MŰEGYETEM 1782

Budapest University of Technology and Economics
Department of Measurement and Information Systems

Bayesian relevance and effect size measures

PhD thesis booklet

Gábor Hullám

Advisor:
György Strausz, PhD (BME)

Budapest, 2016

1 Preliminaries and objectives

Understanding the underlying mechanisms of various phenomena was always a basic goal of scientific research. In each problem domain mechanisms are defined by the often delicate and complex relationships and interactions of variables. Observational studies generally aim to discover these relationships by investigating dependency patterns of variables using diverse methods. In the general case there are one or more selected variables upon which a study focuses on. Typically these are special state descriptors that provide a labeling according to which samples related to the domain can be classified. Hence they are called class or target variables. The analysis of the relationships of variables may bring forth several significant questions:

- How can be relationships characterized?
- Is it sufficient to qualitatively assess whether a relationship exists or a quantitative analysis is also required?
- Is it acceptable if an analysis investigates only univariate relationships (variable pairs)?
- To what extent should multivariate relationships be examined?

The answers mostly depend on the investigated domain, therefore several methods applying different approaches were devised to provide solutions. These methods are collectively called feature subset selection (FSS) methods [KJ97].

Feature subset selection is a widely used technique in several fields such as machine learning and statistics [BL97; RN10]. The overall goal of FSS is to identify relevant, predictive variables with respect to one or more target variables. The result of FSS is a set of relevant variables which can be defined in multiple ways [BW00]. For example, the set can be created by selecting a predefined number of best scoring variables according to some measure of relevance. Another possible option is to apply a previously established threshold on the selected measure [HS97]. Choosing the appropriate FSS method for a specific application can be problematic as there is an abundance of options regarding the applicable measures and selection methods. [GE03]

Another possible approach towards describing relationships between variables is to focus on the detailed characterization of relationships and interactions by using model-based exploratory tools. These methods apply such association measures that, aside from identifying relevant variables, provide additional information on the relationships between variables. The outcome is a refined, knowledge-rich, systems-based model of the investigated domain which however comes at the price of high computational complexity.

The research presented in this dissertation is related to the latter described systems-based approach. The main motivation of this work was to provide methods and solutions for application domains from the fields of biomedicine and genetics that require such level of detailedness. The analysis of genetic association studies became the central focus which requires analysis methods that have systems-based modeling capabilities and provide a consistent handling of statistical results. Bayesian network based methods applied in a Bayesian statistical framework can fulfill those requirements as they provide a detailed characterization of variable relationships based on model averaging and probabilistic inference. The main objectives of this dissertation are: (1) the extension of an existing framework of Bayesian analysis methods (called Bayesian relevance analysis methodology) with novel methods quantifying the effect size of variables, and (2) the description of considerations and settings required to facilitate the application of these methods for the analysis of genetic association studies.

2 Existing methods and approaches

The basic FSS approach focuses on finding the minimal number of relevant variables that allow the prediction of the target variable with a given accuracy [GE03; Inz+00]. This is typically implemented by applying some learning method that creates a classifier based on the data . The two main aspects that are taken into account are classifier accuracy and model complexity. One usually has to accept a trade-off between the two aspects.

The two main building blocks of FSS methods are the feature selection (search) algorithm and a scoring component responsible for classifier learning. Depending on the implementation, and more specifically the integration of these components [LY05], FSS methods can be divided into three groups: (1) filters [HS97; KS96], (2) wrappers [Inz+00] and (3) embedded methods [LY05].

The simplest solutions among filter methods apply a univariate approach, that is a selected measure of relevance is computed separately for each variable, and the highest scoring variables are selected as result [GE03]. Such methods can be efficient due to their simplicity and may identify some of the relevant variables. However, they are unable to analyze the multivariate dependency patterns, i.e. the relationships between variables. Therefore, valuable information can be lost which may hinder the effort to understand the mechanisms of the investigated problem domain.

In several scenarios a comprehensive investigation is needed which requires multivariate modeling methods that enable the analysis of interactions between variables.

2.1 Multivariate modeling methods

Multivariate modeling methods can be categorized in several ways [DGL96; Gel+95; Ste09; Esb+02]. For example, Devroye et al. distinguishes three main groups of methods: conditional modeling, density learning and discriminant function learning [DGL96], whereas Esbensen et al. separates methods according to their main purpose, e.g. classification methods and discrimination methods [Esb+02].

In this section, two selected groups of multivariate methods are highlighted: (1) *conditional modeling based methods* and (2) *systems-based modeling methods.* The new results presented in this dissertation are related to the latter group: systems-based modeling, whose main features are discussed and compared with conditional modeling.

The *conditional modeling* approach typically applies wrapper methods and aims to identify a highly predictive set of variables regardless of their interdependencies and the possible roles of these variables in the causal mechanisms related to the target variable. Even though conditional modeling allows the analysis of interactions of variables, related methods do not provide a refined characterization of relationships [PCB06]. For example, logistic regression, which is a popular conditional modeling method, allows interaction terms to be added to a model, but it does not provide any detail on the relationship type between the target variable and the investigated variables.

Let us consider an example domain in which variables form a dependency pattern as depicted on Figure 1. Let us assume that the variable X_0 is identified as significant with respect to the target variable Y. Using conditional modeling methods, the fact that this is a transitive relationship, i.e. the effect of X_0 on Y is mediated by several other variables ($X_0 \rightarrow X_6 \rightarrow Y$ and $X_0 \rightarrow X_3 \rightarrow X_7 \rightarrow Y$), remains hidden. In an systems-based approach (in which the focus is centered on the interpretation and translation of results) this is a drawback, as its goal is to discover the mechanisms of the domain and thus the role of variables. On the other hand, in other scenarios e.g. in a predictive setup, the result that X_0 is a significant variable can be satisfactory without the need of additional details.

In contrast, *systems-based modeling* methods aim to identify dependency relationships concerning all examined variables (both between targets and predictors, and between targets) [12]. These dependency patterns can be visualized by a directed acyclic graph, using nodes to represent variables and directed edges to represent relationships between them [Lun+00; OLD01]. This graph may coincide with the causal model which describes the mechanisms of the domain. Depending on the relative position according to a selected variable, (which is typically the target variable), relationships can be categorized as direct causes, direct effects, interactions and transitive relationships.

In case of the dependency pattern shown in Figure 1 the target is denoted as Y, and direct causes X_6 and X_7 are depicted as green nodes. These variables directly influence Y, i.e. there are no intermediate variables. Straightforwardly, direct effects (X_9, X_{10} and X_{11} denoted with orange nodes) are directly affected by Y. In contrast, interaction terms (denoted as X_4, X_5 using teal blue nodes) are only conditionally dependent on the target via a common effect, in other words these relationships are mediated by another variable. The other indirect relationship type is called transitive relationship due to the fact that in such a case there is a directed path between the target and the variable (in the model), but the variable is non-adjacent to the target [8]. Two special transitive relationships involve a root cause X_0 and a common effect X_n. The former affects the target and several variables on multiple paths, whereas the latter is influenced by various variables related to the target. Direct causes, direct effects and interactions are relevant from a structural aspect as they shield the target from the direct influence of other variables. On the other hand, transitive relationships can also be relevant in practical aspects, e.g. they might be more accessible in terms of measurement.

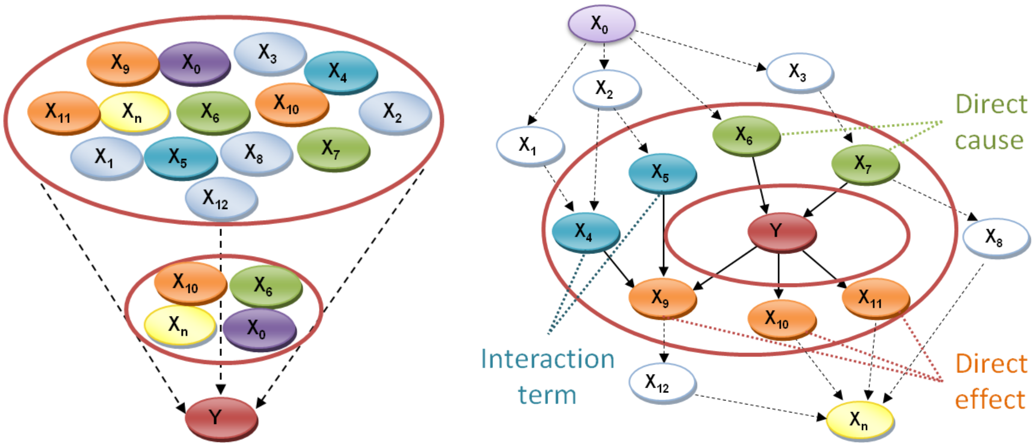

Figure 1: (a) illustration of the conditional modeling approach ignoring structural properties between input variables, (b) illustration of systems-based modeling displaying possible structural relationship types. Y denotes the target, whereas $X_0, X_1, \ldots, X_n$ refer to various measured variables. Relationship types are shown with different colors (1b): X_0 – common cause (purple), X_n – common effect (yellow), X_6, X_7 – direct cause (green), $X_9 - X_{11}$ – direct effect (orange), X_4, X_5 – interaction term (teal blue), $X_1 - X_3, X_8, X_{12}$ – other elements (white). Variables corresponding to nodes that are direct causes, direct effects or interaction terms form a strongly relevant set (see Def. 1) of variables (depicted graphically as a red ring), which statistically isolates the target from other variables.

The drawback of systems-based modeling methods is their computational complexity [Coo90; CHM04] and sample complexity [FY96]. Due to their goal for achieving a refined model of the examined domain they typically require more computational resources than methods related to other approaches. Thus in certain practical scenarios, in which the aim is to find a handful of relevant variables that approximately determine the state of the target, the application of systems-based modeling can be excessive and unnecessary.

2.2 The Bayesian statistical approach

The main challenge regarding systems-based modeling methods is that the identification of a complete model (based on a given data set) is computationally not feasible in most practical cases. The foremost reason of this is the relatively high number of variables with respect to the relatively low number of samples, i.e. insufficient sample size [FY96].

There are two main approaches to alleviate this problem: (1) given a fixed model structure created by experts assess model fitting (with respect to the data), or (2) learn probable models (or parts of models) from data. The former requires a classical statistical approach, that is a hypothesis concerning the structure of the model is required which can be evaluated by the means of statistical hypothesis testing. Structural Equation Modeling (SEM) is a popular methodology in social sciences that follows this paradigm [Pea00].

On the other hand, learning probable models requires a Bayesian statistical approach. This means that instead of evaluating one particular model, several possible models are investigated, i.e. the probability of each model M is assessed based on the data D.

According to the Bayes rule the *a posteriori probability* $P(M|D)$ of a multivariate dependency model can be estimated as [Ber95]:

$$P(M|D) \propto P(D|M) \cdot P(M), \tag{1}$$

where $P(D|M)$ denotes a *likelihood score* which quantifies the probability of (generating) the data given the model M, and $P(M)$ denotes the *prior probability* of the model. The probability of the data which serves as a normalizing term is omitted, for further details see Section 5.4. The consequence of this expression is that a posterior distribution over models can be generated [Mad+96; HGC95].

Furthermore, relying on technique called 'Bayesian model averaging' the common elements (variables) of models can be identified [Mad+96; Hoe+99]. The relevance of a variable is quantified in the form of a posterior probability which is related to its presence in models, e.g. a highly relevant variable is present in most models.

There are several differences between the inherent properties of the classical statistical hypothesis testing paradigm and the Bayesian statistical paradigm. Table 1 summarizes the main points. First of all, the Bayesian approach provides a hypothesis free exploration of the domain, in contrast with the hypothesis testing framework of the classical statistical approach which requires a hypothesis to be tested. This is termed as the alternate hypothesis which is matched against a null hypothesis (typically a worst case model of total independence). Even though Bayesian methods are not hypothesis driven, they allow the incorporation of expert knowledge (i.e. hypotheses) in the form of priors.

Another major difference is related to the method of model validation. In the classical statistical hypothesis testing framework a model is accepted if the related null hypothesis is rejected. This is the case when the p-value corresponding to a computed statistic (i.e. the probability of false rejection) is lower than an arbitrary threshold called significance level (e.g. $\alpha = 0.05$). In the opposite case, the alternate hypothesis is discarded regardless whether the p-value was close to the threshold (e.g. $\mathrm{p-value} = 0.052$) or not (e.g. $\mathrm{p-value} = 0.92$).

Table 1: The comparison of classical statistical and Bayesian approaches based on modeling properties. *Prior knowledge* – the type of a priori information used, *Method of evaluation* – the way of treating results, *Score* – the score used for the evaluation of models, *Result* – the output of modeling, *Variance* – a measure by which variance is defined, *Basis of decision* – the base of deciding on a final model, *Problems* – specific problems of the approach.

Property	**Classical**	**Bayesian**
Prior knowledge	Hypothesis (single model)	Several possible models with prior probabilities
Method of evaluation	Model selection (build your own model)	Model averaging
Score	Statistical test	Bayes factor
Result	p-value (reject or accept null hypothesis)	Posterior probabilities
Variance	Confidence interval	Credible interval
Basis of decision	Significance level	Optimal decision based on expected utility
Problems	Multiple testing problem	Computational complexity

In contrast, the Bayesian framework quantifies belief as posterior probabilities, which is a direct measure of relevance. This allows the probability of models to be compared and enables model averaging. Without discarding any information, all results can be handled consistently.

2.3 The Bayesian network model class

Probabilistic graphical models (PGM) are ideal tools to implement systems-based multivariate modeling as they allow the representation of conditional independencies and dependencies of random variables via a graph structure [12], [FK03; CH92; Mad+96]. The Bayesian network model class is one of the most frequently applied PGMs with a wide variety of application domains including machine learning, computational biology and image processing [Bar12; Mit07]. The three main properties that allow Bayesian networks to be used as versatile modeling tools are: (1) they are able to efficiently represent the joint probability distribution of random variables, (2) they allow the representation of a conditional independency map (i.e. conditional independencies) of random variables, and if a causal interpretation is applicable (3) they are capable of representing directed cause-effect relationships [Pea00]. Bayesian network based methods allow the detection and representation of multivariate dependency relationships, and provide a rich tool set for the detailed characterization of associations [8], [12], [MSL12].

The methods and results described in this dissertation are related to the systems-based multivariate modeling approach. The implementation was based on Bayesian networks applied in a Bayesian statistical framework.

3 Application domains

The dissertation focuses on the following application domains which require systems-based multivariate modeling and pose several new challenges. In addition, due to the high number of possible models, the efficient application of a Bayesian statistical framework is required in order to manage multiple testing and to allow Bayesian model averaging.

3.1 Genetic association studies

In the recent decade, the rapid evolution of biomedical and genetic measurement technologies enabled research concerning the genetic background of multifactorial (i.e. induced by multiple genetic and environmental factors) diseases (e.g. arthritis, depression, and asthma). This new application field requires the capability of modeling complex dependency relationships which is vital for understanding the mechanisms of such illnesses [Ste09]. Genetic association studies (GAS) aimed to identify genetic variables such as single nucleotide polymorphisms (SNP) that influence susceptibility to the investigated disease or affect its severity [Bal07]. A typical GAS consisted of five phases: (1) study design, (2) sample collection, (3) measurement, (4) statistical analysis and (5) interpretation of results.

In the initial period (2000-2005) a simple pairwise association approach was applied, that is statistical dependency was tested between each SNP (or a group of SNPs) and a (typically binary) disease state descriptor. If the distribution of the genotypes (i.e. possible values) of a SNP differs significantly between cases (i.e. patients with disease) and controls (i.e. healthy patients) that indicates that the SNP plays some role in the mechanisms of the investigated disease.

The advent of high-throughput genotyping technologies led to genome-wide association studies (GWAS) which allow the complete measurement of 10^4-10^5 SNPs. In some domains GWAS largely replaced the previously used smaller scale studies in which tens or hundreds of SNPs were examined. The latter is now called candidate gene association study (CGAS).

However, the majority of recent GWAS were only moderately successful. The essential goal of GWAS was to apply a unified approach for statistical analysis, that is to perform the same pairwise analysis for all measured SNPs using the same settings and corrections, instead of selectively analyzing SNPs with various methods. Unfortunately, one of the causes leading to unsatisfactory results was the strict correction for multiple hypothesis testing applied by standard statistical analyses. The correction is required by the hypothesis testing framework to avoid 'by chance' false positive results which are non-negligible in case of thousands of subsequent statistical tests on the same data set. The problem is that the required significance threshold is very low: 10^7-10^8 which poses a considerable limitation on the detectable effect size and the required sample size [GSM08].

The other presumed cause of moderate success is the oversimplified approach of using only simple disease state descriptors while additional environmental and clinical information was neglected. Since multifactorial diseases are largely influenced by environmental variables, recent studies proposed more detailed investigations including such variables [Man+09; EN14]. Hence CGAS came into view again as confirmatory studies using detailed environmental descriptors and phenotypes (i.e. observable features e.g. gender) [PCB13]. However, the previously used univariate methods do not allow the joint analysis of several environmental and genetic variables since that requires multivariate methods.

These obstacles induced an intensive research for new statistical methods. The main requirements can be summarized as follows:

The ability to analyze complex dependency relationships

The 'complex phenotype' approach proposes the joint analysis of genetic, environmental and clinical variables. Therefore, a suitable method should allow multivariate statistical analysis including the detection of interactions [Ste09; Com+00; Sto+04].

Optimal solution for multiple hypothesis testing

The correction for multiple testing is crucial to avoid false positive results, however in its current form (in the hypothesis testing framework) it is overly strict [GSM08]. Moderately significant results are rejected, even though they may be worthy of additional investigation. Furthermore, the correction becomes even more restrictive in case of multiple phenotypes as it requires an increased number of tests and consequently additional correction. New methods should quantify the relevance of moderate or weak results without discarding potentially useful information.

Support for evaluation

Apart from a basic analysis it would be preferable if the new method provided additional tools, i.e. in the form of supplementary measures, that support the visualization and interpretation of results.

Systems-based multivariate modeling relying on Bayesian networks fulfills these requirements as it allows the analysis of dependency relationships, provides a consistent correction for multiple hypothesis testing, and provides a rich tool set for the evaluation and interpretation of results.

3.2 Disease modeling

The aim of disease modeling is to create models with detailed parametrization based on clinical, environmental and genetic variables. The motivation behind building such models is to aid therapy selection, to allow a refined risk assessment, and ultimately to provide medical decision support [LT06; SP78; OLD01; Bel+96; Mik+95]. Although the latter requires the extension of disease models with additional components such as utilities, cost-effectiveness considerations, and other decision theoretic elements. This section only focuses on the creation of models.

There are multiple approaches for building models, each having its advantages and drawbacks.

Model construction based on experts' knowledge

In case of a well known domain with substantial prior knowledge models can be created relying on expert knowledge [Wei+78]. Such a model can be subsequently verified using related data. Straightforwardly, such a methodology can not be applied in completely unexplored domains.

Model learning based on data

When prior knowledge is not available models can be learned relying only on data. In case of Bayesian network based methods a suitable structure learning algorithm can be applied [FK03; CH92]. Then based on the resulting structure a parameter learning method can be executed. A possible drawback of such an approach is that considerably different models may arise depending on the parameterization of the learning algorithm(s). Therefore, a rigorous analysis of parameter settings has to be carried out to validate findings.

Model learning based on data and a priori knowledge

The third option is to use prior knowledge in conjunction with data based model learning. In most practical cases the available a priori knowledge is not sufficient to build a whole model, but it can enhance learning. In the form of informative and non-informative priors most Bayesian methods allow the incorporation of such knowledge. *Informative priors* contain domain specific information, e.g. relevant variables and their relationships. This information can be utilized in structure learning as some variables and relationships can be neglected, while the presence of others can be required [AC08]. *Non-informative priors* consist of more general information, such as the level of direct dependency of variables, the probability of finding relevant variables, or an approximation concerning overall model complexity [Kon+98].

Typically the applied methods implement the third approach, that is relying on some a priori knowledge a model learning is executed. However, in most cases only non-informative priors are used which have a non-trivial connection to the exact domain. The translation between domain knowledge and modeling knowledge is a considerable challenge. For example in case of Bayesian network based modeling if there is a limit on the number of allowed incoming edges related to a node (within the model) then how can that limit be related to a domain specific parameter? There are numerous parameters concerning structure learning which cannot be directly linked to domain specific properties.

Another challenge regarding model learning is related to data sufficiency which mainly depends on the number of variables and the sample size. Furthermore, sufficiency also depends on the cardinality of variables and on the analysis method since the more variables are analyzed jointly, the more data is required to avoid statistically inadequate sample size. Therefore, it is reasonable to have different requirements in terms of modeling in an adequate sample size case than in a low sample size case.

3.3 Intelligent data analysis

Although intelligent data analysis is an integral part of the previously described application fields of genetic association studies and disease modeling [BZ08], this section focuses on a more general perspective.

Intelligent data analysis can be defined as a hypothesis free, data driven statistical analysis with the aim of investigating dependency relationships and providing a detailed description of mechanisms [Kon01; LKZ00; CY08]. However, in some cases the hypothesis free nature can be a drawback, because without a focus it is challenging to detect certain multivariate relationships. Such examples are the following cases:

Contextually relevant variables

One of the main criticism of early genome-wide association studies was their lack of taking appropriate context into account, that is relevant environmental and phenotype descriptors were disregarded in early studies. In new context rich GAS the role of such descriptors is to identify the context in which the genetic and clinical variables have a relevant effect. However, it is plausible that a variable is only relevant in a single context, i.e. only in case of a certain value of a related descriptor, and non-relevant in others [Bou+96]. If the subset of samples corresponding to that context is relatively small compared to the whole set then the relevance of the variable can remain undetected. Especially, if there is no a priori knowledge concerning the context.

Joint effect of variables

In most practical cases the environmental and phenotype descriptors are interdependent and have strong dependencies with the target variable. In contrast, genetic and clinical variables may have considerably weaker dependencies with the target variable, which makes their discovery as relevant variables challenging. Furthermore, if such a variable is relevant only due to its joint effect with another variable and has negligible effect on its own, then its detection can be even more difficult.

A possible solution is to use intelligent exploratory tools which incorporate priors. However, the translation of knowledge conveyed in hypotheses into priors is not a straightforward process, and can prove to be infeasible is some cases. Another possible approach is to create such relevance measures which can be effectively post-processed according to prior knowledge.

4 Objectives

The main objectives of my research can be summarized by the following points:

Objective 1: Relevance measures for intelligent data analysis.
Relevance can be interpreted from various aspects such as parametric, structural and causal aspects. Bayesian network based properties used in a Bayesian statistical framework allow the construction of new relevance measures that combine parametric and structural aspects and provide a more detailed view on relevance relationships.

Objective 2: Investigation of the effect of non-informative priors.
The translation of a priori knowledge into non-informative priors used in Bayesian network based multivariate modeling is a compelling challenge. Particularly in case of parameter priors, whose appropriate setting is essential because they function as complexity regularization variables. In case of several application fields the modeling could be enhanced if the parameter prior can be connected to parametric properties related to the domain.

Objective 3: Application of Bayesian network based multivariate modeling in GAS.
An experimental evaluation of Bayesian relevance analysis, which implements Bayesian network based multivariate modeling, confirmed its applicability for the analysis of genetic association studies. The Bayesian relevance analysis framework in conjunction with its novel extensions (Obj. 1) and new results regarding its parametrization (Obj. 2) could provide a valuable tool set for GAS analysis. Its application could be enhanced by a guideline detailing necessary considerations and recommended settings.

The dissertation is centered around these objectives, and consists of the following results and novel methods.

5 Research method and new results

The foundations of the research presented in this dissertation are related to the research activities of the Computational Biomedicine and Bioinformatics (COMBINE) work group at the Department of Measurement and Information Systems, Budapest University of Technology and Economics. The main research focuses of COMBINE included the development of Bayesian model-based methods utilizing a Bayesian statistical framework and their application in versatile fields such as biomedicine and genetics. The first culmination of this research work was a Bayesian methodological framework called *Bayesian multilevel analysis of relevance* (BMLA), which henceforth will be referred to as Bayesian relevance analysis. This framework was initially developed and applied for ovarian cancer data analysis [Ant07]. Later it was extended and adapted for computational grids by members of COMBINE [12], [26]. Thereafter, collaborations were formed with multiple research groups at Semmelweis University in order to provide Bayesian analysis methods for genetic association studies [9]. Thus genetic association studies became the main application field of Bayesian relevance analysis. Furthermore the methods and results described in this dissertation were motivated by the challenges posed by the analysis of genetic association studies.

My doctoral research began with the adaptation of the Bayesian relevance analysis method to analyze a real-world candidate gene association study (CGAS) data. First, the applicability of the Bayesian relevance analysis method was investigated in a comparative study based on an artificial data set. Its performance was assessed and compared to the performance of several other methods [27]. Second, the Bayesian relevance analysis was applied to an asthma related CGAS data set [3], and the initial methodology for CGAS analysis was developed. Then, during consecutive analysis of several CGAS data related to such fields as rheumatoid arthritis, allergy, leukemia, hypodontia and depression several new challenges emerged. This prompted the extension of Bayesian relevance analysis with novel measures for parametric relevance, and the investigation of the effect of non-informative priors. Finally, the analysis methodology was refined based on new results.

The following section summarizes the new scientific results of the dissertation. First, in Section 5.1 a broad overview of the results is presented according to the objectives discussed in Section 4. Then in Sections 5.2 - 5.4 a detailed description of results is provided, preceded by a brief discussion of related background information.

5.1 Overview of new results

Thesis 1 - overview

Relevance measures typically measure a particular aspect of relevance. Association measures investigate the effect size of variables with respect to a target which allows to assess predictive power. For example, measuring the effect size of a genetic factor with respect to a disease in a case-control study allows assessing whether a variant of that factor increases disease susceptibility. However, this measure will not provide information on the underlying dependency patterns of variables, thus it cannot reveal whether that genetic factor is directly related to that disease or there are mediating factors. In contrast, measures of structural relevance investigate the underlying dependency patterns of variables and utilize Bayesian networks for representation. For example, if the aim of a clinical study is to identify a therapeutic target among the investigated factors, then knowing the role of each factor within the dependency pattern is vital. Influencing a factor that is directly related to (i.e. a potential cause of) the target would yield better results than influencing a transitively related factor. On the other hand, such measures ignore parametric aspects, i.e. they do not assess the quantitative effect of a factor on the target.

In this thesis, I propose a hybrid approach towards relevance by novel Bayesian relevance measures to enable a joint analysis of structural and parametric aspects of relevance. Related publications are the following: [1], [2], [4], [6], [7], [11].

- The proposed hybrid approach requires an *effect size measuring component* quantifying parametric relevance, and a *structural relevance component* providing a qualitative assessment of structural properties (of dependence relationships). The latter component is needed to ensure that the structural aspect of relevance is taken into consideration during effect size computation. For example, a naive hybrid measure could be implemented such that the effect size is computed for all possible dependency structures. However, such a naive measure would not appropriately express structural relevance as those structures would also be considered in which the investigated variable is structurally irrelevant.

- Therefore, I propose a hybrid Bayesian measure of effect size, the *structure conditional Bayesian odds ratio* which only considers those dependency structures in which the investigated variable is structurally relevant. [1], [2], [11].
- I have shown that the *structure conditional Bayesian odds ratio* can be computed using the posterior probabilities of Markov blanket graphs parameterized by the data set. The key notion is that Markov blanket graphs consist of strongly relevant variables.[1] Thus if only such structures are utilized for effect size computation in which the investigated variable is a member of the Markov blanket graph of the target variable, then structural relevance is ensured. This leads to an applicable Bayesian hybrid effect size measure: the *MBG-based Bayesian odds ratio* [2], [7], [11].
- I have extended the MBG-based Bayesian odds ratio measure to a set of variables, allowing the assessment of the joint effect size of multiple factors [2].
- The *MBG-based Bayesian odds ratio* measure was utilized in the analysis of multiple genetic association studies such as the study on the role of folate pathway genes in leukemia [4], the study on the effect of *5-HTTLPR* variants with respect to depression [6], the study investigating genetic and clinical factors of rheumatoid arthritis [7], and a case study exploring the relation between the HTR1A gene and impulsivity [1].

Thesis 2 - overview

The interpretation of scientific results and related decision making are crucial elements of scientific processes. Whether or not a result is plausible (or acceptable) depends on applied benchmarks, common knowledge of a related domain, and additional preferences. Subsequent actions such as publishing results, performing additional experiments, or designing a new study based on the results are also influenced by preferences. The problem is not that such preferences exist, rather that they are not defined formally. A possible solution is the application of a decision theoretic framework with defined preferences. One of the advantages of the Bayesian framework is that it allows the integration of decision theoretic components such as specialized measures that analyze the concordance with the a priori defined preferences. For example, a preference can be a threshold of relevant effect size based on the 'gold standard' (i.e. a generally accepted relevant factor) of a given domain. By applying Bayesian decision theory, the related expected loss defines the optimal course of action, e.g. whether to report the variable as relevant or not.

The investigation of contextual relevance can be another preference implemented by a decision theoretic framework. Contextual relevance means that a variable is relevant only in a specific context, i.e. given a specific value configuration of context forming variables. For example, an environmental variable e.g. 'stress level' (with 'low', 'medium' and 'high' categories) can be a context forming variable, and its 'high' value can be a probable context in which the effect size of stress related factors are relevant. Contextually relevant relationships may not be detected by ordinary measures if the relevant context is rare among the study population. Therefore, the identification of contextually relevant relationships requires a context sensitive decision theoretic measure and candidates for context forming variables. This approach provides relevant results for researchers performing data analysis, particularly in domains with complex dependency patterns.

In this thesis, I propose decision theoretic Bayesian measures which allow a quantitative evaluation of predefined preferences. Related publications are the following: [5], [10], [12].

[1] Strong relevance is a form of structural relevance.

- I propose the effect size conditional existential relevance (ECER) measure which allows the direct application of evaluation specific a priori knowledge. ECER requires the selection of an interval of negligible effect size thereby defining relevant effect sizes [10].
- I propose the contextual extension of ECER (C-ECER) in order to allow the detection of contextually relevant dependency relationships [12].

Thesis 3 - overview

The basic paradigm of the Bayesian approach towards learning dependency structures in the form of Bayesian networks is that the posterior probability $P(G|D)$ of a structure G based on the data D can be estimated based on a likelihood score $P(D|G)$ and a prior probability $P(G)$. The latter components allow the incorporation of various forms of prior knowledge. The utilized prior knowledge can be informative e.g. by assigning higher a priori probability to specific structures, or non-informative e.g. by assuming uniform probability for all possible structures. Apart from such structure priors, the free parameters of the likelihood score can be defined a priori, which are called parameter priors. Informative priors are defined according to domain specific knowledge, however such knowledge is often unavailable. In any case, the prior distribution has to be defined. Non-informative priors are used in such situations when no applicable prior knowledge is available in order to define a neutral a priori assessment of structures and parameters. In case of parameter priors, particularly in case of the frequently applied Dirichlet prior, even the non-informative approach allows multiple options. The free parameter of Dirichlet prior is called the *virtual sample size* which can be viewed as a form of complexity regularization. Even tough virtual sample size is a non-informative prior, it influences Bayesian learning and therefore Bayesian relevance analysis to a significant extent.

* In this thesis, I identified the effects of non-informative parameter priors on Bayesian relevance analysis. Related publications are the following: [2], [3], [7], [8], [9]

- I have derived an expression that connects the virtual sample size parameter of the a priori distribution of parameters related to given variable W to the effect size of variable W [2].
- In a practical approach, this connection can be used to define the parameter prior by setting the virtual sample size according to an expected a priori distribution of effect sizes [2], [7].

Detailed description of new results

5.2 Thesis 1: Bayesian hybrid relevance measures

The relevance of a predictor with respect to a target variable is an essential concept in machine learning, however it can be interpreted in numerous ways, and its relation to the concepts of association, predictive power, and effect size is often not clarified. A general definition of relevance based on conditional probability distributions can be stated as follows [KJ97]:

Definition 1 (Strong and weak relevance). *A feature X_i is strongly relevant to Y, if there exist some $X_i = x_i, Y = y$ and $\mathbf{s_i} = x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ for which $p(x_i, \mathbf{s_i}) > 0$ such that $p(y|x_i, \mathbf{s_i}) \neq p(y|\mathbf{s_i})$. A feature X_i is weakly relevant, if it is not strongly relevant, and there exists a subset of features $\mathbf{S_i'}$ of $\mathbf{S_i}$ for which there exist some x_i, y and $\mathbf{s_i'}$ for which $p(x_i, \mathbf{s_i'}) > 0$ such that*

$p(y|x_i, \mathbf{s_i'}) \neq p(y|\mathbf{s_i'})$. *A feature is relevant, if it is either weakly or strongly relevant; otherwise it is irrelevant.*

In contrast, association is related to unconditional statistical dependence [UC14].

Definition 2 (Association)**.** *An association relationship exists between variables Z and W if $p(Z, W) \neq p(Z)p(W)$, that is Z and W are statistically dependent.*

Measures based on the concept of conditional probability based relevance or on association describe relevance from distinct perspectives. Due to the difference between the approaches, an association between variables Z and W does not necessarily entail that e.g. Z is strongly relevant with respect to W, and vice versa, the fact that W is strongly relevant with respect to Z does not always mean that there is an association between Z and W.

As my main objective was to extend relevance analysis to allow a detailed investigation of multivariate dependency relationships, we examined numerous methods assessing relevance. Based on our findings we distinguished three main approaches:

Association based (parametric) approach.
Relies on effect size measures which quantify the predictive power of one variable on another [Agr02; HTF01]. The nature of the relationship between these variables is not investigated.

Structural approach.
Places the emphasis on investigating the dependency patterns between variables and identifying relationship types. Related methods utilize the structural properties of Bayesian networks by which relationships can be characterized based on data [Nea04; FK03; Hec99; SGS01]. This approach focuses on the qualitative assessment of relationships, i.e. the existence of corresponding structural properties, that is the structural aspect of relevance.

Causal approach.
Assumes a causal (functional) model, which defines cause–effect relationships between variables, and describes the effect of one variable on another. Methods such as structural equation modeling [Pea00] focus on causal relevance, which is existential in the sense that there is either a cause-effect relationship between variables or not, and parametric because it quantifies the change one variable causes in another.

Among these approaches the association based methods are the most frequently applied as they provide simple, univariate effect size measures which quantify relevance from a parametric (predictive) aspect. In contrast, structural uncertainty based methods are used when the investigation of the underlying model and the analysis of multivariate dependency relationships are required. The drawback of these methods is that they focus on a particular aspect of relevance, whereas in practical situations both structural and parametric aspects are uncertain. In such situations an integrated analysis of multiple aspects, e.g. structural and parametric relevance, could provide a more detailed view on the dependency relationships of variables. Bayesian networks as a white-box model class is an ideal candidate for the task as they consist of a structure G in the form of a directed acyclic graph, and a corresponding transparent parametrization layer θ. Existing structure learning methods [Hec99; HMC97; CH92; FGW99; ATS03; Pen+07] allow the analysis of structural relevance, by identifying various structural properties, but they ignore parametric aspects. [2]

[2] More specifically, the parametric layer is analytically averaged out, which can be achieved only if the parameter independence assumption holds [HGC95]. This requires Dirichlet parameter priors (see Section 5.4) and complete data. Otherwise, e.g. if parameter priors are uniquely defined by an expert, or the data is incomplete, then the analytical handling of parameters is no longer possible.

Though the parametrization layer could be used to assess the effect size, i.e. parametric relevance of variables. Therefore,

Thesis 1 ***I propose novel Bayesian relevance measures using a Bayesian network based Bayesian statistical framework to enable the joint analysis of structural and parametric aspects of relevance.***

In this thesis I propose to utilize the structural and parametric properties of Bayesian networks in order to assess the effect size of variables in a Bayesian statistical framework. First, I propose a hybrid Bayesian effect size measure, the structure conditional Bayesian odds ratio (SC-BOR), which combines the structural and parametric aspects of relevance (subthesis 1.1). Second, I propose the application of Markov blanket graphs (MBG) to represent the structural aspect of relevance in the structure conditional effect size measure SC-BOR (subthesis 1.2), and I propose an algorithm for the estimation of the MBG-based Bayesian odds ratio measure (MBG-BOR). Then I propose a multivariate extension of MBG-OR to allow the analysis of the joint relevance of multiple variables (subthesis 1.3).

The results of Thesis 1 are presented in Chapter 4 of the dissertation. Related publications are the following: [1], [2], [4], [6], [7], [11].

5.2.1 Subthesis 1.1: Structure conditional Bayesian effect size (SC-BOR)

I propose a hybrid Bayesian measure of effect size, the structure conditional Bayesian odds ratio $\mathrm{OR}(X_i, Y|\theta, G)$ which relies on both the graph structure G and its parametrization θ of the underlying Bayesian network BN(G, θ).

Proposition 1. *In order to integrate the aspects of structural and parametric relevance I propose to compute the effect size of X_i based on such structures G_j where X_i is structurally relevant, i.e. strongly relevant.*

The Bayesian network based interpretation of strong relevance is related to Markov blanket sets [Pea88]:

Definition 3 (Markov blanket set). *A set of variables $\mathbf{M} \subseteq \mathbf{V} = \{X_1, X_2, ..., X_n\}$ is called a* Markov blanket *set of X_i with respect to the distribution $P(\mathbf{V})$, if $\perp\!\!\!\perp(X_i, \mathbf{V} \setminus \mathbf{M}|\mathbf{M})$, where $\perp\!\!\!\perp$ denotes conditional independence.*

The importance of Markov blanket sets and other structural properties of Bayesian networks is that they are able to encode structural aspects of relevance. The connection between Markov blankets and strong relevance is established by the theorem of Tsamardinos [TA03], which provides the set of conditions under which relevant structural properties are unambiguously represented. Assuming unambiguous representation, for a given structure G all the strongly relevant variables X_i with respect to Y are in the Markov blanket set of Y denoted as $\mathrm{MBS}(Y, G)$.

In other words, $\mathrm{MBS}(Y, G)$ is a strongly relevant set of variables which can be used to define a pairwise relation.

Definition 4 (Markov blanket membership). *The pairwise relation $\mathrm{MBM}(X_i, Y)$ indicating whether X_i is a member of $\mathrm{MBS}(Y, G)$ is called Markov blanket membership.*

The structure conditional Bayesian odds ratio (SC-BOR) can be defined based on the concept of Markov blanket membership as follows:

Proposed definition 1 (Structure conditional Bayesian odds ratio)**.** *Let the indicator function of Markov blanket membership for a given structure G be denoted as $I_{\mathrm{MBM}(X_i,Y|G)}$, which is 1 if $X_i \in \mathrm{MBS}(Y,G)$ and 0 otherwise. Then the structure conditional odds ratio $\mathrm{OR}(X_i,Y|I_{\mathrm{MBM}(X_i,Y|G)})$ is a random variable with a probability distribution $P(\mathrm{OR}(X_i,Y|I_{\mathrm{MBM}(X_i,Y|G)}))$ computed as*

$$P(\mathrm{OR}(X_i,Y|I_{\mathrm{MBM}(X_i,Y|G)})) = \frac{P(\mathrm{OR}(X_i,Y,I_{\mathrm{MBM}(X_i,Y|G)}))}{P(I_{\mathrm{MBM}(X_i,Y|G)})}. \tag{2}$$

This means that the odds ratio is computed for each possible structure G, but only those cases contribute to the distribution $P(\mathrm{OR}(X_i,Y|I_{\mathrm{MBM}(X_i,Y|G)}))$ in which $I_{\mathrm{MBM}(X_i,Y|G)} = 1$, that is X_i is strongly relevant in that structure.

5.2.2 Subthesis 1.2: MBG-based Bayesian effect size (MBG-OR)

SC-BOR can be implemented by applying Bayesian model averaging over structures and parameters using computationally intensive Markov chain Monte Carlo simulation [Mad+96]. However, given the number of possible structures and their parameterizations this computation is highly redundant.

Therefore, instead of learning the whole structure I propose to sample the parameters from the 'relevant part' of the Bayesian network. From the aspect of structural relevance, the structural property which contains all the strongly relevant elements is called the *Markov blanket graph.*

Definition 5 (Markov blanket graph)**.** *"A Markov blanket graph $\mathrm{MBG}(Y,G)$ of a variable Y is a subgraph of a Bayesian network structure G, which contains the nodes of the Markov blanket set of Y, that is $\mathrm{MBS}(Y,G)$ and the incoming edges into Y and its children. Given a target node, which corresponds to the target variable Y, $\mathrm{MBG}(Y,G)$ as a (sub)graph structure consists of nodes that are (1) parents of Y, (2) children of Y or (3) 'other parents' of the children of Y" [26].*

Proposition 2. *The structure conditional Bayesian odds ratio can be computed using the posterior of Markov blanket graphs parameterized by the data set.*

This leads to an applicable Bayesian hybrid effect size measure: the MBG-based Bayesian odds ratio.

Proposed definition 2 (Markov blanket graph based Bayesian odds ratio)**.** *The MBG-based Bayesian odds ratio (MBG-BOR) is computed by averaging over the estimates of odds ratios based on possible MBGs as follows*

$$\text{MBG-BOR}(X_i,Y|D) = \sum_{j=1}^{m} \mathrm{OR}(X_i,Y|\,\mathrm{MBG}_j(Y,G)) \cdot p(\mathrm{MBG}_j(Y,G)|D) \cdot I_{(X_i \in \mathrm{MBG}_j(Y,G))},$$

where m is the number of MBGs with a posterior $p(\mathrm{MBG}_j(Y,G)|D) > 0$. The indicator function $I_{(X_i \in \mathrm{MBG}_j(Y,G))}$ is 1 if $X_i \in \mathrm{MBG}_j(Y,G)$ and 0 otherwise.

The implementation of MBG-BOR is described in Algorithm 1.

Algorithm 1 Calculation of $\text{MBG-BOR}(X_i, Y)$ and its credible interval

```
Require: $n, m, \text{MBG}(Y, G), D$
  for $\text{MBG}_{1\ldots n}$ do
    for $\theta_{1\ldots m}$ do
      draw parametrization $\theta_k = (X_{k1} = x_{k1}, \ldots, X_{kr} = x_{kr})$
      for all $X_k \in \text{MBG}_j$, so that $X_k \neq X_i$.
      estimate $P(Y = 0|X_i = x_i, \theta_k)$
      compute $\text{Odds}(X_i = x_i, \theta_k) = \frac{P(Y=1|X_i=x_i,\theta_k)}{P(Y=0|X_i=x_i,\theta_k)}$
      compute $\text{OR}(X_i, \theta_k) = \frac{\text{Odds}(X_i=x_i^1,\theta_k)}{\text{Odds}(X_i=x_i^0,\theta_k)}$
    end for
    compute $\text{OR}(X_i | \text{MBG}_j) = \sum_{\theta_k=1}^{m} \text{OR}(X_i, \theta_k)$
    update $\text{OR}_{\text{histogram}}(X_i)$
  end for
  $\text{MBG-BOR}(X_i, Y|D) = \sum_{\text{MBG}_j=1}^{n} \text{OR}(X_i, Y | \text{MBG}_j) \cdot p(\text{MBG}_j |D)$
  calculate credible interval for $\text{MBG-BOR}(X_i, Y)$ based on $\text{OR}_{\text{histogram}}(X_i)$
```

5.2.3 Subthesis 1.3: MBG-based Multivariate Bayesian effect size

The MBG-based Bayesian odds ratio measure can be extended to a set of variables, allowing the assessment of the joint effect size of multiple factors.

Proposed definition 3. *Given a set of predictors* $\mathbf{V} = \{X_1, X_2, \ldots, X_n\}$ *the multivariate* MBG-BOR *is calculated as*

$$\text{MBG-BOR}^*(\mathbf{V}, Y) = \sum_{j=1}^{m} \text{OR}(\mathbf{V}, Y | \text{MBG}_j(Y, G)) \cdot p(\text{MBG}_j(Y, G)) \cdot I^*_{(\mathbf{V} \in \text{MBG}_j(Y,G))}, \tag{3}$$

where the indicator function $I^*_{(\mathbf{V} \in \text{MBG}_j(Y,G))}$ *is* 1 *if for any* $X_i \in \mathbf{V}$ *it is true that* $X_i \in \text{MBG}_j(Y, G)$*, and* 0 *otherwise.*

Correspondingly, the MBG-based odds for a set of variables $\mathbf{V}$ is given as

$$Odds^*_{MBG_j(Y,G)}(\mathbf{V}, Y) = \frac{p(Y = 1| \text{MBG}_j(Y, G), X_{n1} = x_{n1}, \ldots, X_{nr} = x_{nr})}{p(Y = 0| \text{MBG}_j(Y, G), X_{n1} = x_{n1}, \ldots, X_{nr} = x_{nr})} \tag{4}$$

, where $x_{n1} \ldots x_{nr}$ are instantiations of variables $X_{ni} \in \mathbf{V}$ that are in $\text{MBG}_j(Y, G)$.

5.3 Thesis 2: A priori knowledge driven Bayesian relevance measures

The classical hypothesis testing framework enables the transformation of a priori knowledge into consistent hypotheses and their subsequent investigation [Bor98]. However, the required correction for multiple hypothesis testing (MHT) prohibits the applicability of this framework in case of domains with complex, multivariate dependency patterns [Nob09]. This is a serious issue in biomedical domains in which a viable hypothesis may provide an essential focus in the data analysis process.

There are several classical [BH95; Dun61] and Bayesian [Sto02] solutions to mitigate the MHT problem. However, multivariate Bayesian methods have a built-in correction for MHT

[GHY12], [NC06]. For example, in case of Bayesian relevance analysis the posterior probability of a structural property is a result of Bayesian model averaging, which takes possible models, i.e. hypotheses into consideration instead of selecting a single best hypothesis. This means that this Bayesian framework allows the exploration of dependency relationships without a specific focus. Furthermore, the correction arises by taking those models into account in which the investigated property is non-relevant.

In cases when a priori knowledge is available, multivariate Bayesian methods can utilize it in the form of various structure and parameter priors [HGC95], [PS12]. However, some aspects of a priori knowledge cannot be transformed into priors because they belong to a different phase of research. More specifically, knowledge related to the phase of evaluation and reporting of results may prove to be inapplicable in the phase of exploratory analysis. In this case, evaluation means the interpretation and further utilization (i.e. translation) of results. For example, results of a study can be compared to previous studies, accepted benchmarks (i.e. 'gold standards') or other domain knowledge present in the corresponding literature. In such cases, the interpretation of results (e.g. expected or unexpected, acceptable or possibly erroneous) depends on the references they are compared to. Subsequently, results are utilized in research related decision making, e.g. whether to publish results, design additional experiments or try an alternative approach. Researchers are frequently faced with such decisions, however these decisions typically remain unformalized.

A possible solution for this is to apply a decision theoretic framework by utilizing loss functions and corresponding measures. Loss functions are special utility functions that can be used to specify the loss for various scenarios of interest [RN10]. For example, given an experimentally validated factor with a known effect size, a loss function can be formulated to use this effect size as a benchmark for relevant effect sizes. In addition to the loss function, this requires a specific measure that can incorporate this evaluation specific prior knowledge. Another aspect of such a priori knowledge is the knowledge of relevant context. In complex mechanisms contextuality can play a relevant role, as some effects only occur in a specific context [WB10], [Bou+96]. In such cases the knowledge regarding contextuality, i.e. the fact that an effect is presumably contextual, and the knowledge of possible factors providing a context is vital, because it may not be possible to detect the effect with general, context insensitive measures. For example, dependencies present only in a subset of the data (i.e. a subpopulation) may be neglected by association measures if the size of the subset is relatively small compared to the whole data based on which the dependencies are evaluated. In order to cope with contextuality a context sensitive loss function and a corresponding measure can be defined, which allow the detection of contextually relevant variables.

The utilization of evaluation specific priors in conjunction with a corresponding decision theoretic framework requires special measures that directly rely on a priori knowledge and are applicable in the established Bayesian framework. Therefore,

Thesis 2 ***I propose a novel decision theoretic Bayesian measure of existential relevance based on parametric properties of Bayesian networks, which allows the application of evaluation specific a priori knowledge.***

In this thesis I propose a novel approach towards measuring existential (structural) relevance. I propose to utilize the Bayesian effect size distribution related to a selected variable and the target variable to measure the existential relevance of the selected variable with respect to the target variable. This novel measure (subthesis 2.1) is called the *effect size conditional existential relevance* (ECER). ECER allows the expression of evaluation specific a priori knowledge by relying on an a priori defined interval (ϵ) of negligible effect size based on experts' knowledge. In order to allow the detection of contextually relevant dependency relationships I propose the contextual extension of ECER called C-ECER (subthesis 2.2). This measure allows the expression of the a priori knowledge related to relevant contexts.

The results of Thesis 2 are presented in Chapter 4 of the dissertation. Related publications are the following: [5], [10], [12].

5.3.1 Subthesis 2.1: Effect size conditional existential relevance (ECER)

I have proposed a Bayesian existential relevance measure, the effect size conditional existential relevance (ECER) which is based on the Bayesian effect size distribution $P(\mathrm{OR}(X_i, Y|\Theta))$. I have shown that ECER allows the direct application of evaluation specific a priori knowledge by allowing the selection of an interval of negligible effect size.

The formalization of ECER requires a C_ϵ interval of negligible effect size which can be defined based on the a priori knowledge of an expert with the constraint that the neutral odds ratio of 1 has to be included in the interval. For example, a C_ϵ interval of size $\epsilon = 0.3$ can be defined symmetrically with respect to the neutral odds ratio such that $C_\epsilon = [0.85, 1.15]$, asymmetrically $C_\epsilon = [0.9, 1.2]$, or unilaterally $C_\epsilon = [1.0, 1.3]$.

Proposed definition 4 (Effect size conditional existential relevance - ECER). *Given an interval of negligible effect size C_ϵ with size $\epsilon \geq 0$, and a distribution of effect size $P(\mathrm{OR}(X_i, Y|\Theta))$ for variable X_i with respect to a selected Y, let $I_{\{ECER_\epsilon(X_i,Y)\}}$ denote that $\mathrm{OR}(X_i, Y|\Theta) \notin C_\epsilon$ that is the effect size of variable X_i is relevant as it is outside the C_ϵ interval. In the Bayesian framework the posterior of effect size conditional existential relevance $ECER_\epsilon(X_i, Y)$ can be defined as $p(I_{\{ECER_\epsilon(X_i,Y)\}})$ [10].*

Note that $\mathrm{OR}(X_i, Y|\Theta)$ denotes the random variable of the Bayesian effect size of X_i with respect to Y, $P(\mathrm{OR}(X_i, Y|\Theta))$ denotes its distribution, and Θ is a random variable of possible parameterizations.

Proposition 3. *The distribution mass of $P(\mathrm{OR}(X_i, Y|\Theta))$ in the intersection of C_ϵ and the credible interval for $\mathrm{OR}(X_i, Y|\Theta)$ quantifies the expert knowledge based parametric irrelevance of variable X_i with respect to Y [10].*

C_ϵ defines the interval of negligible effect size, i.e. parametric irrelevance. If the credible interval for $\mathrm{OR}(X_i, Y|\Theta)$ intersects with interval C_ϵ then X_i is parametrically irrelevant to some extent. The greater the mass of $P(\mathrm{OR}(X_i, Y|\Theta) \in C_\epsilon)$ the less parametrically relevant X_i is (see Figure 2 for an illustration).

If the credible interval of $\mathrm{OR}(X_i, Y|\Theta)$ does not intersect with C_ϵ, that indicates that X_i is parametrically relevant as the effect size distribution $P(\mathrm{OR}(X_i, Y|\Theta))$ only consists of non-negligible values.

Proposition 4. *If the credible interval of $\mathrm{OR}(X_i, Y|\Theta)$ does not intersect with C_ϵ then variable X_i is 'existentially relevant' with respect to Y [10].*

The fact that $p(\mathrm{OR}(X_i, Y|\Theta) \in C_\epsilon) = 0$ means that for a given C_ϵ there is no parametrically encoded independence between X_i and Y. Assuming an underlying Bayesian network $\mathrm{BN}(G, \theta)$ (a graph structure G and its parametrization θ), for which G faithfully represents all conditional independencies (entailed by the Markov assumption [Nea04]), this means that X_i

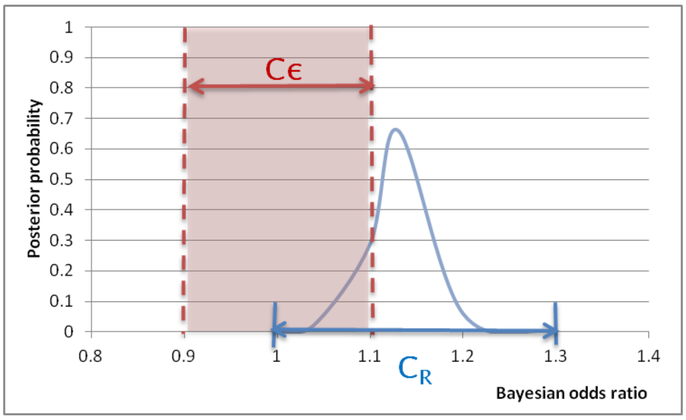

Figure 2: An example for the relationship between the C_ϵ interval of negligible effect size and the credible interval C_R of the posterior distribution of a Bayesian effect size measure of variable X_i. In the presented case C_R intersects with C_ϵ, thus X_i is parametrically irrelevant to some extent given the specific C_ϵ.

and Y are dependent not only on the parametric level, but also on the structural level. However, the exact nature of structural relevance cannot be determined based on $\mathrm{OR}(X_i, Y|\Theta)$ alone, i.e. whether there is a direct relationship or an interaction between X_i and Y. Only the existence of a relationship between X_i and Y can be stated, which is thus called *existential relevance*.

5.3.2 Subthesis 2.2: Contextual extension of ECER

Formerly, the concept of contextuality was defined from the contextual irrelevance perspective [12], i.e. focusing on the independence of variable X_i from the target Y given context $\mathbf{C} = \mathbf{c}$.

I propose to define contextuality from the contextual relevance perspective.

Proposed definition 5 (Contextual relevance). *Let us assume that $X_i \cup \mathbf{C}$ is irrelevant for Y, that is $\perp\!\!\!\perp(Y, (X_i \cup \mathbf{C}))$, and $(X_i \cap \mathbf{C} = \emptyset)$. Then X_i is contextually relevant if there exists some context $\mathbf{C} = \mathbf{c}$ for which $\not\perp\!\!\!\perp;(Y, X_i|\mathbf{c})$ [10].*

Given an a priori defined context $\mathbf{C}$ I propose to extend ECER to C-ECER to detect contextually relevant dependencies.

Proposed definition 6 (Contextual ECER (C-ECER)). *Given an interval of negligible effect size C_ϵ with size $\epsilon \geq 0$, and a distribution of effect size $P(\mathrm{OR}(X_i, \mathbf{C} = \mathbf{c_j}, Y|\Theta))$ for variable X_i with respect to a selected Y, and a context forming set $\mathbf{C}$ with $\mathbf{c_1}, \mathbf{c_2}, \ldots, \mathbf{c_r}$ possible values, let $I_{\{ECER_\epsilon(X_i, \mathbf{C}=\mathbf{c_j}, Y)\}}$ denote that $\mathrm{OR}(X_i, \mathbf{C} = \mathbf{c_j}, Y|\Theta) \notin C_\epsilon$ that is the effect size of variable X_i given context $\mathbf{C} = \mathbf{c_j}$ is relevant as it is outside the C_ϵ interval.*

Then let $I_{\{C-ECER_{\epsilon,\mathbf{C}=\mathbf{c_j}}(X_i,Y)\}}$ denote that X_i is contextually ECER relevant if there exist a context $\mathbf{C} = \mathbf{c_j}$ for which $I_{\{ECER_\epsilon(X_i, \mathbf{C}=\mathbf{c_j}, Y)\}} = 1$.

This means that X_i is C-ECER relevant if there exists a context in which it is ECER relevant: $I_{\{C-ECER_{\epsilon,\mathbf{C}=\mathbf{c_j}}(X_i,Y)\}} = \bigcup_{c_j=c_1}^{c_r} I_{\{ECER_\epsilon(X_i,\mathbf{C}=\mathbf{c_j},Y)\}}$

5.4 Thesis 3: The effect of non-informative parameter priors

The basic paradigm of Bayesian methods is that the posterior probability $P(A|B)$ can be computed using a prior probability distribution $P(A)$ and a likelihood $P(B|A)$ according to the

Bayes' theorem [Ber95]. Let us assume that there is a set of discrete random variables $\mathbf{V} = X_1, ..., X_n$ whose joint probability distribution $P(\mathbf{V})$ can be faithfully represented by a Bayesian network $\mathrm{BN}(G, \theta)$, where G is a directed acyclic graph structure and θ is its parametrization. Structure learning methods aim to identify the most probable structure(s) based on a given data set D, which requires the computation of the posterior probability of each structure G as

$$P(G|D) = \frac{P(G)P(D|G)}{P(D)}. \tag{5}$$

Since $P(D)$ can be neglected as a modeling constant, in order to compute $P(G|D)$ a prior distribution of possible structures $P(G)$ and a likelihood $P(D|G)$ is required.

The term $P(G)$ can express a priori knowledge as some structures can be presumed more probable than others e.g. by a field expert. There are two main approaches towards using priors: informative [AC08] and non-informative [HGC95; GH95; Bun91; CH92]. The informative approach states that a priori knowledge should be used to a full extent e.g. to guide the search for the most probable structure(s). As a result, informative priors greatly influence subsequent analysis and evaluation processes. The non-informative approach, on the other hand, argues that prior knowledge should not be used to pose strict constraints on new research, as it is possible that previous studies or experiments were incomplete or erroneous. Therefore, non-informative priors have limited influence on subsequent analyses, and they are typically uniform, i.e. the prior probability is the same for all possible entities.

Theoretically the effect of a prior gradually diminishes as more data (evidence) becomes available. In real-world cases however, the available sample size is often insufficient to override the effect of a prior. Thus even non-informative priors may have a decisive role in the learning process.

The term $p(D|G)$ is the likelihood score which measures the probability of the data D given a chosen structure G. A popular choice for this scoring metric is the Bayesian Dirichlet (BD) score [Bun91] which is detailed in the following discussion. The so called hyperparameters of this scoring metric can be defined a priori, allowing the incorporation of a priori knowledge in the form of *parameter priors.*

Thesis 3 ***Following a systematic method I identified the effects of non-informative parameter priors on Bayesian relevance analysis.***

In this thesis, I propose the application of non-informative parameter priors based on various criteria. I have derived the connection between the virtual sample size parameter of the Bayesian Dirichlet metric and the a priori effect size distribution (subthesis 3.1). Based on experimental results, I propose to select the virtual sample size parameter according to the expected value of effect size (subthesis 3.2).

The results of Thesis 3 are presented in Chapter 5 of the dissertation. Related publications are the following: [2], [3], [7], [8], [9].

5.4.1 Subthesis 3.1: The connection between the virtual sample size parameter and effect size

In this subthesis I derive an expression that connects the virtual sample size parameter of the a priori distribution of parameters related to given variable W to the effect size of variable W in the form of a log odds [2].

Definition 6 (Prior Dirichlet distribution). *Let W be a discrete random variable with a multinomial distribution, having k possible values. Furthermore, let ν_i denote the probability that W is instantiated with value w_i, i.e. $p(W = w_i)$. Then the a priori probability density function over parameters ν_i for variable W can be defined by a Dirichlet distribution on the Euclidean space with $\mathbf{R}^{k-1}$ dimensions as:*

$$Dir(\nu_1, ..., \nu_{k-1}|\alpha_1, ..., \alpha_k) = \frac{1}{\beta(\alpha)} \cdot \prod_{i=1}^{k} \nu_i^{\alpha_i - 1}, \tag{6}$$

where α_i denotes the virtual sample size corresponding to parameter ν_i, and $\beta(\alpha)$ denotes the multinomial beta function. Furthermore, ν_i is estimated by maximum likelihood estimates $\frac{N_i}{N}$, where N_i is the number of observations in which $W = w_i$ and N is the size of the data set.

Note that in practical cases α_i are assumed to be identical for all ν_i parameters entailing a single virtual sample size parameter α. In order to estimate the a priori distribution of a log odds ratio, the probability density function of the parameters ν_i needs to be transformed according to way the log odds ratio is computed.

Definition 7. *Given a variable W with $k = 2$ possible values and a binary target variable Y the log odds ratio is defined using conditional probability parameters ν' as:*

$$\log \frac{p(Y = y_1|W = w_1)}{p(Y = y_0|W = w_1)} - \log \frac{p(Y = y_1|W = w_0)}{p(Y = y_0|W = w_0)} = \log \frac{\nu'_{1|1}}{\nu'_{0|1}} - \log \frac{\nu'_{1|0}}{\nu'_{0|0}}. \tag{7}$$

Since the target is binary both log odds terms can be simplified such that $\log \frac{\nu'_{1|1}}{1-\nu'_{1|1}}$ and $\log \frac{\nu'_{1|0}}{1-\nu'_{1|0}}$. Thus the required transformation on the a priori probability density function of parameters is $t(\nu) = \log \frac{\nu}{1-\nu}$.

Due to the non-informative application of parameter prior the uniformity of virtual sample sizes $\alpha_1 = \alpha_2 = ...\alpha_k$ can be assumed.

Proposed definition 7. *The general form of the transformed Dirichlet distribution assuming uniform virtual sample size can be given as:*

$$g(z, \alpha) = \frac{Dir(\frac{1}{1+e^{-z}}|\alpha) \cdot e^{-z}}{(1 + e^{-z})^2} \tag{8}$$

Utilizing the fact that the Beta function $\beta(\alpha)$ can be expressed as terms of the $\Gamma(.)$ function as

$$\beta(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k}(\alpha_i))}, \tag{9}$$

the transformed Dirichlet distribution can also be expressed by $\Gamma(.)$ functions.

Proposed definition 8. *The simplified form of the transformed Dirichlet distribution is defined as*

$$g(z, \alpha) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} \cdot (\frac{1}{1 + e^{-z}})^{\alpha+1} \cdot (1 - \frac{1}{1 + e^{-z}})^{\alpha-1} \cdot e^{-z}. \tag{10}$$

This expression defines the a priori distribution of log odds, and shows that the virtual sample size α plays an influential role.

5.4.2 Subthesis 3.2: Selection of virtual sample size parameter

The transformed distribution $g(z, \alpha)$ is analytically intractable in the sense that the exact identification of the high probability density region of the distribution is not feasible. However, the distribution can be sampled using various virtual sample size parameters α, thus allowing the investigation of its effect on the a priori probability density function of log odds [2], [7].

Experimental results shown in Figure 3 indicate an expected effect that as the virtual sample size increases the credible interval of the a priori distribution of log odds decreases. This means that high virtual sample size values express the a priori belief that high odds values are less likely.

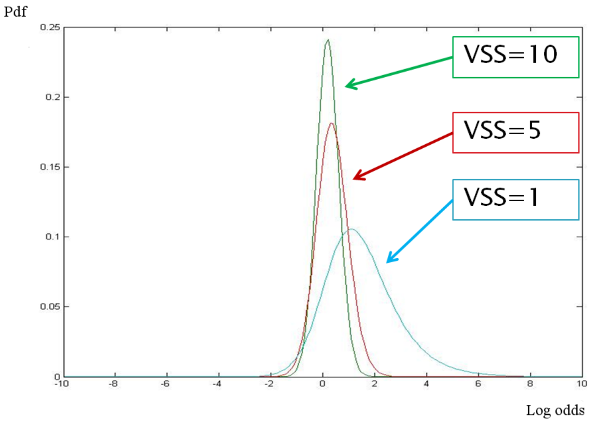

Figure 3: Posterior distributions of log odds corresponding to virtual sample size parameters VSS = 1, 5, and 10.

Proposition 5. *In a practical approach, this connection can be used to define the parameter prior by setting the virtual sample size according to an expected a priori distribution of odds [2], [7].*

6 Applications of new results

This section summarizes the practical applications of the results presented in the PhD dissertation.

6.1 Bayesian relevance analysis of genetic association studies

Bayesian network based Bayesian multilevel analysis of relevance (which is referred to as *Bayesian relevance analysis* in subsequent sections) is a multivariate, model-based Bayesian method which allows the analysis of dependency relationships [26],[25], [12]. It can be utilized just as a general purpose feature subset selection method, but it also allows the refined analysis of complex dependency patterns. Bayesian relevance analysis enables the identification of relevant variables, relevant sets of variables, and interaction models of relevant variables. These are different abstraction levels of relevance which can be analyzed by evaluating the posterior probability of corresponding structural properties of an underlying Bayesian network (e.g. a relevant set of variables correspond to a Markov blanket set). In order to estimate these posterior probabilities a Bayesian model-averaging framework is used.

The results presented in Thesis 3 are closely related to the application of Bayesian relevance analysis, especially for the analysis of candidate gene association studies (CGAS). The novel

Bayesian effect size measures introduced in Thesis 1 and Thesis 2 can be considered as extensions of Bayesian relevance analysis, whose main application field is also the analysis of candidate gene association studies.

Genetic association studies (GAS) investigate the relationship between genetic factors such as single nucleotid polymorphisms (SNP) and various diseases [Hir+02]. Based on the biological samples of healthy and diseased individuals several genetic factors are measured, and the variants are determined. These results are integrated into a data set upon which a statistical analysis is performed. The aim of this analysis is to identify genetic factors which are in a statistical relationship with the target variable, typically a disease descriptor [SB09; Lew02]. In more sophisticated GAS additional sets of clinical, environmental and phenotypic variables (e.g. disease related features and symptoms, or population related features) are also measured and evaluated to aid the discovery of the complex genetic background of multifactorial diseases (e.g. asthma, allergy, rheumatoid arthritis) [EN14].

A previously performed comparative study based on an artificial data set confirmed that Bayesian relevance analysis is an appropriate feature subset selection method, which is capable of the selection of relevant factors when applied for the analysis of genetic association studies [27].

Subsequently, I have applied Bayesian relevance analysis in an asthma candidate gene association study, which was a joint research of the Department of Genetics, Cell- and Immunobiology (DGCI), Semmelweis University and the Department of Measurement and Information Systems, Budapest University of Technology and Economics [3]. Following this successful application, the methodology was applied in several subsequent CGAS in collaboration with the following research groups:

- *Department of Genetics, Cell- and Immunobiology (DGCI), Semmelweis University.* We participated in several genetic association studies related to *asthma* [3] and *leukemia* [4] in which the presented methods were applied. Additionally, we applied Bayesian methodology in studies related to *rheumatoid arthritis* [7], [14], [15], [16].

- *Department of Oralbiology, Semmelweis University.* We participated in two studies investigating the genetic background of hypodontia [17], and parodontitis within the Hungarian population. The presented methods were applied in both studies.

- *MTA-SE Neuropsychopharmacology and Neurochemistry Research Group (MTA-SE-NNRG).* In a joint research effort we aimed to identify relevant environmental and genetic factors related to depression [1]. In a recent study we identified the Galanin (gene)system by the application of Bayesian relevance analysis as a novel therapeutic target for depression [5].

  The complex mechanisms involved in depression require a novel contextual approach in terms of statistical analysis. This prompted the research presented in Thesis 2. and its results were first applied in a depression related study [5], [10].

The results presented in Thesis 1 were applied in a leukemia genetic association study [4] and a rheumatoid arthritis study [7] (joint research with DGCI), furthermore in multiple CGAS investigating genetic and environmental factors that influence depression [6] and impulsivity [1] (joint research with MTA-SE-NNRG).

Based on these applications I have proposed a guideline for the application of Bayesian relevance analysis in genetic association studies, which described recommended parameter settings and practical considerations related to the application of the method [7].

6.2 Development of statistical analysis tools for bioinformatics

The results of Thesis 1 and 3 were applied in the following projects:

- *GENAGRID project.* One of the main goals of GENAGRID was to develop new analysis tools and methodologies for bioinformatic applications. We participated in this project in collaboration with DGCI. Bayesian relevance analysis was tested and applied in related studies [3] and subprojects [27], [11].
- *KOBAK project.* KOBAK aimed to create learning materials [20] [21] [22] [23] [24] and software tools to enhance the learning of biotechnology and bioinformatics. Software tools included analysis tools for various genetic measurements, such as genetic association studies. Bayesian relevance analysis and its extensions were utilized in this project.

Acknowledgements

First and foremost I would like to express my gratitude to Dr. Péter Antal for his guidance, motivation and support of my research work. I am grateful to Dr. György Strausz for his support and encouragement, and for contributing to the stable background that enabled this work. I would like to thank Prof. Dr. György Bagdy the leader of the MTA-SE Neuropsychopharmacology and Neurochemistry Research Group for his support and for the opportunity to use the research group's valuable datasets in order to test the novel methods. I am also grateful to Dr. Gabriella Juhász for her support of my work. My sincere thanks also goes to Prof. Dr. Edit Buzás, Dr. Zsuzsanna Pál, Prof. Dr. Csaba Szalai and their colleagues for past research collaborations. I thank my fellow colleagues of the COMBINE research group for the stimulating discussions both related and unrelated to research. Finally, I would like to thank my family and friends for their limitless patience and their support throughout writing this dissertation.

This research has been supported by the National Development Agency (KTIA_NAP_13-1-2013-0001), Hungarian Brain Research Program - Grant No. KTIA_13_NAP-A-II/14, by the Hungarian Academy of Sciences and the Hungarian Brain Research Program - Grant No. KTIA_NAP _13-2-2015-0001 (MTA-SE-NAP B Genetic Brain Imaging Migraine Research Group), and by MTA-SE Neuropsychopharmacology and Neurochemistry Research Group, Hungarian Academy of Sciences, Semmelweis University.

7 Publication list

Number of peer-reviewed publications:	30
Number of independent citations:	38

7.1 Publications related to the theses

	1	2	3	4	5	6	7	8	9	10	11	12
Thesis 1:	●	●		●		●	●				●	
Thesis 2:					●					●		●
Thesis 3:		●	●				●	●	●			

Journal paper

[1] G. Hullam, G. Juhasz, G. Bagdy, and P. Antal. “Beyond Structural Equation Modeling: model properties and effect size from a Bayesian viewpoint. An example of complex phenotype - genotype associations in depression”. In: *Neuropsychopharmacologia Hungarica* 14.4 (2012), pp. 273–284. doi: 10.5706/nph201212009

[2] G. Hullam and P. Antal. “The effect of parameter priors on Bayesian relevance and effect size measures”. In: *Periodica Polytechnica - Electrical Engineering* 57.2 (2013), pp. 35–48. doi: 10.3311/PPee.2088

[3] I. Ungvari, G. Hullam, P. Antal, P. Kiszel, A. Gezsi, E. Hadadi, V. Virag, G. Hajos, A. Millinghoffer, A. Nagy, A. Kiss, A. Semsei, G. Temesi, B. Melegh, P. Kisfali, M. Szell, A. Bikov, G. Galffy, L. Tamasi, A. Falus, and C. Szalai. “Evaluation of a Partial Genome Screening of Two Asthma Susceptibility Regions Using Bayesian Network Based Bayesian Multilevel Analysis of Relevance”. In: *PLOS ONE* 7.2 (2012), 1–14, e33573. doi: 10.1371/journal.pone.0033573

This paper presents a genetic association study of asthma investigated by Bayesian relevance analysis. This was a joint research of the Dept. of Genetics, Cell- and Immunobiology (DGCI-Semmelweis Univ.) and the Dept. of Measurement and Information Systems (MIT-BUTE), and one of the first applications of Bayesian relevance analysis for genetic association studies. As the co-first author of this paper I performed the Bayesian relevance analysis of the samples and the statistical interpretation of the results. Furthermore, I contributed to the discussion of the Bayesian methodology.

[4] O. Lautner-Csorba, A. Gezsi, D. Erdelyi, G. Hullam, P. Antal, A. Semsei, N. Kutszegi, G. Kovacs, A. Falus, and C. Szalai. “Roles of Genetic Polymorphisms in the Folate Pathway in Childhood Acute Lymphoblastic Leukemia Evaluated by Bayesian Relevance and Effect Size Analysis”. In: *PLOS ONE* 8.8 (2013), 1–13, e69843. doi: 10.1371/journal.pone.0069843

The application of the Bayesian structure conditional odds ratio to allow a detailed characterization of relevance was my contribution to this joint work.

[5] G. Juhasz, G. Hullam, N. Eszlari, X. Gonda, P. Antal, I. Anderson, T. Hokfelt, J. Deakin, and G. Bagdy. “Brain galanin system genes interact with life stresses in depression-related phenotypes”. In: *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 111.16 (2014), E1666–73. doi: 10.1073/pnas.1403649111

This paper presents important results regarding the role of Galanin system genes on depression. My contribution was the application of Bayesian relevance analysis and additional multivariate statistics to investigate the relevance of Galanin system genes.

[6] G. Juhasz, X. Gonda, G. Hullam, N. Eszlari, D. Kovacs, J. Lazary, D. Pap, P. Petschner, R. Elliott, J. Deakin, I. Anderson, P. Antal, K. Lesch, and G. Bagdy. "Variability in the Effect of 5-HTTLPR on Depression in a Large European Population: The Role of Age, Symptom Profile, Type and Intensity of Life Stressors". In: *PLOS ONE* 10.3 (2015), e0116316 15p. DOI: 10.1371/journal.pone.0116316

This paper investigates the role of 5-HTTLPR variants with respect to depression phenotypes. I contributed the Bayesian relevance and effect size analysis, where I applied Bayesian MBG-based odds ratios in order to investigate the interaction of environmental and genetic factors.

Chapter in edited book

[7] G. Hullam, A. Gezsi, A. Millinghoffer, P. Sarkozy, B. Bolgar, S. Srivastava, Z. Pal, E. Buzas, and P. Antal. "Bayesian systems-based genetic association analysis with effect strength estimation and omic wide interpretation: a case study in rheumatoid arthritis". In: *Arthritis Research: Methods and Protocols.* Ed. by S. Shiozawa. Methods in Molecular Biology, vol. 1142. New York: Springer, 2014, pp. 143–176. DOI: 10.1007/978-1-4939-0404-4_14

This chapter discusses the application of Bayesian relevance analysis and a related gene prioritization method for genetic association studies. The guideline for the application of Bayesian relevance analysis, including the necessary preliminary steps, the recommended settings and available analysis and post-processing options was my contribution in the chapter.

[8] P. Antal, A. Millinghoffer, G. Hullam, G. Hajos, P. Sarkozy, C. Szalai, and A. Falus. "Bayesian, Systems-based, Multilevel Analysis of Biomarkers of Complex Phenotypes: From Interpretation to Decisions". In: *Probabilistic Graphical Models for Genetics, Genomics and Postgenomics.* Ed. by C. Sinoquet and R. Mourad. New York: Oxford University Press, 2014, pp. 318–360

This chapter introduces the methodology and related concepts of Bayesian relevance analysis. My contributions included the utilization of structural properties encoding strong relevance, and the interpretation of results based on strong relevance and related relevance subtypes.

[9] P. Antal, A. Millinghoffer, G. Hullam, G.Hajos, C. Szalai, and A. Falus. "A bioinformatic platform for a Bayesian, multiphased, multilevel analysis in immunogenomics". In: *Bioinformatics for Immunomics.* Ed. by M. Davies, S. Ranganathan, and D. Flower. Springer, 2010, pp. 157–185. DOI: 10.1007/978-1-4419-0540-6_11

I contributed to the section describing the application of Bayesian relevance analysis and the interpretation of results.

International conference

[10] G. Hullam and P. Antal. "Towards a Bayesian Decision Theoretic Analysis of Contextual Effect Modifiers". In: *Proceedings of the 7th European Workshop on Probabilistic Graphical Models.* Ed. by L. van der Gaag and A. Feelders. LNAI. Utrecht, The Netherlands: Springer, 2014, pp. 222–237

[11] G. Hullam and P. Antal. “Estimation of effect size posterior using model averaging over Bayesian network structures and parameters”. In: *Proceedings of the 6th European Workshop on Probabilistic Graphical Models*. Ed. by A. Cano, M. Gomez-Olmedo, and T. Nielsen. Granada, Spain: DECSAI, University of Granada, 2012, pp. 147–154

[12] P. Antal, A. Millinghoffer, G. Hullam, C. Szalai, and A. Falus. “A Bayesian View of Challenges in Feature Selection: Feature Aggregation, Multiple Targets, Redundancy and Interaction”. In: *Workshop on New challenges for feature selection in data mining and knowledge discovery (FSDM 2008) at The 19th European Conference on Machine Learning (ECML 2008), Journal of Machine Learning Research - Workshop and Conference Proceedings: FSDM 2008*. Ed. by Y. Saeys, H. Liu, I. Inza, L. Wehenkel, and Y. V. de Peer. Vol. 4. 2008, pp. 74–89

This paper introduces novel concepts including relevant subsets and interaction scores related to Bayesian relevance analysis. I contributed to the analysis of the asthma related data set and the discussion of the results.

7.2 Additional publications

Journal paper

[13] A. Millinghoffer, G. Hullam, and P. Antal. “Statisztikai adat- és szövegelemzés Bayes-hálókkal: a valószínűségektol a függetlenségi és oksági viszonyokig”. In: *Híradástechnika* 60.10 (2005), pp. 40–49

[14] Z. Pal, P. Antal, A. Millinghoffer, G. Hullam, K. Paloczi, S. Toth, H. Gabius, M. Molnar, A. Falus, and E. Buzas. “A novel galectin-1 and interleukin 2 receptor haplotype is associated with autoimmune myasthenia gravis”. In: *Journal of Neuroimmunology* 229.1-2 (2010), pp. 107–111. DOI: 10.1016/j.jneuroim.2010.07.015

[15] S. Srivastava, P. Antal, J. Gal, G. Hullam, A. F. Semsei, G. Nagy, A. Falus, and E. I. Buzas. “Lack of evidence for association of two functional SNPs of CHI3L1 gene(HC-gp39) with rheumatoid arthritis.” In: *Rheumatology International* 31.8 (2010), pp. 1003–1007. DOI: 10.1007/s00296-010-1396-3

[16] Z. Pal, P. Antal, S. Srivastava, G. Hullam, A. F. Semsei, J. Gal, M. Svebis, G. Soos, C. Szalai, S. Andre, E. Gordeeva, G. Nagy, H. Kaltner, N. Bovin, M. Molnar, A. Falus, H. Gabius, and E. I. Buzas. “Non-synonymous single nucleotide polymorphisms in genes for immunoregulatory galectins: association of galectin-8 (F19Y) occurrence with autoimmune diseases in a Caucasian population”. In: *Biochimica et Biophysica Acta-general Subjects* 1820.10 (2012), pp. 1512–1518. DOI: 10.1016/j.bbagen.2012.05.015

[17] G. Jobbagy-Ovari, C. Paska, P. Stiedl, B. Trimmel, D. Hontvari, B. Soos, P. Hermann, Z. Toth, B. Kerekes-Mathe, D. Nagy, I. Szanto, A. Nagy, M. Martonosi, K. Nagy, E. Hadadi, C. Szalai, G. Hullam, G. Temesi, P. Antal, G. Varga, and I. Tarjan. “Complex analysis of multiple single nucleotide polymorphisms as putative risk factors of tooth agenesis in the Hungarian population”. In: *Acta Odontologica Scandinavica* 72.3 (2013), pp. 216–227. DOI: 10.3109/00016357.2013.822547

[18] G. Temesi, V. Virág, E. Hadadi, I. Ungvari, L. Fodor, A. Bikov, A. Nagy, G. Galffy, L. T. L, I. Horvath, A. Kiss, G. Hullam, A. Gezsi, P. Sarkozy, P. Antal, E. Buzás, and C. Szalai. “Novel genes in Human Asthma Based on a Mouse Model of Allergic Airway Inflammation and Human Investigations”. In: *Allergy Asthma and Immunology Research* 6.6 (2014), pp. 496–503. DOI: 10.4168/aair.2014.6.6.496.

[19] A. Gezsi, O. Lautner-Csorba, D. Erdelyi, G. Hullam, P. Antal, A. Semsei, N. Kutszegi, M. Hegyi, K. Csordas, G. Kovacs, and C. Szalai. "In interaction with gender a common CYP3A4 polymorphism may influence the survival rate of chemotherapy for childhood acute lymphoblastic leukemia". In: *Pharmacogenomics Journal* 15.3 (2015), pp. 241–247. doi: `10.1038/tpj.2014.60`

Book chapter

[20] G. Hullam. "Tudásmérnökség, biasok és heurisztikák becsléseknél és döntéseknél". In: *Valószínűségi döntéstámogató rendszerek.* Ed. by A. Antos, P. Antal, G. Hullam, A. Millinghoffer, and G. Hajos. Budapest: Typotex, 2014, pp. 65–85

[21] G. Hullam. "Orvosi döntéstámogatás". In: *Valószínűségi döntéstámogató rendszerek.* Ed. by A. Antos, P. Antal, G. Hullam, A. Millinghoffer, and G. Hajos. Budapest: Typotex, 2014, pp. 128–162

[22] G. Hullam. "Hiányos Adatok". In: *Intelligens adatelemzés.* Ed. by P. Antal, A. Antos, G. Horvath, G. Hullam, I. K. Imre, P. Marx, A. Millinghoffer, A. Pataricza, and A. Salanki. Budapest: Typotex, 2014, pp. 48–57

[23] G. Hullam. "Genetikai asszociációs vizsgálatok standard elemzése". In: *Bioinformatika: molekuláris méréstechnikától az orvosi döntéstámogatásig.* Ed. by P. Antal, G. Hullam, A. Millinghoffer, G. Hajos, P. Marx, A. Arany, B. Bolgar, A. Gezsi, P. Sarkozy, and L. Poppe. Budapest: Typotex, 2014, pp. 90–106

[24] G. Hullam. "Standard analysis of genetic association studies". In: *Bioinformatics.* Ed. by P. Antal, G. Hullam, A. Millinghoffer, G. Hajos, P. Marx, A. Arany, B. Bolgar, A. Gezsi, P. Sarkozy, and L. Poppe. Budapest: Typotex, 2014, pp. 84–100

International conference

[25] A. Millinghoffer, G. Hullam, and P. Antal. "On inferring the most probable sentences in Bayesian logic". In: *Workshop notes on Intelligent Data Analysis in Biomedicine and Pharmacology (IDAMAP-2007), Artificial Intelligence in Medicine (AIME2007).* Ed. by C. Combi and A. Tucker. Amsterdam, The Netherlands, 2007, pp. 13–18

[26] P. Antal, G. Hullam, A. Gezsi, and A. Millinghoffer. "Learning complex bayesian network features for classification". In: *Proc. of Third European Workshop on Probabilistic Graphical Models (PGM06).* Ed. by M. Studeny and J. Vomlel. Prague, Czech Republic, 2006, pp. 9–16

[27] G. Hullam, P. Antal, C. Szalai, and A. Falus. "Evaluation of a Bayesian model-based approach in G[enetic]A[ssociation] studies". In: *Machine Learning in Systems Biology (MLSB2009), Journal of Machine Learning Research - Workshop and Conference Proceedings: MLSB 2009.* Ed. by S. Dzeroski, P. Geurts, and J. Rousu. Vol. 8. 2010, pp. 30–43

Local conference

[28] G. Hullam. "Discovery of causal relationships in presence of hidden variables". In: *13th PhD Mini-Symposium.* Ed. by B. Pataki. Budapest University of Technology and Economics. Budapest, Hungary: Department of Measurement and Information Systems, Feb. 2006, pp. 50–51

[29] G. Hullam. "A first-order Bayesian logic for heterogeneous data sets". In: *14th PhD Mini-Symposium.* Ed. by B. Pataki. Budapest University of Technology and Economics. Budapest, Hungary: Department of Measurement and Information Systems, Feb. 2007, pp. 42–45

[30] G. Hullam. "Bayesian analysis of relevance in presence of missing and erroneous data". In: *15th PhD Mini-Symposium.* Ed. by B. Pataki. Budapest University of Technology and Economics. Budapest, Hungary: Department of Measurement and Information Systems, Feb. 2008, pp. 30–33

References

[AC08] N. Angelopoulos and J. Cussens. "Bayesian learning of Bayesian networks with informative priors". In: *Annals of Mathematics and Artificial Intelligence* 54.1-3 (2008), pp. 53–98.

[Agr02] A. Agresti. *Categorical data analysis.* Wiley & Sons, 2002.

[Ant07] P. Antal. *Integrative Analysis of Data, Literature, and Expert Knowledge.* Ph.D. dissertation, K.U.Leuven, D/2007/7515/99, 2007.

[ATS03] C. Aliferis, I. Tsamardinos, and A. Statnikov. "Causal Explorer: A Probabilistic Network Learning Toolkit for Biomedical Discovery". In: *International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences (METMBS '03).* 2003, pp. 371–376.

[Bal07] D. J. Balding. *Handbook of Statistical Genetics.* Wiley & Sons, 2007.

[Bar12] D. Barber. *Bayesian reasoning and machine learning.* Cambridge University Press, 2012.

[Bel+96] R. Bellazzi, C. Larizza, A. Riva, A. Mira, S. Fiocchi, and M. Stefanelli. "Distributed intelligent data analysis in diabetic patient management". In: *Proc. AMIA Annual Fall Symposium 1996.* 1996, pp. 194–198.

[Ber95] J. M. Bernardo. *Bayesian Theory.* Wiley & Sons, 1995.

[BH95] Y. Benjamini and Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." In: *J. R. Stat. Soc.* 57.1 (1995), pp. 289–300.

[BL97] A. Blum and P. Langley. "Selection of Relevant Features and Examples in Machine Learning". In: *Artificial Intelligence* 97.1-2 (1997), pp. 245–271. DOI: 10.1016/S0004-3702(97)00063-5.

[Bor98] A. Borovkov. *Mathematical Statistics.* Gordon and Breach, 1998.

[Bou+96] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. "Context-Specific Independence in Bayesian Networks". In: *Proc. of the 20th Conf. on Uncertainty in Artificial Intelligence (UAI-1996).* 1996, pp. 115–123.

[Bun91] W. L. Buntine. "Theory Refinement of Bayesian Networks". In: *Proc. of the 7th Conf. on Uncertainty in Artificial Intelligence (UAI-1991).* 1991, pp. 52–60.

[BW00] D. Bell and H. Wang. "A formalism for relevance and its application in feature subset selection". In: *Machine learning* 41.2 (2000), pp. 175–195.

[BZ08] R. Bellazzi and B. Zupan. "Predictive data mining in clinical medicine: Current issues and guidelines". In: *International Journal of Medical Informatics* 77.2 (2008), pp. 81–97.

[CH92] G. F. Cooper and E. Herskovits. “A Bayesian Method for the Induction of Probabilistic Networks from Data”. In: *Machine Learning* 9 (1992), pp. 309–347.

[CHM04] D. M. Chickering, D. Heckerman, and C. Meek. “Large-sample learning of Bayesian networks is NP-hard”. In: *The Journal of Machine Learning Research* 5 (2004), pp. 1287–1330.

[Com+00] D. Comings, R. Gade-Andavolu, N. Gonzaleza, S. Wu, D. Muhleman, H. Blake, F. Chiu, E. Wang, K. Farwell, S. Darakjy, R. Baker, G. Dietz, G. Saucier, and J. MacMurray. “Multivariate analysis of associations of 42 genes in ADHD, ODD and conduct disorder”. In: *Clin Genet* 58.1 (2000), pp. 31–40.

[Coo90] G. F. Cooper. “The computational complexity of probabilistic inference using Bayesian belief network”. In: *Artificial Intelligence* 42 (1990), pp. 393–405.

[CY08] Y. Chen and Y. Yao. “A multiview approach for intelligent data analysis based on data operators”. In: *Information Sciences* 178.1 (2008), pp. 1–20.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, 1996.

[Dun61] O. Dunn. “Multiple Comparisons Among Means.” In: *Journal of the American Statistical Association* 56.293 (1961), pp. 52–64.

[EN14] H. Ehrenreich and K.-A. Nave. “Phenotype-based genetic association studies (PGAS) -towards understanding the contribution of common genetic variants to schizophrenia subphenotypes”. In: *Genes* 5.1 (2014), pp. 97–105.

[Esb+02] K. H. Esbensen, D. Guyot, F. Westad, and L. P. Houmoller. *Multivariate data analysis-in practice: an introduction to multivariate data analysis and experimental design.* 2002.

[FGW99] N. Friedman, M. Goldszmidt, and A. Wyner. “On the Application of the Bootstrap for Computing Confidence Measures on Features of Induced Bayesian networks”. In: *AI&STAT VII.* 1999.

[FK03] N. Friedman and D. Koller. “Being Bayesian about Network Structure”. In: *Machine Learning* 50 (2003), pp. 95–125.

[FY96] N. Friedman and Z. Yakhini. “On the Sample Complexity of Learning Bayesian Networks”. In: *Proc. of the 12th Conf. on Uncertainty in Artificial Intelligence (UAI-1996).* 1996, pp. 274–282.

[GE03] I. Guyon and A. Elisseeff. “An Introduction to Variable and Feature Selection”. In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[Gel+95] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman & Hall, 1995.

[GH95] D. Geiger and D. Heckerman. “A Characterization of the Dirichlet Distribution with Application to Learning Bayesian Networks”. In: *Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence (UAI-1995).* 1995, pp. 196–207.

[GHY12] A. Gelman, J. Hill, and M. Yajima. “Why We (Usually) Dont Have to Worry About Multiple Comparisons”. In: *Journal of Research on Educational Effectiveness* 5 (2012), pp. 189–211.

[GSM08] X. Gao, J. Starmer, and E. Martin. “A Multiple Testing Correction Method for Genetic Association Studies Using Correlated Single Nucleotide Polymorphisms”. In: *Genetic Epidemiology* 32.4 (2008), pp. 361–369. DOI: 10.1002/gepi.20310.

[Hec99] D. Heckerman. “Learning in graphical models”. In: MIT Press, 1999. Chap. A Tutorial on Learning with Bayesian Networks.

[HGC95] D. Heckerman, D. Geiger, and D. Chickering. “Learning Bayesian networks: The Combination of Knowledge and Statistical Data”. In: *Machine Learning* 20 (1995), pp. 197–243.

[Hir+02] J. N. Hirschhorn, K. Lohmueller, E. Byrne, and K. Hirschhorn. “A comprehensive review of genetic association studies”. In: *Genetics in Medicine* 4.2 (2002), pp. 45–61.

[HMC97] D. Heckermann, C. Meek, and G. Cooper. *A Bayesian Aproach to Causal Discovery.* Technical Report, MSR-TR-97-05. 1997.

[Hoe+99] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. “Bayesian Model Averaging: A Tutorial”. In: *Statistical Science* 14.4 (1999), pp. 382–417.

[HS97] M. Hall and L. Smith. “Feature Subset Selection: A Correlation Based Filter Approach”. In: *International Conference on Neural Information Processing and Intelligent Information Systems (1997).* 1997, pp. 855–858.

[HTF01] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data minig inference and prediction.* Springer-Verlag, 2001.

[Inz+00] I. Inza, P. Larrañaga, R. Etxeberria, and B. Sierra. “Feature subset selection by Bayesian network-based optimization”. In: *Artificial intelligence* 123.1 (2000), pp. 157–184.

[KJ97] R. Kohavi and G. H. John. “Wrappers for feature subset selection”. In: *Artificial Intelligence* 97 (1997), pp. 273–324.

[Kon+98] P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald. *A comparison of non-informative priors for Bayesian networks.* 1998.

[Kon01] I. Kononenko. “Machine learning for medical diagnosis: history, state of the art and perspective”. In: *Artificial Intelligence in Medicine* 23.1 (2001), pp. 89–109.

[KS96] D. Koller and M. Sahami. *Toward optimal feature selection.* Technical Report SIDL-WP-1996-0032. 1996.

[Lew02] C. Lewis. “Genetic association studies: design, analysis and interpretation”. In: *Briefings in bioinformatics* 3.2 (2002), pp. 146–153.

[LKZ00] N. Lavrac, E. Keravnou, and B. Zupan. *Intelligent Data Analysis in Medicine.* Technical Report. 2000.

[LT06] P. Lisboa and A. Taktak. “The use of artificial neural networks in decision support in cancer: A systematic review”. In: *Neural Networks* 19.4 (2006), pp. 408–415.

[Lun+00] D. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. “WinBUGS A Bayesian modelling framework: Concepts, structure, and extensibility”. In: *Statistics and Computing* 10.4 (2000), pp. 325–337.

[LY05] H. Liu and L. Yu. “Toward integrating feature selection algorithms for classification and clustering”. In: *Knowledge and Data Engineering, IEEE Transactions on* 17.4 (2005), pp. 491–502.

[Mad+96] D. Madigan, S. A. Andersson, M. Perlman, and C. T. Volinsky. “Bayesian model averaging and model selection for Markov equivalence classes of acyclic digraphs”. In: *Comm.Statist. Theory Methods* 25 (1996), pp. 2493–2520.

[Man+09] T. Manolio et al. “Finding the missing heritability of complex diseases”. In: *Nature* 461.7265 (2009), pp. 747–753.

[Mik+95] S. Miksch, W. Horn, C. Popow, and F. Paky. “Therapy Planning Using Qualitative Trend Descriptions”. In: *Proceedings of 5th Conference on Artificial Intelligence in Medicine Europe (AIME 95), June 25 - 28, 1995, Pavia, Italy.* 1995, pp. 197–208.

[Mit07] A. Mittal. *Bayesian Network Technologies: Applications and Graphical Models: Applications and Graphical Models.* IGI Global, 2007.

[MSL12] R. Mourad, C. Sinoquet, and P. Leray. “Probabilistic graphical models for genetic association studies”. In: *Briefings in bioinformatics* 13.1 (2012), pp. 20–33.

[NC06] A. Neath and J. Cavanaugh. “A Bayesian approach to the multiple comparisons problem”. In: *Journal of Data Science* 4.2 (2006), pp. 131–146.

[Nea04] R. Neapolitan. *Learning bayesian networks.* Vol. 38. Prentice Hall Upper Saddle River, 2004.

[Nob09] W. Noble. “How does multiple testing correction work?” In: *Nature biotechnology* 27.12 (2009), pp. 1135–1137.

[OLD01] A. Onisko, P. Lucas, and M. Druzdzel. “Comparison of Rule-Based and Bayesian Network Approaches in Medical Diagnostic Systems”. In: *S. Quaglini, P. Barahona, and S. Andreassen (Eds.): AIME 2001, Lecture Notes in Artificial Intelligence 2101.* 2001, pp. 283–292.

[PCB06] K. Preacher, P. Curran, and D. Bauer. “Computational Tools for Probing Interactions in Multiple Linear Regression, Multilevel Modeling, and Latent Curve Analysis”. In: *Journal of Educational and Behavioral Statistics* 31.4 (2006), pp. 437–448.

[PCB13] R. Patnala, J. Clements, and J. Batra. “Candidate gene association studies: a comprehensive guide to useful in silico tools”. In: *BMC genetics* 14.1 (2013), p. 39.

[Pea00] J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 2000.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, 1988.

[Pen+07] J. Pena, R. Nilsson, J. Bjrkegren, and J. Tegnér. “Towards Scalable and Data Efficient Learning of Markov Boundaries”. In: *International Journal of Approximate Reasoning* 45 (2007), pp. 211–232.

[PS12] B. Pei and D. Shin. “Reconstruction of biological networks by incorporating prior knowledge into Bayesian network models”. In: *Journal of Computational Biology* 19.12 (2012), pp. 1324–1334.

[RN10] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall, 2010.

[SB09] M. Stephens and D. Balding. “Bayesian statistical methods for genetic association studies”. In: *Nature Review Genetics* 10(10) (2009), pp. 681–690.

[SGS01] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT Press, 2001.

[SP78] P. Szolovits and S. Pauker. “Categorical and Probabilistic Reasoning in Medical Diagnosis”. In: *Artificial Intelligence* 11.1–2 (1978), pp. 115–144.

[Ste09] D. A. Stephens. “Complexity in Systems Level Biology and Genetics: Statistical Perspectives”. In: *Encyclopedia of Complexity and Systems Science.* 2009, pp. 1226–1244.

[Sto+04] J. Stoehlmacher, D. Park, W. Zhang, D. Yang, S. Groshen, S. Zahedy, and H.-J. Lenz. "A multivariate analysis of genomic polymorphisms: prediction of clinical outcome to 5-FU/oxaliplatin combination chemotherapy in refractory colorectal cancer". In: *British Journal of Cancer* 91.2 (2004), pp. 344–354. DOI: 10.1038/sj.bjc.6601975.

[Sto02] J. Storey. "A direct approach to false discovery rates". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.3 (2002), pp. 479–498.

[TA03] I. Tsamardinos and C. Aliferis. "Towards Principled Feature Selection: Relevancy, Filters, and Wrappers". In: *Proc. of the Ninth International Workshop on Artificial Intelligence and Statistics*. 2003, pp. 334–342.

[UC14] G. Upton and I. Cook. *A Dictionary of Statistics 2 rev. ed.* Oxford university press, 2014.

[WB10] S. Wong and C. Butz. "A Comparative Study of Noncontextual and Contextual Dependencies". In: *Foundations of Intelligent Systems*. Vol. 1932. Lecture Notes in Computer Science. 2010, pp. 247–255. DOI: 10.1007/3-540-39963-1_26.

[Wei+78] S. Weiss, C. Kulikowski, S. Amarel, and A. Safir. "A Model-Based Method for Computer-Aided Medical Decision-Making". In: *Artificial Intelligence* 11.1–2 (1978), pp. 145–172.