbf{W}$ given in equation (5.7).

Then let $F_{ij}(t)$ denote the probability of the random walker visiting site j at time t having started from site i. Then we can write

$$P_{ij}(t) = \int_0^t F_{ij}(u)P_{jj}(t-u)du, \tag{5.13}$$

where $P_{ij}(t)$ is the probability for the random walker to move from site i to site j at time t. From this we get through the Laplace transform the following equation

$$\tilde{F}_{ij}(z) = \frac{\tilde{P}_{ij}(z)}{\tilde{P}_{jj}(z)}. \tag{5.14}$$

Now let us take $q(t)$ as the probability of observing a new site, or as the *spreading rate* at time t when the random walker started from site 0:

$$q(t) = \sum_{i=0}^{N-1} F_{0i}(t). \tag{5.15}$$

By taking the *return probabilities* $P_{ii}(t)$ to be the same for all i the equation (5.15) can be written in the following form

$$\tilde{q}(z) = \frac{1}{\tilde{P}_{00}(z)} \sum_{i=0}^{N-1} \tilde{P}_{0i}(z) = \frac{1}{z\tilde{P}_{00}(z)}. \tag{5.16}$$

Having this, the quantity of interest is the average number of distinct sites visited, which is obtained by integrating the probability of observing a new site, $q(t)$, over time t:

$$Q(t) = \int_0^t q(t')dt'. \tag{5.17}$$

Using the above formulation of equation (5.16), $Q(t)$ is obtained by the inverse Laplace transform of the function $\tilde{q}(z)/z$.

Spreading simulation

In spite of the strong argument for the scaling exponent α to be most likely 2, Jasch and Blumen [35] found in their numerical simulations of small–world networks a value $\alpha = 1.85$. In the simulations they had chosen $N = 50000$ by taking an average over 500 random walkers for each of the 100 small–world networks and they varied p in the interval $0.01 \leq p \leq 0.1$. It was established by the author of this dissertation, Kertész and Kaski [45] that the intuitive $\alpha = 2$ relation is correct as the limiting value when $N \to \infty$.

The equation (5.3), as is usual in scaling theory, is valid only asymptotically and in this case the scaling limit is $N \to \infty, t \to \infty$ and $p \to 0$. The scaling regime can be estimated from the variation of the mean vertex distance ℓ as a function of p [63], it turns out that the distribution of $\ell k/N$, where k is the degree of the lattice in the small–world network, has a scaling function with the argument $x = pkN$ and which is of sigmoidal shape. This curve suggests that one cannot expect a good scaling for the above mentioned crossover, if $pkN \gg 100$. Therefore, it seems likely that in [35] the investigated values of p were not small enough to assure the proper scaling behaviour (in fact Jasch and Blumen had $p_{\min}kN = 1000$ which is perhaps not large enough [35]).

For this reason the simulations must be carried out with considerably smaller values of p. In order to do so, the system size must be increased as well. A more proper choice is $k = 2$, $N = 10^5$ and varying the p as $p = 10^{-4}, 10^{-3.5}, 10^{-3}, 10^{-2.5}$. In order to estimate the average of $Q(t)$, 100 realizations and 100 random walkers per realization results in an adequate statistics, i.e., $p_{\min}kN = 20$. The average number of distinct visited sites $Q(t)$ as functions of t and p, is depicted in figure 5.4. In this plot it is seen that for the two largest values of p saturation of $Q(t)$ has set in.

Figure 5.5 shows a scaling plot of the results on $Q(t)$ where $Q(t)/\sqrt{t}$ is plotted as a function of tp^{α}. The scaling was found to be optimal with the choice of $\alpha = 2$. For comparison the same plot with $\alpha = 1.85$, which is the value found in [35], is also shown. The results clearly support the simple scaling picture discussed above, i.e., $\alpha = 2$.

Return probability

The return probability stands for the probability of the random walker to return to the initial site. Also this quantity shows a transition when the parameter p is varied.

In the case of the small–world graphs the return probability P_{ii} is independent of the choice of i. This is known to decay as $1/\sqrt{t}$ for the $p = 0$ case while an exponential decay is expected for large p. A scaling form similar to equation (5.3)

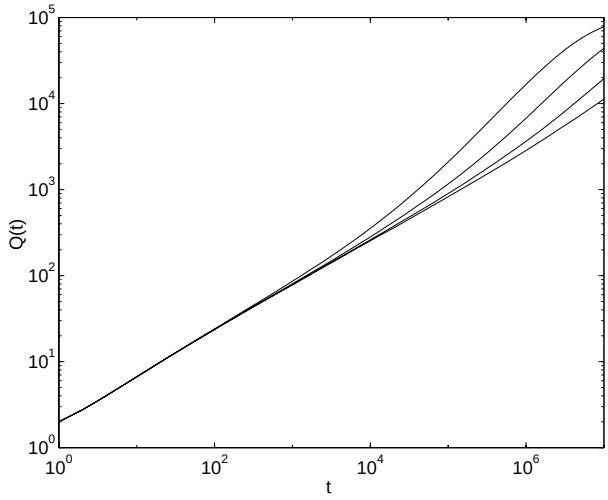

Figure 5.4: Raw data for the average number of distinct sites visited $Q(t)$ of the quenched system as a function of the number of time steps t and the probability values $p = 10^{-4}, 10^{-3.5}, 10^{-3}, 10^{-2.5}$ plotted from the lowest to the highest respectively. For large p the saturation due to the finite size $N = 10^5$ of the systems starts to become visible.

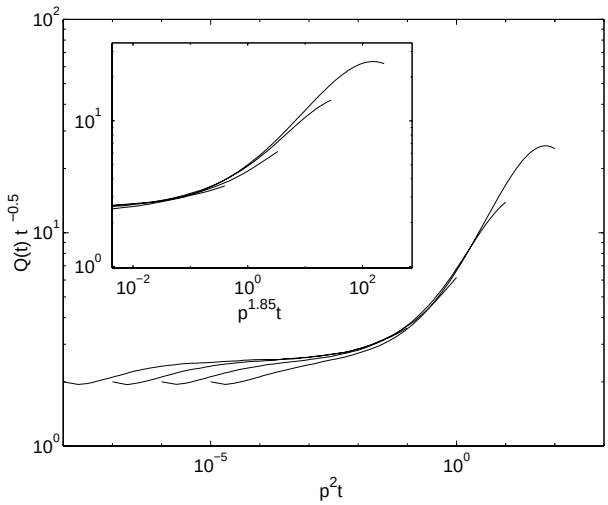

Figure 5.5: Scaling plot of the data of figure 5.4 with $\alpha = 2$. The inset presents a scaling with the exponent of the reference [35] $\alpha = 1.85$.

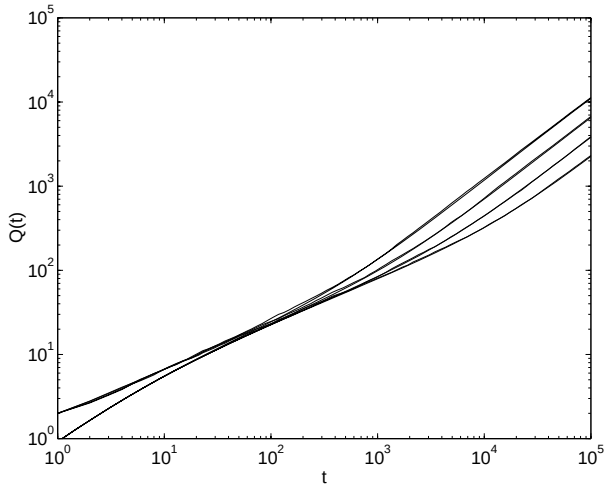

Figure 5.6: Simulation results for the mean number of distinct sites visited $Q(t)$ of the annealed system for long range jump probabilities $p = 10^{-4}$, $10^{-3.5}$, 10^{-3}, and $10^{-2.5}$ plotted from the lowest to the highest, respectively. These curves start from $S(0) = 1$. Analytical results, which start from $Q(0) = 0$ are also shown.

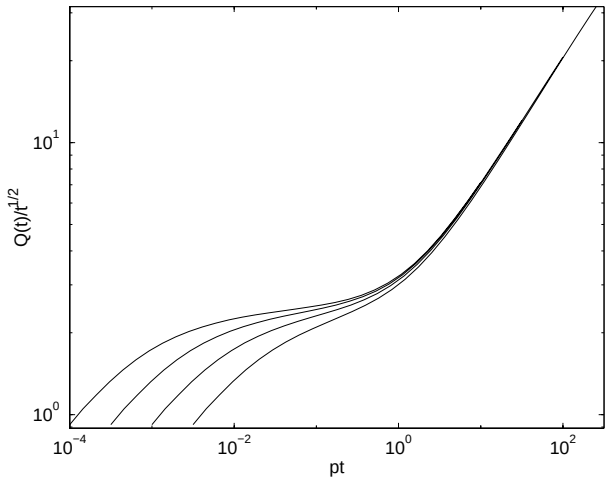

Figure 5.7: Scaled spreading $(Q(t)/\sqrt{t})$ of the annealed system against the scaled time (pt) for long range jump probabilities $p = 10^{-4}$, $10^{-3.5}$, 10^{-3}, and $10^{-2.5}$ plotted from the lowest to the highest, respectively.

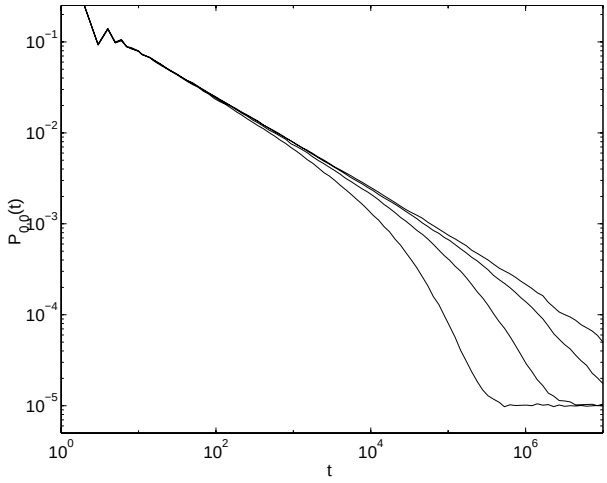

Figure 5.8: Raw data for the return probability P_{00} of the quenched system. The p values are the same as for figure 5.4, now increasing from top to bottom. The whole time interval was binned by 100 bins of equal sizes on the logarithmic scale.

should be also valid for P_{ii}, which Scala et al. have shown in [76]:

$$P_{ii}(t) = t^{-1/2}\kappa_P(tp^{\alpha}), \tag{5.18}$$

where $\kappa_P(x)$ is a rapidly decaying scaling function with the limit $\kappa_P(x) = \text{const.}$ for $x \ll 1$. However, the argument of κ_P should be the same as in equation (5.3). Also Jespersen et. al [36] gives a form for the scaling of the return probability and report that sometimes the transition occurs earlier than $n \sim p^{-2}$.

In order to get an even higher accuracy for the results there are 10 times more runs for the averages. In order to minimise the effect due to the finite size of the samples, i.e., the $n \to \infty$ limit of $1/N$ is subtracted from the measured values. Figure 5.8 shows the raw data of the return probability P_{00} and Figure 5.9 the scaling plots. Again, it can be seen that the scaling with the intuitively expected $\alpha = 2$ is superior to the one obtained by Jasch and Blumen [35]. Figure 5.10 shows a plot of the return probabily of the annealed model having a very similar form as in the quenched case. The corresponding scaling plots are shown in figure 5.11. From this figure it is apparent that the quenched system obeys the scaling extremely well.

5.2.1 Self–consistent model

It was noted earlier that the model of annealed disorder is independent of the previous history of the walker and has a different scaling exponent. However, the

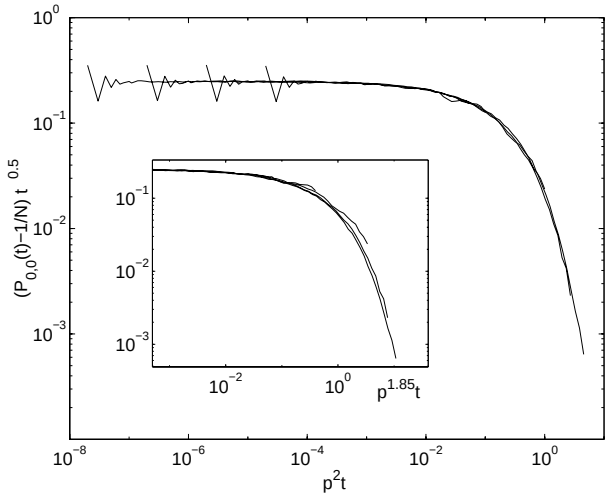

Figure 5.9: Scaling plot of the data of figure 5.8 using $\alpha = 2$. For minimizing the finite size effects the asymptotic value $1/N = 10^{-5}$ was subtracted from P_{00}. For comparison, the inset shows the scaling plot with the $\alpha = 1.85$, which was the exponent of the reference [35].

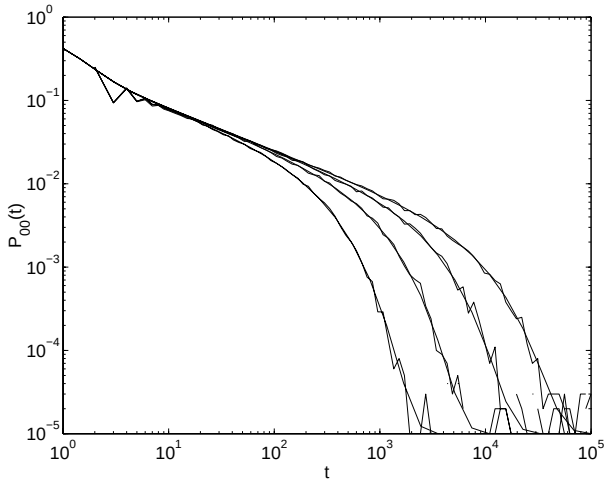

Figure 5.10: Simulation results for the return probability for long range jump probabilities $p = 10^{-4}$, $10^{-3.5}$, 10^{-3}, and $10^{-2.5}$ (uneven line). The smooth curves show the results of the analytical theory.

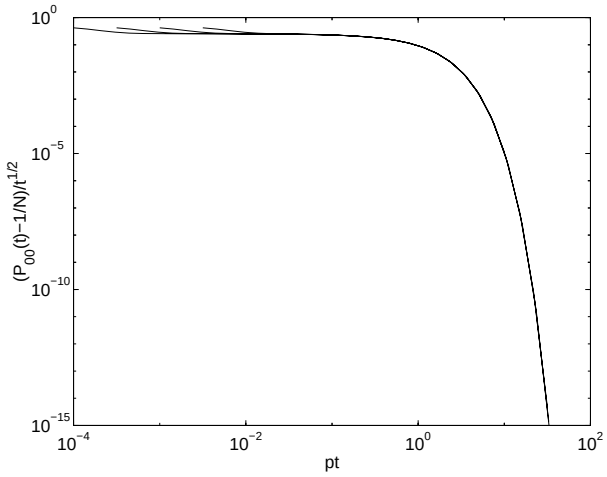

Figure 5.11: Scaled return probabilities against scaled time for long range jump probabilities $p = 10^{-4}$, $10^{-3.5}$, 10^{-3}, and $10^{-2.5}$.

exponent can be transformed to correspond to the quenched case when the new transition is made dependent on the history. Then the crossover is shifted such that $\alpha = 2$ as in the case of quenched disorder.

Since the scaling of the transition occurs in the quenched system later (as $\sim p^2 t$) we replaced the multiplier p of $\mathbf{W}^{(L)}$ in equation (5.7) with $p \cdot q(t)$ to simulate the situation where the random walker has a probability of making a long range leap only when visiting a previously unseen site. Now the transition matrix reads as follows:

$$\mathbf{W} = (1 - p)\mathbf{W}^{(S)} + p \cdot q(t)\mathbf{W}^{(L)}. \tag{5.19}$$

This then means that the corresponding master equation cannot be solved explicitly but it can still be estimated to arbitrary accuracy with iteration. In figure 5.12 it is shown that the resulting time dependent behaviour of the random walk spreading for our self-consistent model and simulated quenched system are very similar. Apart from the short times the agreement between these results seem to be quite good. Figure 5.13 presents again the scaling of the data in figure 5.12, indicating that the scaling is proper, but different from the quenched case.

Figure 5.13 shows the scaling plot with $\alpha = 2$ for the self-consistent model. Apart from early times the scaling seems to hold once again. Hence it can be concluded that the numerical results justify the choice of the equation (5.19). This reflects the fact that in the quenched model a 1–dimensional random walk has to be carried out between two long jumps.

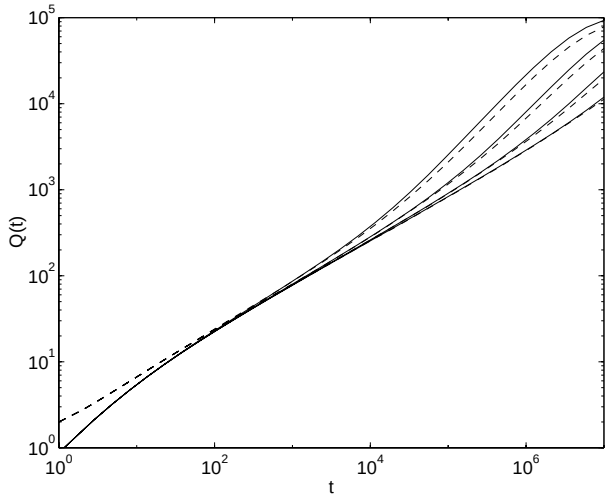

Figure 5.12: Results for the spreading as a function of time of the self–consistent annealed model obtained from the analytical theory (solid line) and from the quenched simulations (dashed line) for long range jump probability $p = 10^{-4}$, $10^{-3.5}$, 10^{-3}, and $10^{-2.5}$.

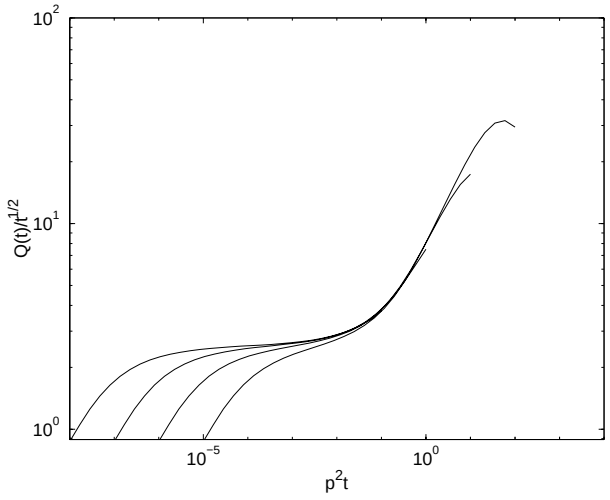

Figure 5.13: Scaled spreading of the self-consistent annealed model $Q(t)/\sqrt{t}$ against scaled time p^2t for long range jump probabilities $p = 10^{-4}$, $10^{-3.5}$, 10^{-3}, and $10^{-2.5}$.

5.3 Discussion

In this chapter the basic models of random graphs were discussed: the Erdős & Rényi, the Watts & Strogatz, and the Barabási networks. The attention was focused on the spreading on the Watts–Strogatz type networks. It was established that the distribution of the number of visited sites has a transition with the power–law exponent $\alpha = 2$. Also the distributions of the number of visited sites and the return probabilities with annealed disorder was shown to have qualitatively similar properties as with the case of quenched disorder. However, in order to make the transition to have the same exponent the transition probabilities of the annealed walker had to be made time dependent.

Chapter 6

Self-organised criticality

Another dynamical system of wide current interest is the model of sandpiles [3]. Although this model is a considerable simplification of the corresponding natural phenomenon, it is expected to provide some insight to similar events in various systems, like the breakdown of an electrical power grid and the collapse of communication networks. The original model assumes a regular lattice of sites with capacity to hold "grains". When the load of grains at a single site exceeds a predefined limit then part of the load is transferred to its neighbours, which in turn may overflow and thus initiate a cascade process, i.e. an avalanche. In this thesis the sandpile model in one dimensional small–world networks is shown to have many interesting non–trivial properties. The distributions of the key characteristics of the system have transitions similar to the spreading dynamics, that are explainable through some kind of a competition of two mechanisms [47].

The term which often appears in the context of sandpiles is referred to as *self–organised criticality*. *Self–organisation* means that the system attains through a dynamic process some form without outside input. In the case of the model of sandpiles this form is that all sizes of the avalanches occur. *Criticality* in turn refers to a characteristic of the system to make a transition from one form to some other completely different form. Critical phenomena are analogous to phase transitions in materials experiencing a change of conditions, such as temperature.

The sandpile model has been investigated in many kinds of graph topologies, including those of small–world networks, but in higher than one dimensions. The reason why 1–dimensional systems are not generally considered interesting is that Bak, Tang and Wiesenfeld have shown that there is no self–organised criticality [4, 38], which means that avalanches of all sizes occur. Recently however, Kulkarni et al. [40] have investigated the activity of specific sites on small-world networks with the Bak–Sneppen model, which is a model similar to the sandpile model with self–organised criticality. In the case of a 2–dimensional sandpile model with random long range connections, i.e. a system with the small–world property,

Arcangelis and Herrmann [19] have demonstrated that for the distribution $N(s)$ of the avalanche size s an approximative scaling relation of the following form holds: $sN(s) \sim \kappa(sp^{0.65\pm0.1})$, where p is the small-world parameter and κ is the scaling function. Also Moreno et al. [58] have analysed the Bak–Sneppen model but on the scale-free networks of Barabási and Albert [5], and they have found that the model approximately obeys the mean field exponential law $N(s) \sim s^{-3/2}$ and that the scale-free model lacks a critical threshold. In addition, Lee et al. [49] have presented an analysis of the sandpile model on scale–free networks, proving a relationship between the distributions of size and duration of an avalanche, and the power–law exponent of the graph connectivity. In a recent study by Lahtinen et al. [47] showed that despite the fact that the orginal 1–dimensional sandpile model does not exhibit self–organised criticality, this property does appear when the long range connections of the small–world network are added.

In this chapter the possible effect of long range connections in a 1–dimensional network topology on self-organised criticality is investigated. In this model the long range connections are formed in two alternative ways. In the first way each long range connection is formed temporarily by choosing a distant site for the grain to jump randomly and independently of previous jumps. This is called *annelead* disorder. In the second way a fixed graph topology with randomly chosen long range connections are generated before the process is started. This is called *quenched* disorder. In both of these cases an avalanche has local as well as global character, being in competition.

6.1 Model

The 1-dimensional sandpile model can be considered as a linear chain of m sites or bins that are numbered $1, \ldots, m$, as depicted in Figure 6.1. In the beginning of the process the chain is considered empty and the process is started by dropping grains randomly to the sites of the system. If the number of grains in a site exceeds 2 an *avalanche* is initiated by *toppling* grains from it. In each toppling a site i having more than 2 grains is chosen at random and then 2 of its grains are moved to the immediate neighbours $i-1$ and $i+1$, provided that $1 < i < m$. If on the other hand $i = 1$ or $i = m$ one grain is removed from that site and at the same time from the system altogether and another grain is moved to the neighbour 2 or $m-1$, respectively (see Figure 6.1). This corresponds to a system which is open from both ends. In addition to these basic moves of grains, long range jumps are introduced by using two different policies. On one hand the long range connections are generated before the process such that from each site i a single permanent connection is created to another site $j \in \{1, \ldots, i-2, i+2, \ldots, m\}$ randomly with probability p. Then if during the process the site i being toppled has such a connection, one more grain is moved from i to j. This policy

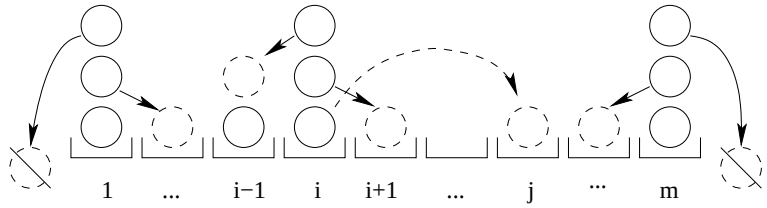

Figure 6.1: An illustration of possible single grain topplings in our sandpile model. There are two kinds of short range jumps (solid arrows), i.e. those in the middle and those in the open ends of the system being removed from the system. The long range jump (dashed arrow) occurs with probability p.

is called *quenched randomness* and it essentially corresponds to a sandpile model on a Watts–Strogatz type random graph, where a few random connections have been added to an otherwise regular lattice [88]. On the other hand long range temporary connections can be added dynamically during the process such that in each toppling with probability p one grain is moved to a randomly chosen site $j \in \{1, \ldots, i-2, i+2, \ldots, m\}$. This policy is called *annealed randomness*, and it can be related to the stochastic sandpiles considered previously by Manna [52].

6.1.1 System without long range connections

First the situation in which no long range moves of grains are possible, i.e. $p = 0$, is scrutinised. In this case the system becomes strictly one dimensional and there is no distinction between the quenched and annealed distorder. Thus the system does not show self–organised criticality as already noted by Bak et al. [4, 38]. However, since this system is open from both ends, rather than only from one end, it reacts differently to grain-additions than the traditional sand-pile model [3]. In this system when a grain is dropped in the middle of a string of n critical sites, i.e. sites with 2 grains each, the resulting avalanche will have size $s = n$ (i.e. the number of sites toppled). A grain is added to site i, counting from the left end of the string of critical sites, with site $i = 1$ being the first critical site. Once the avalanche is completed the sites of the critical string remain critical except one site, with only one grain, located at the point $n - i + 1$. Now the duration of the avalanche, denoted by t, can be expressed as follows

$$t = in - i(i-1), \qquad 1 \leq i \leq n. \tag{6.1}$$

The solutions of this Diophantine equation are determined by the possible integer values of t, n, and i. With the fact that $s = n$ this equation can be written for the avalanche size s as functions of time t and location i:

$$s(t, i) = t/i + i - 1. \tag{6.2}$$

For this set of functions the envelope function can be determined through differentiation with respect to i, resulting in the following equation

$$s_{env}(t) = 2\sqrt{t} - 1, \tag{6.3}$$

which is also the lower bound for the avalanche size s for the given time t.

When the system with $p = 0$ is simulated for a sufficiently long time all the sites will be filled with two grains except for one site, called here as gap, at location r that has only one grain. Now the avalanche size s given t depends on the location of the gap r. From the above described process it can be seen that the gap r appears randomly with equal probability at locations $\{0, \ldots, m\}$, where zero implies that there is no gap in the system. R denotes random variable corresponding to the avalanche size and S the random variable of the gap location, respectively. The probability of the random variable S given r is as follows

$$\mathsf{P}\{S = s|r\} = \frac{r}{m}\delta(s - r) + (1 - \frac{r}{m})\delta(s - (m - r)). \tag{6.4}$$

In this formula the first term describes the probability of dropping a grain to the area of size $s = r$, left from the gap, and the second term correspondingly to the area right from the gap. Here it is assumed that the system size is large enough to ignore the unit size of the gap. When the joint probability $\mathsf{P}\{S = s, R = r\} = \mathsf{P}\{S = s|r\}\mathsf{P}\{R = r\}$ is marginalised over R, the probability of an avalanche of size s is obtained:

$$\mathsf{P}\{S = s\} = \sum_{r=0}^{m} \mathsf{P}\{S = s|r\}\mathsf{P}\{R = r\} = \frac{1}{m+1}(\frac{s}{m} + (1 - \frac{m-s}{m})) = \frac{2s}{m(m+1)}. \tag{6.5}$$

From this equation it can be seen that the avalanche size distribution is linear in s, and thus the system does not exhibit self–organised criticality.

Let us then consider the distribution of the avalanche duration, which turns out to be quite complex. However, for $p = 0$ the average avalanche duration of given size can be determined. From equation (6.1) it can be seen that the possible values of i for given t are the integer divisors of t. Then the distribution of the random variable T of the avalanche duration is $\mathsf{P}\{T = t|t \leq m\} \sim \nu(t)$ where $\nu(t)$ is the number of integer divisors of t, provided that $t \leq m$. On the other hand if $t > m$ the distribution falls because in equation (6.1) i and n are limited from above. With these limitation the avalanche duration has the following maximum value $t_{max} = mi' - i'(i' - 1)$, with $i' = \lceil \frac{m}{2} \rceil$ (i.e. rounded up to the nearest integer) such that after t_{max} the distribution is zero. Now the average duration of an avalanche of a given size s is obtained from the equation (6.1) as follows

$$\begin{aligned} \langle t \rangle = \mathsf{E}\{T|s\} \quad &= s^{-1} \textstyle\sum_{i=1}^{s} [is - i(i-1)] = s^{-1} \sum_{i=1}^{s} [i(s+1) - i^2] \\ &= s^{-1}[\tfrac{1}{2}(s^2 + s)(s+1) - \tfrac{1}{6}(2s^3 + 3s^2 + s) \\ &= \tfrac{1}{6}s^2 + \tfrac{1}{2}s + \tfrac{1}{3}, \end{aligned} \tag{6.6}$$

which indicates quadratic and linear dependence on the avalanche size. It is evident, however, that for realistic avalanche sizes s the quadratic dependence is dominating.

6.1.2 System with long range connections

When a system has long range connections, i.e. $p > 0$, the avalanche dynamics has both local and global character. In this the connections to the two nearest neighbours, like in a system without long range connections, give rise to the local avalanches. In turn the long range connections give rise to two phenomena, on one hand by removing grains from the local avalanche and on the other hand by facilitating an initiation of another local avalanche at the other end of the long range connection. The grain removal causes the local avalanche to relax and thus halt quicker, i.e. damping down the avalanche activity, while the long range jumps tend to increase the avalanche activity, i.e. nucleating new local avalanches. Thus these two processes are competing.

In this system when p is varied, we can expect that there is a transition in the distribution of the avalanche duration. The avalanche loses momentum after long range jumps take effect, which should happen after p^{-1} trials for an occurrence of a long range jump. Therefore in the annealed case the number of trials is equal to the number of time steps and thus the distribution of the duration has a transition at $\hat{t}$:

$$\hat{t} \propto p^{-1}. \tag{6.7}$$

In the quenched case, however, the number of trials is proportional to the size of the local avalanche. The transition in the size distribution of a quenched system, denoted now by $\bar{s}$, is thus

$$\bar{s} \propto p^{-1}. \tag{6.8}$$

From equations (6.2) and (6.3) it is apparent that the the corresponding transition in the annealed case for the avalanche size has a power–law relation:

$$\hat{s} \propto p^{-\alpha}, \tag{6.9}$$

where the exponent $\frac{1}{2} \leq \alpha \leq 1$. When $p > 0$ it can be expected that due to increased probability of grains reaching the ends of the chain and leaving the system, there will be fewer grains contributing to avalanches. This in turn will reduce the size of uniform strings of critical sites i.e. sites with 2 grains, thus reducing the size of local avalanches. Since the local avalanches are smaller one could expect the size of the global avalanche for given duration to follow closer equation (6.3), which when combined with equation (6.7) would suggest $\alpha = 1/2$. Furthermore, combining the dominant relation between the avalanche duration and avalanche size indicated in equation (6.6) i.e. $\langle t \rangle \sim s^2$, and the relation in equation (6.7), i.e., $\hat{t} \propto p^{-1}$, lends also some support to $\alpha = 1/2$. This result

seems analogous to the relation obtained in [46], indicating that the transition to self–organised criticality in the quenched system takes place slower than in the annealed system.

6.2 Simulation results

Now we turn our attention is turned to computational studies, and first describe the simulation set–up. In computer simulations one usually faces the problems of finite system size and sufficiency of the statistics in relation to the available computing time and the speed of computers. In small discrete systems their discrete characteristics, such as saturation effects, are always distinctly visible in the statistics of the simulations. Thus one wants to increase the system size in order to better correspond to an infinite system at the thermodynamic limit. As a compromise the system sizes here are chosen as $m = 100, 316, 1000$, for both the annealed and quenched systems. In the quenched case for the probability parameter p a number of values between zero and one are chosen as follows: $\log_{10} p = 0, -1/8, -2/8, -3/8, \ldots, -4$. The annealed case has a similar set of p–values ranging from $\log_{10} p = 0$ to $\log_{10} p = -5$. For each p–value a sufficient number of time steps are used such that for the annealed system we had at least 1000 avalanches per simulation run, and, for reasons of longer computational time involved, the quenched system was simulated such that at least 100 avalanches per simulation runs occur. For sufficient statistics results were obtained as averages over 100 runs using different random number sequences for both the annealed and quenched cases.

Figure 6.2 (a) shows the histograms of the avalanche sizes $N(s)$ in the annealed system of size $m = 1000$ for 41 different values of p. For small values of p the avalanche size distribution $N(s)$ grows first monotonically to reach a maximum after which it decreases. When p increases the maximum moves to smaller s-values more or less linearly in the logarithmic scale. This implies that there is a power law dependency of the maximum of $N(s)$ vs. p. In Figure 6.2 (b) are the corresponding histograms of the quenched case. Here the behaviour seems quite similar to the annealed case but now the maximum is less distinct and it seems to move slower as a function of p than in the annealed case. This in turn indicates that in the quenched case the power law exponent is smaller than in the annealed case.

Next the power law behaviour of the avalanche size distribution is investigated in more detail, by using the same scaling approach as de Arcangelis and Herrmann [19]. Here it is assumed that $N(s)$ scales as follows:

$$sN(s) = \kappa(sp^{-\alpha}), \tag{6.10}$$

where κ is the scaling function, and α is the scaling exponent. In Figure 6.3 is

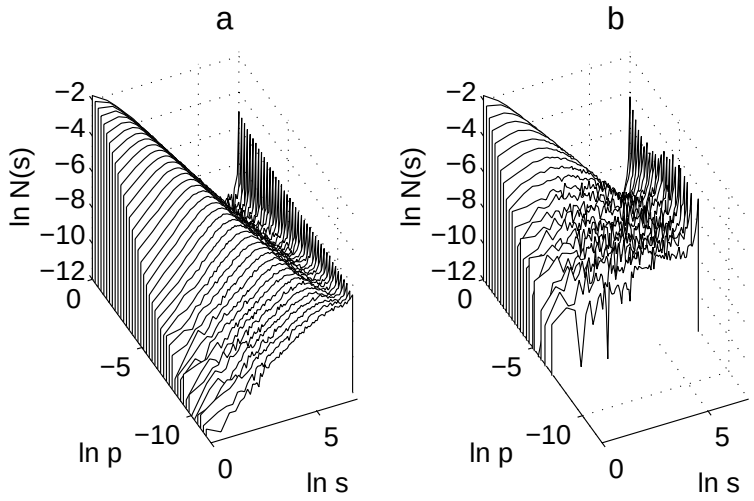

Figure 6.2: Histograms of the avalanche sizes in systems of size 1000 for different values of p. The left figure (a) is the annealed case, and the right (b) is the quenched.

plotted $\log sN(s)$ as a function of $\log p^{1/2}s$ for the annealed system in order to test whether the scaling conjecture with the exponent $\alpha = 1/2$ is valid, as indicated above (see equation (6.9) and related discussion). This exponent is different from the one obtained by de Arcangelis and Herrmann [19], i.e., when $p \to 1$ the exponent $\alpha \approx 3/2$. As the scaling reflects the turning point on the distribution the saturation of the histograms has been cut out. This cut has been done in such a way that when $p = 0$ the 200 last values of s are not plotted, and for each step of decreasing p–value 10 more points have been removed from the plot. As is evident from this figure, for large avalanche sizes s there seems to be data collapse and the scaling seems to hold thus confirming the analysis in Section 6.1. There does not appear to be a good data collapse for the small avalanches, but the intermediate avalanche sizes show a decreasing tendency following an approximate relation

$$N(s) \sim s^{-3/2}, \tag{6.11}$$

which is the power law behaviour found for the standard sandpile model in higher dimensions (≥ 2) [3]. This can be seen best in the inset of Figure 6.3 for the case of $p = 1$, corresponding to the curve with the highest point at $\log s = 0$ and decreasing the fastest for the group of curves.

In the quenched system there is a scaling with an exponent twice as large, i.e. $\alpha = 1$ seems to hold better, as evident in figure 6.4. The small values of $p \leq 10^{-1}$ have been omitted, as the scaling holds in the limit $s \longrightarrow \infty$ and the finite size of the system prevents the avalanches with these small p–values

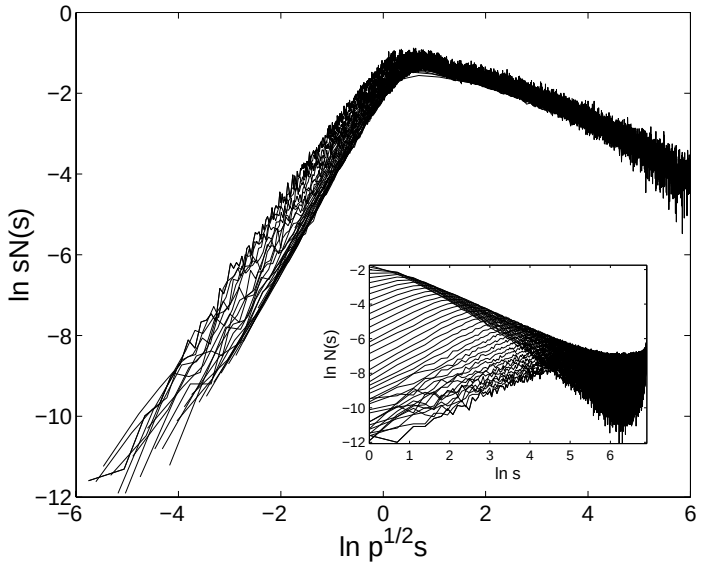

Figure 6.3: Scaled histograms of the size of the annealed avalanches in the system of size 1000. In the insert are the unscaled plots. When $s = 1$ the values of p decreases from the highest line, $p = 1$, to the lowest, $p = 10^{-5}$.

from reaching the turning point before the saturation takes place. This scaling for large p–values is in accordance with the equation (6.8), thus confirming our analysis for the quenched system, discussed above. As pointed out earlier the overall behaviour of $N(s)$ in the annealed and the quenched systems are similar, as is evident by comparing the inserts of Figure 6.3 and Figure 6.4, respectively. This similarity extends also to the scaling of $N(s)$ for intermediate avalanche sizes, i.e. the equation (6.11) with the power law exponent $3/2$ holds also in the quenched system and is most evident for $p = 1$ curve in Figure 6.4.

Next the probability of an avalanche to go through the entire system is considered. This probability is called the *traversal probability* and is denoted here by P_{tr}. This quantity can be simply estimated by using the ratio of occurrences of the maximal avalanches to the number of all observed avalanches. First the $p = 0$ case must be investigated in which the annealed and quenched system are the same. For three different system sizes $m = 100, 316, 1000$, we obtain from equation (6.5) that $\log P_{tr} = -3.92, -5.07, -6.22$, respectively. As for $p \neq 0$ Figure 6.5 shows the traversal probability estimates both for the annealed and quenched systems and for three different system sizes. The quenched system behaves qualitatively similarily with the annealed system, but suffers from more noisy data. Thus there is first an increase due the increased probability of the avalanche crossing a single gap (a site with only one grain), then it remains con-

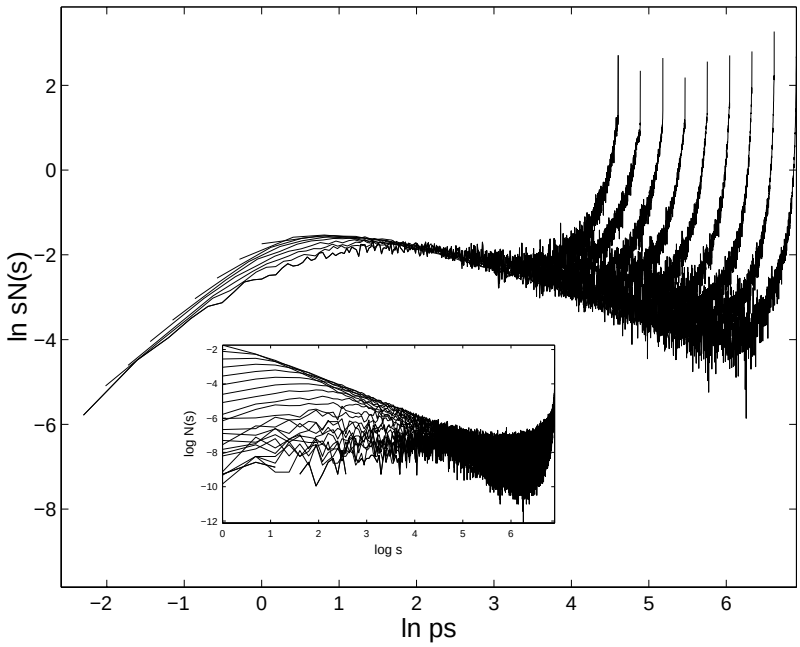

Figure 6.4: Scaled histograms of the size of the quenched avalanches in the system of size 1000. The insert shows the unscaled plots. Only values of $p > 10^{-1}$ are included.

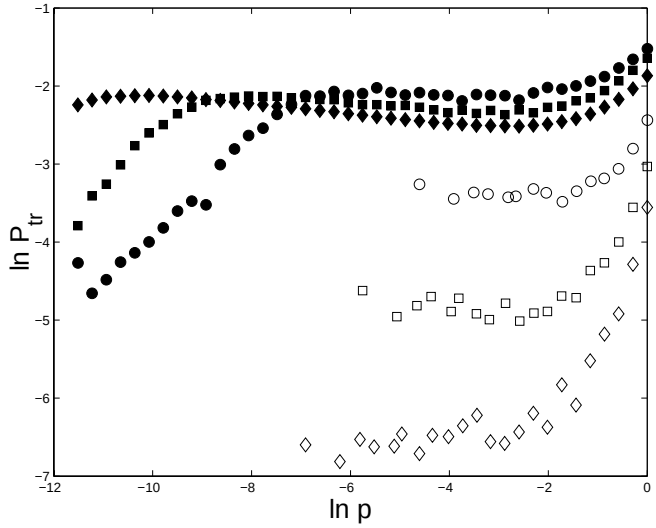

Figure 6.5: Traversal probabilities, where circles stand for the system size $m = 100$, squares for $m = 316$ and diamonds for $m = 1000$. The empty symbols stand for the annealed system, and the filled ones correspond to the quenched case.

stant, or decreases slightly, until increasing again when p approaches unity. The behaviour for the intermediate p–values is explained as a regime where the local avalanche relaxation, discussed above in section 6.1, is more dominant. The final increase in P_{tr} for increasing p is explained by the increase in the number of local avalanches. Both the annealed and the quenched systems show a tendency of P_{tr} to decrease as the system size increases.

Next we look at the *filling factor*, which characterizes the amount of grains that can be added to the system without starting avalanches. This stands for the inverse of the density of grains in the system and can be expressed as follows

$$g = 1 - \frac{\rho}{2m}, \tag{6.12}$$

where ρ is the number of grains in the system. In Figure 6.6 one can see that the filling factor $\langle g \rangle$ averaged over separate runs and avalanches increases with p, at an approximate rate proportional to p^{β}. It is evident from these log-log plots that for quite a wide range of p-values and independently of the system size the power–law exponents turn out to be $\beta = 0.60$ and $\beta = 0.90$ for the annealed and quenched cases, respectively. On the other hand when p is very small, g converges to $\frac{1}{2m}$ which for increasing system size m approaches zero and makes the curves collapse to single lines with which the power–law fits coincide. Also it is worth noting that with both systems the filling factor seems to converges to $1/e$

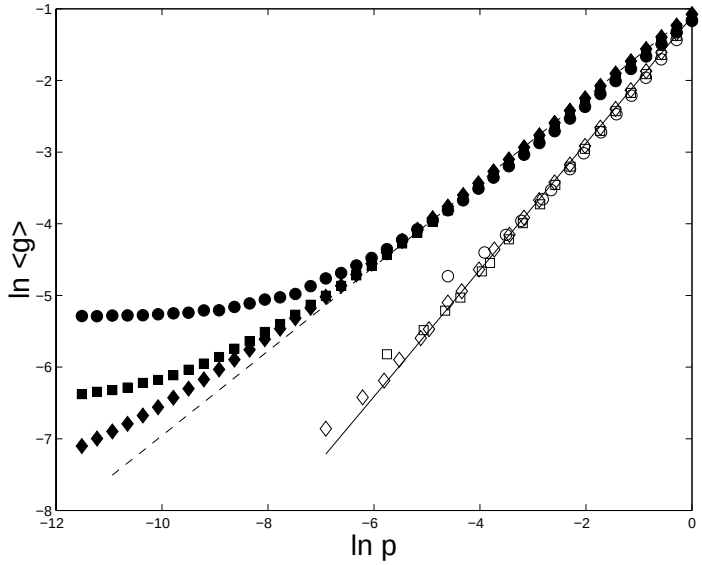

Figure 6.6: The average filling–factor, where circles stand for the system size $m = 100$, squares for $m = 316$, and diamonds for $m = 1000$. The empty symbols correspond to the annealed system, and the filled ones to the quenched case. The dashed line represents the power law of the annealed system $g \sim p^{0.60}$ and the solid line the one of the quenched case $g \sim p^{0.90}$.

when p approaches 1.

Finally we examine the distribution of the duration of avalanches is examined, and the numerical results are depicted in Figure 6.7. Panel (a) shows the results for the annealed system for $p \in [10^{-5}, 1]$ and in panel (b) for the quenched system for $p \in [10^{-3}, 1]$. When $p = 1$ the two systems behave in a similar manner, as they also do for very small values of p, i.e. in the annealed case for $p = 10^{-5}$ and in the quenched case for $p = 10^{-3}$. For the intermediate p-values the avalanches in the quenched system are sometimes almost twice as long as the avalanches in the annealed case. In both cases the behaviour of the duration distributions are so complex that they do not seem to conform to any simple scaling law. The reason why the duration distribution for the quenched avalanches have much longer tails is caused by formation of loops of fixed connections, such that some part of the grains will always return to the same location.

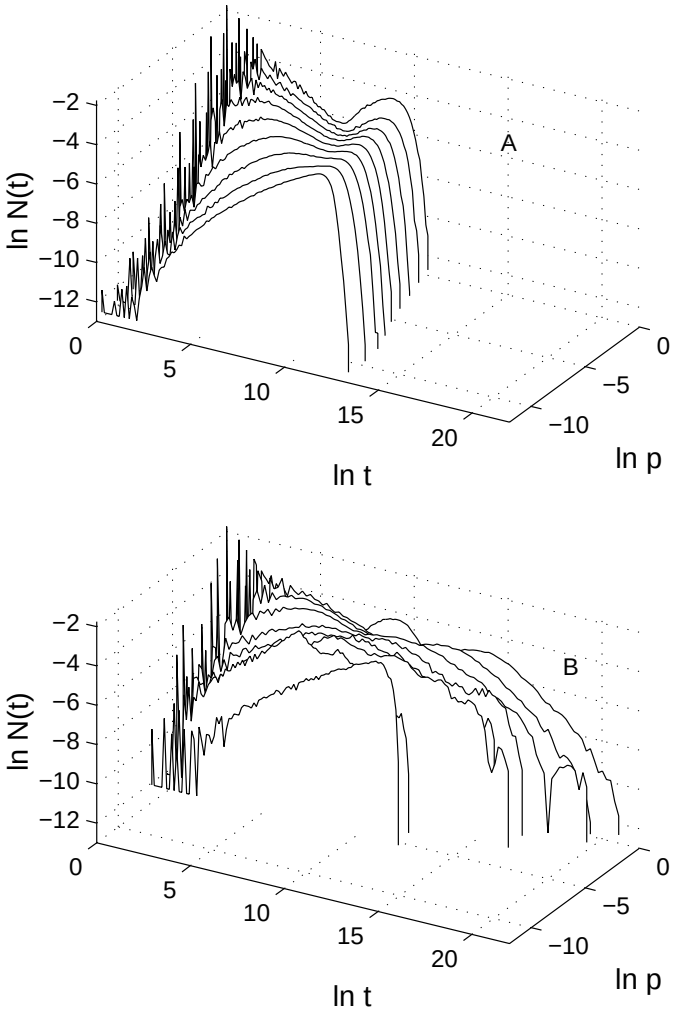

Figure 6.7: Histograms of the avalanche durations with systems of size 1000 for different values of p. In (a) are the results for the annealed system and in (b) for the quenched system. In (b) the curve with $\ln p = 0$ is similar in shape with the corresponding ($\ln p = 0$) curve in (a), albeit only partly visible in the plot.

6.3 Discussion

In this section an extension to the 1-dimensional sandpile model was examined analytically and with simulations for two alternative forms of small–world randomness: annealed and quenched. It was shown that the avalanche size distribution exhibits non-trivial transition from non–critical regime of small avalanches to the critical regime of large avalanches. This behaviour can be explained by a competition between two mechanisms: the avalanche nucleation and local relaxation. At higher dimensions (≥ 2), however, the self–organised criticality cannot be explained by the competition of these two mechanisms, primarily due to the lack of a sufficiently compact local neighbourhood. This happens because the local avalanches have more space for expansion and the long range jumps do not provide commensurate relaxation to significantly dampen them. An approximate scaling was established for the avalanche size distribution as a function of the small–world parameter p. The competition between the two mechanisms is most evident in the behaviour of the traversal probability, such that for small p–values the local relaxation mechanism dominates while for p approaching unity avalanche nucleation becomes more dominant. In addition it was found that the filling factor or the density of grains in the system shows power–law behaviour as a function of long range connection probability both for the annealed and the quenched systems converging unexplainably to $1/e$ at $p = 1$, but with different exponents.

Chapter 7

Conclusions

This thesis dealt on two topics, statistical inference and random graph simulations. In statistical inference the main contribution was the introduction of a new metric on probability distributions namely the transformation discrepancy. This discrepancy measure was applied in a modelling setting called *indirect inference*, where the likelihood function was not computable. With random graphs the focus was on their dynamical properties, and in such setting the two kinds of disorder, *annealed* and *quenched*.

In Chapter 2 the basic concepts of Bayesian statistical analysis were reviewed and applied to an example of a fault diagnostics system. It was demonstrated that it is possible to estimate the parameters and states of Poisson mixture processes containing a transition between states at unknown time using the Reversible Jump MCMC method. The estimation becomes more difficult when the transition has occurred close to the end of the total time, in which case the counted events of the device only exhibit behaviour of the initial states. In this estimation the availability of data for purely intact devices, and presence of more than one counter to record events, is critical.

In Chapter 3 it was shown that in modelling problems the models can be effectively compared and *ipso facto* selected by a discrepancy measure determined as the sum of pairwise costs. This leads to a metric measure on sample sets, which are sample sequences drawn at random from the models under consideration. The convergence of this measure is also guaranteed under proper assumptions concerning the underlying cost function of the individual pairs of elements. This metric was then applied in Chapter 4 to compute the posterior probability in cases where the likelihood functions are difficult to compute. The analytic and experimental studies show that the transformation method based on information theoretic foundations is a valid addition to the field of Bayesian modelling.

Both the transformation and the kernel estimates make an approximation of the principle assumption that points close in space to each other are also close in

probability. When one is interested in the probability of a given sample the kernel estimate is almost the only choice, but when the problem is about model selection the transformation metric can be utilised. It turns out that the transformation method does improve the estimate, at least in the examples, from the kernel estimate with the expense of more computational time required. This method also removes the problem of choosing the kernel and the bandwidth, but by adding one parameter of its own, and using more computational time, which may in time critical cases mean that the kernel method must be used.

In Chapter 5 of this thesis we looked at random graph models of small–world networks, and their dynamic behaviour with the spreading phenomenon of random walks. It was shown that for sufficiently small probabilities of long range links the proper scaling variable for the average number of distinct sites visited by a random walker and also for the return probability is np^2, i.e., the natural power–law exponent $\alpha = 2$ holds for the small–world networks. Also it was established that the annealed random walk model with rarely occuring long range jumps reflects some aspects of the dynamics in quenched small–world networks. In the simplest case, with time independent transition probabilities, the model can be solved analytically. However, as expected, only qualitative agreement between the quenched and the annealed models can be observed. With properly chosen time-dependent transition probabilities even the proper crossover exponent $\alpha = 2$, or p^{-2} dependence is obtained. Thus the random walker spreading in a quenched system can be estimated by an annealed model.

In Chapter 6 an extension to the 1-dimensional sandpile model was investigated analytically and with computer simulations for two alternative forms of small–world randomness: annealed and quenched. It was shown that the avalanche size distribution exhibits a non-trivial transition from a non–critical regime of small avalanches to the critical regime of large avalanches. This behaviour is caused by a competition between two mechanisms: the avalanche nucleation and local relaxation. However, in higher dimensions, the self–organised criticality cannot be explained by the competition of these two mechanisms, primarily due to the lack of a sufficiently compact local neighbourhood topology.

In this study we have also established an approximative scaling of the avalanche size distribution as a function of the probability of long range links p. The competition between the two mechanisms turned out to be most evident in the behaviour of the traversal probability, such that for small p–values the local relaxation mechanism dominates while for p approaching unity avalanche nucleation becomes more dominant. In addition, it was found that the filling factor, or the density of grains in the system, shows power–law behaviour as a function of long range link probability (p) both for the annealed and the quenched systems converging unexplainably to $1/e$ at $p = 1$, but with different exponents. The duration distribution of avalanches was also studied and it was found that avalanches in the quenched system are longer living and in both cases so complex that there was no simple

scaling law behaviour.

Chapter 8

Publications

This monograph is based on the following articles:

1. *Reversible jump MCMC for two–state multivariate Poisson mixtures* [48], written with prof. Jouko Lampinen and published in *Kybernetika*, vol. 39, 3, p. 307–315, 2003. This article contains the information presented in section 2.2. In this paper the contribution of the author of this thesis was the development of the model, doing the simulations and analysing the results.

2. *Transformation Discrepancy* [43], has the introduction of the transformation metric of chapter 3. This paper will be submitted for publication, with the author of this thesis as the sole author.

3. *Inference over uncomputable likelihoods* [44], also to be submitted for publication. In this article the author and Dr. Jukka Heikkonen applied the conversion metric to the problem of uncomputable likelihoods as explained in section 4. In this paper the contribution of the author of this thesis was the derivation of the theory and its analysis.

4. *Scaling of random spreading in small world networks* [45], published in *Physical Review E*, vol. 64, p. 057105(3), 2001, and written in collaboration with cooperation with prof. János Kertész and prof. Kimmo Kaski. In this paper the inaccuracies of publication by Jasch and Blumen [35] was corrected as explained in section 5. Here the contribution of the author of this thesis consisted of, jointly with the other authors, developing the analytical theory, and then on his own building up the simulations and the analysis.

5. *Random spreading phenomena in annealed small world networks* [46], published in *Physica A*, vol. 311/3-4, p. 571–580, 2002, and also written in collaboration with prof. János Kertész and prof. Kimmo Kaski. Here analysis of annealed random graphs was introduced and used for analysing the

spreading dynamics as a continuation of the previous study (published in Physical Review E). This is contained at the end of the section 5. Here the contribution of the author of this thesis consistet of doing the simulations, and the analysing all the results.

6. *Sandpiles on Watts–Strogatz type small–worlds* [47], accepted for publication in *Physica A*, 2004, and written again together with prof. János Kertész and prof. Kimmo Kaski. This paper analyses the effect of the small–world topology on the self–organising sandpile model in one dimension. This is presented in section 6. The author of this dissertation introduced the model of sandpiles to the small–world networks, did the analytical theory, simulated the models and analysed the results.

References

[1] E. Aarts and J. Korst. *Simulated Annealing and Boltzmann Machines*. John Wiley & Sons, 1989.

[2] R. Albert, A.L. Barabási, and H. Jeong. Scale-free characteristics of random networks: The topology of the world wide web. *Physica A*, pages 69–77, 2000.

[3] P Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: an explanation of 1/f noise. *Physical Review Letters*, 59:381–384, 1987.

[4] P Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality. *Physical Review E*, 38:364–371, 1988.

[5] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[6] D. ben Avraham and S. Havlin. *Diffusion and Reactions in Fractals and Disordered Systems*. Cambridge UP, 2000.

[7] C.H. Bennett, P. Gács, M. Ling, P.M.B. Vitányi, and W.H. Zurek. Information distance. *IEEE Trans. Inform. Theory*, 44(4), 1998.

[8] D.A. Berry. *Statistics: A Bayesian Perspective*. Duxbury Press, 1996.

[9] D.P. Bertsekas. A distributed algorithm for the assignment problem. Technical report, Laboratory for Information and Decision Systems, March 1979.

[10] B. Bollobás. *Random Graphs*. Cambridge UP, 2001.

[11] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. CRC press, 1984.

[12] R. Bubley. *Randomized Algorithms: Approximation, Generation, and Counting*. Springer Verlag, 2001.

[13] B.A. Carreras, D.E. Newman, I. Dobson, and A.B. Poole. Evidence for self-organized criticality in a time series of electric power system blackouts. *IEEE Trans. on Cir. and Sys. -I*, pages 1733–1740, Sept. 2004.

[14] M.-H. Chen, Q.-M. Shao, and J. Q. Ibrahim. *Monte Carlo Methods in Bayesian Computation*. Springer, 2000.

[15] B. Cheng and D.M Titterington. *Neural Networks: A Review from a Statistical Perspective*, volume 9. Institute of Mathematical Statistics, 1994.

[16] A.C. Chiang. *Fundamental Methods of Mathematical Economics*. McGraw-Hill, 1984.

[17] W. Cook and A. Rohe. Computing minimum-weight perfect matchings. *INFORMS Journal on Computing*, 11:138–148, 1999.

[18] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley & Sons, 1991.

[19] L. de Arcangelis and H.J. Herrmann. Self-organized criticality on small world networks. *Physica A*, 308:545–549, 2002. cond-mat/0110231.

[20] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer Verlag, 2001.

[21] P.J. Diggle and R.J. Gratton. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society, series B*, 46:193–212, 1984.

[22] S.N. Dorogovtsev and J.F.F. Mendes. Evolution of networks. *Advances in Physics*, 51:1079–1187, 2002.

[23] S.N. Dorogovtsev and J.F.F. Mendes. *Evolution of Networks. From Biological Nets to the Internet and WWW*. Oxford University Press, 2003.

[24] M. Faloutsos, P. Faloutsos, and Faloutsos C. On power-law relationships of the internet topology. *Comput. Commun. Rev.*, 29:251–262, 1999.

[25] I.J. Farkas, I. Derényi, A.L. Barabási, and T. Vicsek. Spectra of "real-world" graphs: Beyond the semi circle law. *Phys. Rev. E*, 64:026704, 2001.

[26] A. Fronczak, P. Fronczak, and J.A. Holyst. Mean-field theory for the clustering coefficients in Barabási–Albert networks. *Phys. Rev. E*, 68:046126, 2003.

[27] D. Gamerman. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall, 1997.

[28] A. Gelman, J.B. Carlin, H.S. Stern, and D.R. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 1995.

[29] A. Gelman, X. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.

[30] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(2):721–741, 1984.

[31] C. Gourieroux, A. Monfort, and E. Renaulta. Indirect inference. *Journal of Applied Econometrics*, 8:pp. S85–Sl18, 1993.

[32] P. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[33] M.H. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, 1996.

[34] B. Hughes. *Random walks and random environments*. Oxford UP, 1995.

[35] F. Jasch and A. Blumen. Target problem on small-world networks. *Phys. Rev. E*, page 041108, 2001.

[36] S. Jespersen and A. Blumen. Small-world networks: Links with long-tailed distributions. *cond-mat/0009082*, 2000.

[37] S. Jespersen, I. M. Sokolov, and A. Blumen. Small-world rouse networks as models of cross-linked polymers. *J. Chem. Phys.*, page 7652, 2000.

[38] L. P. Kadanoff, S. R. Nagel, Wu. L., and S-M. Zhou. Scaling and universality in avalanches. *Phys. Rev. A*, 39:6524–6537, 1989.

[39] J. Kleinberg. The small-world phenomenon: an algorithm perspective. In F. Yao and E. Luks, editors, *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170. ACM, 2000.

[40] R.V. Kulkarni, E. Almaas, and D. Stroud. Evolutionary dynamics in the Bak–Sneppen model on small–world networks. cond/mat/9905066.

[41] S. Kullback. *Information Theory and Statistics*. Dover, 1968.

[42] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Math. Stat.*, 22:76–86, 1951.

[43] J. Lahtinen. Transformation discrepancy. *IEEE Transactions on Information Theory*, 2005. (submitted).

[44] J. Lahtinen and J. Heikkonen. Inference over uncomputable likelihoods, 2005.

[45] J. Lahtinen, J. Kertész, and K. Kaski. Scaling of random spreading in small world networks. *Physical Review E*, 64:057105, 2001.

[46] J. Lahtinen, J. Kertész, and K. Kaski. Random spreading phenomena in annealed small world networks. *Physica A*, 311/3-4:571–580, 2002.

[47] J. Lahtinen, J. Kertész, and K. Kaski. Sandpiles on Watts–Strogatz type small–worlds. *Physica A*, (Accepted for publication), 2004.

[48] J. Lahtinen and J. Lampinen. Reversible jump mcmc for two–state multivariate poisson mixtures. *Kybernetika*, 39(3):307–315, 2003.

[49] D.-S. Lee, K.-I. Goh, B. Kahng, and D. Kim. Sandpile avalanche dynamics on scale-free networks. *Physica A*, 338:84–91, 2004.

[50] M. Li and P. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer Verlag, 2. edition, 1997.

[51] M. Magdon-Ismail. No free lunch for noise prediction. *Neural Computation*, 12:547–564, 2000.

[52] S.S. Manna. Large-scale simulation of avalanche cluster distribution in sandpile model. *J. Stat. Phys.*, 59:509–521, 1990.

[53] S.C. Manrubia, J. Delgado, and B. Luque. Small-world behaviour in a system of mobile elements. *Europhysics Letters*, 54:693–699, 2001.

[54] A.D. Marrs. An application of Reversible-Jump MCMC to multivariate spherical Gaussian mixtures. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems 10*. MIT Press, 1998.

[55] S. Milgram. The small world problem. *Psychology Today*, pages 61–67, 1967.

[56] J. S. Milton and J.C. Arnold. *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences*. McGraw-Hill, 1995.

[57] R. Monasson. Diffusion, localization and dispersion relations on "small-world" lattices. *Eur. Phys. J. B*, pages 555–567, 1999.

[58] Y. Moreno and A. Vasquez. The Bak–Sneppen model on scale–free networks. *Europhysics Letters*, 57:765–771, 2002.

[59] C.F. Moukarzel. Spreading and shortest paths in systems with sparse long-range connections. *Phys. Rev. E*, 60(6):R6263–R6266, 1999.

[60] R.M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, September 1993.

[61] M.E.J. Newman, C. Moore, and D.J. Watts. Mean-field solution of the small-world network model. *Phys. Rev. Lett.*, 84(14):3201–3204, 1999.

[62] M.E.J. Newman, C. Moore, and D.J. Watts. Mean–field solution of the small-world network model. *Phys. Rev. Lett.*, page 3201, 2000.

[63] M.E.J. Newman and D.J. Watts. Scaling and percolation in the small-world network model. *Phys. Rev. E 60, 7332-7342*, 1999.

[64] H. Niederreiter. *Random number generation and Quasi Monte Carlo methods*. SIAM, Philadephia, 1992.

[65] S.A. Pandit and R.E. Amitkar. Random spread on the family of small-world networks. *Phys. Rev. E*, 63:041104, 2001.

[66] G. Parisi and M. Raiéville. On the finite size corrections to some random matching problems, 2002. cond-mat/0204595.

[67] R.K. Pathria. *Statistical Mechanics*. Pergamon Press, 1977.

[68] J. Puzicha, T. Hofmann, and J. Buhmann. Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, page 267. IEEE Computer Society, 1997.

[69] S. Redner. How popular is your paper? an empirical study of the citation distribution. *Eur. Phys. J. B*, 4:131–134, 1998.

[70] A. Reka and A. Barabasi. Topology of evolving networks: local events and universality. *Phys. Rev. Lett.*, 85:5234–5237, 2000.

[71] S. Richardson and P.J. Green. On the bayesian analysis of mixtures with an unknown number of components. *J. of Royal Stat. Soc., Series B*, 59:731–792, 1997.

[72] J. Rissanen. Lectures on statistical modeling theory. Technical report, Tampere University of Technology, 1999. Available at http://www.cs.tut.fi/∼rissanen/.

[73] C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, 1999.

[74] W. Rudin. *Principles of Mathematical Analysis*. McGraw–Hill, 1976.

[75] L. Savage. *The Foundations of Statistics*. Dover, 1972.

[76] A. Scala, L.A.N. Amaral, and M. Barthélémy. Small-world networks and the conformation space of a short lattice polymer chain. *Europhysics Letters*, pages 594–600, 2001.

[77] R.J. Schalkoff. *Pattern Recognition: Statistical, Structural and Neural Approaches*. John Wiley & Sons, 1992.

[78] M.F. Schlesinger, G.M. Zaslavsky, and U. (Eds.) Frisch. *Lévy Flights and Related Topics in Physics*. Springer, 1995.

[79] H.G. Schuster. *Deterministic Chaos*. Wiley-VCH, 3rd edition, 1995.

[80] D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, 1992.

[81] C. Shannon. *Collected Papers*. IEEE Press, 1992.

[82] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[83] I.M. Sokolov. Lévy flights from a continuous-time process. *Phys. Rev. E*, page 011104, 2001.

[84] A. Vasquez. Knowing a network by walking on it: emergence of scaling, 2000. cond-mat/0006132.

[85] A. Vehtari and J. Lampinen. Bayesian input variable selection using posterior probabilities and expected utilities. Technical Report B28, Helsinki University of Technology, 2002.

[86] V. Viallefont, S. Richardson, and P. Green. Bayesian analysis of Poisson mixtures. *Journal of Nonparametric Statistics*, 14(1-2):181–202, 2002.

[87] D.J. Watts. *Small Worlds*. Princeton UP, 1999.

[88] S. Watts and S.H. Strogatz. Collective dynamics of 'small–world' networks. *Nature*, 393:440–442, 1998.

ISBN 951-22-7489-2 (printed)
ISBN 951-22-7490-6 (PDF)
ISSN 1455-0474