Department of Information and Computer Science

Testing the Significance of Patterns with Complex Null Hypotheses

Niko Vuokko

A!

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 11/2012

Testing the Significance of Patterns with Complex Null Hypotheses

Niko Vuokko

Doctoral dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the School of Science for public examination and debate in Auditorium T2 at the Aalto University School of Science (Espoo, Finland) on the 18th of February 2012 at 12 noon.

Aalto University
School of Science
Department of Information and Computer Science

Supervisor
Heikki Mannila

Instructor
Petteri Kaski

Preliminary examiners
Martti Juhola, University of Tampere, Finland
Myra Spiliopoulou, Otto-von-Guericke-Universität Magdeburg, Germany

Opponent
Tapio Elomaa, Tampere University of Technology, Finland

Aalto University publication series
DOCTORAL DISSERTATIONS 11/2012

ISBN 978-952-60-4494-1 (printed)
ISBN 978-952-60-4495-8 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Unigrafia Oy
Helsinki 2012

Finland

The dissertation can be read at http://lib.tkk.fi/Diss/

441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Niko Vuokko

Name of the doctoral dissertation
Testing the Significance of Patterns with Complex Null Hypotheses

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 11/2012

Field of research Computer and Information Science

Manuscript submitted 20 September 2011 **Manuscript revised** 17 November 2011

Date of the defence 18 February 2012 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

In data mining large amounts of data are searched through for useful information, pieces of which are called patterns. Significance testing is an important part of this task as the found patterns need to be assessed for their relevance and significance before further actions.

Advances in science have brought along the need to evaluate the significance of complicated data patterns within complicated datasets. Significance testing has been historically conducted with specialized methods that cannot be adapted to new applications and many of these methods have problems with their theoretical justification.

This thesis suggests using the framework of property-based randomization for building reliable and flexible significance testing tools that can be adapted and extended for a wide variety of applications. The concepts of representation-based randomization and iterative pattern mining are also discussed as ways to enlarge the scope of these tools.

The final chapter of the thesis makes a review of the use of these general ideas in various applications such as databases and time series collections. The publications of the thesis are discussed along with selected introductions to other randomization methods that have been proposed.

Keywords data mining, significance testing, randomization, null hypothesis, null model, MCMC, frequent pattern, clustering, classification, time series

ISBN (printed) 978-952-60-4494-1 **ISBN (pdf)** 978-952-60-4495-8

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Espoo **Location of printing** Helsinki **Year** 2012

Pages 178 **The dissertation can be read at** http://lib.tkk.fi/Diss/

A! Aalto-yliopisto

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tiivistelmä

Tekijä
Niko Vuokko

Väitöskirjan nimi
Monimutkaisten nollahypoteesien käyttö tietohahmojen merkitsevyyden arvioinnissa

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietojenkäsittelytieteen laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 11/2012

Tutkimusala Informaatiotekniikka

Käsikirjoituksen pvm 20.09.2011 **Korjatun käsikirjoituksen pvm** 17.11.2011

Väitöspäivä 18.02.2012 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Tiedonlouhinnassa käydään läpi suuria tietomääriä ja etsitään niistä hyödyllistä informaatiota. Merkitsevyystestaus on keskeinen osa tätä tehtävää, sillä löydettyjen tiedonjyvien oleellisuus ja merkitsevyys tulee tarkistaa ennen jatkotoimenpiteitä.

Tieteen kehittyessä on muodostunut tarpeelliseksi etsiä entistä monimutkaisempia rakenteita tietojoukoista, joiden koko on samalla kasvanut. Historiallisesti merkitsevyystestaamista varten on usein käytetty erikoistuneita menetelmiä, joita on ollut vaikea sovittaa uusiin ongelmiin. Lisäksi monia näistä menetelmistä on vaikeaa perustella teoreettisesti.

Tämä väitöskirja ehdottaa yleistä merkitsevyystestauksen mallia, jossa tietolähteen ominaisuuksille perustuvista nollamalleista rakennetaan erikoistuneita, mutta samalla luotettavia ja erityisen hyvin muokattavia merkitsevyystestausmenetelmiä. Tämän yleisen mallin käyttömahdollisuuksia laajennetaan vielä lisäksi esityspohjaisen satunnaistuksen ja iteratiivisen merkitsevyystestauksen ratkaisuilla.

Työn viimeisessä osassa tämän yleisen mallin toimintaa esitellään monipuolisella joukolla sovelluksia esimerkiksi tietokannoille ja aikasarjakokoelmille. Väitöskirjan julkaisut esitellään yleisellä tasolla yhdistämällä niiden sisältö muihin ehdotettuihin satunnaistusmenetelmiin.

Avainsanat tiedonlouhinta, merkitsevyystestaus, satunnaistus, nollahypoteesi, nollamalli, MCMC, usein toistuva hahmo, ryvästys, luokittelu, aikasarjat

ISBN (painettu) 978-952-60-4494-1 **ISBN (pdf)** 978-952-60-4495-8

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Espoo **Painopaikka** Helsinki **Vuosi** 2012

Sivumäärä 178 **Luettavissa verkossa osoitteessa** http://lib.tkk.fi/Diss/

Contents

List of Publications

This thesis consists of a summary part of 5 chapters and the following 5 publications:

I Markus Ojala, Niko Vuokko, Aleksi Kallio, Niina Haiminen and Heikki Mannila. Randomization Methods for Assessing Data Analysis Results on Real-Valued Matrices. *Statistical Analysis and Data Mining*, 2(4), pages 209–230, 2009.

II Sami Hanhijärvi, Markus Ojala, Niko Vuokko, Kai Puolamäki, Nikolaj Tatti and Heikki Mannila. Tell Me Something I Don't Know: Randomization Strategies for Iterative Data Mining. In *KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 379–388, 2009.

III Niko Vuokko and Petteri Kaski. Testing the Significance of Patterns in Data with Cluster Structure. In *ICDM '10: Proceedings of the 10th IEEE International Conference on Data Mining*, pages 1097–1102, 2010.

IV Niko Vuokko and Petteri Kaski. Significance of Patterns in Time Series Collections. In *SDM '11: Proceedings of the 11th SIAM International Conference on Data Mining*, pages 676–686, 2011.

V Aleksi Kallio, Niko Vuokko, Markus Ojala, Niina Haiminen, and Heikki Mannila. Randomization Techniques for Assessing the Significance of Gene Periodicity Results. *BMC Bioinformatics*, 12:330, 2011.

This numbering of the publications is used throughout the thesis.

Preface

The work presented in this thesis has been carried out at the Department of Information and Computer Science of Aalto University School of Science from 2007 to 2011. I am thankful for the funding of my work provided mainly by the Finnish Doctoral Programme in Computational Sciences (FICS) and the Academy of Finland funded Finnish Centre of Excellence for Algorithmic Data Analysis Research (Algodan). The research scholarship grant from the Nokia Foundation is also gratefully acknowledged.

I wish to thank my supervisor Professor Heikki Mannila for all the advice and high level vision on how to think about problems in data mining and how to conduct research on it. I also thank my advisor Petteri Kaski for all the effort he put in helping me finalize my work and compile this thesis. Furthermore, I owe special gratitude to my colleague and friend Markus Ojala for our mutual support.

Collaboration with my co-authors and colleagues has always been a pleasure and I have gained a lot of understanding in all the discussions I have had with the great personalities of Jaakko Hollmén, Kai Puolamäki, Evimaria Terzi, Jefrey Lijffijt, Panagiotis Papapetrou, Jaakko Talonen and many others. I thank all the people at the department for together creating a wonderful atmosphere, which has been a great motivator for me.

I am forever thankful for the support and example given to me by my parents and Vaija, in life and in pulling through my graduate studies. Finally, for all this and much more, I thank Leena and Lauri for all the fun in life.

Turku, December 2011

Niko Vuokko

Chapter 1

Introduction

1.1 Overview and Motivation

During the past decades, advances in computer technology have enabled a fundamentally new, data-driven approach to research. Instead of conducting meticulous theoretical work to come up with new ideas and hypotheses, it is possible to gather and analyze unprecedented amounts of experimental data with the intent of letting the data point out new ideas instead of just verifying them. This "discovery science" [1, 2] paradigm brought the need for *data mining*, that is, using computational methods to explore large datasets for significant and interesting pieces of information called *patterns*. This knowledge can then be used for the better understanding of the underlying phenomena or for making predictions.

Example 1.1. *A supermarket gathers information about their customers through the use of loyalty cards that can be used for tracking customer behavior. In addition to modeling the sociological phenomenon and finding new marketing techniques from the models, the supermarket can simply use the collected data to see what is actually happening and optimize its marketing directly based on what works best in computer simulations.*

Finding new information from data is not enough. Large amounts of data always contain also large amounts of details occurring merely by random chance. Instead of using theory to generate hypotheses then verified by data, the data-centric process requires theoretical work to verify the hypotheses generated by data. However, testing the significance of patterns in a correct manner has proven to be surprisingly difficult [3, 4]. In addition to the philosophical problems in significance testing, there are also pitfalls in formulating the *null hypothesis* that defines the particular claim evaluated and in testing the significance of the null hypothesis.

With time there has been a growing interest in seeking out a wider range of different types of patterns from data and avoiding standard assumptions such as gaussianity in cases where they cannot be reasonably validated. Assessing the probability of a pattern arising by random chance in such a setting requires good modeling of the data source incorporating this model to the significance test. This work describes a framework suitable for situations where the data source is difficult to model. In principle, this is done by letting the practitioner choose high level properties that he wishes to preserve in the data samples and then it is the duty of the framework to establish a resampling scheme in accordance with these wishes.

The work of this thesis emphasizes also a transition towards data-centric thinking in significance testing. Just as in the discovery science paradigm where the findings from data are evaluated with theory, this work develops significance testing methods around what the data contains and how it behaves instead of estimating these traits from theory, which is the common earlier approach to significance testing. Transitioning to a data-centric setting in significance testing simultaneously shifts the problem closer to computer science and data mining and further from traditional statistics.

Example 1.2. *The supermarket has identified an interesting pattern in their data and now wishes to test whether this pattern is significant and worth putting to use. Instead of evaluating the probability of the pattern to arise from some model*

of customer behavior, the supermarket can evaluate the probability that the pattern exists in datasets that are similar to the original data.

With the data source properties gathered, we can state a *property-based null model* for our testing which, given a dataset and possible parameters, defines a probability distribution over all datasets in the sample space. In basic terms, a property-based null model is simply a collection of properties and rules on how they need to be preserved when choosing the data samples. Property-based null models are defined and studied more closely in Section 3.2.

The samples from the null model are produced using *Markov chain Monte Carlo* methods. These methods gradually perturb a dataset with local operations and control the process by managing the movements of the Markov chain. However, many data properties are difficult to measure and manage effectively, complicating their use in significance testing, especially in cases where these properties are an integral part of the character of the data. This problem can, in some cases, be overcome with the use of data representations that expose these properties and suitably separate them from the rest of the data.

The ability to include diverse and multiple properties in a null model makes it easier also to iteratively modify the null model and compare the significance results. With such measures it is then possible to understand better which data patterns depend on or are implied by certain data properties. Different null models correspond to different contexts for the original data, for the patterns tested and for the results of significance tests. These results might or might not be similar for different contexts.

The use of data mining has gradually expanded from the relatively simple applications for static database data such as in Examples 1.1 and 1.2 to a wide variety of complex data sources such as social networks, industrial process monitoring or gene behavior. These complexities on the data source level are then combined with mining increasingly convoluted patterns, requiring high adaptability also from the significance tests that are

used.

Example 1.3. *Electroencephalography (EEG) and functional Magnetic Resonance Imaging (fMRI) are two technologies used for monitoring brain activity. The properties of the ensuing datasets are however rather different. For example, the poor and convoluted spatial resolution of EEG is compensated by its excellent temporal resolution, whereas for fMRI the situation is reversed. The complications in these methods are fundamental and can be dealt with only once the data is analyzed. In addition to these complications, significance testing for these datasets may require taking into account specific features such as physiological time delays, prior knowledge of brain structure and comparisons or aggregates across multiple test subjects.*

The purpose of this work is to highlight some of the present problems in significance testing and find ways to make the process more automated for wider use while providing reliable significance assessment. The work has two main themes. First, one core element of the work is the use of property-based null models that can be used to reduce the amount of application-specific theoretical work needed. Second, providing better access to the components of a null model through the use of data representations is explored. These main themes of the work are accompanied by significance testing methods for specific applications to demonstrate practical use cases.

The rest of the thesis is organized as follows. Chapter 2 discusses the basics of statistical significance testing and defines the core aspects. Chapter 3 discusses the randomization approach to significance testing and introduces the concepts of property-based null models and representation-based randomization. Applications of these ideas are discussed in Chapter 4, where problems in testing the significance of complex null hypotheses are demonstrated within various fields of science. Finally the work is concluded in Chapter 5.

1.2 Summary of the Publications and Contributions of the Author

Publication I discusses the problem of randomizing real-valued matrices that have comparable rows and comparable columns. The paper considers two separate tasks where either the means and variances or the full value distributions of each row and column are approximately preserved during randomization. Two main algorithms are introduced for the problem and these algorithms are then combined with various local modification operations and error measures. The first published version [5] of the paper considers only the simpler task of preserving the means and variances.

In the first published version [5] of Publication I, the methods were developed jointly with Markus Ojala. The current author developed most of the theory while the experiments were mostly performed by Ojala. The text was written jointly by all authors. Most additional work in Publication I was made by Ojala.

Publication II considers the problem of finding independent patterns from data. The paper proposes a general scheme of randomization where previously found patterns are iteratively added to the null model and fixed during randomization, allowing patterns not explained by the earlier findings to be found. The approach is tested with binary matrices for which row and column sums, itemset frequencies and cluster structure are used as different pattern types.

Forming the original idea and the theory for the work and designing and interpreting the experiments were conducted jointly by the authors. Hanhijärvi and Ojala implemented the algorithms and conducted the experiments.

Publication III proposes a solution for testing the significance of patterns against the null hypothesis "structure in the data is a result of the cluster structure". The proposed method bases itself on the theoretical connections between principal component analysis and the k-means algorithm.

The paper discusses applications in exploring the relations of results in unsupervised and supervised learning for a dataset and in finding patterns that are independent from the cluster structure of the dataset. The paper is briefly expanded to include multiple other forms of structure in data.

Forming the original idea and theory and running experiments was done by the current author. Verifying the theory, planning experiments and writing the paper was jointly done with Kaski.

Publication IV proposes a wavelet-based null model for the significance testing of time series collections and introduces a randomization method that is compatible with this null model. The null model is built on three constraints on the perturbation in time, frequency and across the individual time series in the collection. The model is compared with multiple other (implicit) null models often used for the task. Experimental results for the different randomization methods are compared with each other and reflected against the theoretical analysis.

The current author was responsible for the original idea, theory and running experiments. Verifying the theory, planning experiments and writing the paper was jointly done with Kaski.

Publication V proposes a new approach for testing the significance of gene cyclicity by comparing the observed expression levels to the others in the dataset. The main advantage of this approach is its suitability for the inherently low quality data that also contains unnatural substance arising from the experimental setting. Theoretical analysis and experimental results are used for measuring the success of the approach and for comparing it with other methods. The general process of the problem is also discussed in detail and best practice recommendations are made.

The design and development of the methods was done by all authors. The current author carried out the mathematical analysis, whereas Kallio implemented the method and carried out the experiments. The manuscript was written jointly by the current author and Kallio.

CHAPTER 2

Significance Testing of Patterns in Data Mining

Data mining is the general concept for exploring large datasets on the lookout for interesting *patterns*, that could be of some use [6]. The large amount of data to be processed is often construed to be one of the conditions in qualifying an activity as data mining. This condition sets limits on the intricacy of the analysis tools, but also implies that the data is usually filled with a myriad of various details any of which could be deemed interesting in someone's opinion. Therefore the process of data mining is split into a set of different analyses that explore different aspects of the data.

A data mining algorithm f is used for searching the patterns of the pattern collection $\mathcal{P}_f$. Given a dataset $\mathcal{X}$ as input, the algorithm outputs a set of patterns $f(\mathcal{X}) = \mathcal{P} \subseteq \mathcal{P}_f$. However, not all of these patterns deserve a second look and it is important to prune out the trivial patterns from $\mathcal{P}$ to improve the efficacy, cost and quality of the data mining process. The main source of these trivial patterns is the fact that although the data mining algorithm f is of general purpose, no dataset is ever just "generic". Any dataset has its set of special features intrinsic to the data source, features that are considered trivial for the data coming from this source, but not in

the general case.

Example 2.1. *Depending on the dataset and our prior knowledge and expectations on its structure, correlations in data can be called interesting in some case if they are strongly positive or negative, close to zero or anything there between. Suppose then that we are exploring national statistics of law enforcement and restrict our interest in strong correlations between variables. Despite this restriction, high correlations between city size, the level of law enforcement funding and the amount of crime are certainly not interesting, but the correlation between funding per capita and the amount of crime may be interesting, regardless of its value.*

To improve the data mining results we thus face the problem of automatically recognizing the reported patterns that might or might not deserve further analysis [7]. To this end we begin with an assumption of the ground truth that includes our prior information on the features of the data source. We then conduct *statistical significance testing* [8], where we call a pattern *significant* and accept it for further analysis if we find it unlikely that the pattern could have arisen by sheer chance from the ground truth.

The decision on the significance of a pattern can of course never be deterministically correct as long as we live in a probabilistic world. Therefore the best we can hope for is to provide probabilistic bounds on our degree of confidence in the significance of a pattern. Nonetheless, there is also another, bigger problem in assessing the significance of patterns. As defined above, significance is always a relative term. This means that the quality of significance testing results is a direct consequence of our ability to specify the ground truth both correctly. Deficiencies in specifying the features of the ground truth result in spurious patterns declared significant, qualitative errors that are often very difficult to detect even in closer analysis. The difficulties in understanding the ground truth are additionally highlighted in the present tendency towards more complex null hypotheses.

2.1 Examples of Patterns in Data Mining

A pattern in data can be any type of extractable piece of information that might be of value in the analysis of the data. The use of the word "information" is intentional, because a pattern can be of use only if it delivers non-trivial information of the dataset in the sense of information theory and Shannon entropy [9].

Generally patterns can be divided into two classes based on the type of information they provide. The first class contains descriptive patterns that give new knowledge on what types of non-random structures the data contains. The second class contains predictive patterns that can be used for inference. In the usual workflow of data analysis [7] we start with pattern mining in the sense of "nugget searching": exploring the dataset for descriptive patterns. After this has been achieved, our enhanced understanding of the data allows us to move on to doing machine learning: seeking predictive patterns and utilizing them for data-driven decision making in operative real world applications. Thus the technical descriptive patterns provide an initial stepping stone for devising practical solutions to real world problems.

The borderline between the two pattern classes is somewhat vague, but in general a descriptive pattern tells us something that is certain and reduces the corresponding entropy to zero whereas a predictive pattern reduces the uncertainty and residual entropy (limited by the mutual information). For example, "all sheep have four legs" is a descriptive pattern, but "most sheep are white" is a predictive pattern. Relative to the pattern classes, there is a dual separation of data mining tasks into unsupervised methods that seek out descriptive patterns and supervised methods that attempt to find the most useful predictive patterns for a given task.

Simple examples of descriptive patterns are correlations of variables, frequent itemsets, subgroups of data points or temporal dependency and variations. Despite their predictive nature, association rules are also descriptive patterns that enumerate a certain class of possibly useful infer-

ence rules. Similarly possible clustering in data is also a descriptive pattern, which can be used for associating points with each other. More specialized examples of descriptive patterns include behavioral profiles of traffic flows [10] and biological [11] or computer viruses [12], functionalities of a gene [13] or the seriation of paleontological data [14].

Supervised methods generate predictive patterns that attempt to estimate the values of some output function given the inputs in data. In classification the output function designates data points with pre-determined labels. In regression tasks there are no necessary constraints to the nature of the target function. For example, a face recognition system may conduct the task of classifying subjects to men and women, but also conduct regression to estimate their age.

2.2 Statistical Significance of Data Mining Results

The meaning and nature of statistical significance remains a somewhat controversial topic in statistics [15]. In this work, however, we take a practical view of the matter and work along the lines generally accepted for use in the data mining community.

The ground truth on which our significance testing is based is called the *general hypothesis*. This hypothesis contains all the fundamental properties of the data source and also any data-specific properties that we wish to treat as an integral part of the data. The realism of these properties is an essential part of the realism of our significance testing results.

The significance of a pattern P is tested with a pair of null hypothesis and alternative hypothesis. The *null hypothesis* H_0 corresponds to the claim

> *" Pattern P is a consequence of the general hypothesis "*

In other words, it corresponds to the case of finding no significance in pattern P. The *alternative hypothesis* is the complementary claim of declaring the pattern P significant. Statistical significance testing assesses whether

the null hypothesis can be accepted or rejected with some required level of confidence.

Example 2.2. *A supermarket gives its customers loyalty cards, which can be used to get discounts for purchases. This function allows the supermarket to record each customer's shopping history to a common matrix of customers and products, each entry showing the number of bought items by the customer. The supermarket has identified some interesting behavioral patterns among its customers, but before proceeding with these findings it wants to see whether the discoveries are significant and worth pursuing or mere coincidence.*

The general hypothesis in this significance testing scenario must, in the least, acknowledge the highly sparse nature of the data as no customer can ever make purchases of a significant subset of all the 10 000 different products offered. Additionally the general hypothesis may assume properties on how many transactions are approximately conducted per month, how popular each item is in general or how often different customers come to shop and how much they buy in item count or in dollar amount.

All these properties and many more need to be evaluated before conducting significance testing to ensure that the results reflect whatever was the purpose of the assessment.

In mathematical terms, assuming the null hypothesis H_0 corresponds to limiting the probability space Ω and modifying its probability measure Pr to suit the general hypothesis. The assessment on the null hypothesis is based on comparing the values of a certain test statistic t, given by some measure of interest that indicates the strength of pattern P in a given dataset. Choosing the measure of interest has very few requirements. Any function $t : \mathcal{D} \to \mathbb{R}$ is admissible, where $\mathcal{D}$ is the space of the datasets.

Example 2.3. *Let pattern P be a certain correlation between two data features. In this case the value of the correlation can be chosen as the measure of interest, indicating the strength of P in a dataset.*

Example 2.4. *The supermarket of Example 2.2 has found out that customers who buy eggs and milk often have also sugar in their cart. To test the significance of this association rule* $\{eggs, milk\} \rightarrow sugar$*, commonly used measures like* support *(number of transactions which contain all three products),* confidence *(how often sugar is bought in cases where eggs and milk are) or* lift *(comparison to the case of random recombination) can be utilized as the test statistic.*

However, in many cases the choice of the measure is not as simple to make and the analysis must take into account the effect of the measurement choice in the significance testing results. For example, there is no simple way to measure the strength of cluster structure in data. Different choices may range from eigenvalue-related measures [16] or clustering errors with given methods and parameters to computing conductance [16] or the Dunn index [17] or to general measures such as the Bayesian Information Criterion [18].

Once the test statistic t has been chosen, the strength of pattern P in dataset D as measured by t is computed. Pattern P is assessed more significant the more the value $t(D)$ deviates from what one could expect to observe under the null hypothesis. In a *one-tailed test* only exceptionally large (or small) values of $t(D)$ are considered interesting. In this case the significance of P is a monotonically increasing function of $t(D)$. If we on the other hand are interested in both large and small values of $t(D)$, the test is called a *two-tailed test*.

Example 2.5. *When assessing the significance of correlations both a one-tailed or a two-tailed test can be utilized, depending on the specific goals of the analysis. Using a two-tailed test on the value of the correlation reports both strongly negative or strongly positive correlations as significant. The same can be accomplished by utilizing a one-tailed test to the absolute correlation value. On the other hand, testing the original correlation value with a one-tailed test will test only for exceptionally strong positive correlations.*

For the ease of presentation, let us now consider only the case of a one-tailed test. The null hypothesis is rejected with *level of confidence* α if

$\Pr(t(X) \geq t(D)|H_0, D) \leq \alpha$. In other words, the probability of observing a value at least as extreme as $t(D)$ should be at most α under the null hypothesis.

Definition 2.6. *The probability* $\Pr(t \geq t(D)|H_0)$ *is called the* p-value *of the pattern P that the test statistic t tests for. To compute the p-value, the distribution of the values of t under* H_0 *needs to be estimated. This distribution is called the* null distribution *and the underlying probabilistic model in the sample space is called the* null model *[19, 20].*

In most cases it is not possible to explicitly state the null distribution as either the data source is not known well enough or the null hypothesis is too complex [21, 22]. In *analytical significance testing* the null distribution is explicitly known. In this case the p-values can be exactly computed from the cumulative distribution function. Typical analytical null distributions include the χ^2, binomial, Student's t-distribution and the Fisher's F-distribution [23]. In many cases more complicated null distributions can also be simplified with the central limit theorem. Additionally, also the use of general probability bounds such as the Chebyshev and Chernoff bounds or the Azuma-Hoeffding inequality is considered analytical significance testing.

Example 2.7. *Acme Inc. has recently received complaints claiming that the company discriminates obese people in recruitment. Foregoing the complexities arising from gender, age and ethnicity among other factors, the company acquires public research results on the body mass index averages of the general population and then conducts an anonymous survey for its workforce.*

Assuming that both data follow normal distributions of equal variance, the Student's t-test can be used for testing whether the data means differ in significant extent. In this case the test statistic is the difference between the two data means normalized by sample size and variance. The final p-value used for assessing the validity of the claims is then the probability that a random draw from the relevant Student's t-distribution would have a lower value than the value of the

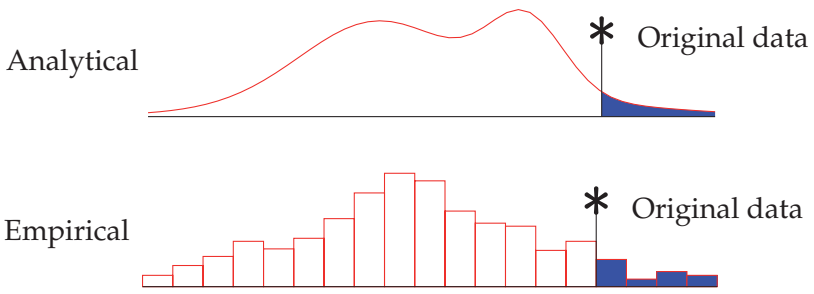


Figure 2.1: Illustration of analytical and empirical significance testing. The filled blue areas indicate the p-value, in other words, the cumulative probability of test statistic values exceeding the value for the original dataset. In the analytical test this value of the cumulative distribution function can be directly computed whereas in empirical testing it needs to be estimated from a histogram.

test statistic. This can be done by a simple evaluation of the cumulative distribution function.

In the usual case of not being able to compute exact p-values from an analytical null distribution, *empirical significance testing*, also called *Monte Carlo testing*, is used to provide *empirical p-values* [21]. This may be necessary in a case where the probability density function (pdf) of the null distribution is known, but it is not possible to integrate over it. In other cases the pdf is known only in proportion without knowledge of the normalization constant. In some cases the pdf is in its entirety unknown. However, empirical significance testing has been found useful also in cases where analytical testing is feasible, but cumbersome [21]. Figure 2.1 gives an illustration of the difference between analytical and empirical testing.

A special advantage of empirical significance testing is that it allows the use of almost arbitrary test statistics, whereas the user of analytical significance testing is constrained to those test statistics whose probabilistic

behavior is known. The only general requirement for conducting empirical significance testing is the capability to sample from the null distribution. The sampling process makes empirical significance testing usually considerably slower than analytical significance testing procedures.

Empirical significance testing employs the law of large numbers to obtain reliable estimates on the p-value [24]. Let us denote the true p-value with $p^* = \Pr(t \geq t(D)|H_0)$. The null distribution is sampled for N independent and identically distributed (i.i.d.) samples D_i, for each of which the event "$t(D_i) \geq t(D)$" follows the Bernoulli distribution with success probability p^*. Combined with the original dataset D for which the null hypothesis is assumed true, there are $N + 1$ i.i.d. samples which can be used for estimating the p-value. The sum of their test statistic values follows the binomial distribution $\mathrm{Bin}(N + 1, p^*)$. Therefore the unbiased empirical estimator of the true p-value p^* can be written as

$$p = \frac{1}{N+1}\Big(1 + \sum_{i=1}^{N} \mathbb{I}(t(D_i) \geq t(D))\Big).$$

The variance of this estimate is $p^*(1 - p^*)/N$ denoting how the quality of an empirical p-value improves as the number of samples N grows.

Example 2.8. *A biology student comes up with a hypothesis that the trees in a nearby little grove are not randomly distributed but rather form clusters. She estimates the shape and size of the grove and measures the distances from each tree to its closest neighbor, letting the sum of these distances to form her test statistic. Then she generates random samples by placing the same amount of trees uniformly at random to a same-shaped grove on her computer. Finally she tests her hypothesis by comparing the original sum of distances to those obtained from the random samples.*

There are few differences between analytical and empirical significance testing. In general, analytical testing can be seen as an aspiration and empirical testing as the reality check needed in the majority of cases. Re-

gardless of the type, the process of statistical significance testing goes as follows:

1. Choose the general hypothesis, the "ground truth".
2. Choose the null hypothesis H_0, the claim whose truth value we are evaluating.
3. Choose a test statistic t that can be used to evaluate H_0.
4. Choose the required level of confidence α for rejecting H_0.
5. Estimate or compute $p = \Pr(t \geq t(D)|H_0)$ and reject H_0 if $p \leq \alpha$.

2.3 Multiple Hypothesis Testing

In exploratory data analysis there is never only one single data pattern that we are looking for. Rather, the analysis will give us a large collection of different patterns that we then need to validate for significance. Testing the significance of all these patterns separately is equivalent to having a separate null hypothesis for each of the tested patterns. This simultaneous significance testing is called *multiple hypothesis testing* [25, 26].

Testing the significance of multiple hypotheses is even more complicated than the case of a single null hypothesis. Additionally, there is no single correct way to conduct it. Instead, the user needs to decide what he or she wants to accomplish and correspondingly choose an appropriate method for the purpose.

Example 2.9. *Suppose we are mining a matrix of size $m \times n$ for significant correlations between attributes (columns) and use a statistical test of level α for significance testing. The total number of attribute pairs and thus of null hypotheses is $\binom{n}{2}$ and each true null hypothesis is declared false (Type I error) with probability α. This however means that even if the data is fully random and all null hypotheses are true, on expectation $\alpha\binom{n}{2}$ null hypotheses are declared false and the probability of having no false positives shrinks to $(1-\alpha)^{\binom{n}{2}} \approx e^{-\alpha\binom{n}{2}}$.*

To analyze the setting of multiple hypothesis testing we express all the different possible outcomes in a two-way table as shown in Table 2.1 [27]. In the case of multiple null hypotheses the concept of significance becomes more complicated [27, 28, 29]. Should we minimize the number of false negatives T, false positives V or both? Or should maximize the "signal-to-noise" ratio S/R of true positives to the total count of significant findings? These questions do not have a definitive answer and each application requires its own consideration of what are sufficient and necessary requirements for the results of significance testing. This consideration involves striking a balance between risks and costs for acting on different types of error [29, 30].

Table 2.1: Common table representation used for analyzing multiple hypothesis testing. The entries in the table correspond to the following terms: U: True negative, V: False positive (Type I error), T: False negative (Type II error) and S: True positive.

	Declared non-significant	Declared significant	Total
True hypotheses	U	V	m_0
False hypotheses	T	S	$m - m_0$
Total hypotheses	$m - R$	R	m

Example 2.10. *Filtering incoming e-mail for spam or analyzing criminal evidence to indicate the guilty requires a strict control of false positives V. On the other hand, in detecting computer viruses or terrorists attempting to sneak a bomb somewhere it is preferable to cause some false positives to maximize the true positive count.*

A statistical test $\mathcal{A}$ is called more *conservative* than test $\mathcal{B}$ if $\mathcal{A}$ outputs less false positives than $\mathcal{B}$. Conversely, $\mathcal{B}$ is called more *optimistic* than

$\mathcal{A}$. In addition to these attributes, multiple hypothesis tests are qualified also based on how they behave in situations where a varying proportion of the null hypotheses are true or false. A statistical test provides *strong control* if it successfully controls the Type I errors for an arbitrary number of true or false hypotheses, that is, for any value of m_0 in Table 2.1. A test has *weak control* on the other hand if it can reliably control the Type I errors only under the complete null hypothesis $m_0 = m$. In general, weak control without any other safeguards is unsatisfactory for decision making. See [29] for a more comprehensive review of these terms and concepts.

Data mining is an exploratory data analysis task whose results are used as input to more elaborated analysis phases. The main objective of such *screening* activity is to substantially reduce the set of tested hypotheses (m in Table 2.1) while carrying over all the true positive cases to the next analysis phase [31].

Example 2.11. *In drug discovery large amounts of chemical compounds are screened for effectiveness. A false negative in this process carries the cost of lost profit on a new drug while a false positive incurs only the cost of running additional analysis on the compound. This skewed cost structure promotes using some version of the S/R ratio for significance testing.*

There are two main control measures of significance in multiple hypothesis testing. The *familywise error rate* (FWER) [32, 33, 34] is defined as the probability $\Pr(V > 0)$ of reporting any false positives. Controlling the FWER assigns high costs on any false positives. As discussed above, high hypothesis counts m create problems with the FWER since it becomes increasingly difficult to control $\Pr(V > 0)$ even if the ratio R/m does not change. The Holm-Bonferroni procedure [33] is the most commonly used method for the strong control of the FWER. With threshold α this procedure starts by sorting all the p-values into descending order $p_1 \geq p_2 \geq \cdots \geq p_m$ and multiplying each p-value with its index: $q_i \leftarrow ip_i$. After this, the null hypotheses are evaluated in the increasing order of the

numbers q_i. Hypothesis i is rejected if $q_i \leq \alpha$ and all hypotheses prior to i have also been rejected. If null hypothesis i is accepted, all hypotheses $j > i$ are also accepted without further evaluation.

The other common measure of significance is the *false discovery rate* (FDR) [35, 27]. Assuming $R > 0$, the FDR is defined as the expected ratio $\mathbb{E}[V/R]$ of false positives from all subjects reported as significant. The FDR is clearly less conservative than FWER although controlling FDR does provide weak control on FWER. However, this also means that the FDR is much better at preventing false negatives, which is usually more important in data mining. The classical procedure for controlling the false discovery rate is the Benjamini-Hochberg method [27], which is valid for hypothesis families that are pairwise independent or positively correlated. This procedure, originally introduced in [34] for the weak control of FWER, proceeds similarly to the Holm-Bonferroni method discussed above except that the sorted p-values are multiplied by m/i instead of i, that is, $q_i \leftarrow mp_i/i$. The Benjamini-Hochberg-Yekutieli method [36] is the extension of this procedure that is valid under arbitrary dependency structures.

The FDR ratio $\mathbb{E}[V/R]$ is not well defined if $\Pr(R = 0) > 0$. Additionally, controlling the FDR becomes overly optimistic when $\Pr(R = 0)$ is high. A solution to this problem, called the *positive FDR* (pFDR) and defined as $\mathbb{E}[V/R|R > 0]$, is introduced in [31]. As a distinct theoretical advantage, the pFDR is equal to the Bayesian posterior probability of a null hypothesis being true when modeling the test statistic as a probabilistic mixture of cases where the null hypothesis is true or false.

As a final note, whereas multiple hypothesis testing is not the topic of this work, it is an important part of analyzing the significance of patterns in data mining and has been widely used in the publications of this thesis.

CHAPTER 3

Randomization in Significance Testing

Sampling a probability distribution of unknown character is a general problem widely faced not only in data mining and machine learning, but also, for example, in physics [37]. Lacking knowledge of the exact null distribution, it is necessary to estimate it. Considering the expanse of possible probability distributions, it is often easier to start with the null model that describes the null distribution in more general terms. Although the null model and the specific null distribution are simply two different views of the same probabilistic entity, the higher level view and the lesser amount of detail makes the null model a better initial approach when the underlying data source is not fully understood.

It is not obvious how to conduct empirical significance testing without defining an explicit null distribution. Such testing task is often done by taking the original dataset and manipulating it with certain operations until the result is somehow determined to be independent of the original data. This procedure is then repeated multiple times to attain the needed test samples. It is important to ensure that the final manipulation result of the *randomization* process meets the requirements set in the null model.

The next two sections 3.1 and 3.2 discuss two different approaches to randomization. Explicit null models describe the null distribution in a detailed sense, but leave some of the specifics to be decided, for example, based on the actual dataset. In practice this often means choosing the family of the null distribution upfront and then fixing the parameterization of the null distribution by fitting it to the dataset that is tested.

Property null models, on the other hand, are used when little confidence can be placed on assumptions about the details of the null distribution. Instead, property null models describe the properties that the null distribution must possess without directly dictating the way this requirement materializes in the distribution. In practice, one way to convert these requirements to an actual null distribution is to let the required properties vary based on normal distributions. However, the underlying null distribution in the sample space is usually not known explicitly. Sections 3.2 and 3.3 will discuss this in more detail.

3.1 Explicit Null Models

A null model is called *explicit* if it explicitly defines the corresponding null distribution. In the typical case the available knowledge about the data source and the test statistic are used for choosing a class of probability distributions as the explicit null model. The estimated null distribution is then constructed from this class of parameterized distributions based on parameter values inferred from data. The whole process of explicit null model based randomization is depicted in Figure 3.1.

Example 3.1. *The son of a famous statistician wants to test whether a coin his father gave him is fair, that is, with equal odds. His null hypothesis states that the probability of the coin ending up as 'heads' is 0.5 and also rules out any other outcomes than 'heads' and 'tails'. This provides him an explicit null model where the number of 'heads' H resulting from a row of N tosses follows the binomial distribution: $H \sim Bin(N, 0.5)$. The boy then proceeds to toss the coin for, say,*

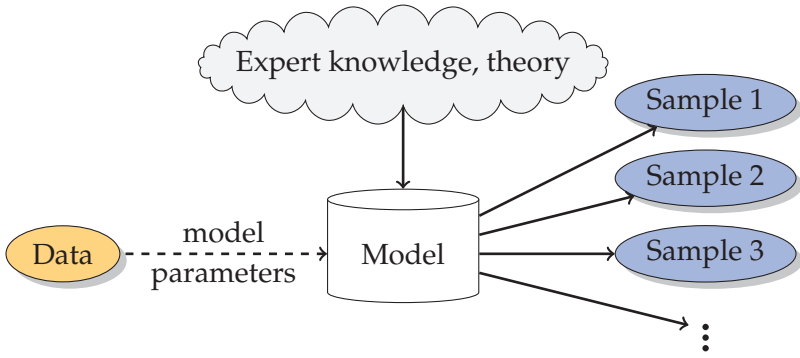
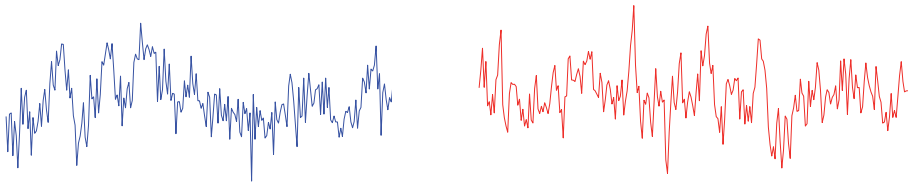


Figure 3.1: A diagram explaining the process of explicit null model randomization

a hundred times ($N = 100$). Once this is done, he evaluates the surprise value of the number of 'heads' h he received by computing its two-tailed p-value: $p = \Pr(|H - N/2| \geq |h - N/2|)$.

Usually when testing explicit models with Monte Carlo methods, the dataset to be tested is given and the null distribution needs to be approximated by sampling it. However, in Example 3.1, the null distribution H is analytical and known in beforehand whereas the data to be tested needs to be generated by sampling, that is, by tossing the coin. Thus, despite its common nature, Example 3.1 depicts an unusual instance of testing explicit null models.

Example 3.2. *The brain activity of a patient is recorded with a single electroencephalography (EEG) sensor fixed to the scalp. The data received from the sensor is the chaotic looking time series shown in Figure 3.2 (a). We construct a null model for the data by fitting an autoregressive (AR) model of order 1 to the data. This assumes that each data point is dependent only of the previous data point. The randomized samples of the data, one of which is shown in Figure 3.2 (b), are generated by simulating this AR model.*



(a) Original data (b) Randomized data

Figure 3.2: Example EEG data before and after explicit model based randomization, see Example 3.2.

Example 3.3. *A supermarket wishes to analyze the shopping patterns of its loyal customers based on their transaction histories. The analyst assumes that each customer visits the supermarket at random times and intervals. Therefore he chooses an explicit null model that models each customer as a Poisson process and infers the rate parameter of the process based on the transaction history of the customer. The random data samples for each customer are then generated by simulating the corresponding Poisson process for the duration of the customer's history.*

Example 3.4. *Some datasets contain significant cluster structure whose formation is well understood and it is desirable to nullify the effects of this structure from the significance testing results. One simple approach for this task, based on an explicit null model for this task, is to fit a mixture model of Gaussian distributions to the data. This requires making decisions on the mixture component count and component covariance type (identity, diagonal, or full), based on either expert information or data-centric model selection tools.*

3.2 Property Null Models

In contrast with the examples of Section 3.1, sometimes either the data source or the null hypothesis is so complex that it is not possible to choose

any explicit null model that realistically models the actual null distribution. Therefore there is a need to find alternative ways to construct randomization methods that realistically sample the null distribution.

This process begins with identifying specific properties of the data source and of the null hypothesis. The properties of the data source dictate the general type of data we are handling and what is or is not possible to occur in data coming out of this source. The properties of the analyzed null hypothesis on the other hand limit the scope of patterns assessed as significant.

Example 3.5. *We are evaluating interesting patterns in a social network dataset. It is reasonable to pick the degrees of the nodes or their distribution as a data property, because the degrees in these networks tend to follow the power law. Depending also on the pattern analyzed, we might choose the number of common neighbors for node pairs or the number of second neighbors as a property in the null model.*

Example 3.6. *Suppose our dataset is a collection real-valued numbers. If we wish to test the dependence of data patterns and the data mean, we should include the mean as a property in the null model. Similarly we may end up with properties such as data variance, skewness, kurtosis or the shape of the data histogram.*

In property model based randomization the goal is to generate random samples, most often uniformly, from the collection of datasets that fulfill these property requirements. The main difference to explicit model based randomization is that the null distribution of a property model is only implicitly determined and never directly sampled from. The general process of conducting property null model based randomization is shown in Figure 3.3.

A *property null model* is a model of the data source that, given a dataset and possibly parameters, defines a probability distribution over all the data samples covered by the general hypothesis. This resulting distribution determines the effect of the given requirements to the randomization results and is used for computing the empirical p-values.

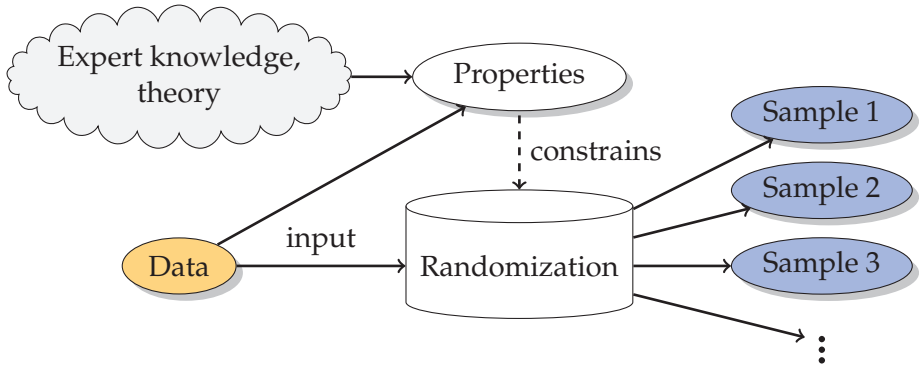


Figure 3.3: A diagram explaining the process of property null model randomization

In strict terms, the property null model is a probability space $(\Omega, \mathcal{F}, \Pr)$, where $\Omega = \prod_{\mathcal{D},\theta} \Omega_{\mathcal{D},\theta}$ is a product of probability spaces specific for a dataset $\mathcal{D}$ and parameters θ. The probability measure Pr is the product measure of the individual spaces and $\mathcal{F}$ is the standard σ-algebra defined similarly to a Tikhonov topology. Once a dataset $\mathcal{D}_0$ is given, the null model is restricted to the subset $\Omega_{\mathcal{D}_0,\cdot}$. If the behavior of the null model parameters is understood well enough, we may give a prior probability $\Pr(\theta)$ for the parameters. In this case the p-values for a test statistic f are defined as

$$p = \int \Pr(f(X) \geq f(\mathcal{D}_0) \mid \Omega_{\mathcal{D}_0,\theta}) \Pr(\theta) d\theta.$$

Usually this is not feasible and a point estimate θ_0 (using Dirac delta as prior) is used, resulting in the common version

$$p = \Pr(f(X) \geq f(\mathcal{D}_0) \mid \mathcal{D}_0, \theta_0)$$

with more convenient notation.

In a typical resampling application, however, all this underlying complexity can be waived. In fact, a property null model can be seen as a simple compilation of the given requirements for the data source and the null

hypothesis. As a property null model may not define an explicit probability distribution, extensive care needs to be used not only in the description of the properties of the null model, but also in constructing a randomization method that outputs samples from the correct null distribution. Whereas the null model describes the requirements set for the resampling task, it is the responsibility of the randomization method to have a mechanism to enforce them. Randomization methods that sample from the null distribution of a null model are *compatible* with the null model.

Example 3.7. *Suppose that we are exploring binary datasets for significant correlations. We combine the binary property of the data source with the null hypothesis*

> " Correlations in the data are explained by the total number of 1s in the data "

Random samples from this null model can be generated by uniformly at random permuting all the entries of the data matrix. Even though the resulting random samples are certainly from the null distribution, it is not necessary to state the actual distribution from which we are sampling.

Example 3.8. *Permutation methods have been widely used for testing the capabilities of binary classifier methods on a specified problem at least since the 1980s [38]. The simplest test compares the classification performance on the original data to the performance results run on data where the input data is retained as original, but the labels of each data point are randomly permuted [39].*

This randomization implies a uniform null distribution of all possible binary classifiers whose output for the given data (and only for this data) includes exactly as many labels of each type as does the output of the original classifier. Once again, the null model and distribution are simple to describe, but still the actual null distribution cannot be described in closed form.

The use of randomization with learning methods is discussed more broadly in Section 4.3

Both examples above illustrate simple *permutation tests* [24] that have historically been the main approach to empirical significance testing. The permutation methods however are often used without any reference to the null model they impose [40, 41], resulting in concerns over the realism of such methods and their compatibility with the null hypotheses actually tested. Such use of randomization implies using an implicit null hypothesis that assumes that no dependencies or behavioral differences exist. In some cases the resulting significance testing results may have little relevance to the null hypothesis that was intended to be tested. For examples on how to evaluate and compare null models of different significance testing methods, see [42, 43] and Publication IV.

One of the main purposes of property null models is to enable the inclusion of relatively complex properties in the null model. This makes it also somewhat difficult to both 1) explore all the coherent datasets and 2) exclude all undesirable datasets. On the other hand, property null models are also used when the required properties of the null model are difficult to state with precision, which translates to a rather wide grey area between what is and what is not a good randomized data sample. Therefore property null models are often (but not always, see Example 3.9) sampled using soft constraints that can be seen as ranking the elements of the sample space based on how well they fit the property model instead of using hard constraints that classify the samples strictly into sets of acceptable and unacceptable ones.

Example 3.9. *Sampling binary matrices that have given sums of entries in each row and column is an old statistics problem, discussed in more detail in Section 4.1. The common approach with an explicit model [44] uses soft constraints for the sums and can be interpreted as a maximum entropy solution to the problem [45, 46]. The common approach with a property model [47] on the other hand uses hard constraints, allowing no deviation from the original sum values.*

The use of soft constraints is not purely an implementation issue. Considering the original dataset, in most cases it is not a product of some de-

terministic process but rather a consequence of a combination of probabilistic events. For example, the transaction history of a supermarket customer shows signs of daily decisions such as "which supermarket should I visit?", "which brand and flavor of yoghurt should I buy?" and "should I get some ice cream for evening snack or not?". To model this uncertainty in the original data, we associate the property null model constraints with probability distributions and fix, for instance, the mode or mean of the distributions to the values of the constraints in the original data.

We rarely have knowledge on what this probability distribution for the constraints of the null model should be. Lacking further information and in the interest of simplicity and implementability we thus choose to use the normal distribution and fix the original value of the constraint as the mean of the distribution. We will come back to the topic of choosing the variance later.

Combining multiple constraints to a single null model is common. For example, the null model of Publication I contains $2(m + n)$ constraints for an $m \times n$ input data matrix, namely mean and variance constraints for each row and column. It is obvious that in this example, and in practically any other application too, the constraints are not independent from each other. Combining constraints together for modeling their dependencies would not only be a difficult modeling task, but might also render the sampling of the model overly difficult, slow or impossible.

The variances allowed for each constraint need to be handled both in a general level and per constraint to ensure a balanced and sufficient consideration of the constraints. The relative variances of the constraints are chosen based on the relative importance and behavior of each constraint function. The general level of error allowed in the model constraints is regulated with a coefficient w that acts on all the constraint variances. Large values of w penalize more strongly deviating from the properties of the original dataset, but make sampling slower and increase the risk for false negatives. Smaller values of w act in reverse, encouraging stronger randomization with less attention to the null model constraints and higher

risk of false positives. Tuning w effectively for each null model and application remains difficult despite some theoretical work in Publication I and additional modeling of the behavior of the randomization is needed to make this task more reliable.

Denoting the original data with $\mathcal{D}$, each constraint function with f_i, $1 \leq i \leq n_c$, and data samples with X, the combination of multiple constraints with Gaussian distributions results in a null distribution with density

$$\Pr(X = x) \quad \propto \quad \exp\left(-w \sum_{i=1}^{n_c} \frac{\left(f_i(x) - f_i(\mathcal{D})\right)^2}{2\sigma_i}\right), \tag{3.1}$$

where σ_i is the variance related to the ith constraint distribution. The density of Equation (3.1) is reminiscent of the Boltzmann distribution in simulated annealing [48], the computer science equivalent of thermodynamics. In this analogy, the error in the null model constraints gives the energy of the system and w is the inverse temperature. The whole system on the other hand attempts to simulate the data source and the conditions that generated the original dataset. However, instead of seeking the energy minimum of this system with some cooling schedule that regularly increases w, the system states are sampled and used as such with constant temperature. These samples are then assumed to be comparable samples with the original dataset, generated from a similar data source under similar conditions.

In summary, property null models allow significance testing for null hypotheses whose null distributions are too difficult to be modeled directly. This flexibility, however, brings with itself mounting difficulties in ensuring the realism of the null model and the compatibility of the randomization method with the null model.

3.3 Markov Chain Monte Carlo Methods for Sampling Property Models

3.3.1 Markov Chain Monte Carlo

Permutation methods that directly generate samples from a property null model are typically not usable with all but the most stripped-down property models. The property constraints in the null model often limit the number of acceptable samples to such a small fraction of the space of datasets that we are able to sample at all that constructing acceptable random samples from scratch becomes impossible. Nevertheless, samples can be generated by adjusting a prior acceptable sample so that the result still fulfills the null model constraints. This can be accomplished with a Markov chain whose state space is some easily managed superset of the set of acceptable samples.

A standard solution to this task is the use of *Markov chain Monte Carlo* (MCMC) methods [49, 50]. MCMC methods draw random samples from a given distribution by using a Markov chain whose equilibrium distribution is the target distribution. This approach, however, brings with it all the general problems of sampling from the equilibrium distribution of a Markov chain, including the convergence of the chain [51], correlation of the samples and heavy computational requirements arising from these two problems. Although many solutions [51] exist for these problems such as coupling methods [52], diameter and ergodicity estimation, Gelman-Rubin diagnostic [53] and autocorrelation estimation [54], these methods are not generally usable in randomization applications, where datasets are large and high-dimensional and the number of possible values in each dimension may range from thousands up to the continuum.

MCMC methods were visibly promoted for randomization purposes in [22], where examples were given for sampling explicit null models with MCMC methods. The paper introduces two statistical methods for generating exact empirical p-values with MCMC sampling. These methods and

related topics are discussed in detail in Section 3.3.2 for their critical role in the feasibility of using MCMC methods for significance testing.

In this work we concentrate in the use of MCMC for sampling property null models. The discussion of Section 3.2 and Equation (3.1) indicate the need to sample a distribution which we know only in proportion. The *Metropolis-Hastings* algorithm [49, 55] was introduced in the 1950s for sampling the Boltzmann distribution of Equation (3.1) and later generalized by Hastings.

The Metropolis-Hastings algorithm defines a Markov chain that has the desired equilibrium distribution by weighting the transition probabilities in the chain. In particular, suppose we have an arbitrary irreducible Markov chain $\mathcal{X}$ in a space that we wish to sample with a density proportional to $P(x)$. As a general idea we need to promote transitions to states for which the density P is higher than the equilibrium distribution of the original chain $\mathcal{X}$ and respectively discourage transitions to states whose presence we wish to reduce. Let us now denote by $T_{ij} = \Pr(\mathcal{X}(k+1) = x_j \mid \mathcal{X}(k) = x_i)$ the transition probabilities of $\mathcal{X}$. We may now construct a new Markov chain $\mathcal{Z}$ with the desired equilibrium distribution by accepting each transition $x_i \to x_j$ to occur only with probability

$$\Pr(\mathcal{Z}(k+1) = x_j \mid \mathcal{Z}(k) = x_i) = \min\left(\frac{P(x_j)T_{ji}}{P(x_i)T_{ij}}, 1\right).$$

It is important to note that this construction of the chain $\mathcal{Z}$ requires no knowledge about the original equilibrium distribution of $\mathcal{X}$ and additionally the target equilibrium distribution of $\mathcal{Z}$ is needed only in proportion, without the normalizing constant.

What still remains a problem in the use of MCMC methods is the vast size of the sampling space. An MCMC method that always simply moves to a neighboring state can reasonably sample only space whose size is significantly smaller than the number of steps undertaken. In applications,

however, it is impossible to ever sample more than trivial amounts of the sample space.

Example 3.10. *Resampling the hard constraint case of binary matrices with fixed row and column sums in Example 3.9 can be done with an MCMC method, described in detail in Section 4.1. However, the sample space of this Markov chain grows very rapidly as a function of the matrix size. In fact, looking only at $n \times n$-matrices where each row and column sum is equal to k, the number of these matrices exceeds $(n!)^k/(k!)^n \approx (n/k)^{nk}$ [56, 57].*

In applications where $n = 500$ and $k = 50$ this number is approximately 10^{24500}. Thus it is impossible to sample this space in any meaningful amount.

3.3.2 Validity and Nature of Empirical p-values

Traditionally conducting empirical significance testing requires generating independent and identically distributed (i.i.d.) samples from the null distribution. This however is very difficult to implement with MCMC methods, as subsequent samples in general depend on each other. Furthermore, general methods for exact sampling of Markov chains, such as coupling from the past [52], cannot be used due to the huge state space.

In [22] it is noted that independence of the samples is not actually necessary in the limited use of MCMC methods for significance testing. Instead, the simpler requirement of *exchangeability* of the samples is sufficient for producing exact (unbiased) and correct p-values. As originally defined by de Finetti and strongly advanced by Diaconis [58, 59] and others, a sequence $S_1, S_2, S_3, \ldots$ is called *exchangeable* if the joint distribution P of any n variables S_i (datasets in our case) for any n is invariant under any permutation π of $\{1, 2, \ldots, n\}$:

$$P(S_1, S_2, \ldots, S_n) = P(S_{\pi(1)}, S_{\pi(2)}, \ldots, S_{\pi(n)}).$$

Two different methods for producing exchangeable samples from a Markov chain $\mathcal{Z}$ were introduced in [22] and illustrated in Figure 3.4.

Starting from the tested dataset $\mathcal{Z}(0) = X_0$, the *parallel method* of Figure 3.4 (a) first runs the chain backwards for N steps to reach a state $\mathcal{Z}(-N) = \mathcal{R}$. Each data sample X_i is then generated separately by running the chain forwards from $\mathcal{R}$ for N steps.

Lemma 3.11 (Besag and Clifford [22], given there without full proof)**.** *The randomized samples $X_1, X_2, \ldots, X_n$ resulting from the parallel method and the original dataset X_0 are exchangeable.*

Proof. Let us denote the forward and backward transition densities of $\mathcal{Z}$ with P and Q respectively and let π be the equilibrium distribution of $\mathcal{Z}$. Note that we may assume that under the null hypothesis X_0 is sampled from π. Basic properties of Markov chain time reversal [51] then state that

$$\pi(x)P^n(x,y) = \pi(y)Q^n(y,x),$$

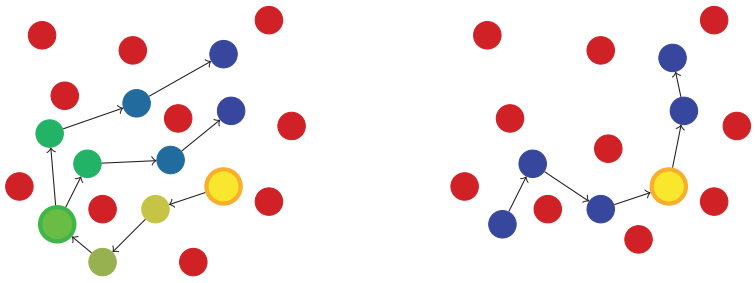
which we will here employ in form

$$\Pr(X_0 = a_0)\Pr(\mathcal{R} = y \mid X_0 = a_0) = \Pr(\mathcal{R} = y)\Pr(X_0 = a_0 \mid \mathcal{R} = y).$$

Therefore the joint distribution of $X_0, X_1, \ldots, X_n$ allows the decomposition

$$\begin{aligned}
&\Pr(X_0 = a_0, X_1 = a_1, \ldots, X_n = a_n) \\
&= \int \Pr(X_0 = a_0)\Pr(\mathcal{R} = y \mid X_0 = a_0)\prod_{i=1}^{n}\Pr(X_i = a_i \mid \mathcal{R} = y)\, dy \\
&= \int \Pr(\mathcal{R} = y)\Pr(X_0 = a_0 \mid \mathcal{R} = y)\prod_{i=1}^{n}\Pr(X_i = a_i \mid \mathcal{R} = y)\, dy. \qquad (3.2)
\end{aligned}$$

The last expression in Equation (3.2) is symmetric for $X_0, X_1, \ldots, X_n$ and the claim follows since the distributions $\Pr(X_i|\mathcal{R})$, $0 \leq i \leq n$ are all equal, because $X_0, X_1, \ldots, X_n$ are contemporaneous, that is, generated from the same chain with the same number of steps starting from the same state. □



(a) Parallel method

(b) Serial method. The placement of the original data in the chain is chosen at random.

Figure 3.4: Illustrations of the parallel and serial method of [22]. The large yellow circle depicts the original data and the blue circles are the randomized data samples.

In the *serial method* (Figure 3.4 (b)) the total number n of generated samples is first chosen and integer r is drawn uniformly at random from the set $\{0, 1, 2, \ldots, n\}$. The original dataset is then designated as data sample X_r. For each $0 \leq i \leq n, i \neq r$, the randomized data sample X_i is generated by running the chain either

1) backwards from X_r for $(r - i)N$ steps (when $0 \leq i < r$) or

2) forwards from X_r for $(i - r)N$ steps (when $r < i \leq n$).

We refer the reader to the original papers [22, 60] for details of the serial method and discussion on both of these methods.

The two methods produce results of similar quality. However, the strong parallel nature of the parallel method can give it a clear advantage in modern computing and thus it has been used in Publications I to IV of the thesis.

Despite the need for ensuring the correctness of the empirical p-values in the particularly difficult case of MCMC randomization, most earlier

publications on the field have leaned on the assumption that long enough runs of the chain will produce samples that are independent enough [61, 47]. This assumption is given a reason through experiments that show the behavior of test statistics to have strong symmetry and homogeneity throughout the sample space. This assumption is not only convenient, but also necessary since the general theory of Markov chains is typically not useful here. For example, the mixing time of the Markov chain in Example 3.10 is calculated in [62]. As expected, the resulting required running time of the chain is not practicable in applications and also clearly more than what is heuristically considered to stabilize the chain's behavior in [47].

The same experimental observations of symmetries can also be used to suggest that it is not actually necessary to sample the whole sample space. As the test statistic t maps the high-dimensional sample space to a real number, it also folds together multiple symmetric areas of the space. For example, suppose our test statistic t is the correlation between two variables. Now any common permutation of the target variables in the original dataset generates a large number of new samples throughout the sample space. However, the local topology of t is similar around each of these new samples and around the original data. However, no theoretical proofs or evidence has been given to this conjecture and this remains a central problem in the future research of MCMC-based randomization.

By using the parallel method (or the serial method) we can avoid the problem of needing independent samples as the samples generated by this method automatically produce correct p-values. However, they are correct only in the context provided by the implicit general hypothesis, defined by the underlying Markov chain and the step count N. To see this, consider Figure 3.5. Based on the parametrization of the Markov chain and the step count N, the chain can reach different subsets of the whole sample space with the chain parameters determining the shape and the step count N determining the size of the subset. These subspaces are the context of the significance test as the significance of patterns in the original dataset is

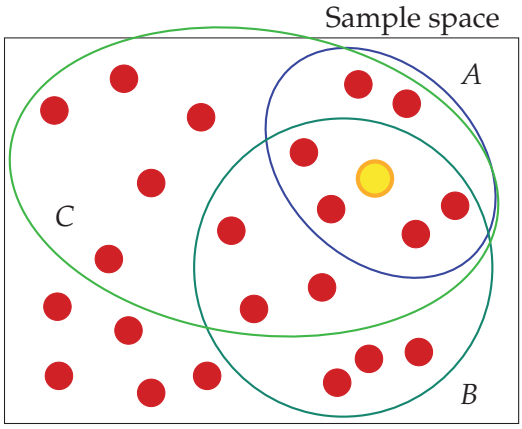


Figure 3.5: Illustration of different contexts for significance testing. The big yellow circle depicts the original data and the ellipses A, B and C depict different areas of the sample space that are sampled.

tested only relative to the samples within this context.

Therefore the purpose of running the MCMC chain long enough is not to produce correct p-values but merely to choose a broader context for the significance test. Usually a balance is needed between having a reasonable running time and a wider context. Another possible way to improve the sampling is to use parallel tempering [63], which combines the advantages of good mixing in high temperatures and stricter constraints in low temperatures. Measuring the changes in test statistics can be used for evaluating the sampling context when the chain is run for longer periods or when otherwise tuning the parameters.

3.4 Representation-based Randomization

As discussed in sections 3.2 and 3.3, sampling property models is not as straightforward and easy as it is with explicit models. Although the gen-

eral "squared error in constraints" model (3.1) can accommodate almost any types of constraints in the null model, using complex constraints may require extensive computing time. Additionally, the smoothness of the error function (3.1) is important for the MCMC sampler to efficiently traverse the state space, a matter that may incur problems with complex constraints.

Example 3.12. *Consider the topic of Section 4.3.2, randomizing data while preserving its cluster structure. One simplistic way to conduct such task could be to build a property model that constrains the clustering error from a given clustering algorithm such as k-means. Despite recent advances in the efficient computation of a k-means clustering [64, 65], evaluating the clustering error in each step of the MCMC procedure is not practical. Due to the nature of clustering, the minimums of the error function (corresponding to the most relevant areas for resampling) are represented in the state space by numerous and widely distributed sharp peaks of high density but low density mass. This makes proper MCMC sampling of this null model difficult.*

To improve upon these two problems it is necessary to somehow extract the constraints of the null model from data into more accessible form. In Publication III the use of invertible representations for datasets is discussed for this purpose. The rationale of this is shown in Diagram (3.3).

$$\begin{array}{lll} \text{Original} & \mathcal{D} \xrightarrow{\textit{represent}} & f(\mathcal{C}, \mathcal{R}) \\ \Big\downarrow & & \Big\downarrow{\scriptstyle \textit{randomize } \mathcal{R}} \\ \text{Randomized} & \widehat{\mathcal{D}} \xleftarrow{\textit{combine}} & f(\mathcal{C}, \widehat{\mathcal{R}}) \end{array} \tag{3.3}$$

Here the constraints $\mathcal{C}$ of the property null model are separated from the data $\mathcal{D}$ to make randomization easier. When directly randomizing the original data, in this setting it may be necessary to repeatedly invert f to evaluate the effect of randomization on the constraints. On the other hand, if successful, representing data in a different form can significantly

simplify managing the constraints of the null model during randomization. In addition, representations open new possibilities of iterative pattern mining, discussed in more detail in Section 4.6.

In any event, the original form of data is never more than one subjective view, not necessarily the optimal one for any single purpose. For example, the Box-Cox transformation [66] is a very common tool in statistics and econometrics for manipulating the data to a more suitable form. Thus finding the best representation of the dataset should always be a part of the analysis process. Many useful representations of data are robust against elementary transformations. For example, the PCA decomposition remains invariant under affine transformations (dilation, rotation, and translation), which may be desirable.

The concept of representation-based randomization was first defined in Publication III, but it has been implicitly used already earlier for handling time series data in Fourier and wavelet space [67]. See Section 4.4 for more details.

Chapter 4

Randomization Methods

The need for more advanced significance testing methods has recently arisen for various reasons. First, many application fields that handle complicated types of data have seen research reports warning about negligence and carelessness in evaluating pattern significance [68, 4, 3, 69, 70, 71, 72]. Second, the increase in computational resources has resulted in interest for larger and larger datasets and for more complex data patterns that are not only more descriptive and harder to compute, but also decidedly more difficult to assess for significance. Third, advances in pattern mining have built up interest in analyzing sets of patterns and their joint behavior, resulting in more complex null hypotheses that cannot be tested with standard statistical tools.

Example 4.1. *Suppose that we are comparing two text corpora, one containing all the publications ever published in a certain data mining journal and another containing the publications of some journal focused on machine learning. Now let us ask whether the vocabularies used in the two collections of publications differ significantly from each other or not.*

Answering this question requires not only wading through the huge dataset and comparing the frequencies of words, but considering also things like the distributions of the words across publications, probably even across time. Also lin-

guistic information such as associations of words could be incorporated to a more rigorous analysis.

Complex hypotheses arise when the specifications of a data source are less simple, but also when there are additional constraints on how to interpret the information in the dataset. A simple example of such a case is when the data contains some pattern that we have already identified and analyzed. In such a case we might wish to brush aside this pattern and its effects from the data and see what other meaningful patterns exist. This line of thought brings us to the concept of *iterative data mining*, discussed in Section 4.6. Property models are a useful tool with more complex null hypotheses since modifying them is relatively straightforward.

While Chapter 3 focused in building a general framework for using randomization with complex null hypotheses, this chapter looks at some of the accomplishments in this field in more detail. We begin with property null models for specific types of datasets. Sections 4.3 and 4.4 demonstrate the use of representations for accessing complex null hypotheses and finally Section 4.6 examines the case of pattern sets and their iterative significance testing.

4.1 Significance Testing for Databases

Data mining began as a practice of exploring databases for their efficient use in knowledge discovery. Due to their relative simplicity and practical abundance, relational databases have ever since remained a basic theme in data mining. The simplest task in mining relational databases is the analysis of a single database table, whereby the table is usually expressed as a matrix of its entities and attributes. When discussing pattern mining from databases, the database tables are usually represented as simple binary matrices.

Resampling of binary matrices has throughout time revolved around the idea of preserving the sum of values in each row and column. This

problem arose from the need in community ecology to resample *contingency tables* [61]. Contingency tables are nowadays central not only in ecology, but also in topics such as paleontology [73], social sciences [74], psychology [44] and marketing [75].

The early method for resampling contingency tables is given in [44] by Rasch. The method samples each entry (i, j) of the matrix independently from a Bernoulli distribution with probability of a '1' entry equal to

$$\frac{\exp(-r_i - c_j)}{1 + \exp(-r_i - c_j)},$$

where r_i and c_j are the corresponding row and column sums, respectively. This seemingly naive method is also the maximum entropy solution to the problem [45], but preserves the row and column sums of the matrix only in expectation.

Swap randomization [61, 47] is a commonly used MCMC-based algorithm for exactly preserving the row and column sums in the data samples. An illustration of a single step of the method is given in Figure 4.1. The method continues swapping quartets of matrix elements until the target measure whose significance is tested has converged. Based on experiments, Gionis *et al.* suggest running the chain for a total of $5L$ iterations, where L is the number of ones in the data matrix. The swapping method is significantly slower than the Rasch method as the exactness requirement prevents the use of immediate sample generation. However, with suitable index structures the swapping algorithm can be made very efficient and even matrices with tens of millions of elements (with 10 % matrix fill rate) can typically be randomized in a few seconds.

Mining patterns from graphs can be seen as a similar problem to that with databases, since standard unweighted graphs can be represented as binary adjacency matrices. Especially the randomization of graph data can be seen as an analogous problem to that of database tables, the only significant exception is that in an undirected graph the upper and lower triangle parts of the adjacency matrix need to remain synchronized during

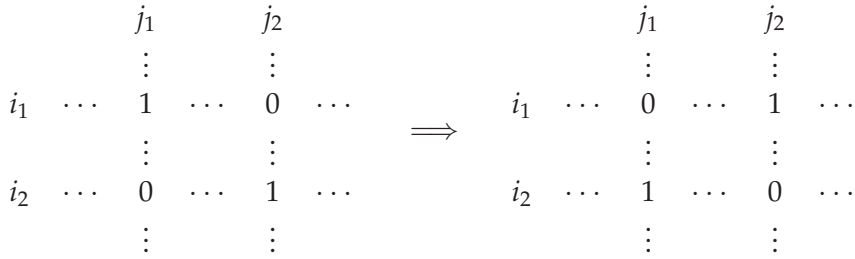


Figure 4.1: An example of a single step in swap randomization. The rows i_1 and i_2 (and columns j_1 and j_2) may be arbitrarily ordered. The row and column sums are preserved in each step.

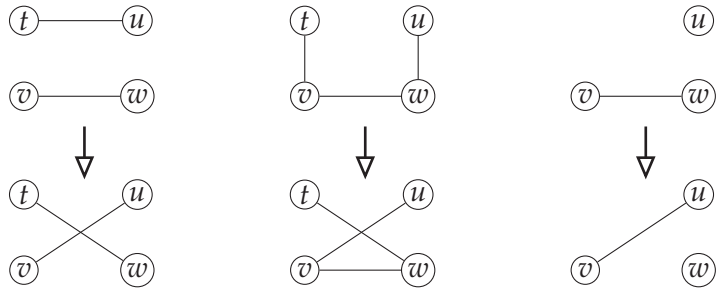


Figure 4.2: Examples of the graph randomization methods introduced in [76]. The rightmost method has an additional requirement that the degrees of nodes u and w must differ by exactly one.

randomization. Such modifications of the standard swap randomization were introduced in [76] by Hanhijärvi *et al*. The three suggested graph operations are illustrated in Figure 4.2.

It is possible to also conduct significance testing with explicit models for graphs. The baseline to start from in case is to use a graph model [77] such as Erdös-Rényi [78] (uniform graph structure), Watts and Strogatz [79]

or Barabási–Albert [80, 81, 82] (power-law/scale-free structure).

Analyzing tables in a database separately may easily miss the wealth of information a database as a whole portrays. One possibility to counter this problem is to combine several tables into one with Cartesian products. This however easily leads to huge datasets. Using the results of previous data analyses can help in replacing the indiscriminate Cartesian products with selective join operations that not only can cut down the data size considerably, but also may help in detecting the relevant data patterns in a more focused analysis [83, 84].

Significance testing of patterns spanning multiple database tables [84] is considered by Ojala in [85]. The paper suggests using swap randomization not only to selective database tables but also to the natural join relations between tables. The latter operation can be seen as a simple permutation of the foreign keys that determine the join operation in question. This additional randomization choice enables iterative testing of null hypotheses that assess whether a given pattern in the database is a consequence of some specific table or a subset of tables or of the way two types of data (tables) in the database are connected with each other.

The problem of mining interesting patterns from databases and assessing their significance has seen lots of successful research. With the problem of finding the relevant patterns from data partially solved, the problem of distilling this information to a compact form for human users has seen increasing attention more lately [86]. Especially the problem of finding the most significant or most explanatory k patterns and the problem of finding independently significant patterns have been considered and remain still without an exhaustive answer. The latter one of these problems is discussed in more detail in Section 4.6.

4.2 Randomization Methods for Real-valued Data

The use of computational methods in various fields of research has seen significant proliferation during the past decade. With this shift, the predat-

ing focus of data mining research into binary data and databases has been supplemented with various types of real-valued data. Although intervals can always be used for discretizing data, a great deal of information is lost in the process. Therefore new methods and approaches are needed for this task.

The richness of information in continuous data makes the process of null model formulation more difficult than in a discrete case. With real-valued data it is not enough to merely fix the sum of elements in each of the columns and rows as is done with binary matrices. For example it is not sensible to compare the set of customer ages $\{13, 44, 60\}$ in the database to a random sample $\{-100, 0, 217\}$ or perhaps even to $\{39, 39, 39\}$. Therefore a minimal requirement for the realistic resampling of real-valued data is the preservation or regulation of the variance in data.

Constrained randomization of real-valued matrices was studied in [5], where the means and variances of each row and column of the data were constrained. The work based on extending the idea of a Markov chain of local operations from [47]. Two local operations suitable for real-valued matrices were suggested, *SwapRotation* of Figure 4.3 (a) and *MaskAddition* of Figure 4.3 (b). In the extended version Publication I of this paper three additional local operations were tested: *ElementReplacement* (Figure 4.3 (c)), *ElementResampling* (Figure 4.3 (e)) and *ElementPerturbation* (Figure 4.3 (d)).

In running the Markov chain of [5], the error in the property model constraints between the original matrix and the current state of the chain is measured. For each row and column the squared error in the first and second moments is separately computed and combined as the error function

$$E(A, X) = w_r \sum_{i=1}^{m} \left(|r_i^A - r_i^X|^2 + w_s |R_i^A - R_i^X|^2 \right) + \sum_{j=1}^{n} \left(|c_j^A - c_j^X|^2 + w_s |C_j^A - C_j^X|^2 \right), \quad (4.1)$$

$$\begin{array}{ccccccc} & & j_1 & & j_2 & & \\ & & \vdots & & \vdots & & \\ i_1 & \cdots & \alpha & \rightarrow & \beta & \cdots & \\ & & \uparrow & & \downarrow & & \\ i_2 & \cdots & \gamma & \leftarrow & \delta & \cdots & \\ & & \vdots & & \vdots & & \end{array}$$

(a) *SwapRotation*

$$\begin{array}{ccccccc} & & j_1 & & j_2 & & \\ & & \vdots & & \vdots & & \\ i_1 & \cdots & +\alpha & \cdots & -\alpha & \cdots & \\ & & \vdots & & \vdots & & \\ i_2 & \cdots & -\alpha & \cdots & +\alpha & \cdots & \\ & & \vdots & & \vdots & & \end{array}$$

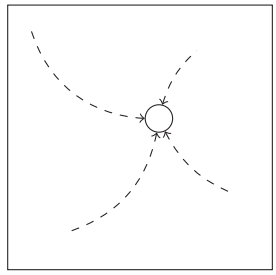
(b) *MaskAddition*

$$\begin{array}{ccc} & j_1 & \\ & \vdots & \\ i_1 \quad \cdots & U(\min(A), \max(A)) & \cdots \\ & \vdots & \end{array}$$

(c) *ElementReplacement*: Replace with random number from uniform distribution

$$\begin{array}{ccc} & j_1 & \\ & \vdots & \\ i_1 \quad \cdots & +\alpha & \cdots \\ & \vdots & \end{array}$$

(d) *ElementPerturbation*



(e) *ElementResampling*: Replace entry with one random original matrix entry

Figure 4.3: Local operations used with the randomization Markov chain in [5] and Publication I for original data matrix A. In figures (b) and (d) the value of perturbation α is chosen randomly from the uniform distribution $U(-s, s)$ for some s.

where A is the original data matrix, X is the current state of the Markov chain and

$$r_i^M = \sum_{j=1}^{n} M_{ij}, \qquad c_j^M = \sum_{i=1}^{m} M_{ij},$$
$$R_i^M = \sum_{j=1}^{n} M_{ij}^2, \qquad C_j^M = \sum_{i=1}^{m} M_{ij}^2$$

are the row and column moments for matrix M with elements M_{ij}. This error function of Equation (4.1) is then inserted into Equation (3.1) and the Metropolis-Hastings method is used to generate random samples. Additionally the work of Besag and Clifford discussed in Section 3.3.2 is used to generate exact p-values.

To weigh equally both moment types in rows and columns, the parameters of the null model in Equation (4.1) were chosen as $w_r = m/n$ and $w_s = 1$ in [5]. Other parameters needed in the randomization are the "inverse temperature" w of the randomization process in Equation (3.1) and the number of steps to run the chain. Tuning these parameters is not straightforward, requiring further research despite some theoretical results and recommendations in Publication I.

Another approach to randomizing real-valued matrices is to discretize the matrix values with suitable granularity. The `SwapDiscretized` method discussed in [5] and Publication I discretizes the values of the matrix separately for rows and columns, runs a Markov chain of *SwapRotation* operations and accepts the state transitions if and only if the row and column distributions of the discretized values remain unchanged in the operation. Therefore the method approximately preserves not only the means and variances of the data, but also the full row and column value distributions.

The capabilities of the `SwapDiscretized` method were substantially extended in [87] to include handling of dissimilar data attributes, sparse data and missing values. Additionally, this paper introduces the use of the *Kolmogorov-Smirnov test* [88] for evaluating the quality of randomization results. The Kolmogorov-Smirnov test evaluates whether it is reasonable

to say that the rows and columns of both the original and randomized data follow the same distributions.

For further details on the `SwapDiscretized` method we refer the reader to the original papers [5, 87] and Publication I. The advantages of the method include its speed and ease of use. However, the artificial value boundaries in discretization make it slightly awkward when all the data attributes are of truly continuous type.

Including only means and variances of variables (columns) or entities (rows) in the property null model constraints has in experiments shown to prefer a Gaussian distribution for these sets of values. Although this fits well with the maximum entropy principle [46], the known properties of the data source or the null hypothesis itself may require a stricter control over the null distribution. This necessitates departure from the Gaussian construct to more precise data modeling for conducting better significance testing.

Example 4.2. *A significant share of observation data from nature or from man-made sources follow a so-called* $1/f$-type distribution *[89, 90] (see Section 4.4 for additional notes) whose spectrum is biased towards the lower frequencies. In bioinformatics, gene expression data has been shown to follow a power law distribution [91]. Sparsity of data is another phenomenon that commonly is an essential part of how a data source behaves.*

In Publication I the work of [5] was extended to allow value distribution constraints in the null model. The paper uses the L_1-distance of unnormalized cumulative distribution functions [92] for evaluating the error in each row and column distribution. Consider an original dataset A of size $m \times n$. Denote by $A_{i\cdot}$ the vector of elements in a single row i sorted ascending and by $\widehat{A}_{i\cdot}$ its randomized counterpart also sorted ascending. The error between the distributions of $A_{i\cdot}$ and $\widehat{A}_{i\cdot}$ is defined as

$$E(A_{i\cdot}, \widehat{A}_{i\cdot}) = \sum_{j=1}^{n} |A_{ij} - \widehat{A}_{ij}|. \tag{4.2}$$

The total error between two matrices is defined as the simple sum of errors in all rows and columns and the randomization is conducted as in [5].

However, updating the distribution errors in each row and column during the randomization is computationally demanding. Updating the error when a single element $\widehat{A}_{ij}$ is moved needs $\mathcal{O}(l + \log n)$ time, where l is the number of elements in the sorted vector $\widehat{A}_{i\cdot}$ between the old and new values of $\widehat{A}_{ij}$ as for these elements the sum terms of Equation (4.2) need to be re-evaluated. In practice, the value l is a linear fraction of n to facilitate sufficient perturbation, leading to slow randomization with large datasets. To counter this, Publication I suggests using histogram approximations for each row and column. With this modification the update times for a single row or column drop down to $\mathcal{O}(b)$ for a histogram approximation of b bins. This time needed is not anymore related to data size, but it may depend on the complexity of data as more varied distributions require more bins for sufficient modeling. In contrast, the `SwapDiscretized` method and the methods of [5] that constrain only means and variances of rows and columns reach constant update times per iteration.

4.3 Learning Methods and Randomization

4.3.1 Evaluating Supervised Learning Methods

Learning from data is a central goal in data mining. Patterns in data that enable effective classification or regression are not only sought out, but their characteristics are also analyzed to improve the learning results. Traditionally tasks like feature or model selection and validation are conducted with methods such as cross-validation, regularization and various *minimum description length* (MDL) [93] or *Vapnik-Chervonenkis theory* (VC-theory) [94] based approaches. Permutation tests are a relatively new arrival to evaluating learning methods. Although used in different forms since 1980s [38], an important landmark was the paper by Mukherjee [39] where permutation tests are used to evaluate the following null hypothe-

sis:

> *" A given family of classifiers cannot learn to accurately predict the labels of a test point given a training set. "*

The test of [39] is based on breaking the connection between the data attributes and labels by permuting the labels. This idea was more recently extended for the general evaluation of generalization error by Magdon-Ismail and Mertsalov in [95]. In both publications the validity of the approach is proven with VC-theory.

Permutation tests have been utilized also for feature selection in [96, 97] with the general idea of comparing the prediction performance before and after permuting the values of a given feature. The work of François *et al.* [97] is closely related to the general process of empirical significance testing discussed in Section 2. In [97] the significance of each data feature is separately tested by using the mutual information of the feature and the target output values as the test statistic. Once the significance testing is concluded, features assessed non-significant are pruned out from the data.

In addition to analyzing the trustworthiness of machine learning results, randomization methods have also been used for testing the properties of learning methods and their suitability for a given dataset. In [98] permutation tests are used for assessing whether a learning method is able to use interdependencies between data attributes to improve classification results. This test uses the null hypothesis

> *" Data attributes are mutually independent given the class label. "*

The test of [98] independently permutes the values of all data attributes within each class label separately. The learning method conducts classification for the resulting randomized dataset not unlike the naive Bayes method which computes the likelihoods of each attribute value separately and then combines them.

4.3.2 Patterns in Clustered Data

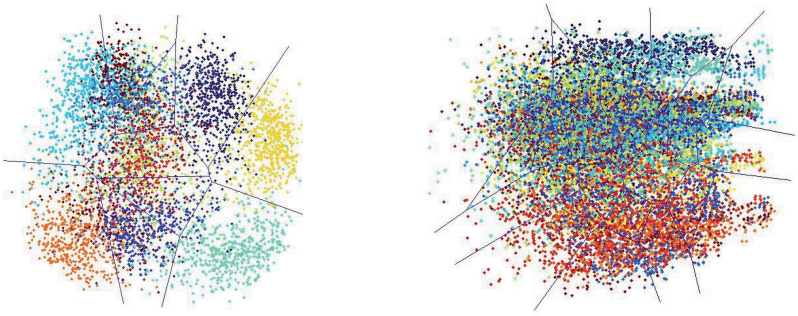
Clustering, despite its most vague definition, is perhaps the most prominent of unsupervised learning concepts. The strong visual character of cluster structure also means that it is a global pattern of a dataset, affecting the results of almost any data analysis that considers the dataset as a whole. For example, a cluster in data, that is, a group of points situated in the same region of space, is also a group of points in which several data attributes correlate strongly with each other. More importantly, clusters are often useful patterns when learning from data.

Considering this pervasive nature of cluster structure among the family of data patterns, we may ask to what extent does the cluster structure of a dataset predefine the results of a data analysis task. This question is studied in Publication III with the null hypothesis:

> *" Patterns in data exist independently from the cluster structure of the data. "*

The paper introduces a new algorithm `PCARand` that combines the ideas of Publication I and [99] with representation-based randomization (see Section 3.4). By operating in the space of principal components the algorithm gains access to the cluster structure of data, making it possible to preserve it during randomization. Although the underlying theory from [99] applies directly only to cluster structure as defined by the k-means method, other clustering concepts can be easily used with the kernel trick [100].

Compared with the approach of Example 3.4, where a Gaussian mixture model is used as an explicit null model for cluster structure preserving randomization, the `PCARand` algorithm does not need any input parameters that fix the cluster count. This is notable, as such a parameter prevents a randomization method from preserving any cluster structure that does not fit the pre-chosen model. Additionally, the experiments in Publication III suggest that randomizing data based on the Gaussian mixture model may easily fail to actually randomize the data to a sufficient extent.



(a) UCI OCR dataset (b) UCI LETTER dataset

Figure 4.4: Visualizations of two UCI [101] datasets. Data points are colored differently based on their class labels. Straight blue lines depict Voronoi diagrams for k-means clustering results. Data is mapped to the plane by using its first two principal components.

Publication III considers also the problem of relating unsupervised and supervised classification results for a dataset. The `PCARand` method of the paper can be directly used for this purpose as the chosen classification structure can be seen as a data pattern among others. As an example of the relation between these two different classification types, see Figure 4.4. In the figure the classification structure of two datasets [101] is shown with the coloring of the data points and a single instance of k-means clustering on the datasets is shown as a Voronoi diagram. In Figure 4.4 (a) it appears that the clusters in the dataset follow the class labels quite closely whereas in Figure 4.4 (b) this appears not to be the case. This difference indicates how differently the cluster structure may interact with the class labels depending on the underlying dataset.

The case of Figure 4.4 can be demonstrated also with the numbers in Table 4.1 that show the multi-class classification accuracies reached with support vector machines (SVM) [94] for original and randomized datasets. These accuracies show how, despite the similar SVM accuracy for the orig-

Table 4.1: Support vector machine classification accuracies before and after randomization for the two datasets of Figure 4.4. `GeneralMetropolis` is an MCMC-based algorithm from Publication I, which preserves the row and column distributions of data.

Randomization	UCI OCR	UCI LETTER
Original data	99.4 %	97.9 %
`PCARand` (Publication III)	90.0 %	50.2 %
`GeneralMetropolis` (Publication I)	69.9 %	32.7 %

inal datasets, the two datasets in question contain a very different relation between unsupervised and supervised classification structure. With the UCI OCR dataset a good accuracy of 90 % is still reached after randomizing the data with `PCARand`, effectively removing all but the cluster structure from the data. This means that the different classes of the dataset form clearly distinct and round clusters (in the Euclidean sense) in the data and there is little other structure in the data relevant for classification. With the UCI LETTER dataset, on the other hand, the SVM accuracy drops sharply down to 50 % although there exists clear class structure in the original data. Therefore the classes cannot be distinguished with round and equal-sized clusters provided, for example, by the k-means algorithm. However, this does not mean that the different classes in the data were not separated as clusters of some form. Comparison with the results of the `GeneralMetropolis` algorithm shows how only a small part of the cluster and classification structure in both of the datasets can be seen to be dependent of the row and column distributions alone.

The work of Publication III concentrates in the relation of unsupervised and supervised learning in real-valued data, but the approach can be transplanted also to the case of binary data. A combined property model based randomization method that uses k-means clusters and fre-

quent itemset supports as model constraints was introduced in Publication II. The preservation of cluster structure in binary data is a simpler problem and the methods of this paper can be easily used for conducting similar tests as in Publication III. Additionally, it might be interesting to test the relation of association rules and classification results in binary data such as document collections. For this task, however, there are no ready-to-use solutions, although the work of Publication II can probably be extended to this task with limited effort.

4.4 Randomizing Time Series Data

Time series are a ubiquitous type of data that arise in various applications such as finance, medicine, climate research and industrial monitoring. Usually time series data contain a large numbers of samples and the data sources describe a collection of synchronous time series that measure different aspects of the same or related phenomena. In addition, sources of time series data are often not so well understood making them difficult to model.

Added to the above mentioned complexities in time series data, the inherent autocorrelation of nearby observations in time series makes the formalization of realistic null models and thus the task of significance testing for this data type especially complicated.

Significance testing for time series data was historically conducted with analytical Gaussian noise models for the data itself or its differences [102, 103, 104]. Analytical testing has been used also with more elaborate modelling for functional Magnetic Resonance Imaging (fMRI) data [105]. Several bootstrapping and autoregression schemes have been suggested [40, 106, 107, 108] and model-based resampling based on ideas such as physical attractors [109] and maximum entropy [110] have received attention more lately.

Harmonic methods such as Fourier and wavelet [111] analysis have clear advantages in analyzing time series and their use has grown steadily,

promoted especially by the research community of medical imaging that has been a strong contributor to time series research [112]. Increasing attention to temporal variation in non-stationary time series has caused the earlier resampling methods based on Fourier analysis [113, 114, 115, 116] to give way to wavelet methods [117, 118, 119, 67, 120]. Additionally, the special implications of multiple hypothesis testing for using wavelet methods has been studied in [121] by Şendur *et al*. While wavelets have been used in fields outside medical research such as finance [122, 123], there has been no substantial work in conducting wavelet-based significance testing in these research areas.

An additional special advantage in the use of wavelets for significance testing is their nature of optimally decorrelating signals with a $1/f$-type spectrum [124, 90, 125]. This decorrelation of data points has been an important factor in arguing the validity of the rather crude permutation-based resampling methods in wavelet domain [117, 118, 119].

As already mentioned, time series data comes often not alone but in a collection of simultaneous measurements of the same or related phenomena. Such data requires significance testing to take into account not only the temporal and frequency information in the data, but also the dependency structures that link the various time series of the data collection together. This type of work has been conducted especially in the field of medical time series such as electroencephalography (EEG) and functional Magnetic Resonance Imaging (fMRI) research [113, 119] and more lately in finance [126]. In these applications there are several channels generating synchronized time series data having strong inter-dependencies. Wavelets have been widely accepted as the solution of choice in both application areas, since they are very good at localizing patterns simultaneously in the temporal and frequency domains. Especially, the wavelet methods used with stand-alone time series have been exported to the problem of time series collections through multi-dimensional wavelet decompositions [119, 127, 128].

Despite the significant efforts spent in finding suitable wavelet repre-

sentations for time series significance testing, the common method used for the definitive resampling has usually always been some variation of simply permuting the wavelet detail coefficients. Among others, permuting coefficients in each frequency scale independently or in tandem [117], shifting the coefficients circularly [118] and permuting blocks of coefficients [118] have been suggested. Another approach was used in [128], where the signs of wavelet packet transform detail coefficients were randomly reassigned. The improvements of these more recent resampling methods over the historical bootstrapping methods have been due to using the properties of the wavelet transform.

The alternative approach to improving the quality of significance testing by concentrating in replacing simple bootstrapping with something more elaborate has seen much less attention than the harmonic analysis based approach. The early work of [129] suggests conducting property-based randomization with the Metropolis-Hastings algorithm (see Section 3.3.1). The paper suggests using a randomization constraint that preserves the values of the autocorrelation function

$$C_p(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} x_n x_{(n-\tau) \bmod N}$$

for all delays $0 \leq \tau \leq N/2$, where N is the length of the time series. This approach, however, leaves open the problem of temporal variation in the data, as it constrains only the global correlative properties in the data.

A method for combining the benefits of wavelet analysis and modern resampling methods was introduced in Publication IV. The paper studies a specific setting of homogeneous time series collections, where the time series are, *a priori*, assumed to be indistinguishable from each other. In other words, the formalized null model assumes that there is no prior information on the data source that is specific to any fixed subset of the time series.

The null model of Publication IV is built on a model of the data source where external events induce responses of varying breadth, strength, de-

lay and duration in the time series. These events are assumed to be intrinsic to the data source, that is, natural and not determined significant. Therefore these properties of the data source need to be modeled and incorporated to the null model to ensure that the significance test will not report spurious patterns.

Example 4.3. *Recording EEG data from a patient involves attaching an array of electrodes (few tens up to hundreds) to the scalp. If we assume no prior medical knowledge of the functions of a brain, the collection of the simultaneous EEG recordings from the electrodes is a homogeneous time series collection.*

The external events in this example are related to, for example, asking the patient to tap his left hand index finger, showing him pictures or inflicting pain. The null model of Publication IV attempts to preserve the characteristic effects related to these events while being indifferent of the exact timing and locality (among the different electrodes) of the events and their effects shown in the EEG signals.

Examples of patterns that could be found interesting would include cases where a certain EEG channel seems to react especially strongly and with higher frequency than the others to a certain event or an occasion of no meaningful response from a few channels by the time of a broad general response.

In summary, the null model allows controled mixing to occur through time, because there is no prior knowledge on the timings of the external events. Additionally the homogeneity assumption implies that controlled mixing may be conducted also between the different time series. However, mixing between different frequency bands should not be allowed since this would affect the expected distribution of the types of the external events. In addition to this, studies of Breakspear *et al.* in [118] have shown that perturbation over different frequency bands causes large distortions to the linear properties of the time series, which then leads to excessive Type I errors.

For an illustration of the properties of this null model, consider Figure 4.5 that shows an example of a pattern assessed as significant by the

null model. The null model enforces two properties on each frequency band. First, at each point in time (that is, for each k), the *spatial profile* of the resampled collection of the time series must agree with that of the original data. In particular, in Figure 4.5 each time point must have two series with activity and two series with no activity. Second, for each time series (that is, for each i), the *temporal profile* of the resampled collection must agree with the original data. Thus in Figure 4.5 each series must have two time points with activity and two time points without activity.

To bring this conceptual null model into practice, Publication IV poses three requirements on how to manage the temporal, spatial and frequency domain information during randomization. These requirements are approached in Publication IV through the use of *scale-wise* matrices of wavelet detail coefficients. For a time series collection X of M time series, the scale-wise matrix of the jth wavelet scale (frequency band) is

$$\mathcal{W}_X^j = \begin{bmatrix} w_{1,j,1} & w_{1,j,2} & \cdots & w_{1,j,K_j} \\ w_{2,j,1} & w_{2,j,2} & \cdots & w_{2,j,K_j} \\ \vdots & \vdots & \ddots & \vdots \\ w_{M,j,1} & w_{M,j,2} & \cdots & w_{M,j,K_j} \end{bmatrix},$$

where K_j is the number of detail coefficients on scale j and $w_{i,j,k}$ is the kth detail coefficient on scale j of the ith time series in X. In other words, the $\mathcal{W}_X^j$ matrices collect together all the wavelet detail coefficients of all the time series for each scale separately.

Based on the $\mathcal{W}_X^j$ matrices, the precise null model formulation of Publication IV states the following:

REQUIREMENT 1. (PERTURBATION IN FREQUENCY)
Scale-specific detail coefficient sets $w_{\cdot,j,\cdot}$ must be resampled separately. In other words, the randomization of each $\mathcal{W}_X^j$ matrix needs to be conducted independently from the others.

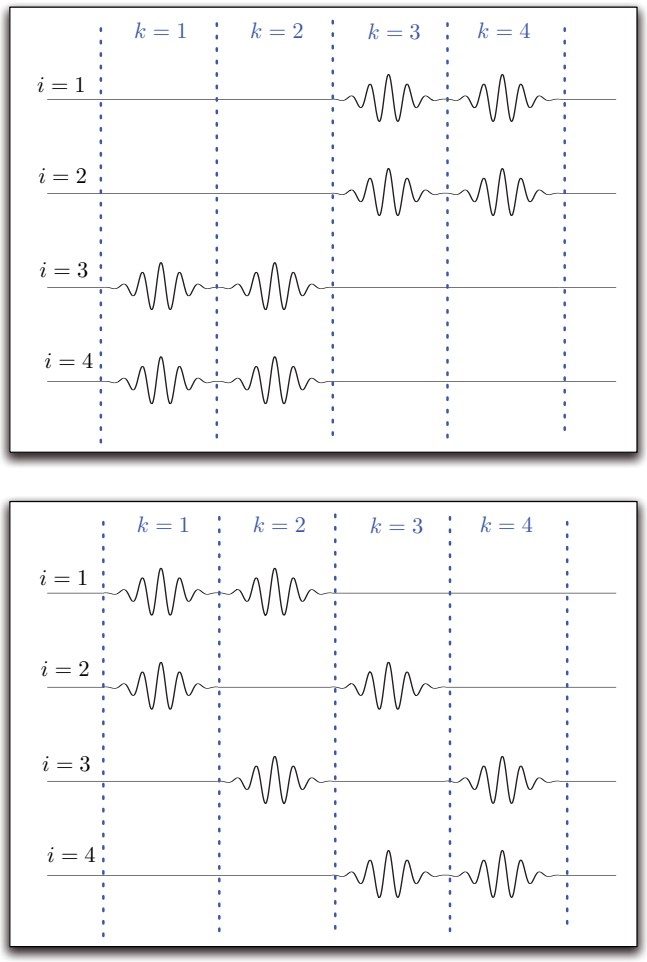


Figure 4.5: An intuitive illustration of a pattern in data considered significant by the null model of Publication IV (top) versus a resampled dataset from the null distribution (bottom). A fixed band of frequencies is shown for a collection of four time series indexed by i, with the temporal dimension indexed by k.

REQUIREMENT 2. (PERTURBATION IN TIME)
The distribution of coefficients $w_{\cdot,j,k}$ must be preserved for each frequency. In other words, the column distributions of $\mathcal{W}_X^j$ for each j are constraints of the property null model.

REQUIREMENT 3. (PERTURBATION IN SPACE)
The distribution of coefficients $w_{i,j,\cdot}$ must be preserved for each i and j. In other words, the row distributions of $\mathcal{W}_X^j$ for each j are constraints of the property null model.

These requirements were derived from an analysis of the time series collection randomization problem. However, it is readily seen that the required properties of this null model fit directly into the framework of Publication I, which considers randomizing real-valued matrices while preserving their row and column distributions. The proposed new randomization method of Publication IV thus centers around randomizing each $\mathcal{W}_X^j$ separately with the methods of Publication I.

4.5 Significance Testing with Low Quality Data

Time series are a standard example of data whose modeling is difficult due to the complexity and high amount of data. However, datasets of low quality have their own problems, as the lack of sufficient data or the lack of confidence on the data requires the significance test to be careful in its conclusions. This section reviews the significance testing research conducted in two major areas of low quality data: uncertain data and data that is missing values. A specific contribution given by Publication V in the domain of bioinformatics is also discussed.

Missing values are probably the most common cause for datasets of low quality. They can occur when, for example, the data acquisition process is imperfect, the data gets corrupted or data points are intentionally erased to anonymize data. There exists a wide set of methods, called *impu-*

tation methods, for filling in missing values in data [130]. Of these methods, especially the multiple imputation technique [131] deserves attention for its Monte Carlo nature. Multiple imputation creates multiple imputed datasets for each of which the missing values are filled in differently, although usually using the same imputation method. Mixed use of different methods for this task was considered by Jörnsten *et al.* in [132].

Manipulating the original dataset with imputations affects also the patterns that are found from the data and their significance. Therefore there has also been research done to evaluate the impact of various imputation methods for the significance testing results, although the research has mainly concentrated around regression problems. An analytical solution that combines the separate significance testing results for each imputed dataset into one p-value while also taking into account the unknown variability in the missing values was introduced in [133, 134]. More recently [132] introduced an ensemble method that attempts to optimize the combination of results from multiple separate imputation methods. The results of the paper indicated that this strategy not only improves the quality (reduces the error) of imputed data points, but also the quality of the eventual p-values.

As a further complication to the idea of multiple imputation, many survey conducting organizations publish their data to researchers in a *multiply-imputed* form. In such datasets missing values are first imputed, but after this the values of attributes that are more sensitive or identifiable are fully imputed, in other words replaced with resampled values. Extending previous research to conduct significance testing with multiply-imputed datasets was done in [135, 136].

A special problem concerning missing intervals in time series and their effect to the values of the cross-correlation function is discussed in [137]. The paper suggests two new Fourier transform based significance testing methods for the problem. Both of the methods begin with estimating the spectrum of the original data. On the general level the methods then conduct some randomization in the spectral space and reconstruct ran-

domized signals with the inverse Fourier transform. Finally the intervals missing from the original data are removed also from these generated data samples and the significance of cross-correlation values is then evaluated.

Let us now move on to the topic of uncertainty in data. Acquiring data that is noisy either inherently or due to the acquisition process requires associating the data points with probabilities of its possible values. Another source of uncertainty arises when results of predictive patterns are stored in a database, because such derived data in general always contains some uncertainty even if the underlying data were fully known and deterministic.

Example 4.4. *Satellite photos in the visible wavelength are collected and stitched together to form continuous view of the earth's surface. However, radiation from the cosmic rays deteriorates the image quality at the time of its taking and also as it is stored in the satellite [138].*

Additionally, producing useful satellite imagery requires automatically recognizing areas of cloud cover from the photos so that unobstructed data can be provided throughout the view. This recognition task is probabilistic and recording the cloudiness of each pixel in the photos results in a large database of uncertain data.

The topic of uncertain data is a rather recent one and there has been no definite research on significance testing for this data type. The most relevant work discuss the problem of accurately estimating the support of itemsets in an uncertain database [139]. In [140] Calders *et al.* show how this problem can be efficiently solved with simple probability tools such as the central limit theorem and Hoeffding's inequality, assuming that the database elements are pairwise independent. The case of more complex probabilistic models for the data remains open.

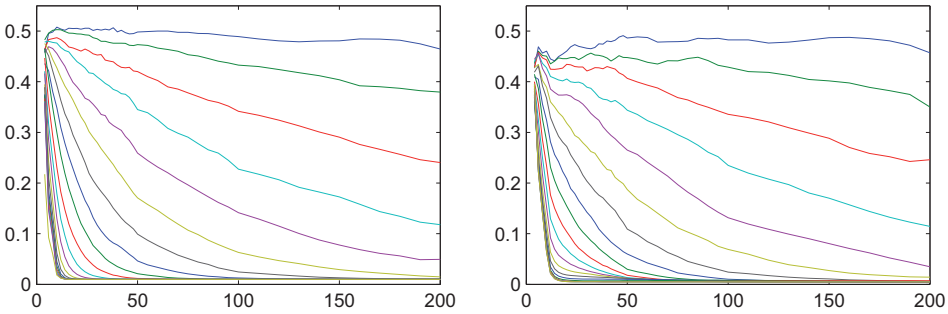
Another case of low quality data, but of rather different type, arises in bioinformatics. Genes that relate to DNA replication and to the cell cycle (phases of cell division) are sought out by finding genes whose activity levels present substantial cyclical nature [141, 142]. Since the search for

these *cyclical genes* begun, there has been a number of experiments suggesting large sets of genes to be cyclical, but the results have had surprisingly little in common. Reasons such as high noise levels and variation and bias in experiment arrangements [41], but also lapses in conducting proper significance testing [71], have been cited as the cause.

Publication V studies the problems arising from high noise and experiment arrangements. The paper contains theoretical analysis for the behavior of some randomization methods most commonly used for detecting significant periodicity in genes. This analysis especially highlights the changes in randomization results when the number of samples in the data changes. The theoretical results are also reflected with experiments on the behavior of p-values. The results of the analysis, as seen in Figure 4.6, indicate that increasing the sampling frequency from the common 1 sample per 10 minutes may substantially improve the quality and also the reliability of the periodicity testing results.

One of the problems in recognizing cyclical genes is the need to synchronize all the cells of the experiment before the testing period by, for example, limiting their nutrient supply or altering the temperature away from normal levels [143]. These procedures may put the cells to a stressed state, causing them to act abnormally and produce false results in finding cyclical genes [71]. To counter this practical problem that causes abnormal periodic behavior in the genes, Publication V introduces a simple randomization method that tests the cyclicity of the genes against each other instead of a static background of pure noise. This method splits the data series into parts every half-cycle and randomly recombines the latter half-cycles with the former ones.

The null model of this method assumes that the majority of the genes in the data do not express cyclical behavior and so their expression levels can be used for constructing an approximate null distribution. On the other hand, preserving the order of time points and the original aggregate expression levels at each time point successfully handles cases of global patterns such as the problematic stress response. As a comparison,



(a) Permutation methods: Permute time points randomly

(b) Autoregressive method: Fit an AR(1) model to each gene separately

Figure 4.6: Examples of how the sample count (x-axis) affects the p-values (y-axis) reported by two commonly used randomization methods. In both figures, each separate curve corresponds to a standard sine wave at interval $[0, 4\pi]$, with noise added so that the sine signal accounts for 5 %, 10 %, ..., 100 % of the total signal energy.

the autoregression-based resampling method suggested by Futschik and Herzel [71] cannot factor in such global erratic behavior as their method resamples the expressions of each gene separately and considers only the average variation across the whole time of the experiment. Additionally, on expectation the method produces samples where the general expression level of each resampled gene is a direct function of the first measured value, which is somewhat arbitrary.

The periodicity at frequency w in the time-wise expressions $E(g,t)$ of a gene g is computed using the commonly used Fourier score [141, 142]

$$F(g) = \sqrt{\left(\sum_t \sin(wt)E(g,t)\right)^2 + \left(\sum_t \cos(wt)E(g,t)\right)^2}.$$

The automatic analysis of the genes with these methods is naturally only the first step in assessing the properties of the genes. Therefore false positives are preferred to false negatives and the significant periodicities are reported by controlling the false discovery rate at level 0.05. Whereas Publication V conducts this using the Benjamini-Hochberg method, Futschik and Herzel [71] use a heuristic method to find a suitable periodicity score threshold using which results in an approximately correct false discovery rate. Such estimates, however, tend to produce too optimistic significance testing results, see [144] and references within.

4.6 Iterative Pattern Mining

As we have already noticed in the earlier sections, data sources are often modeled as a compilation of various properties. Additionally, complex null hypotheses arise often from using multiple different types of constraints on the resampled data. These constraints are built up when, in the progress of data analysis, new properties of data are learned and these have to be taken into account also in significance testing.

Previously unknown properties of the data source are however not the only class of constraints that may be included to the null hypotheses. By limiting the scope of data analysis to the specific dataset under review, any pattern $\mathcal{P}$ that exists in the dataset can be seen as a feature of the data source. Adding a constraint that forces the presence of $\mathcal{P}$ to the null hypothesis enables the data analysis to discount the effects of $\mathcal{P}$ from any other analysis results, making it possible to conduct a more precise analysis. This process of adding more patterns to the null hypothesis as they are found is called *iterative pattern mining*.

Example 4.5. *Suppose that we are analyzing the gene expression levels of a yeast cell in varying environments. We have discovered that gene g has an exceptionally high expression level in the cold and dry environment e. This data point* (g, e),

however, affects also the analyses of, for example, the general expression level of all genes in e and the general expression level of g and its variance.

If this property of the gene g is then considered normal behavior from g, we should constrain the resampling process to reflect this piece of information. On the other hand, even if this behavior is assessed as surprising and specific to this tissue sample, adding the constraint to the null model allows us to better evaluate which other patterns in the data exist in their own right and not simply as a consequence of the surprising level in (g, e).

The problem of iteratively finding more patterns is closely related to the problem of finding the "top-k patterns" from binary or discrete data that have the highest importance by some measure [145, 146, 86]. This problem has been studied extensively in the contexts of comparing explanatory data models [38, 147], pruning of hypotheses [148] and using joint entropy [149, 150] or changes in partitioning of the transactions in data [151]. Also several approaches to adjusting p-values when encountering new patterns have been suggested [152, 153]. The top-k task is NP-hard to solve [154], but in some cases where the importance measure is submodular, the greedy approach has a proven approximation ratio of $(e-1)/e$ [154].

In any case, the goal is to provide a minimal amount of information to the actual user of data analysis while maximizing the usefulness of the results. This can be seen as a refinement of the goal in significance testing to provide the human expert evaluating the results with only those patterns that bear the most importance.

Publication II introduces a new randomization-based approach to finding out only the most important patterns. Instead of the more common approach which prunes out patterns from the set of all discovered patterns, the algorithm of Publication II starts with an empty set of patterns and whereas earlier work such as [152] by Gallo *et al.* use null models of independent variables or transactions, the work of Publication II enables the use of arbitrary null models with iterative mining through the use of

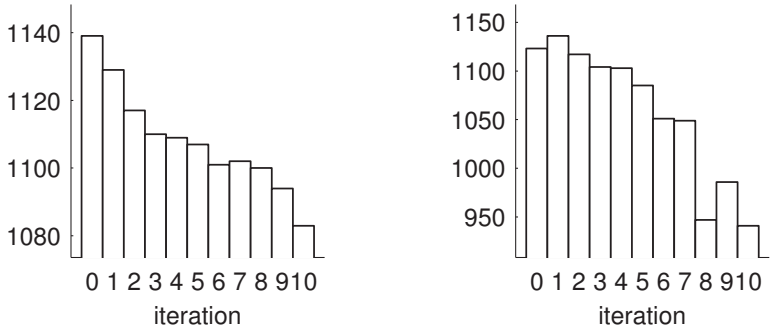
property-based null models.

The general idea of the algorithm in Publication II is to run in iterations, with iteration choosing the pattern of lowest p-value and adding it as a constraint to the property model. For example, suppose that we have found a highly frequent itemset $\mathcal{I} = \{$ flowers, chocolate, milk $\}$ with frequency 8 192 from our data. As in Equation (3.1), we add a Gaussian constraint that penalizes data samples where the frequency of this itemset differs significantly from this original frequency. Therefore this pattern and patterns that depend on it will not show up during the next iterations of the process.

The progress in the number of patterns found can be seen in Figure 4.7. In the figures, itemsets of size 2 and 3 are iteratively mined and with each iteration the itemset of the smallest p-value is added as a constraint to the null model. It is evident that there are numerous dependencies between the found patterns in the two datasets as most iterations substantially shift around the number of significant patterns. It is also useful to note how the pattern count can also increase instead of decreasing. This occurs when there is anti-correlation between patterns, that is, they exhibit "either or" type of behavior and co-exist less often than what random combining would produce.

The iterative process may be continued, for example, until no patterns of some pre-chosen significance are found anymore or until a sufficient amount of patterns is discovered. A particular advantage in using property null models for this task is that the patterns that are sought out and constrained are not limited to one fixed pattern type, but can be more freely chosen. In Publication II this is demonstrated by mixing patterns of fixed row or column margins and frequent itemsets, whose supports in the original dataset are used as the property model constraints.

Publication II uses also k-means clusters as constraints to its null model. This constraint, however, is not as flexible as the later work in Publication III and cannot be included in a property-based null model. Instead, Publication II modifies the standard swap randomization method (see Sec-



(a) COURSES data (b) PALEO data

Figure 4.7: Number of significant itemsets in each iteration while adding the most significant itemset as a constraint to the null model. The COURSES data is a "student - courses completed" dataset gathered in University of Helsinki. The PALEO data contains information on what species fossils are found in given palaeontological sites [155].

tion 4.1) so that swaps may be conducted only within the clusters of a given clustering solution.

As already discussed in Section 3.2, it is debatable whether hard of soft constraints should be used for a given problem. For the case of constraining itemset frequencies, however, this is not the case. Publication II proves that successful randomization using simultaneous hard constraints for the row and column sums and for the frequencies of some itemsets would imply the claim "**RP** = **NP**". Therefore this cannot be done, at least to the current knowledge.

Chapter 5

Conclusion

Modern science is increasingly leaning towards data-driven solutions for finding new ways to understand the surrounding world. Significant patterns in large swaths of data can not only verify theoretical conjectures, but also stimulate new discoveries. However, in many cases the pattern leading to a discovery can only be detected by analyzing extensive amounts of data and automatically rejecting spurious patterns.

This difficulty builds great needs for providing applicable and reliable data analysis tools. Not only does the hidden nature of some patterns in data make their sensitive detection difficult, but the need for this sensitivity increases the risks for reporting false patterns. Additionally, this trade-off between the risk of missing real patterns (false negatives) and the risk of reporting false patterns (false positives) is always specific to every individual data analysis and dataset.

The modern "discovery science" [2, 156] way of doing research needs efficient, reliable, sensitive and especially flexible significance testing tools. This work has concentrated in building a framework upon property-based null models that can address these issues. Implementing and using the general method has been discussed, along with emphasis for the correctness of the significance testing context and the ensuing p-values. The latter

part of the work focused on representing use cases and solutions for specific applications such as real-valued matrices and time series collections. Additionally attention has been given to using data representations for improved data modeling and to conducting iterative significance testing with the framework.

Whereas the introduced MCMC-based randomization scheme is highly customizable and flexible, its power makes it also more computationally demanding than the historical analytical or model-based significance tests. However, randomization is inherently a strongly parallel task and the current and the foreseeable trend in advances of computing abates this deficiency.

The significance testing methodology discussed in this work is still in an early phase in many ways and additional research to improve both its scope and easy applicability is needed. For example, better evaluation of the convergence of the randomization process is needed. The current work should be seen more as a foundation or a template which can be used for developing specialized significance testing setups rather than as a compilation of tools ready for use. Future work in improving the methodology includes theoretical work in understanding better the role of randomization parameters in controlling the process and assessing further the capabilities of using representations for randomization.

In summary, testing the significance of complex null hypotheses with property null models and MCMC methods has shown to be a valid approach in many cases and I believe that they will be increasingly used in the future once their theory is better understood and as their advantages become better known among the researchers in need of these tools.

Bibliography

[1] Herbert A. Simon. Machine discovery. *Foundations of Science*, 1(2): 171–200, June 1995.

[2] Herbert A. Simon, Raúl E. Valdés-Pérez, and Derek H. Sleeman. Scientific discovery and simplicity of method. *Artificial Intelligence*, 91 (2):177–181, April 1997.

[3] John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, August 2005.

[4] Sholom Wacholder, Stephen Chanock, Montserrat Garcia-Closas, Laure El Ghormli, and Nathaniel Rothman. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, 96(6):434–442, 2004.

[5] Markus Ojala, Niko Vuokko, Aleksi Kallio, Niina Haiminen, and Heikki Mannila. Randomization of real-valued matrices for assessing the significance of data mining results. In *SDM '08: Proceedings of the Eighth SIAM International Conference on Data Mining*, pages 494–505, 2008.

[6] David J. Hand, Heikki Mannila, and Padhraic Smyth. *Principles of Data Mining*. The MIT Press, 2001.

[7] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, pages 1–34. 1996.

[8] Erich L. Lehmann. *Testing Statistical Hypotheses*. Springer, 2nd edition, January 1997.

[9] David J. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, 2002.

[10] Roland Chrobok, Joachim Wahle, and Michael Schreckenberg. Traffic forecast using simulations of large scale networks. In *Proceedings of the 2001 IEEE Intelligent Transportation Systems Conference*, pages 434–439, 2001.

[11] Jeroen de Ridder, Anthony Uren, Jaap Kool, Marcel Reinders, and Lodewyk Wessels. Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens. *PLoS Computational Biology*, 2(12):e166, December 2006.

[12] Jeffrey O. Kephart and William C. Arnold. Automatic extraction of computer virus signatures. In *Proceedings of the 4th Virus Bulletin Internation Conference*, pages 178–184, 1994.

[13] Uwe Scherf *et al.* A gene expression database for the molecular pharmacology of cancer. *Nature Genetics*, 24(3):236–244, 2000.

[14] Mikael Fortelius, Aristides Gionis, Jukka Jernvall, and Heikki Mannila. Spectral ordering and biochronology of European fossil mammals. *Paleobiology*, 32(2):206–214, March 2006.

[15] Raymond S. Nickerson. Null hypothesis significance tests: A review of an old and continuing controversy. *Psychological Methods*, 5(2): 241—-301, 2000.

[16] Satu E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1): 27–64, 2007.

[17] Joseph C. Dunn. Well separated clusters and optimal fuzzy-partitions. *Journal of Cybernetics*, 4:95–104, 1974.

[18] Gideon E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[19] Nicholas J. Gotelli and Gary R. Graves. *Null Models in Ecology*. Smithsonian Institution Press, 1996.

[20] Nicholas J. Gotelli and Brian J. McGill. Null versus neutral models: What's the difference? *Ecography*, 29:793–800, 2006.

[21] Julian Besag and Peter J. Diggle. Simple Monte Carlo tests for spatial pattern. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 26(3):327–333, 1977.

[22] Julian Besag and Peter Clifford. Generalized Monte Carlo significance tests. *Biometrika*, 76(4):633–642, 1989.

[23] David R. Cox and David V. Hinkley. *Theoretical Statistics*. Chapman and Hall/CRC, 1979.

[24] Phillip Good. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer, 2000.

[25] John W. Tukey. The problem of multiple comparisons. Unpublished manuscript, 1953. In The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948–1983, pages 1–300.

[26] Yosef Hochberg and Ajit C. Tamhane. *Multiple comparison procedures*. John Wiley & Sons, Inc., New York, NY, USA, 1987.

[27] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[28] Juliet P. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46:561–584, 1995.

[29] Sandrine Dudoit, Juliet P. Shaffer, and Jennifer C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18 (1):71–103, 2003.

[30] John D. Storey and Robert Tibshirani. Statistical significance for genomewide studies. *PNAS: Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[31] John D. Storey. The positive false discovery rate: a Bayesian interpretation and the q-value. *Annals of Statistics*, 31(6):2013–2035, 2003.

[32] K.R. Gabriel. Simultaneous test procedures–some theory of multiple comparisons. *The Annals of Mathematical Statistics*, 40(1):224–250, 1969.

[33] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[34] R.J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, December 1986.

[35] Branko Şoric. Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*, 84(406):608–610, 1989.

[36] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29:1165–1188, 2001.

[37] Mark E. J. Newman and Gerard T. Barkema. *Monte Carlo Methods in Statistical Physics*. Oxford University Press, 1999.

[38] David Jensen. Knowledge discovery through induction with randomization testing. In *KDD '91: Proceedings of the 1991 Knowledge Discovery in Databases Workshop*, pages 148–159, 1991.

[39] Sayan Mukherjee, Polina Golland, and Dmitry Panchenko. Permutation tests for classification. In *AI Memo: Artificial Intelligence Laboratory, Massachusetts Institute of Technology*, 2003.

[40] Edward Carlstein. Resampling techniques for stationary time-series: some recent developments. In *IMA Volumes in Mathematics and its Applications: New Directions in Time Series Analysis I*, volume 45, pages 75–85. Springer - Verlag, 1992.

[41] Matthias Futschik and Toni Crompton. Model selection and efficiency testing for normalization of cDNA microarray data. *Genome Biology*, 5:1–20, 2004.

[42] Marti J. Anderson. Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences*, 58(3):626–639, 2001.

[43] John P. Boyd and Kai J. Jonas. Are social equivalences ever regular? permutation and exact tests. *Social Networks*, 23:87–123, 2001.

[44] George Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press, 2nd edition, 1980.

[45] Kleanthis-Nikolaos Kontonasios and Tijl de Bie. An information-theoretic approach to finding informative noisy tiles in binary databases. In *SDM '10: Proceedings of the 10th SIAM International Conference on Data Mining*, 2010.

[46] Edwin T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, September 1982.

[47] Aristides Gionis, Heikki Mannila, Taneli Mielikäinen, and Panayiotis Tsaparas. Assessing data mining results via swap randomization. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 167–176. ACM Press, New York, NY, USA, 2006.

[48] Scott Kirkpatrick, Charles D. Gelatt, and Mario P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[49] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Mici Teller, and Edward Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[50] Persi Diaconis. The Markov chain Monte Carlo revolution. *Bulletin of the American Mathematical Society*, 46(2):179–205, November 2008.

[51] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, December 2008.

[52] James G. Propp and David B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9:223–252, August 1996.

[53] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, November 1992.

[54] Ishwar V. Basawa. Estimation of the autocorrelation coefficient in simple Markov chains. *Biometrika*, 59(1):85–89, April 1972.

[55] W. Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[56] Wan-Di Wei. The class $\mathcal{A}(r,s)$ of (0,1)-matrices. *Discrete Mathematics*, 39(3):301–305, 1982.

[57] Bo-Ying Wang and Fuzhen Zhang. On the precise number of $(0,1)$-matrices in $U(R,S)$. *Discrete Mathematics*, 187:211–220, 1998.

[58] Persi Diaconis and David A. Freedman. Finite exchangeable sequences. *Annals of Probability*, 8(4):745–764, 1980.

[59] Persi Diaconis. Recent progress in de Finetti's notions of exchangeability. In *Bayesian Statistics 3: Proceedings of the 3rd Valencia International Meeting, June 1–5, 1987*, pages 111–125. Oxford University Press, 1988.

[60] Julian Besag. Markov chain Monte Carlo methods for statistical inference. 2004. URL `http://www.ims.nus.edu.sg/Programs/mcmc/files/besag_tl.pdf`.

[61] George W. Cobb and Yung-Pin Chen. An application of Markov chain Monte Carlo to community ecology. *The American Mathematical Monthly*, 110(4):265–288, April 2003.

[62] Ivona Bezáková, Nayantara Bhatnagar, and Eric Vigoda. Sampling binary contingency tables with a greedy start. In *SODA '06: Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 414–423. SIAM, 2006.

[63] Charles J. Geyer. Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pages 156–163, 1991.

[64] David Arthur and Sergei Vassilvitskii. *K*-means++: The advantages of careful seeding. In *SODA '07: Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007.

[65] Greg Hamerly. Making *k*-means even faster. In *SDM '10: Proceedings of the 10th SIAM International Conference on Data Mining*, pages 130–140, 2010.

[66] George E. Box and David R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26(2):211—-252, 1964.

[67] Angela Laird, Baxter Rogers, and Elizabeth Meyerand. Comparison of Fourier and wavelet resampling methods. *Magnetic Resonance in Medicine*, 51(2):418–422, February 2004.

[68] Christophe Ambroise and Geoffrey J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS: Proceedings of the National Academy of Sciences*, 99(10):6562–6566, 2002.

[69] Andrew Gelman and Hal Stern. The difference between "significant" and "not significant" is not itself statistically significant. *American Statistician*, 60(4):328–331, 2006.

[70] Raymond Hubbard and R. Murray Lindsay. Why *p*-values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, 18(1):69–88, 2008.

[71] Matthias E. Futschik and Hanspeter Herzel. Are we overestimating the number of cell-cycling genes? the impact of background models for time series data. *Bioinformatics*, 24(8):1063–1069, 2008.

[72] Neal S. Young, John P. A. Ioannidis, and Omar Al-Ubaydli. Why current publication practices may distort science. *PLoS Medicine*, 5 (10):e201, October 2008.

[73] Kai Puolamäki, Mikael Fortelius, and Heikki Mannila. Seriation in paleontological data using Markov Chain Monte Carlo methods. *PLoS Computational Biology*, 2(2), February 2006.

[74] Erling B. Andersen. *Discrete Statistical Models with Social Science Applications*. Elsevier Science Ltd, 1979.

[75] William S. Shields and Roger M. Heeler. Analysis of contingency tables with sparse values. *Journal of Marketing Research*, 16(3):382–386, August 1979.

[76] Sami Hanhijärvi, Gemma C. Garriga, and Kai Puolamäki. Randomization techniques for graphs. In *SDM '09: Proceedings of the 9th SIAM International Conference on Data Mining*, pages 780–791, 2009.

[77] Béla Bollobás. *Random Graphs (Second Edition)*. Cambridge University Press, 2001.

[78] Paul Erdös and Alfréd Rényi. On random graphs I. *Publicationes Mathematicae*, 6:290–297, 1959.

[79] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[80] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[81] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, January 2002.

[82] Albert-László Barabási and Eric Bonabeau. Scale-free networks. *Scientific American*, 288(5):50–59, May 2003.

[83] Shiby Thomas and Sharma Chakravarthy. Performance evaluation and optimization of join queries for association rule mining. In *DaWak '99: Proceedings of the First International Conference on Data Warehousing and Knowledge Discovery*, pages 241–250, 1999.

[84] Wim Le Page. *Mining Patterns in Relational Databases*. PhD thesis, University of Antwerp, 2009.

[85] Markus Ojala, Gemma C. Garriga, Aristides Gionis, and Heikki Mannila. Evaluating query result significance in databases via randomizations. In *SDM '10: Proceedings of the 10th SIAM International Conference on Data Mining*, pages 906–917, 2010.

[86] Geoffrey I. Webb. Discovering significant patterns. *Machine Learning*, 68(1):1–33, 2007.

[87] Markus Ojala. Assessing data mining results on matrices with randomization. In *ICDM '10: Proceedings of the 10th IEEE International Conference on Data Mining*, pages 959–964, 2010.

[88] Frank J. Massey Jr. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[89] Richard F. Voss. $1/f$ (flicker) noise: Brief review. In *Proceedings of the 33rd Annual Symposium on Frequency Control*, pages 40–46, 1979.

[90] Gregory W. Wornell. Wavelet-based representations for the $1/f$ family of fractal processes. *Proceedings of the IEEE*, 81:1428–1450, October 1993.

[91] Jose C. Nacher and Tatsuya Akutsu. Sensitivity of the power-law exponent in gene expression distribution to mRNA decay rate. *Physics Letters A*, 360(1):174–178, 2006.

[92] Sung-Hyuk Cha and Sargur N. Srihari. On measuring the distance between histograms. *Pattern Recognition*, 35(6):1355–1370, June 2002.

[93] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, September 1978.

[94] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[95] Malik Magdon-Ismail and Konstantin Mertsalov. A permutation approach to validation. In *SDM '10: Proceedings of the 10th SIAM International Conference on Data Mining*, pages 882–983, 2010.

[96] Predrag Radivojac, Zoran Obradovic, A. Keith Dunker, and Slobodan Vucetic. Feature selection filters based on the permutation test. In *ECML '04: Proceedings of the 15th European Conference on Machine Learning*, pages 334–346, 2004.

[97] Damien François, Vincent Wertz, and Michel Verleysen. The permutation test for feature selection by mutual information. In *ESANN '06: Proceedings of the 2006 European Symposium on Artificial Neural Networks*, pages 239–244, 2006.

[98] Markus Ojala and Gemma C. Garriga. Permutation tests for studying classifier performance. In *ICDM '09: Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, pages 908–913, 2009.

[99] Chris Ding and Xiaofeng He. *K*-means clustering via principal component analysis. In *ICML '04: Proceedings of the 21st International Conference on Machine Learning*, pages 225–232, 2004.

[100] Chris Ding, Xiaofeng He, and Horst D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM '05: Proceedings of the 5th SIAM International Conference on Data Mining*, pages 606–610, 2005.

[101] A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL `http://archive.ics.uci.edu/ml`.

[102] Gerhard Tintner. On tests of significance in time series. *The Annals of Mathematical Statistics*, 10(2):139–143, June 1939.

[103] Geoffrey H. Moore and W. Allen Wallis. Time series significance tests based on signs of differences. *Journal of the American Statistical Association*, 38(222):153–164, June 1943.

[104] R. L. Anderson. Tests of significance in time-series analysis. In *Statistical Inference in Dynamic Economic Models*, pages 352–355. John Wiley & Sons, 1950.

[105] K. J. Friston, A. P. Holmes, J-B. Poline, P. J. Grasby, S. C. R. Williams, R. S. J. Frackowiak, and R. Turner. Analysis of fMRI time-series revisited. *NeuroImage*, 2(1):45–53, 1995.

[106] Jens-Peter Kreiss and Jürgen Franke. Bootstrapping stationary autoregressive moving-average models. *Journal of Time Series Analysis*, 13:297–317, July 1992.

[107] Edward Bullmore, Michael Brammer, Steve C. R. Williams, Sophia Rabe-Hesketh, Nicolas Janot, Anthony David, John Mellers, Robert Howard, and Pak Sham. Statistical methods of estimation and inference for functional MR image analysis. *Magnetic Resonance in Medicine*, 35:261–277, February 1996.

[108] Dimitris N. Politis. The impact of bootstrap methods on time series analysis. *Statistical Science*, 18:219–239, 2003.

[109] Silvia Golia and Marco Sandri. A resampling algorithm for chaotic time series. *Statistics and Computing*, 11(3):241–255, July 2001.

[110] Hrishikesh D. Vinod. Maximum entropy ensembles for time series inference in economics. *Journal of Asian Economics*, 17:955–978, December 2006.

[111] Stéphane Mallat. *A wavelet tour of signal processing*. Academic Press, New York, NY, USA, 2nd edition, 1999.

[112] Metin Akay. Wavelet applications in medicine. *IEEE Spectrum*, 34 (5):50–56, May 1997.

[113] Dean Prichard and James Theiler. Generating surrogate data for time series with several simultaneously measured variables. *Physical Review Letters*, 73(7):951–954, August 1994.

[114] James Theiler and Dean Prichard. Constrained-realization Monte-Carlo method for hypothesis testing. *Physica D*, 94(4):221–235, July 1996.

[115] Thomas Schreiber and Andreas Schmitz. Improved surrogate data for nonlinearity tests. *Physical Review Letters*, 77(4):635–638, July 1996.

[116] Claudia Kirch. Resampling in the frequency domain of time series to determine critical values for change-point tests. *Statistics & Decisions*, 25:237–261, 2007.

[117] Ed Bullmore, Chris Long, John Suckling, Jalal Fadili, Gemma Calvert, Fernando Zelaya, T. Adrian Carpenter, and Mick Brammer. Colored noise and computational inference in neurophysiological (fMRI) time series analysis: Resampling methods in time and wavelet domains. *Human Brain Mapping*, 12(2):61–78, February 2001.

[118] Michael Breakspear, Michael J. Brammer, and Peter A. Robinson. Construction of multivariate surrogate sets from nonlinear data using the wavelet transform. *Physica D: Nonlinear Phenomena*, 182:1–22, August 2003.

[119] Michael Breakspear, Michael J. Brammer, Edward T. Bullmore, Pritha Das, and Leanne M. Williams. Spatiotemporal wavelet resampling for functional neuroimaging data. *Human Brain Mapping*, 23:1–25, 2004.

[120] Ivo D. Dinov, John W. Boscardin, Michael S. Mega, Elizabeth L. Sowell, and Arthur W. Toga. A wavelet-based statistical analysis of fMRI data: I. motivation and data distribution modeling. *NeuroInformatics*, 3(4):319–342, December 2005.

[121] Levent Şendur, John Suckling, Brandon Whitcher, and Ed Bullmore. Resampling methods for improved wavelet-based multiple hypothesis testing of parameter maps in functional MRI. *Neuroimage*, 37(4): 1186–1194, October 2007.

[122] W. Goffe. *Wavelets in Macroeconomics: An Introduction*, pages 137–149. Kluwer Academic, 1994. In Computational Techniques for Econometrics and Economic Analysis.

[123] James B. Ramsey. The contribution of wavelets to the analysis of economic and financial data. *Philosophical Transactions of the Royal Society of London. Series A*, 357(1760):2593–2606, 1999.

[124] Patrick Flandrin. Wavelet analysis and synthesis of fractional brownian motion. *IEEE Transactions on Information Theory*, 38(2):910–917, March 1992.

[125] Urs E. Ruttimann, Michael Unser, Robert R. Rawlings, Daniel Rio, Nick F. Ramsey, Venkata S. Mattay, Daniel W. Hommer, Joseph A. Frank, and Daniel R. Weinberger. Statistical analysis of functional MRI data in the wavelet domain. *IEEE Transactions on Medical Imaging*, 17:142–154, April 1998.

[126] Jussi Nikkinen, Seppo Pynnönen, Mikko Ranta, and Sami Vähämaa. Cross-dynamics of exchange rate expectations: a wavelet analysis. *International Journal of Finance & Economics*, 15, 2010.

[127] Brandon Whitcher. Wavelet-based bootstrapping of spatial patterns on a finite lattice. *Computational Statistics & Data Analysis*, 50(9): 2399–2421, 2006.

[128] Rajan S. Patel, Dimitri Van De Ville, and F. DuBois Bowman. Determining significant connectivity by 4D spatiotemporal wavelet packet resampling of functional neuroimaging data. *NeuroImage*, 31 (3):1142–1155, 2006.

[129] Thomas Schreiber. Constrained randomization of time series data. *Physical Review Letters*, 80(10):2105–2108, March 1998.

[130] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley-Interscience, September 2002.

[131] Donald B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434):473–489, 1996.

[132] Rebecka Jörnsten, Hui-Yu Wang, William J. Welsh, and Ming Ouyang. DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics*, 21(22):4155–4161, 2005.

[133] Kim-Hung Li, Trivellore E. Raghunathan, and Donald B. Rubin. Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86(416):1065–1073, 1991.

[134] Kim-Hung Li, Xiao-Li Meng, Trivellore E. Raghunathan, and Donald B. Rubin. Significance levels from repeated *p*-values with multiply-imputed data. *Statistica Sinica*, 1:65–92, 1991.

[135] Jerome P. Reiter. Significance tests for multi-component estimands from multiply imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131(2):365–377, 2005.

[136] Jerome P. Reiter. Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika*, 94(2):502–508, 2007.

[137] David Simpson, Antonio F. Infantosi, and Daniel A. Botero Rosas. Estimation and significance testing of cross-correlation between cerebral blood flow velocity and background electroencephalograph activity in signals with missing samples. *Medical and Biological Engineering and Computing*, 39:428–433, 2001.

[138] Fabian Gieseke, Gabriel Moruz, and Jan Vahrenhold. Resilient *K*-d trees: *K*-means in space revisited. In *ICDM '10: Proceedings of the 10th IEEE International Conference on Data Mining*, pages 815–820, 2010.

[139] Charu C. Aggarwal, Yan Li, Jianyong Wang, and Jing Wang. Frequent pattern mining with uncertain data. In *KDD '09: Proceeding of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 29–38, 2009.

[140] Toon Calders, Calin Garboni, and Bart Goethals. Approximation of frequentness probability of itemsets in uncertain data. In *ICDM '10: Proceedings of the 10th IEEE International Conference on Data Mining*, pages 749–754, 2010.

[141] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.

[142] Ulrik de Lichtenberg, Lars Juhl Jensen, Anders Fausbøll, Thomas S. Jensen, Peer Bork, and Søren Brunak. Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics*, 21(7):1164–1171, 2005.

[143] Penny K. Davis, Alan Ho, and Steven F. Dowdy. Biological methods for cell-cycle synchronization of mammalian cells. *Biotechniques*, 30 (6):1322–1331, June 2001.

[144] Shuo Jiao. *Detecting differentially expressed genes while controlling the false discovery rate for microarray data*. PhD thesis, University of Nebraska - Lincoln, 2010.

[145] Stefan Wrobel. An algorithm for multi-relational discovery of subgroups. In *PKDD '97: Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 78–87, 1997.

[146] Jiawei Han, Jianyong Wang, Ying Lu, and Petre Tzvetkov. Mining top-*k* frequent closed patterns without minimum support. In *ICDM '02: Proceedings of the 2nd IEEE International Conference on Data Mining*, pages 211–218, 2002.

[147] Szymon Jaroszewicz. Interactive HMM construction based on interesting sequences. In *LeGo '08: Proceedings of Local Patterns to Global Models Workshop at PKDD '08: the 12th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 82–91, 2008.

[148] Tobias Scheffer and Stefan Wrobel. Finding the most interesting patterns in a database quickly by using sequential sampling. *Journal of Machine Learning Research*, 3:833–862, March 2003.

[149] Arno J. Knobbe and Eric. K. Ho. Maximally informative *k*-itemsets and their efficient discovery. In *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 237–244, 2006.

[150] Arno J. Knobbe and Eric K. Ho. Pattern teams. In *PKDD '06: Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 577–584, 2006.

[151] Björn Bringmann and Albrecht Zimmermann. The chosen few: On identifying valuable patterns. In *ICDM '07: Proceedings of the 7th IEEE International Conference on Data Mining*, pages 63–72, 2007.

[152] Arianna Gallo, Tijl De Bie, and Nello Cristianini. MINI: Mining informative non-redundant itemsets. In *PKDD '07: Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases*, pages 438–445, 2007.

[153] Geoffrey I. Webb. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. *Machine Learning*, 71:307–323, 2008.

[154] Taneli Mielikäinen and Heikki Mannila. The pattern ordering problem. In *PKDD '03: Proceedings of the 7th European conference on Principles and Practice of Knowledge Discovery in Databases*, pages 327–338, 2003.

[155] Mikael Fortelius. Neogene of the old world database of fossil mammals (NOW), 2006. URL `http://www.helsinki.fi/science/now/`.

[156] Catherine Blake and Meredith Rendall. Scientific discovery: A view from the trenches. In *DS '06: Proceedings of the 9th International Conference in Discovery Science*, pages 41–52, 2006.

DISSERTATIONS IN INFORMATION AND COMPUTER SCIENCE

TKK-ICS-D17 Savia, Eerika.
Mutual Dependency-Based Modeling of Relevance in Co-Occurrence Data. 2010.

TKK-ICS-D18 Liitiäinen, Elia.
Advances in the Theory of Nearest Neighbor Distributions. 2010.

TKK-ICS-D19 Lahti, Leo.
Probabilistic Analysis of the Human Transcriptome with Side Information. 2010.

TKK-ICS-D20 Miche, Yoan.
Developing Fast Machine Learning Techniques with Applications to Steganalysis Problems. 2010.

TKK-ICS-D21 Sorjamaa, Antti.
Methodologies for Time Series Prediction and Missing Value Imputation. 2010.

TKK-ICS-D22 Schumacher, André
Distributed Optimization Algorithms for Multihop Wireless Networks. 2010.

Aalto-DD99/2011 Ojala, Markus
Randomization Algorithms for Assessing the Significance of Data Mining Results. 2011.

Aalto-DD111/2011 Dubrovin, Jori
Efficient Symbolic Model Checking of Concurrent Systems. 2011.

Aalto-DD118/2011 Hyvärinen, Antti
Grid Based Propositional Satisfiability Solving. 2011.

Aalto-DD136/2011 Brumley, Billy Bob
Covert Timing Channels, Caching, and Cryptography. 2011.

In data mining large amounts of data are searched through for useful information, pieces of which are called patterns. Significance testing is an important part of this task as the found patterns need to be assessed for their relevance and significance before further actions. Advances in science have brought along the need to evaluate the significance of complicated data patterns within complicated datasets. This thesis suggests using the framework of property-based randomization for building reliable and flexible significance testing tools that can be adapted and extended for a wide variety of applications. Additional concepts are also discussed as ways to enlarge the scope of these tools. Finally, examples of using these general ideas in applications such as databases and time series collections are given.

Aalto-DD 11/2012

ISBN 978-952-60-4494-1
ISBN 978-952-60-4495-8 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Information and Computer Science
www.aalto.fi