HYPOTHESIS TESTING AND FEATURE SELECTION IN SEMI-SUPERVISED DATA

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2015

By
Konstantinos Sechidis
School of Computer Science

Contents

Word Count: 45434

List of Tables

List of Figures

Abbreviations

CMB	Candidate Markov Blanket
CMI	Conditional Mutual Information
FNR	False Negative Rate
FPR	False Positive Rate
IAMB	Incremental Association Markov BLanket
JMI	Joint Mutual Information
KL	Kullback-Leibler
MAR-C	Missing At Random Class dependent
MAR-F	Missing At Random Feature dependent
MB	Markov Blanket
MCAR	Missing Completely At Random
MIM	Mutual Information Maximization
MNAR	Missing Not At Random
MRMR	Max-Relevance Min-Redundancy
PU	Positive-Unlabelled
SCAR	Selected Completely At Random
SS	Semi-Supervised

Notation

α	Probability of committing a type I error
β	Probability of committing a type II error
$\mathbf{X}$	feature-set/joint random variable: $\mathbf{X} = \{X_1 ... X_d\}$
$\mathbf{x}$	a realization of $\mathbf{X}$: $\mathbf{x} = [x_1 ... x_d]$
X	random variable describing a single feature
x	a realization of X
$\mathbf{X}_\theta$	feature-set/joint random variable of the selected features
$\mathbf{X}_{\widetilde{\theta}}$	feature-set/joint random variable of the unselected features
$\mathbf{X}_{\mathrm{MB}}$	feature-set/joint random variable of the Markov blanket of Y
$\mathbf{X}_{\mathrm{CMB}}$	feature-set/joint random variable of the candidate Markov blanket of Y
S	random variable that describes the labelling mechanism
$s = 1$	a labelled example
$s = 0$	an unlabelled example
Y	class variable, corrupted by missingness (partially observed)
$y = 1$	a positive example
$y = 0$	a negative example
$\widetilde{Y}_m$	surrogate variable, fully observed, taking value m when Y has missing values.
$\widetilde{y}_m = 1$	a positively labelled example ($y = 1, s = 1$)
$\widetilde{y}_m = 0$	a negatively labelled example ($y = 0, s = 1$)
$\widetilde{y}_m = m$	an unlabelled example ($s = 0$)
$\widetilde{Y}_0$	surrogate variable, fully observed, taking value 0 when Y has missing values.
$\widetilde{y}_0 = 1$	a positively labelled example ($y = 1, s = 1$)
$\widetilde{y}_0 = 0$	a negatively labelled or an unlabelled example
$\widetilde{Y}_1$	surrogate variable, fully observed, taking value 1 when Y has missing values.
$\widetilde{y}_1 = 0$	a negatively labelled example ($y = 0, s = 1$)
$\widetilde{y}_1 = 1$	a positively labelled or an unlabelled example

Abstract

Hypothesis Testing and Feature Selection in Semi-Supervised Data

Konstantinos Sechidis

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy, 2015

A characteristic of most real world problems is that collecting unlabelled examples is easier and cheaper than collecting labelled ones. As a result, learning from partially labelled data is a crucial and demanding area of machine learning, and extending techniques from fully to partially supervised scenarios is a challenging problem. Our work focuses on two types of partially labelled data that can occur in binary problems: semi-supervised data, where the labelled set contains both positive and negative examples, and positive-unlabelled data, a more restricted version of partial supervision where the labelled set consists of only positive examples. In both settings, it is very important to explore a large number of features in order to derive useful and interpretable information about our classification task, and select a subset of features that contains most of the useful information.

In this thesis, we address three fundamental and tightly coupled questions concerning feature selection in partially labelled data; all three relate to the highly controversial issue of when does additional unlabelled data improve performance in partially labelled learning environments and when does not. The first question is what are the properties of statistical hypothesis testing in such data? Second, given the widespread criticism of significance testing, what can we do in terms of effect size estimation, that is, quantification of how strong the dependency between feature X and the partially observed label Y? Finally, in the context of feature selection, how well can features be ranked by estimated measures, when

the population values are unknown? The answers to these questions provide a comprehensive picture of feature selection in partially labelled data. Interesting applications include for estimation of mutual information quantities, structure learning in Bayesian networks, and investigation of how human-provided prior knowledge can overcome the restrictions of partial labelling.

One direct contribution of our work is to enable valid statistical hypothesis testing and estimation in positive-unlabelled data. Focusing on a generalised likelihood ratio test and on estimating mutual information, we provide five key contributions. (1) We prove that assuming all unlabelled examples are negative cases is sufficient for independence testing, but not for power analysis activities. (2) We suggest a new methodology that compensates this and enables power analysis, allowing sample size determination for observing an effect with a desired power by incorporating users prior knowledge over the prevalence of positive examples. (3) We show a new capability, supervision determination, which can determine a-priori the number of labelled examples the user must collect before being able to observe a desired statistical effect. (4) We derive an estimator of the mutual information in positive-unlabelled data, and its asymptotic distribution. (5) Finally, we show how to rank features with and without prior knowledge. Also we derive extensions of these results to semi-supervised data.

In another extension, we investigate how we can use our results for Markov blanket discovery in partially labelled data. While there are many different algorithms for deriving the Markov blanket of fully supervised nodes, the partially labelled problem is far more challenging, and there is a lack of principled approaches in the literature. Our work constitutes a generalization of the conditional tests of independence for partially labelled binary target variables, which can handle the two main partially labelled scenarios: positive-unlabelled and semi-supervised. The result is a significantly deeper understanding of how to control false negative errors in Markov Blanket discovery procedures and how unlabelled data can help.

Finally, we present how our results can be used for information theoretic feature selection in partially labelled data. Our work extends naturally feature selection criteria suggested for fully-supervised data, to partially labelled scenarios. These criteria can capture both the relevancy and redundancy of the features and can be used for semi-supervised and positive-unlabelled data.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

The Author

Konstantinos Sechidis holds an MSc in Communications and Signal Processing with distinction from the Imperial College London (UK), and another MSc degree with distinction (first in the course) from Aristotle in Information Systems from the University of Thessaloniki (Greece). He completed his undergraduate studies in Electrical and Computer Engineering in the Aristotle University of Thessaloniki, where he obtained a 5-year diploma degree with a specialization in telecommunications.

Since 2011 he is a doctoral student in the EPSRC funded Center of Doctoral Training in Computer Science, and in 2012 he started his PhD in the Machine Learning and Optimization group having Dr. Gavin Brown as his main supervisor and Dr. Robert Stevens as his co-supervisor. His main research area is feature selection in partially labelled data (Sechidis and Brown, 2015a,b; Sechidis et al., 2014a), which is the focus of this thesis, while he also published a work in feature selection in multi-label data (Sechidis et al., 2014b) and submitted a work in the field of distributed feature selection (Bolón-Canedo et al., 2015).

Before starting his PhD, Konstantinos published two works: the first one in the stratification of multi-label data (Sechidis et al., 2011) and the second in photo annotation through multi-label algorithms (Spyromitros-Xioufis et al., 2011).

Konstantinos has received various awards and fellowships. In 2014 his paper "Statistical Hypothesis Testing in Positive-unlabelled Data" won the best student paper award in the European Conference of Machine Learning and the runner-up in the Research Symposium of Computer Science Deparment in University of Manchester. From 2011 to 2013 he was a Fellow of Propondis Foundation. Konstantinos has participated in the following conferences: ICML 2012, ECML 2014, ECML 2015, S+SSPR 2014, and schools: Machine Learning Summer School (MLSS) 2012 , Gaussian Processes Winter School 2014 and Machine Learning for Personalized Medicine Summer School 2015.

Acknowledgements

Firstly, I would like to express my deep gratitude to my supervisor Dr. Gavin Brown. I was extremely fortunate to receive his guidance, his inspiration and his endless support all these years. He taught me how to do research, how to become a good teacher and he introduced me to the world of academia. He is a great example for me.

My sincere thanks to my co-supervisor Dr. Robert Stevens, his encouragement was invaluable. Also, I am grateful to Dr. Mikel Lujan and Dr. Borja Calvo for their support throughout these years.

I would like to thank all members of staff at the University of Manchester for providing me with a perfect working environment all these years. Special thanks to the Centre for Doctoral Training (CDT) in Computer Science, which is funded by an Engineering and Physical Sciences Research Council (EPSRC) grant [EP/I028099/1]. A great thanks to all my CDT fellows: Ed, Aitor, Colin, Dave, James, Michele, Rob, Fardeen, Tom, Nico, Ilia, Dimitris and Matt, and to the organisers Dr. Barry Cheetham, Dr. Jonathan Shapiro and Dr. Steve Furber.

I am grateful to all the people of the Public Welfare Foundation "Propondis", which partially funded my PhD. Special thanks goes to the director Yannis Baveas for honouring me with this competitive scholarship, as well as for his constant moral support.

I would like to thank the people of my research group: Nikos, Sarah, Henry, Adam, Ming-Jie, Nara, Richard, Veronica, Diego, Ketzi, Anna and Michalis. Their daily support, friendship and our discussions helped me to move forward. I would like to thank all the members of the MLO lab, specially Joe, Jon, Andrew, Fabio, Vassilis, Richard, Fan-Lin and Ubai. Special thanks to Çiğdem for providing help to get a better insight in statistics, and to my maths teacher John Chalidis for his endless encouragement, support and advice. A big thank to all my friends, old and new, who have made the last few years so enjoyable.

Last but not least, I would like to acknowledge from the bottom of my heart the endless love and support of my family, specially: my parents, Viki and Aris; my sister's family, Aris, Konstantinos, Natasa, Antonis and the baby on its way; and my partner, Idoia. They all kept me going and without them I would not be capable of anything.

Ευχαριστώ πολύ – Eskerrik asko

In loving memory of my uncle Lazaros.

Chapter 1

Introduction

Many real world applications generate huge amounts of data which benefit from the application of advanced machine learning techniques. There are two main challenges associated with these so called *big data*: the large amount of examples and the large number of dimensions. Following the terminology introduced by Zhai et al. (2014), we refer to the first concept as *big instance size*, and to the latter one as *big dimensionality*. The challenge of big instance size is intimately tied to the problems of partially labelled learning as we will explain in the next section. Big dimensionality, on the other hand, can be tackled by ignoring the irrelevant and redundant information that the data sets contain.

An important research direction is to transfer techniques and methodology from supervised learning to *partially labelled* situations. An easy solution is simply to *ignore* the unlabelled data. That said, whether unlabelled data in fact may help, is an interesting and controversial question (Singh et al., 2009). This thesis explores the feature selection challenge in the context of partially labelled data.

1.1 Exploring the Instances: Partial Labelling

Nowadays the collection of large amounts of data is a cheap and easy procedure. Terabytes of data are collected every second by simply storing information from diverse sources such as Twitter messages or images received from telescopes. While collecting these data is an automatic procedure, the procedure of labelling them is expensive both in terms of time and in terms of resources. So it is crucial to explore techniques that can handle partially labelled data.

Our work focuses on two types of partially labelled data that can occur in

binary problems: *semi-supervised data*, where the labelled set contains *both* positive and negative examples, and *positive-unlabelled data*, a more restricted version of partial supervision where *only* positive examples are labelled. Under the traditional *semi-supervised* framework (Smith and Elkan, 2007), the labelled set is assumed to be an unbiased sample from the population, which makes the analysis relatively straightforward. But when this assumption does not hold, a spectrum of new challenges arises from the different possible labelling mechanisms. An example is learning from *positive-unlabelled* data, a special case of partially labelled learning, where we have a small number of examples from the positive class, and a large number of unlabelled examples which could be either positive *or* negative. This type of problem is common in text mining, bioinformatics, and computer vision. For example, a typical application is text classification. Given a number of query documents belonging to a particular class (e.g. academic articles about machine learning), plus a corpus of unlabelled documents, the task is to classify new documents as relevant to the query or not.

In order to tackle the challenges of partial labelling in a principled framework we must take into account the probabilistic *mechanism* behind the reason why the labels are missing. An appropriate 'language' for representing probabilistic situations is that probabilistic of graphical models. However, using graphical models in the presence of missing data is not straightforward. We make use of an alternative graphical representation, a recent contribution by (Mohan et al., 2013), called *missingness graphs* (m-graphs). These graphs encode the dependencies between the mechanisms that are responsible for the missing information and the measured variables in our dataset. Figure 1.1 shows an example of how a missingness graph can describe a partially labelled scenario, where the labelled set is an unbiased sample from the population. Using statistics terminology, this is the scenario in which the labels are *missing completely at random* (MCAR). The use of m-graphs can benefit our analysis in many ways. Firstly, by using them we make the assumptions under the generation of partial labelling explicit. These assumptions will play a crucial role for the rest our analysis. Secondly, through these graphs we capture all available information contained in the partially labelled datasets. Finally, m-graphs are a tool for dealing with missing data in an inference-free manner, which is very important in filter feature selection approaches, which is the focus of this thesis.

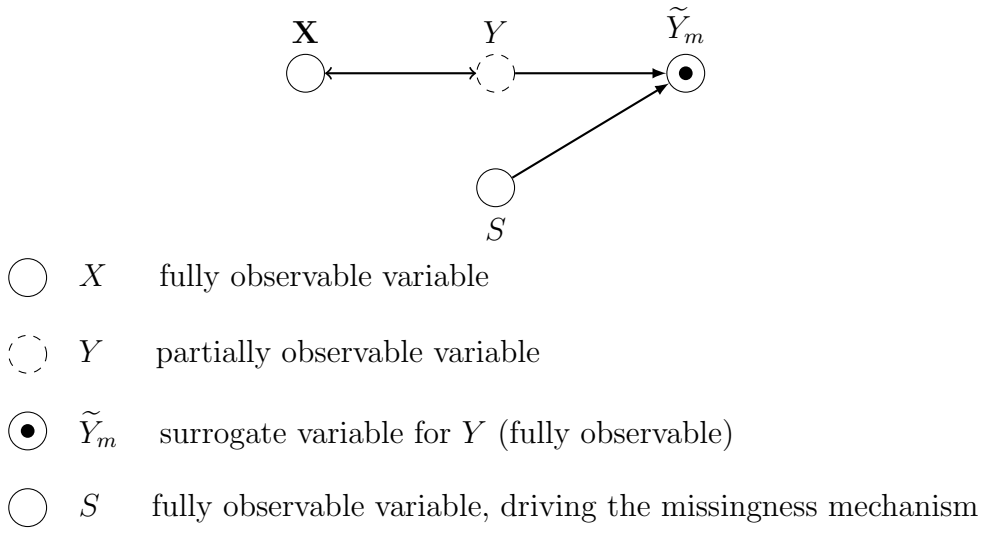


Figure 1.1: *Example of an m-graph describing partially labelled data where solid circles ○ represent fully observed variables; dashed circles ◌ represent variables with missing values Y and the hollow circle with the dot ⊙ represents the fully observed proxy variable $\widetilde{Y}_m$, which takes the values of Y for the labelled examples and the token m for the examples with missing labels.*

1.2 Exploring the Dimensions: Feature Selection

Given a prediction tasks, features can be categorised in three ways: relevant, irrelevant and redundant. We need to identify and select the relevant features, since they contain most of the useful information and to do so we use measures of dependence. Our work focuses on *information theoretic* measures, and we tackle the problem of *feature selection* (Guyon et al., 2006), focusing on information theoretic solutions (Brown et al., 2012).

However, the term feature selection is associated with *three intimately related questions*, often conflated, and as we will see, it is beneficial to disentangle them. The three questions regard the *testing*, *estimation*, and *ranking* of features, in relation to a class label. These are the three main applications of the measures of dependance (Reimherr and Nicolae, 2013)[1].

[1]The actual wording followed by Reimherr and Nicolae (2013) is *detection* and *quantification*, instead of testing and estimation. We prefer the latter one, which is the wording followed by Lehmann (1966).

Question 1 – Testing: *"Is feature X significantly correlated to label Y?"*

Question 2 – Estimation: *"How strong is the dependency between X and Y?"*

Question 3 – Ranking: *"Using a finite sample of data, can we recover a ranking of features that will be close to what we would obtain if we had access to the full data distribution?"*

A valid answer to Q1 would be simply 'yes' or 'no', using standard hypothesis testing methodology. Following this, Q2 exists because in many scenarios, answering just yes/no is not enough – fairly arbitrary decision thresholds can promote bad research practices[2]. A valid answer to Q2 is a real-valued point estimate (or an interval estimate) of the true dependency, also known as *effect size* in the statistical literature. A valid answer to Q3 depends on what we mean by being "close" to the true ranking — we will cover this in more detail later. Finally, it is important to note that a good answer to Q2 will naturally lead us to a good answer for Q3, *but not the reverse*, i.e. answering Q3 well does not solve Q2. For example, Recursive Feature Elimination with an SVM (Guyon et al., 2006, Chapter 5) can produce a good ranking of features, but does not produce an explicit measure of dependency for individual features. Our goal is to understand the dynamics of feature selection in these scenarios, with the main issue being *how to make best use of all the available information in both labelled and unlabelled data.*

1.3 Motivating example — why is a label missing?

When confronted with a partially labelled dataset, we can adopt several different assumptions regarding the labelling procedure. Exploring the mechanism which underlies a missing label is very important for the rest of our analysis, since it will determine the way we should use the unlabelled examples for testing, estimation and ranking activities. In this section we will motivate the different assumptions about the labelling of the data with an illustrative example.

[2]Such as the suppression of "negative" results, where $p = 0.051$ is "not significant", but $p = 0.049$ is for some reason, "significant", described by many authors e.g. Wainer and Robinson (2003).

Let us assume that we are working for an advertising company that wants to explore the profile of people that like watching the TV series "Game of Thrones". Particularly, we want to see if there is an association between the professional status (employed/unemployed/student) of a person and whether they like the series or not. To find out if someone likes watching the series, the company calls the person and asks a number of questions. Let's say that we have the phone numbers of 2000 people, and we call all of them. In any real world scenario we would only manage to get answers from a subset of people, and thus only in this fraction of the 2000 people we will be able to justify if they like the series or not. The remainder of the people are regarded as *missing data*, but can be missing for various reasons, articulated below.

Missing Completely At Random (MCAR) When the missingness mechanism is an entirely random process – independent of either features or class label. This can happen, for example, if we did not get a response due to a random event not associated with our survey (e.g. the telephone line was broken, or engaged)

Missing At Random, Feature-dependent (MAR-F) When the missingness mechanism depends directly only on the feature values. For example, if we phone at a mid-morning time, we might expect to have more responses from those unemployed or students, and fewer responses from people employed.

Missing At Random, Class-dependent (MAR-C) When the missingness mechanism depends directly only on the class values. In our example, if our telephone survey happens exactly at the time when the series is broadcast, we are unlikely to get responses from people who enjoy watching it.

Missing Not At Random (MNAR) The most complex situation is when the missingness depends on both the features *and* class values. This might happen for example if we change our problem, to whether someone likes a TV show broadcast at 11am – if we phone at exactly that time, we are unlikely to get responses from either those who like the show or those employed.

In our work, we classify the different kinds of missing labels in the these four scenarios, which are different from the traditional classification in statistics

(MCAR/MAR/MNAR) (Little and Rubin, 2002). In Section 3.3 we connect our classification with the established literature in statistics and machine learning.

Let us now imagine, that we do not make a phone call, but instead a postal survey, asking people to mail back what they like/don't like about the TV show. It may well be the case that *only* people who actually like the show make the effort to reply, and then only a fraction of them, hence we will have responders *only from the positive class*. This is known as *positive-unlabelled data*. In this case, since we observe data only from one class, the missing mechanism depends on the label. In other words, the probability of labelling a negative example in the positive-unlabelled case is zero. Table 1.1 summarises the different partially labelled environments and the different labelling assumptions that generate them.

	Labelled examples		Unlabelled examples	Possible missing data mechanisms
	Positive	Negative		
Fully Supervised	✓	✓	✗	none
Positive-unlabelled (PU)	✓	✗	✓	MAR-C MNAR
Semi-supervised (SS)	✓	✓	✓	MCAR MAR-F MAR-C MNAR

Table 1.1: *Fully supervised and different types of partially labelled data.*

1.4 Research Questions

The literature around learning from partially labelled data is rich (Chapelle et al., 2006), and the same holds for feature selection activities in fully-supervised data (Brown et al., 2012). But there is a lack of principled approaches that tackle both challenges together. With our work we explore these challenges by decomposing the feature selection problem and exploring how testing, estimation and ranking can be performed in partially labelled data, and with an inference-free manner. These three activities generate the three questions we will try to answer in the theoretical part of our work. The first question is *"Can we test independence despite the partial supervision and control the possible errors over our statistical*

decision?" Then we answer the question *"Can we estimate the degree of dependence despite the partial supervision?"* While the last question is *"Can we rank the features according to the dependence in such a way that we will be close to the true ranking despite the partial supervision?"*

Answering these three questions is very useful in its own sake and enables extensions of our theoretical findings in a lot of different applications in machine learning. First, we focus on the experimental design domain, and answer a question that many practitioners working with partially labelled data are interested in: *"How many examples do we need to collect, and how many of them should be labelled in order to control the possible errors of our statistical decision over independence?"* The next application comes from the structure learning domain. We explore how we can learn Markov blankets around partially labelled targets in a model-independent/inference-free manner. The importance of Markov blanket discovery algorithms is two-fold: they constitute the main building block in constraint-based structure learning of Bayesian network algorithms and as a technique to derive the optimal set of features in filter feature selection approaches. The question we answer is *"How can we use the conditional tests of independence in order to derive a Markov blanket discovery algorithm for partially labelled data?"* In the final domain, we explore the information theoretic feature selection, where we answer the question *"Can we construct feature selection algorithms that take into account the relevancy and the redundancy despite the partial supervision?"*

1.5 Contribution of this Thesis

This thesis focuses on information theoretic feature selection in partially labelled datasets, and on how we can use the tool of m-graphs to capture all the available information that the data contain, and use it in a model-independent/inference-free manner. A summary of the contributions is provided here, while we give more detailed description of them in the conclusions chapter (Chapter 10).

- We derive different ways to test independence in partially labelled data in an inference-free manner, by using proxy variables under different missingness scenarios (Chapter 4).

- We suggest different ways to take into account the unlabelled examples in

order to produce consistent estimators for the mutual information and try to reduce its bias/variance. Furthermore, we suggest ways to overcome the problems caused by the partial supervision, through incorporating prior knowledge (Chapter 5).

- We suggest ways to rank the features according to their relevance with the partially observed target variable, by exploring how to use the unlabelled examples and our prior knowledge in an efficient way (Chapter 5).

- In positive-unlabelled data we offer a methodology to determine the sample size and/or the labelled-set size by incorporating prior knowledge (Chapter 7), and we extend this methodology to semi-supervised data.

- We suggest algorithms for Markov blanket discovery in partially labelled data, by incorporating prior knowledge in a completely inference-free manner (Chapter 8).

- We derive different feature selection criteria, that generalise the criteria used in fully-supervised data and can capture both the relevancy and the redundancy of the features (Chapter 9).

1.6 Structure of this Thesis

In Chapter 2, we present the background material on information theoretic feature selection, by exploring the three activities that are closely related with it: testing, estimation and ranking. Furthermore, we present a generalised framework that unifies many different feature selection criteria, while we also give some background material on Markov blanket discovery algorithms.

Chapter 3 reviews the literature related to the mechanisms behind the generation of partially labelled data. Furthermore, we present the main tools to analyse each scenario, the m-graph, and we connect it to the established literature.

Our first theoretical contribution is presented in Chapter 4, where we explore how we can test independence by using surrogate proxy variables instead of the unobservable target. We compare the different approaches in terms of their validity and informedness. These two terms, which will be explained in details later, are related with the two possible statistical errors: false positive and false negative.

Then we focus on estimating mutual information in partially labelled data, and Chapter 5 presents different ways to derive consistent estimators of the mutual information under the different missingness scenarios. By incorporating the unlabelled examples in a clever way we can derive consistent estimators despite the partial supervision, while on the missingnes scenarios that this is not feasible, we explore different ways to incorporate prior knowledge to overcome the limitation caused by the partial supervision.

The last theoretical contribution is presented in Chapter 6, where we focus on how to rank the features when the target is partially observed. Again, the usage of unlabelled examples is crucial to obtain rankings equivalent with the true ranking. Furthermore, we explore efficient ways to use our prior knowledge.

Chapter 7 presents an extension of our results in the area of experimental design. Particularly, we suggest how we can use our analysis for testing (presented in Chapter 4) in order to decide the minimum number of examples we need to collect to have an informed decision when testing independence. Our analysis allows us the novel capability of supervision determination, where we can decide the minimum number of examples we need to label in order to have an informed decision.

Chapter 8 presents another application of our work. We suggest an algorithm for Markov blanket discovery (which combines the results of our theoretical analysis of Chapters 4 and 6), when we have positive-unlabelled data. Then, we move to semi-supervised data, and we show how to incorporate prior knowledge and what kind of knowledge we need. At the end we show that our approach outperforms the previously suggested ones in the semi-supervised environment, when the class prior change.

Chapter 9 presents an application in the area of feature selection in partially labelled data. We explore how our theoretical results from Chapters 5 and 6, can be used in order to select features that take into account both the relevancy and the redundancy.

Chapter 10 reviews the results of this thesis, and provides a guide for practitioners on hypothesis testing, effect size estimation and feature selection in semi-supervised and positive-unlabelled data. We also suggest a number of future directions in the areas of model selection and evaluation.

1.7 Publications

The work presented in this thesis has resulted into two publications, with one further journal paper under review:

Sechidis et al. (2014a): Konstantinos Sechidis, Borja Calvo, and Gavin Brown. Statistical hypothesis testing in positive unlabelled data. In *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 66–81. Springer Berlin Heidelberg, 2014a

Sechidis and Brown (2015a): Konstantinos Sechidis and Gavin Brown. Markov blanket discovery in positive-unlabelled and semi-supervised data. In *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 351–366. Springer Berlin Heidelberg, 2015a

Sechidis and Brown (2015b): Konstantinos Sechidis and Gavin Brown. Hypothesis testing and feature selection in semi-supervised data. *Under review*, 2015b

Chapters 4, 5, 6 and 9 are expanded versions of Sechidis and Brown (2015b). Chapter 7 is an updated version of Sechidis et al. (2014a), while Chapter 8 was presented in Sechidis and Brown (2015a).

Other published/under-review work not relevant to this work

In collaboration with other members of the Machine Learning and Optimization group we published one work and one other is under review:

Sechidis et al. (2014b): Konstantinos Sechidis, Nikolaos Nikolaou, and Gavin Brown. Information theoretic feature selection in multi-label data through composite likelihood. In *Structural, Syntactic, and Statistical Pattern Recognition (SSPR)*, pages 143–152. Springer Berlin Heidelberg, 2014b

Bolón-Canedo et al. (2015): Veronica Bolón-Canedo, Konstantinos Sechidis, Noelia Sánchez-Maroño, Amparo Alonso-Betanzos, and Gavin Brown. Some guidelines for distributed feature ranking. *Under review*, 2015

Chapter 2

Literature Review: Feature Selection in Categorical Data

This chapter will give an overview of testing and estimation using information theoretic quantities, and how we can select useful features via testing and estimation activities. Our work focuses on filter methods and – more specifically – on information theoretic feature selection approaches (Guyon et al., 2006, Chapter 6). Under this approach the features X are selected by quantifying the information that they share with the class variable Y.

In information theoretic feature selection the main challenge is to estimate the mutual information, one of the most common measures of dependence used in machine learning (more details in Section 2.1). Answering whether two random variables are independent or not requires us to threshold the value of the estimated mutual information. To derive such a threshold we will use the asymptotic distribution of the estimator and a hypothesis testing procedure, which takes the form of an independence test (more details in Section 2.2).

By ranking the estimated mutual information values for the different features, we can select the most informative features, according to their relevancy with the class variable and their redundancy with the other selected features (more details in Section 2.3). A different, but closely related, approach to select features uses conditional tests of independence to derive the most informative features (more details in Section 2.4).

2.1 Estimating Effect Sizes

In our work we will focus on two ways of measuring the common information between random variables, which are widely used in both machine learning and statistics: the mutual information and the squared-loss mutual information.

2.1.1 Mutual Information

In fully supervised data, the features X are ranked and the ones selected are those that have the highest *mutual information* with the class label Y. The population value of the mutual information (or Shannon's mutual information) between two categorical random variables is defined as

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}, \tag{2.1}$$

where $p(x,y) = \Pr\{X = x, Y = y\}$ is the probability mass function of the joint distribution when the random variable X takes on the value x from its alphabet $\mathcal{X}$ and Y takes on $y \in \mathcal{Y}$, while $p(x) = \Pr\{X = x\}$ and $p(y) = \Pr\{Y = y\}$ are the probability mass functions of the marginal distributions. The mutual information is the *Kullback-Leibler divergence* between the joint distribution $p(x,y)$ and the product of the marginals $p(x)p(y)$, i.e. $I(X;Y) = D_{KL}(p(x,y)||p(x)p(y))$ (Cover and Thomas, 2006).

When we have sample data, $\{(x^i, y^i)|i = 1, ..., N\}$, we can derive a point estimate of the mutual information through the *maximum-likelihood* estimates of the probabilities

$$\widehat{I}(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \widehat{p}(x,y)\ln\frac{\widehat{p}(x,y)}{\widehat{p}(x)\widehat{p}(y)}. \tag{2.2}$$

This estimate, sometimes known as a "plug-in" estimate of the mutual information (Paninski, 2003), can be seen as a measure of effect size when we want to quantify the dependency between random variables, and has several nice properties. Firstly, it is a non-negative quantity which takes its minimum zero value when the random variables are independent. Furthermore, it can be associated with both upper and lower bounds on the Bayes error (Fano, 1961; Hellman and Raviv, 1970; Zhao et al., 2013). Brown et al. (2012) present an extensive discussion of this in the context of feature selection, including various heuristics

which provide approximations for high dimensional data, resulting in a unifying theoretical framework derived from a simple probabilistic model.

However, something Brown et al. (2012) did not consider is the *distribution* of the estimator in equation (2.2). Together with point estimates, it is a good practice to give an *interval* estimate, a range of possible values that the mutual information can take.

The maximum likelihood estimator of the mutual information is known to asymptotically follow a normal distribution (Brillinger, 2004; Kojadinovic, 2005).

Theorem 2.1 (Sampling Distribution of the Mutual Information)**.**
The maximum likelihood estimator of the mutual information is asymptotically normally distributed with the following parameters:

$$\widehat{I}(X;Y) \sim \mathcal{N}\left(\mu, \sigma^2\right) \textit{ with } \begin{cases} \mu = I(X;Y) \\ \sigma^2 = \frac{1}{N}\left(\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\left(\ln\frac{p(x,y)}{p(x)p(y)}\right)^2 - I(X;Y)^2\right) \end{cases}$$

The proof can be found in appendix A.

However, this holds tightly only for strong dependencies; that is, for relatively large values of the mutual information. For weakly dependent variables a scaled non-central χ^2 distribution is a tighter fit (Goebel et al., 2005). By using the fact that the maximum likelihood estimator of the mutual information is essentially the likelihood ratio statistic for testing independence (more details over the link between testing and estimation in Section 2.2), we can use distribution results available for the latter to derive the distribution of the estimator (Brillinger, 2004).

Theorem 2.2 (Distribution of the Mutual Information for Small Effects)**.**
The maximum likelihood estimator of the mutual information follows a scaled non-central χ^2 distribution

$$\widehat{I}(X;Y) \sim \frac{1}{2N}\chi^2\left(\nu,\lambda\right) \textit{ with } \begin{cases} \nu = (|\mathcal{X}|-1)(|\mathcal{Y}|-1) \\ \lambda = 2NI(X;Y) \end{cases},$$

where we used the notation: $\chi^2\left(\nu,\lambda\right)$ for a non-central χ^2 distribution with ν degrees of freedom and non-centrality parameter λ.
Proof sketch can be found in appendix A.

The discussion here can be long — answering whether or not the estimator follows normal or scaled non-central χ^2 is equivalent to questioning the distribution of the likelihood ratio test statistic[1]. More details over this issue can be found in Chun and Shapiro (2009). As a rule of thumb we can say that the closer we are to the independence assumption the non-central χ^2 approximation is better. On the other hand, when we have strong effects in terms of mutual information, the normal approximation is better.

2.1.2 Squared-loss Mutual Information

It is important to note that estimation of (2.1), due to the logarithm operation, can be computationally expensive, and also subject to numerical instabilities. One way to avoid this is, instead of the KL-divergence, to use the *Pearson divergence* between the joint and the product of the marginals (Pearson, 1900) — this is a second-order Taylor approximation of the Shannon's mutual information. The population value of this divergence, also known as *squared-loss mutual information*, is

$$I_2(X;Y) = \frac{1}{2}\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\frac{\left(p(x,y)-p(x)p(y)\right)^2}{p(x)p(y)}. \tag{2.3}$$

Calculating squares instead of the logarithms makes the estimator computationally less expensive and more numerically stable, and turns out to have useful properties in our analysis of feature ranking. Sugiyama (2012) offers an in depth discussion on the squared loss mutual information, and its applications in machine learning including hypothesis testing and feature selection. In order to derive sampling distributions of $\widehat{I}_2(X;Y)$, we can use the same methodologies as for $\widehat{I}(X;Y)$. It can be proved that the squared loss mutual information (2.3) is a second order approximation of Shannon's mutual information. The closer we are to independence, in other words, the smaller the effects are, the better the approximation will be. Small effects are those of main interest, since when we have larger effects, differentiating between them is more trivial.

In statistics literature, two widely used measures of association between categorical variables are the ϕ coefficient and Cramer's-V coefficient (or Cramer's ϕ_c-coefficient) (Cramér, 1999, Chapter 21). It is interesting to note here that

[1]The natural relationship between testing independence and estimating the mutual information will be explored in Section 2.2.

there is a natural relationship between these coefficients and the squared loss mutual information. For example when both X and Y are binary the ϕ coefficient is used, which can be written as $\phi = \sqrt{2\widehat{I}_2(X;Y)}$. When we have categorical variables with more than two categories the Cramer's-V coefficient is used, which can be written as: $V = \sqrt{\frac{2\widehat{I}_2(X;Y)}{\min(|\mathcal{X}|-1,|\mathcal{Y}|-1)}}$. To the best of our knowledge, this is the first time that the relationship between the squared loss mutual information and ϕ/Cramer's-V coefficient is explored.

Using squared loss mutual information will turn out to be particularly important in the context of *feature ranking* in partially labelled data, more details in Chapter 6.

Answering whether two random variables are independent or not requires us to threshold the value of the estimated mutual information. To derive such a threshold, we will use the asymptotic distribution of the estimator and a hypothesis testing procedure. By following this procedure we will have an informed decision and control over the two possible errors: concluding independence, where in fact there is a dependence (a false negative, or type-II error), or the opposite, concluding dependence, where in fact there is none (a false positive, or type-I error). This is the focus of the following section.

2.2 Testing Independence

According to Guyon et al. (2006, Chapter 2) there are two complementary views of feature selection. In the *machine learning* view, the features are ranked according to a score and then the top-k features are selected (where the parameter k is predefined or can be set through a cross-validation procedure). In the *statistical view*, we select the features by hypothesis testing – this methodology will be our focus in this section.

To detect a dependency between a feature and a target variable we need a *test of independence* procedure. The three main approaches for testing a hypothesis in statistics are: the *Fisherian* (also known as significance testing) using p-values, the *Neyman-Pearson* using fixed error probabilities and the *Bayesian* using posterior error probabilities through Bayes factors. There is a long discussion over the advantages and disadvantages of each approach, but this is beyond the scope of this thesis. Berger (2003) presents these approaches and summarizes the main criticisms for each one. In a recent article, Nuzzo (2014) criticises the usage of

p-values and one of the possible alternatives that he suggests is to combine them with the usage of effect sizes and confidence intervals. In our work we follow this approach and we combine it with the *Neyman-Pearson* framework for testing independence.

In this framework, a core activity is a-priori power analysis, which is mainly used for *sample size determination* — while not so common in machine learning, this is widely used in biosciences, clinical trials, social sciences, and many more fields. It allows a researcher to determine the minimum sample size, subject to specified constraints on the type-I (i.e. false positives, which means falsely deciding dependence) and type-II errors (i.e. false negatives, which means falsely deciding independence). In our analysis of partially labelled data, the sample size will play a crucial role, since we will try to explore if and how we can benefit by adding unlabelled data to our labelled sample. Furthermore, with our work we will introduce a set of methodologies for *labelled-set size determination* in partially labelled data.

In order to detect dependencies between categorical features and a partially labelled target, we will explore two widely used tests of independence: the G-test and the X^2-test (Cressie and Read, 1989). Both of them have been used widely in machine learning, for example in structure learning of Bayesian networks (Tsamardinos and Borboudakis, 2010; Spirtes et al., 2001, Section 5.5). Furthermore, these tests are extensively used in life sciences (Samuels et al., 2012), behavioral sciences (Gruijter and Kamp, 2007) and biology (Sokal and Rohlf, 1995), and our work can be very relevant to the experimental design for partially labelled data in these domains. In the following sections we will introduce the two tests and how these tests can be used for power analysis and sample size determination.

2.2.1 G-test of Independence

The G-test is a generalised likelihood ratio test (Woolf, 1957), where the test statistic can be calculated from sample data counts arranged in a contingency table. Denoted by $o_{x,y}$, the observed count of the number of times the random variable X takes on the value x from its alphabet $\mathcal{X}$, while Y takes on the value $y \in \mathcal{Y}$. By $o_{x,\cdot}$ and $o_{\cdot,y}$ we denote the marginal counts. The estimated expected frequency of (x, y), assuming X, Y are independent, is given by $e_{x,y} =$

$\widehat{p}(x)\widehat{p}(y)N = \frac{o_{x,\cdot}o_{\cdot,y}}{N}$. The G-statistic can now be defined as:

$$G(X;Y) = 2\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} o_{x,y}\ln\frac{o_{x,y}}{e_{x,y}} = 2N\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\widehat{p}(x,y)\ln\frac{\widehat{p}(x,y)}{\widehat{p}(x)\widehat{p}(y)} = 2N\widehat{I}(X;Y). \quad (2.4)$$

From this expression we see the relationship between the G-statistic and the mutual information, so the latter can be seen as the natural unit of *effect size* for the G-test (Rosenthal, 1994). Under the null hypothesis that X and Y are statistically independent, the G-statistic is known to be asymptotically χ^2-distributed, with $\nu = (|\mathcal{X}|-1)(|\mathcal{Y}|-1)$ degrees of freedom (Agresti, 2013). For a given dataset, we calculate (2.4) and check to see whether it exceeds the critical value defined by a significance level α read from a standard statistical table giving the CDF of the χ^2-distribution. If the critical value is not exceeded, the variables are judged to be independent. An equivalent way to test the null hypothesis is by using the p-value, which is equal to $1 - F(G(X;Y))$, where F is the CDF of the χ^2-distribution. The p-value represents the probability of obtaining a test statistic equal or more extreme than the observed one, given that the null hypothesis holds. After calculating this value, we check to see whether it exceeds a significance level α. If p-value $\leq \alpha$, we reject the null hypothesis of independence.

2.2.2 X^2-test of Independence

Another popular way to test independence between categorical random variables is by using the X^2-test (Pearson, 1900). This test is closely associated with the G-test, since the X^2-statistic is the second order Taylor approximation of the G-statistic, and has the form:

$$X^2(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\frac{(o_{x,y}-e_{x,y})^2}{e_{x,y}} = N\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\frac{\left(\widehat{p}(x,y)-\widehat{p}(x)\widehat{p}(y)\right)^2}{\widehat{p}(x)\widehat{p}(y)} = 2N\widehat{I}_2(X;Y).$$

Because of this relationship between the X^2-statistic and the squared-loss mutual information, the latter one can be seen as the natural effect size for the X^2-test (Rosenthal, 1994). Again, under the null hypothesis that X and Y are statistically independent, the X^2-statistic has the same asymptotic distribution as the G-statistic. Thus, when we want to test independence, we compare the X^2-statistic to the same critical value as earlier. The various relations can be summarised in the Figure 2.1.

$$
\begin{array}{ccc}
G\ (X;Y) & = & 2N\widehat{I}\ (X;Y) \\
{\scriptstyle 2^{nd} \text{ order Taylor approximation}} \downarrow & & \downarrow {\scriptstyle 2^{nd} \text{ order Taylor approximation}} \\
X^2(X;Y) & = & 2N\widehat{I}_2(X;Y)
\end{array}
$$

Figure 2.1: *Links between mutual information, squared loss mutual information and statistical tests.*

2.2.3 Power Analysis

While the user specified significance level defines the probability of committing *type I error* (α), which is the probability that the test will falsely reject the null hypothesis, in order to explore the probability of committing a *type II error* (β), we should perform a *power analysis* (Cohen, 1988).

The *power* of a test is the probability that the test will reject the null hypothesis when the alternative hypothesis is true – or in practical machine learning terms, the probability of correctly selecting a relevant feature. This is also known as the *true positive rate*, or the probability of *not* committing a type II error. However, to do this we need a test statistic with a known distribution under the alternative hypothesis.

Historically, as McManus (1991) mentions, the first attempt to define a proper limiting distribution under the alternative hypothesis is introduced by Neyman (Neyman et al., 1965) by using the tool of *local power analysis* and *local alternatives*. This power is also known as *Pitman's limiting power* (Pitman, 1979, Chapter 7), while the local alternatives are sometimes known as *population drift* (Chun and Shapiro, 2009). The main idea behind the local power analysis and the local alternatives approach is that we assume a sequence of alternative situations which converge to the null hypothesis as the sample size increases.

Following this approach allows us to derive a simple form for the distribution of the likelihood ratio statistic under the alternative hypothesis. Although this is a strong mathematical assumption, it is valid when we want to observe small effects. Small effects are the challenging ones to observe, since when the effects are large, it is easier to discriminate between the null and the alternative (Chun and Shapiro, 2009).

Under the alternative hypothesis of dependence, it is known that the G-statistic has a large-sample *non-central* χ^2 distribution, with the same degrees

of freedom as in the null distribution. The non-centrality parameter $\lambda_{G(X;Y)}$ has the same form as the G-statistic, but with sample values replaced by population values (Agresti, 2013, Section 6.6.4). In other words, the non-centrality parameter under the alternative hypothesis is given by $\lambda_{G(X;Y)} = 2NI(X;Y)$. Thus $\lambda_{G(X;Y)}$ is a parameter, and the G-statistic is a random variable following the distribution defined by $\lambda_{G(X;Y)}$:

$$G(X;Y)\text{-statistic} \sim \chi^2\left(\nu, \lambda_{G(X;Y)}\right) \text{ with } \begin{cases} \nu = (|\mathcal{X}|-1)(|\mathcal{Y}|-1) \\ \lambda_{G(X;Y)} = 2NI(X;Y) \end{cases}. \quad (2.5)$$

It is also known that the X^2-statistic asymptotically follows a non-central χ^2 distribution under the alternative hypothesis, with a non-centrality parameter $\lambda_{X^2(X;Y)} = 2NI_2(X;Y)$ (Agresti, 2013, Section 6.6.4), and it holds that:

$$X^2(X;Y)\text{-statistic} \sim \chi^2\left(\nu, \lambda_{X^2(X;Y)}\right) \text{ with } \begin{cases} \nu = (|\mathcal{X}|-1)(|\mathcal{Y}|-1) \\ \lambda_{X^2(X;Y)} = 2NI_2(X;Y) \end{cases}. \quad (2.6)$$

Given this context, we can proceed with one of the most important tools in power analysis – sample size determination.

2.2.4 Sample Size Determination

Sample size determination is a core a-priori power analysis activity. In this prospective procedure we specify the significance level of the test (e.g. $\alpha = 0.05$), the desired power (e.g. power $= 0.99$ or the probability of committing a false negative to be $\beta = 1 - \text{power} = 0.01$) and the desired effect size described in terms of $I(X;Y)$ — from this, we can determine the minimum number of examples required to detect that effect.

The effect size (ω) was defined by Cohen (1988) as the square root of the non-centrality parameter divided by the sample size. So it turns out that the effect size of the G-test can be naturally expressed as a function of the *mutual information*, since $\omega = \sqrt{2I(X;Y)}$. In our work, for the effect sizes we followed the three levels introduced by Cohen (1988); small ($\omega = 0.10 \Leftrightarrow I(X;Y) = 0.005$), medium ($\omega = 0.30 \Leftrightarrow I(X;Y) = 0.045$) and large ($\omega = 0.50 \Leftrightarrow I(X;Y) = 0.125$). The effect size of the X^2-test can be expressed as a function of the *squared-loss mutual information*, since $\omega = \sqrt{2I_2(X;Y)}$.

2.2.5 Conditional Tests of Independence

As we will see later in this chapter, in order to derive relevant, but not redundant features, apart from an unconditional test of independence, we will also need to explore the conditional independence of X and Y given a subset of features $\mathbf{Z}$. We denote as $O_{x,y,\mathbf{z}}$ the observed count of the number of times the random variable X takes on the value x from its alphabet $\mathcal{X}$, Y takes on $y \in \mathcal{Y}$ and $\mathbf{Z}$ takes on $\mathbf{z} \in \boldsymbol{\mathcal{Z}}$, where $\mathbf{z}$ is a vector of values when we condition on more than one variable. We furthermore denote by $O_{x,.,\mathbf{z}}$, $O_{.,y,\mathbf{z}}$ and $O_{.,.,\mathbf{z}}$ the marginal counts. The estimated expected frequency of $(x, y, \mathbf{z})$, assuming X, Y are conditional independent given $\mathbf{Z}$, is given by $E_{x,y,\mathbf{z}} = \frac{O_{x,.,\mathbf{z}} O_{.,y,\mathbf{z}}}{O_{.,.,\mathbf{z}}} = \widehat{p}(x|\mathbf{z})\widehat{p}(y|\mathbf{z})O_{.,.,\mathbf{z}}$. To calculate the G-statistic we use the following formula:

$$G(X;Y|\mathbf{Z}) = 2\sum_{x,y,\mathbf{z}} O_{x,y,\mathbf{z}} \ln \frac{O_{x,y,\mathbf{z}}}{E_{x,y,\mathbf{z}}} = 2\sum_{x,y,\mathbf{z}} O_{x,y,\mathbf{z}} \ln \frac{O_{.,.,\mathbf{z}} O_{x,y,\mathbf{z}}}{O_{x,.,\mathbf{z}} O_{.,y,\mathbf{z}}} = \tag{2.7}$$

$$= 2N \sum_{x,y,\mathbf{z}} \widehat{p}(x,y,\mathbf{z}) \ln \frac{\widehat{p}(x,y|\mathbf{z})}{\widehat{p}(x|\mathbf{z})\widehat{p}(y|\mathbf{z})} = 2N\widehat{I}(X;Y|\mathbf{Z}),$$

where $\widehat{I}(X;Y|\mathbf{Z})$ is the maximum likelihood estimator of the conditional mutual information between X and Y given $\mathbf{Z}$ (Cover and Thomas, 2006).

Under the null hypothesis that X and Y are statistically independent given $\mathbf{Z}$, the G-statistic is known to be asymptotically χ^2-distributed, with $\nu = (|\mathcal{X}| - 1)(|\mathcal{Y}| - 1)|\boldsymbol{\mathcal{Z}}|$ degrees of freedom (Agresti, 2013). Knowing that and using (2.7) we can calculate the p-value as $1 - F(G)$, where F is the CDF of the χ^2-distribution and G the observed value of the G-statistic. After calculating this value, we check whether it exceeds a significance level α. If $p-$value $\leq \alpha$, we reject the null hypothesis, otherwise we fail to reject it.

Under the alternative hypothesis (i.e. when X and Y are dependent given $\mathbf{Z}$), the G-statistic follows a large-sample *non-central* χ^2 distribution (Agresti, 2013, Section 16.3.5). The non-centrality parameter (λ) of this distribution has the same form as in the G-statistic, but with sample values replaced by population values, $\lambda = 2NI(X;Y|\mathbf{Z})$. The effect size of the G-test can be naturally expressed as a function of the *conditional mutual information*, since according to Cohen (1988) the effect size is the square root of the non-centrality parameter divided by the sample, thus we have $\omega = \sqrt{2I(X;Y|\mathbf{Z})}$.

2.3 Feature Selection Through Estimation: Information Theoretic Approaches

So far we have seen how we can use mutual information to test independence and measure the effect size between a feature and the target variable. But, when we have a set of features it is also useful to order them according to their relevance with the target variable, a procedure known as *ranking* (Reimherr and Nicolae, 2013). Ranking features according to their dependency with the target variable provides very important and useful information. Applications of this principle range from model and feature subset selection to decision tree construction.

Feature subset selection is a special case of feature extraction, in which we select the features instead of combining to extract new features. Guyon et al. (2006) categorizes the feature selection techniques in three groups: filters, wrappers and embedded.

Filters are independent of the classifier and they define a scoring criterion (or relevance index) by which they produce a ranking of the features. *Wrappers* are classifier dependent; they use an evaluation measure to check the performance of the different subsets of features with a particular classifier and they choose the subset with the best performance. Finally, *embedded* methods are again classifier dependent, since they are part of the learning algorithm and the feature selection is applied in the training procedure.

From the above descriptions we can find the strengths and the weaknesses of each approach. Filters are classifier independent; they are the fastest method and they are less likely to overfit, but on the other hand their performance is worse than the classifier specific methods (some of the filters may underfit to the data). Embedded methods are classifier specific. They cannot be used generally, since they use a particular model and they are slower than the filters, but they may have better performance, than the filters and they tend to overfit less than the wrapper methods. Wrappers, being classifier dependent, may achieve better performance but on the other hand, are computational intensive and tend to overfit more than the other techniques (Guyon et al., 2006; Brown et al., 2012).

2.3.1 Filter Feature Selection via Mutual Information: A Unifying Framework via Maximizing Conditional Likelihood

In our work we focus on filter methods and, more particularly, we will be discussing information theoretic feature selection techniques. In filter methods, firstly we rank the features and then we select the ones that contain most of the useful information. By ranking the features on their mutual information with the class independently of each other, we derive a ranking that takes into account the *relevancy* with the class label. Choosing the features according to this ranking corresponds to a widely used feature selection heuristic, the *Mutual Information Maximization* (MIM) criterion (Lewis, 1992); where the score of each feature X_k is given by:

$$J_{mim}(X_k) = I(X_k; Y). \tag{2.8}$$

This approach does not take into account the *redundancy* between the features. By using more advanced techniques, e.g. Fleuret (2004), we can take into account both the relevancy and the redundancy between the selected features, *without* having to compute very high dimensional distributions. For example, one of the criteria that controls both relevance and redundancy and provides a very good trade-off in terms of accuracy, stability and flexibility (Brown et al., 2012) is the *Joint Mutual Information* (JMI) criterion. This criterion ranks the features according to the score:

$$J_{jmi}(X_k) = \sum_{X_j \in \mathbf{X}_\theta} I(X_k X_j; Y), \tag{2.9}$$

where $\mathbf{X}_\theta$ is the set of already selected features (Yang and Moody, 1999).

Brown et al. (2012) suggested a unified framework for many information theoretic heuristic criteria, by starting from a clearly specified objective function: the conditional likelihood of the class given the feature. This demonstrated that several common heuristics are approximate iterative maximisers of this objective. This analysis lead to the *Conditional Mutual Information* (CMI) criterion, which

ranks the features according to the score:

$$J_{cmi}(X_k) = I(X_k \mathbf{X}_\theta; Y). \tag{2.10}$$

Apart from MIM and JMI, many other criteria can be derived from optimising this function, among which Markov blanket discovery algorithms (such as IAMB, presented in the following section). This unified framework focused only in fully supervised data — *our work serves to naturally extend this to semi-supervised scenarios.*

All criteria need an estimate of the mutual information between a feature or a feature set and the target variable, which is derived from finite data sets. For that reason the *accuracy* of the estimator plays a crucial role in the ranking of the features. To measure the accuracy we use the *Mean Squared Error* (MSE), which can be expressed via the bias-variance decomposition as:

$$\begin{aligned} MSE\left(\widehat{I}(X;Y)\right) &= \mathbb{E}\left[\left(\widehat{I}(X;Y) - I(X;Y)\right)^2\right] \\ &= bias\left(\widehat{I}(X;Y)\right)^2 + var\left(\widehat{I}(X;Y)\right). \end{aligned}$$

The bias can be written explicitly, by making use of the fact that the estimator follows a non-central χ^2 distribution:

$$\begin{aligned} bias\left(\widehat{I}(X;Y)\right) &= \mathbb{E}\left[\widehat{I}(X;Y)\right] - I(X;Y) \\ &= \frac{(|\mathcal{X}|-1)\,(|\mathcal{Y}|-1)}{2N}. \end{aligned}$$

This expression is known in the literature as the *Miller-Madow bias correction factor* (Miller, 1955). To the best of our knowledge, this is the first time that the connection between the scaled non-central χ^2 distribution, and this correction factor, has been noted. The variance can be written:

$$\begin{aligned} var\left(\widehat{I}(X;Y)\right) &= \mathbb{E}\left[\left(\widehat{I}(X;Y) - \mathbb{E}\left[\widehat{I}(X;Y)\right]\right)^2\right] \\ &= \frac{(|\mathcal{X}|-1)\,(|\mathcal{Y}|-1)}{2N^2} + \frac{2I(X;Y)}{N}. \end{aligned}$$

More details on the bias-variance of the plug-in estimator can be found in Paninski (2003). As we see, the mean square error of the estimator, which captures both

the bias and the variance, depends on the sample size, the characteristics of the features and on the actual effect size, which does not guarantee that the ranking using an estimator is the same as the population ranking. In the limit of data, $N \to \infty$, the estimated ranking converges to the population, since the maximum likelihood estimator is consistent.

2.3.2 Semi-supervised Feature Selection: Filter, Wrapper and Embedded Approaches

In the literature of semi-supervised feature selection there is a lack of principled information theoretic filter methods. However they have been suggested different kinds of filter, wrapper and embedded approaches. In this section, for completeness, we will present some of these approaches.

Wu and Flach (2002) suggested a semi-supervised wrapper type approach where they used a χ^2 statistic for a goodness-of-fit test to select features based on both the labelled and the unlabelled set. Ren et al. (2008) introduced another wrapper approach, by proposing an iterative procedure, where at each step, unlabelled examples receive labels from the classifier and then a wrapper-based feature selection is performed. In a recent work, Bellal et al. (2012) presented another semi-supervised wrapper approach based on random forests. They showed that the way internal estimates are used to measure variable importance in random forests is also applicable to feature selection in semi-supervised learning, exploiting the information of both labeled and unlabeled data.

Xu et al. (2010) suggested a semi-supervised embedded approach, by formulating the feature selection as a convex-concave optimization problem. The proposed method selects features through maximizing the margin between different classes, while at the same time exploiting the geometry of the probability distribution that generates the data. Helleputte and Dupont (2009) suggested a method for semi-supervised embedded feature selection using linear models and including regularization to enforce sparsity.

Finally, Zhao et al. (2008) suggested a semi-supervised filter algorithm based on manifold learning and spectral graph theory. The local geometrical structure and the discriminant structure in the data are captured by two graphs and the scores of the features are characterized by their degree of preserving these two graph structures.

2.4 Feature Selection Through Testing: Markov Blanket Discovery

In this section we will introduce the notation and the background material on Markov blanket discovery algorithms.

2.4.1 Markov Blanket: Notation and Definitions

Assume that we have a binary classification dataset $\mathcal{D} = \{(\mathbf{x}^i, y^i)|i = 1, ..., N\}$, where the target variable Y takes the value $y = 1$ when the example is positive, and $y = 0$ when the example is negative. The feature vector $\mathbf{x} = [x_1...x_d]$ is a realization of the d-dimensional joint random variable $\mathbf{X} = X_1...X_d$. With a slight abuse of notation, in the rest of our work, we interchange the symbol for a set of variables and for their joint random variable. Following Pearl (1988), we have the following definitions.

Definition 2.3 (Markov blanket — Markov boundary)**.**
The Markov blanket of the target Y is a set of features $\mathbf{X}_{\text{MB}}$ with the property $Y \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}_{\text{MB}}$ for every $\mathbf{Z} \subseteq \mathbf{X}\backslash\mathbf{X}_{\text{MB}}$. A set is called Markov boundary if it is a minimal Markov blanket, i.e. non of its subsets is a Markov blanket.

In probabilistic graphical models terminology, the target variable Y becomes conditionally independent from the rest of the graph $\mathbf{X}\backslash\mathbf{X}_{\text{MB}}$ given its MB $\mathbf{X}_{\text{MB}}$. Figure 2.2 shows a toy Bayesian network. The MB of the target variable Y is the feature set that contains the *parents* (X_4 and X_5), *children* (X_9 and X_{10}) and *spouses* (X_7 and X_8, which are other parents of a child of Y) of the target.

Learning the Markov blanket for each variable of the dataset, or in other words, inferring the local structure, can naturally lead to *causal structure learning* (Pellet and Elisseeff, 2008). Apart from playing a huge role in the structure learning of a Bayesian network, Markov blanket is also related to another important machine learning activity: *feature selection.*

Koller and Sahami (1996) published the first work about the optimality of the Markov blanket in the context of feature selection. Recently, Brown et al. (2012) introduced a unifying probabilistic framework and showed that many heuristically suggested feature selection criteria, including Markov blanket discovery algorithms, can be seen as iterative maximizers of a clearly specified objective function: the conditional likelihood of the training examples.

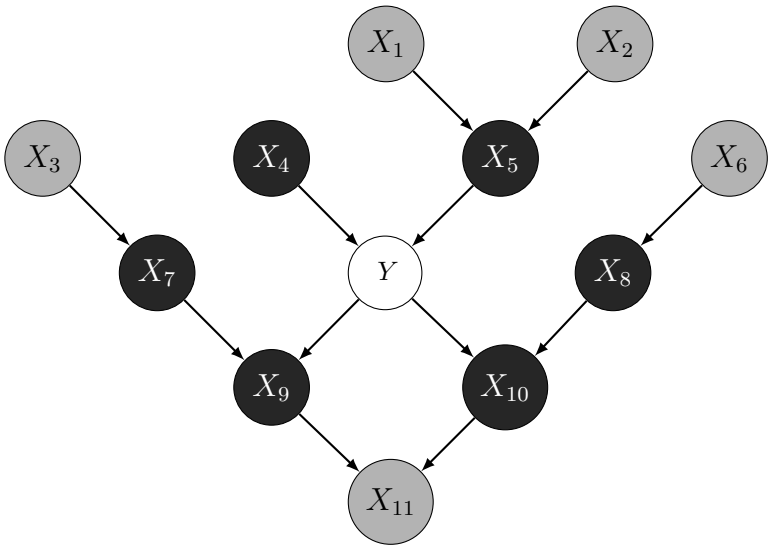


Figure 2.2: *Toy Markov blanket example where: white nodes represent the target variable, black ones the features that* belong *to the MB of the target and grey ones the features that* do not belong *to the MB.*

2.4.2 Supervised Markov Blanket Discovery Algorithms

Margaritis and Thrun (1999) introduced the first theoretically sound algorithm for Markov blanket discovery, the Grow-Shrink (GS) algorithm. This algorithm consists of two-stages. The first stage is the *growing*, where we add features to the Candidate Markov Blanket (CMB) set until the point that the remaining features are independent with the target given the candidate blanket. The second stage is the *shrinkage*, where we remove potential false positives from the CMB. Tsamardinos and Aliferis (2003) suggested an improved version of this approach, the Incremental Association Markov Blanket (IAMB), which can be seen in Algorithm 1. Many measures of association have been used to decide which feature will be added to the candidate blanket during the growing phase (Alg. 1 - Line 4), with the main being the *conditional mutual information* (Pocock et al., 2012). But, Yaramakala and Margaritis (2005) suggested the use of the *significance of the conditional test of independence*, which is more appropriate in statistical terms than the raw conditional mutual information value. Finally, there is another class of algorithms that tries to control the size of the conditioning set in a two-phase procedure: first identify parents and children, then identify spouse nodes (i.e.

nodes with which Y shares one or more common children). The most representative algorithms are the HITON (Aliferis et al., 2010) and the Max-Min Markov Blanket (MMMB) (Tsamardinos et al., 2003). All of these algorithms assume faithfulness of the data distribution. As we already saw, in all Markov blanket discovery algorithms, the conditional test of independence (Alg. 1 - Lines 5 and 11) plays a crucial role. Furthermore, to choose the most strongly related feature in Line 4, we evaluate the p-values for the conditional tests and choose the feature with the smaller one.

Algorithm 1: Incremental Association Markov Blanket (IAMB)

Input : Target Y, Features $\mathbf{X} = X_1...X_d$, Significance level α
Output: Markov Blanket: $\mathbf{X}_{\text{CMB}}$

1 Phase I: forward — growing
2 $\mathbf{X}_{\text{CMB}} = \emptyset$
3 **while** $\mathbf{X}_{\text{CMB}}$ has changed **do**
4 Find $X \in \mathbf{X} \backslash \mathbf{X}_{\text{CMB}}$ most strongly related with Y given $\mathbf{X}_{\text{CMB}}$
5 **if** $X \not\perp\!\!\!\perp Y | \mathbf{X}_{\text{CMB}}$ using significance level α **then**
6 Add X to $\mathbf{X}_{\text{CMB}}$
7 **end**
8 **end**
9 Phase II: backward — shrinkage
10 **foreach** $X \in \mathbf{X}_{\text{CMB}}$ **do**
11 **if** $X \perp\!\!\!\perp Y | \mathbf{X}_{\text{CMB}} \backslash X$ using significance level α **then**
12 Remove X from $\mathbf{X}_{\text{CMB}}$
13 **end**
14 **end**

2.4.3 Semi-Supervised Markov Blanket Discovery

To the best of our knowledge, there is only one algorithm for deriving the MB of semi-supervised targets: BASSUM (BAyesian Semi-SUpervised Method) (Cai et al., 2011). BASSUM follows the HITON approach, finding first the parent-children nodes and then the spouses, and tries to take into account both labelled and unlabelled data. BASSUM makes the "traditional semi-supervised" assumption that the labelled set is an unbiased sample of the overall population (i.e. the labels are MCAR), and uses the unlabelled examples in order to improve the reliability of the conditional independence tests. For example, to estimate the G-statistic, in equation (2.7), it uses both labelled and unlabelled data for the

observed counts $O_{.,.,\mathbf{z}}$ and $O_{x,.,\mathbf{z}}$. This technique is known in statistics as *available case analysis* or *pairwise deletion*, and is affected by the ambiguity over the definition of the overall sample size, which is crucial for deriving standard errors and the sampling distribution of the statistics (Allison, 2001, page 8). This can lead to unpredictable results. For example, there are no guarantees that the G-statistic will follow χ^2 distribution after this substitution. Another weakness of BASSUM is that it cannot be applied in partially labelled environments where we have the restriction that the labelled examples come only from one class, such as the positive-unlabelled data. In order to explore the Markov blanket of this type of data, we should explore how to test conditional independence in this scenario; this is the focus of Chapter 4. Before that, we will formally introduce the partially labelled data scenarios in the following chapter.

2.5 Chapter Summary

In this chapter, we have thoroughly reviewed the dynamics of dealing with information theoretic measures in fully supervised data. First, we analysed testing and estimation, and then we connected these two activities with feature selection. For estimation, we focused on presenting Shannon's mutual information, and the squared loss mutual information, two of the most widely ways in machine learning to quantify the shared information between random variables. Then we explored the links between these two measures and two widely used tests of independence: the G-test and the X^2-test. Finally, we discussed approaches to feature selection using information theoretic heuristics or Markov blanket discovery.

The main objective of this thesis is to extend the feature selection methodologies on partially labelled scenarios, by deriving valid ways to test independence and obtaining consistent estimates of the mutual information. Before we proceed to this, we will formally present the background on partially labelled data and a recent associated graphical representation.

Chapter 3

Literature Review: Partially Labelled Data

While the collection of unlabelled data is an automatic procedure, the procedure of labelling them is expensive in terms of both time and resources. As a result, in order to deal efficiently with large amount of data, we should explore techniques that are able to handle partially labelled data. In the current section, we will set up the framework and the notation that we will follow to explore the dynamics of feature selection in partially labelled data.

There are two important issues that we should explore when we are dealing with missing labels. The first deals with how the data are collected. Section 3.1 presents the two main assumptions over the generation of partially labelled datasets. The other important issue is to explore why a label is missing. As we presented in the introduction, there are four main assumptions over why a label is missing: MCAR, MAR-F, MAR-C and MNAR. Section 3.2 introduces the m-graphs, which constitute a key tool to explore these four assumptions. Section 3.3 connects the m-graphs to the established literature on partially labelled data. Finally, Section 3.4 explores the challenges of feature selection in partially labelled data.

3.1 Different Scenarios for the Generation of Partially Labelled Data

A partially labelled dataset $\mathcal{D}$ is a combination of two samples; the labelled $\mathcal{D}_L$ and the unlabelled $\mathcal{D}_U$, so the full dataset is $\mathcal{D} = (\mathcal{D}_L, \mathcal{D}_U)$. We will assume that we have N examples, out of which N_L belong to the labelled set, while N_U to the unlabelled. For the labelled set we have knowledge about the class labels $\mathcal{D}_L = \{(\mathbf{x}^i, y^i) | i = 1, ..., N_L\}$, while for the unlabelled set we record only the feature vector $\mathcal{D}_U = \{(\mathbf{x}^i) | i = (N_L + 1), ..., N\}$, with $N = N_L + N_U$.

There are two subtly different scenarios and the assumptions behind each one will guide us in solving the feature selection problem. Our analysis will focus on one scenario but it can be easily generalised to the other, since both are equivalent under certain conditions. By pointing out the differences between these two scenarios, we can get more insight into the challenges of partially labelled data.

Following Smith and Elkan (2007), we name the first scenario *single-training set*. In this case, first we sample the whole dataset $\mathcal{D}$, then we label some of the examples to form the labelled set $\mathcal{D}_L$, and the remaining ones form the unlabelled set $\mathcal{D}_U$. A convenient way to analyse this scenario is by introducing a binary random variable S to describe if an example is labelled ($s = 1$) or not ($s = 0$). So the single-training set scenario assumes that the training data $\mathcal{D}$ are drawn randomly from $p(\mathbf{X}, Y, S)$, and for each tuple $<\mathbf{x}, \mathrm{y}, \mathrm{s}>$ that is drawn, $<\mathbf{x}, \mathrm{s}>$ is recorded and when $s = 1$ we also record the value of y, otherwise it is labelled as "missing". In this way, the labelled dataset $\mathcal{D}_L$ comes from the joint distribution $p(\mathbf{x}, y | s = 1)$, while the unlabelled dataset $\mathcal{D}_U$ comes from the distribution $p(\mathbf{x} | s = 0)$. In this scenario the probability $p(s = 1)$ can be estimated via $\frac{N_L}{N}$, and we can also directly estimate $p(\mathbf{x}, y | s = 1)$ and $p(\mathbf{x} | s = 0)$ from the $\mathcal{D}_L$ and $\mathcal{D}_U$ respectively. A recent work that follows this scenario is by Fox-Roberts and Rosten (2014).

Another setting assumes that the unlabelled set $\mathcal{D}_U$ is sampled *independently* from the labelled set $\mathcal{D}_L$ (Seeger, 2002). We name this scenario as *labelled-background*. In this case, the unlabelled examples are drawn separately from the background distributions $p(\mathbf{x})$. As before, we can directly estimate $p(\mathbf{x}, y | s = 1)$ and $p(\mathbf{x})$ from $\mathcal{D}_L$ and $\mathcal{D}_U$ respectively, but we *cannot* estimate the probability of selection $p(s = 1)$ and thus the probability $p(\mathbf{x} | s = 0)$.

The main difference is that $p(s = 1)$ and $p(\mathbf{x}|s = 0)$ can only be estimated in the first scenario. Of course, if we are given the $p(s = 1)$ in the labelled-background, the two scenarios are equivalent (Hein, 2009, Section 3.4). This can be seen from the expression:

$$p(\mathbf{x}) = p(\mathbf{x}|s = 1)p(s = 1) + p(\mathbf{x}|s = 0)p(s = 0),$$

where by knowing $p(\mathbf{x}|s = 1)$, $p(\mathbf{x})$ and $p(s = 1)$, we can calculate $p(\mathbf{x}|s = 0)$. While the distinction between these two scenarios seems artificial, Hein (2009, Section 3.4) presents two examples in order to show that both of them can occur in practice. The curious reader will get more insight into these two scenarios by comparing the following two works in the positive-unlabelled context: Elkan and Noto (2008), who follow the single-training-set setting, and Li et al. (2011), who follow the labelled-background setting. Hein (2009, Section 3.4) presents these two scenarios in the context of learning under sample selection bias using labelled and unlabelled examples.

In our work, we follow the single-training set scenario to collect the partially labelled dataset $\mathcal{D}$. But how the labelled $\mathcal{D}_L$ and the unlabelled $\mathcal{D}_U$ datasets are generated has to do with the assumptions over why a label may be missing or not. As we presented in Section 1.1, these assumptions can be categorized as MCAR, MAR-F, MAR-C and MNAR. This set of assumptions can be explored in a natural way under the single-training set scenario, as we will see in our work, and by using the missingness graphs.

3.2 Formal Notation: Missing Labels and Missingness Graphs

Following the notation of Smith and Elkan (2007), we assume that the partially labelled data are sampled from the joint distribution of $p(\mathbf{X}, Y, S)$ where the three random variables take the following values:

- $\mathbf{x} = [x_1...x_d] \in \mathcal{X}$ and $\mathbf{X} = X_1...X_d$ is the joint random variable of the d categorical features or covariates, where $\mathcal{X}$ is a finite subset of $\mathbb{R}^d$.
- $y \in \mathcal{Y} = \{0, 1\}$ and Y represents the binary class variable. We use $y = 1$ when the example is positive, and $y = 0$ when the example is negative.

- $s \in \mathcal{S} = \{0, 1\}$ and S is a further binary random variable indicating an example as labelled ($s = 1$) or unlabelled ($s = 0$).

The random variable Y is not observed for all the examples of the partially labelled dataset. When $s = 1$ the actual value Y is observed, while when $s = 0$ we have a missing value. From now on, we will use a token m to represent the missing values. So what we actually observe is not the variable Y, but a "surrogate" variable $\widetilde{Y}_m$, which takes the following values:

$$\widetilde{Y}_m = \begin{cases} 1, & \text{if } \; s = 1, \; y = 1 \\ 0, & \text{if } \; s = 1, \; y = 0 \\ m, & \text{if } \; s = 0 \end{cases} .$$

Note that this is essentially identical to Y, except that whenever there is a missing value, we substitute a token, m. The procedure of introducing the variable $\widetilde{Y}_m$ to deal with missing data is similar to *dummy variable adjustment* or *missing indicator method* (Allison, 2001), a procedure used in statistics when feature values are missing.

In order to further understand this, we will use the formalism of *missingness graphs*, introduced in a series of recent papers, e.g. Mohan et al. (2013); Mohan and Pearl (2014). Example m-graphs can be seen in Figure 3.1, where the notation is as follows.

- ◯ X fully observable variable
- ◌ Y partially observable variable
- ◉ $\widetilde{Y}_m$ surrogate variable for Y (fully observable)
- ◯ S fully observable variable, driving the missingness mechanism

In an m-graph, associated with every partially observed variable Y there are two additional variables: the surrogate $\widetilde{Y}_m$, and S, a fully observed variable which controls the underlying mechanism of whether a value is missing. If $s = 0$, the value of Y for this observation will be missing. However, contrary to conventional use of a missingness indicator, S is treated as a 'driver' of equality between Y and $\widetilde{Y}_m$. A quote from the original paper explains their utility eloquently:

> *"Since every d-separation in the graph implies conditional independence in the distribution, the m-graph provides an effective way of representing the statistical properties of the missingness process and, hence, the potential of recovering the statistics of variables [..] from partially missing data"* Mohan et al. (2013)

This captures the aim of this thesis – to *recover the statistics of variables from partially missing data*. In the m-graphs of Figure 3.1, the bidirectional arc between X and Y means that the causality can be either way, although in our work we make no causal assumptions. More details on the causal/anti-causal setting for semi-supervised data can be found in Schölkopf et al. (2013). In the next section, we will analyse the four different missingness scenarios presented earlier — MCAR, MAR-F, MAR-C and MNAR — by using m-graphs and connecting them to the established literature on partially and semi-supervised data.

3.3 Literature of Partially Labelled Data in the Language of m-graphs

The mechanism behind missing labels plays a crucial role in the analysis of partially labelled data and is closely connected with the concept of *sampling bias*. In the traditional semi-supervised scenario it has been assumed that the labelled set was an unbiased sample of the overall population (Smith and Elkan, 2007). However, during recent years, scenarios have been explored where labelled sample bias occurs (Lafferty and Wasserman, 2007; Plessis and Sugiyama, 2012).

Chawla and Karakoulas (2005) published the first work on sampling bias in partially labelled data. In generic machine learning, *sample bias* occurs when the training examples do not represent the population distribution. This happens because the selection process for each training example depends on the features and/or on the target variable (Storkey, 2009)[1]. In partially labelled settings, sample bias occurs when the distribution of the labelled examples does not represent the population, or equivalently when the distribution of the labelled set is different than the distribution of the unlabelled set. The problem of sample

[1]Sample bias is commonly associated with *dataset shift* (Quionero-Candela et al., 2009). The techniques presented in the current work can be extended in scenarios where we have *covariate shift* or *prior probability shift*, which are the two main types of dataset shift according to Moreno-Torres et al. (2012).

selection bias received a great attention in econometrics, and Heckman (1979) in his nobel-prize winning work, suggests one of the first procedures of correction the sample selection bias. The first works dealing with the sample-selection bias in machine learning are (Zadrozny, 2004; Smith and Elkan, 2004; Fan and Davidson, 2007). Chawla and Karakoulas (2005) state that to deal with sample bias in semi-supervised data, we should model the underlying missing data mechanism, and this is the approach that we follow in our work by using m-graphs.

In the literature of partially labelled data, we can find works that follow different assumptions over the missingness mechanism. Our work categorises these assumptions in the four scenarios presented earlier: MCAR, MAR-F, MAR-C and MNAR. As we already mentioned, to model the missing data mechanism we use m-graphs: Figure 3.1 presents the four graphical models.

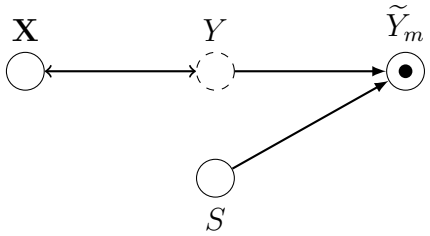
In the following subsections we will examine each of these four scenarios, and see how the m-graph formalism can be used to encode them and connect them with the literature. Following this, we will present a full analysis of each scenario, showing how estimation of quantities like mutual information can be achieved, and how we can make optimal use of the unlabelled data. Table 3.1 presents the sections of this thesis, where each scenario is analysed.

Scenario	m-graph	Background	Full analysis for feature selection tasks
MCAR	Fig. 3.1a	Section 3.3.1	Sections 4.2, 5.2, 6.2
MAR-F	Fig. 3.1b	Section 3.3.2	Sections 4.3, 5.3, 6.3
MAR-C	Fig. 3.1c	Section 3.3.3	Sections 4.4, 5.4, 6.4
MNAR	Fig. 3.1d	Section 3.3.4	possible only with model-dependent approaches

Table 3.1: *Sections where the different missingness mechanism scenarios are analysed.*

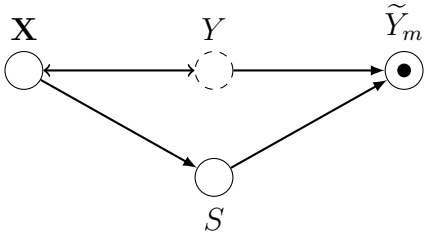
3.3.1 Labels Missing Completely at Random (MCAR)

Under this assumption, the examples for the labelled set are selected completely at random, or in other words, the probability of labelling an example is independent of its feature or class value. As a result we do not have any sample selection bias in the labelled set. According to Smith and Elkan (2007), this is the assumption that the traditional semi-supervised learning scenario makes, and was used in the earliest works on semi-supervised data (Seeger, 2002). Figure 3.1a presents the m-graph that captures the assumption that the labels are MCAR, where we see



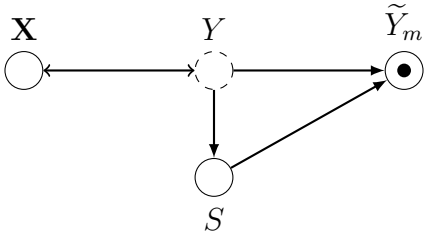
$$\frac{\mathbf{X}, Y \perp\!\!\!\perp S}{\begin{array}{c} p(s|\mathbf{x}, y) = p(s) \\ p(\mathbf{x}, y|s) = p(\mathbf{x}, y) \end{array}}$$

(a) *MCAR: the missingness mechanism S does not depend directly on features X or on target Y.*



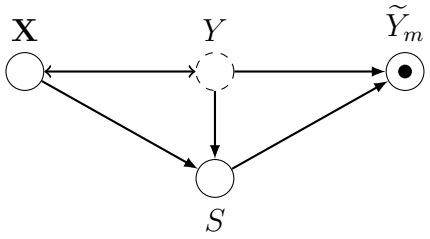
$$\frac{Y \perp\!\!\!\perp S|\mathbf{X}}{\begin{array}{c} p(s|\mathbf{x}, y) = p(s|\mathbf{x}) \\ p(y|\mathbf{x}, s) = p(y|\mathbf{x}) \end{array}}$$

(b) *MAR-F: the missingness mechanism S depends directly only on the features.*



$$\frac{\mathbf{X} \perp\!\!\!\perp S|Y}{\begin{array}{c} p(s|\mathbf{x}, y) = p(s|y) \\ p(\mathbf{x}|y, s) = p(\mathbf{x}|y) \end{array}}$$

(c) *MAR-C: this missingness mechanism S depends directly only on the class variable.*



No independence between $\mathbf{X}, Y, S$

(d) *MNAR: the missingness mechanism S depends directly both on features and on the target variable.*

Figure 3.1: *m-graph for the different missingness scenarios occurred in partially labelled data: (a) data are missing completely at random (MCAR), (b) data are missing at random feature dependent (MAR-F), (c) data missing at random class dependent (MAR-C) and (d) data missing not at random (MNAR).*

that the variable S does not have a direct connection with $\mathbf{X}$ or Y. We also highlight some of the independence properties that can be read directly from the graph. A recent work that follows this assumption in the semi-supervised context is by Fox-Roberts and Rosten (2014).

If we work only with the labelled data, and ignore the unlabelled examples (a procedure known as *listwise deletion* in statistics), we can have valid tests and unbiased estimates, but we will lose statistical power because of the smaller sample size.

3.3.2 Labels Missing at Random Feature Dependent (MAR-F)

This labelling bias is generated when the missingness mechanism depends directly only on the features, or in other words, if the labelling of an example is conditionally independent of the class, given the feature values. This scenario is known in missing data literature as missing at random (MAR) (Moreno-Torres et al., 2012). The traditional MAR, introduced by Little and Rubin (2002), requires conditional independence in the *event level*. Instead of this, in our work we follow the recent literature (Mohan et al., 2013; Mohan and Pearl, 2014; Van den Broeck et al., 2015; Tian, 2015; Thoemmes and Mohan, 2015) and we assume conditional independence in the *random variable level*, which is known in statistics literature as MAR+ (Potthoff et al., 2006). Furthermore, in our work, in order to stress the fact that the missingness depends directly on the features, we will name this scenario as *missing at random feature dependent* (MAR-F), and Figure 3.1b presents the m-graph that captures this assumption.

As we see, under this assumption, the posterior probability $p(y|\mathbf{x})$ can be calculated from the labelled set $p(y|\mathbf{x}, s = 1)$, or in other words, the class boundaries can be derived from the labelled set. This is similar to the clustering assumption from the semi-supervised literature (Chapelle et al., 2006). For that reason, this is widely used in semi-supervised learning and the importance of this assumption is also presented in the framework of semi-supervised regression by Lafferty and Wasserman (2007). This bias is also called *learnable* (Smith and Elkan, 2007), since the labelling mechanism can be learnt from the observed variables.

Furthermore, the conditional distribution on the labelled set is equal to the true conditional distribution $p(y|\mathbf{x}, s = 1) = p(y|\mathbf{x})$, while the marginals are

different $p(\mathbf{x}|s = 1) \neq p(\mathbf{x})$. So this type of bias produces *covariate shift*, which is a special case of dataset shift, between the distribution in the labelled set and the true distribution (Hein, 2009). In our work we will explore the dynamics of testing, estimation and ranking in MAR-F semi-supervised data, and our results can be used in order to explore this activities under covariate shift scenarios.

3.3.3 Labels Missing at Random Class Dependent (MAR-C)

This labelling bias is generated when the missingness mechanism depends directly only on the class label, or in other words, when labelling one example is conditionally independent of the features given the class. Following Moreno-Torres et al. (2012), we name this scenario as *missing completely at random class dependent* (MAR-C) and Figure 3.1c presents the m-graph that captures this assumption. We should mention at this point that in the missing data literature (Little and Rubin, 2002), this scenario is classified as *missing not at random*, since the missingness mechanism depends directly on the partially observed variable. For example, Thoemmes and Mohan (2015, Figure 3(a)) present this scenario as a simple version of a missing not at random scenario.

It is interesting to mention that in the MAR-C scenario it is not possible to directly estimate $p(y)$ (Smith and Elkan, 2004). But, having prior knowledge over it, the bias introduced by this sampling mechanism can be corrected (Hein, 2009). Our work explores *how we can incorporate this prior knowledge in testing, estimation and ranking.*

In the semi-supervised setting, there are many works that followed this assumption (Rosset et al., 2004; Zou et al., 2004; Lawrence and Jordan, 2006; Plessis and Sugiyama, 2012). A practical application where we can use this assumption is in *class-prior-change* scenario (Plessis and Sugiyama, 2012), which occurs when the class balance in the labelled set does not reflect the population class balance: $p(y = 1|s = 1) \neq p(y = 1)$.

Furthermore, the class conditional distribution in the labelled set is equal to the true class conditional distribution $p(\mathbf{x}|y, s = 1) = p(\mathbf{x}|y)$, while the probability of the class is different $p(y|s = 1) \neq p(y)$. So this type of bias produces *prior-probability drift*, which is a special case of dataset shift, between the distribution in the labelled set and the true distribution (Hein, 2009). In our work

we will explore the dynamics of testing, estimation and ranking in MAR-C semi-supervised data, and our results can be used in order to improve these activities under prior-probability drift scenarios.

A more restricted version of this model, where we observe examples only from the positive class, generates *positive-unlabelled* data under the widely used *selected completely at random assumption* (SCAR) (Elkan and Noto, 2008). In this case, there is conditional independence at the *event level*: $\mathbf{X} \perp\!\!\!\perp S|y = 1$, while there are no negatively labelled examples $p(s = 1|y = 0) = 0$. A common approach to solve this problem is to assume unlabelled examples as the negative class. For example, Blanchard et al. (2010), focusing on the labelled-background setting, use that approach and prove that semi-supervised novelty detection can be reduced to Neyman-Pearson binary classification, using, in their terminology, the nominal and unlabelled samples as the two classes.

3.3.4 Labels Missing Not at Random (MNAR)

Finally, when the selection process depends directly both on the features and on the class we have *complete bias* or *arbitrary bias*. We denote this scenario as *missing not at random* (MNAR) and Figure 3.1d presents the m-graph of this scenario. Heckman (1979), in his Nobel prize-winning work, suggested one of the first procedures for correcting this sample selection bias. Heckman introduced a two-step procedure using linear models: a feature-set is used to build the regression model for predicting Y and a different feature set to build the binary probit selection model for predicting S, while these two models were correlated (Chawla and Karakoulas, 2005).

When the MNAR holds, there are no independencies between $\mathbf{X}, S, Y$, and the analysis is impossible without a model-based technique. As a result, we will not focus on this scenario, since we are interested in exploring classifier independent ways for filter feature selection and we give more details about our approach in the following section.

3.4 The Main Challenges of Filter Feature Selection in Partially Labelled Data

The main ways for handling missing data can be categorised as follows (Allison, 2001):

- Deletion based methods, such as list-wise deletion (or complete case analysis) and pairwise deletion (available case analysis).
- Maximum likelihood (ML) based methods, where a general way of obtaining the estimates is through expectation-maximization (EM) algorithm.
- Imputation based methods, which is used in conjunction with ML methods and substitutes with each missing value with an inferred value.

The first class of methods is model-independent, while the last two are model-based, and require inference. In the partially labelled data, the missing values are in the labels, and as a result, inference over them is equivalent to solving the classification problem.

On the other hand, the main characteristic of filter feature selection approaches is that they are *classifier or model independent* as opposed to wrapper/embedded methods, which are classifier dependent (Guyon et al., 2006; Brown et al., 2012). Thus, our main objective is to select features without any model dependent technique and without performing inference. For that reason we will use the tool of m-graphs to represent the causal mechanisms responsible for missing labels and solve the feature selection problem in a model-independent manner. Apart from this desirable property, Van den Broeck et al. (2015) summarise some further advantages of this framework: it provides consistent parameter estimates, and the estimates are computable in closed-form with a single pass over the data. Both of them are very important when the data set is very large, i.e. in "Big Data" scenarios. Now, we will start with our theoretical analysis and the following chapter will explore hypothesis testing in partially labelled data.

3.5 Chapter Summary

This chapter concludes the background material of this thesis. We formally introduced the problem of partially supervision, and we also presented m-graphs, a

useful tool for analysing the assumptions behind the missing labels. By exploring the literature of partially labelled data we categorised these assumptions in four groups: MCAR, MAR-F, MAR-C and MNAR. Then, we presented the m-graph for each one and connected them to the established literature on partially labelled data. Finally, we explored the challenges of filter feature selection in partially labelled data, and we pointed out the necessity of inference-free approaches for dealing with these challenges.

After setting up the framework of partially labelled data, we now move to the main part of this thesis: a theoretical analysis of testing, estimation and ranking despite partial supervision and in an inference-free manner.

Chapter 4

Theoretical Analysis of Hypothesis Testing in Partially Labelled Data

In the previous chapters we presented hypothesis testing in fully supervised data, and we introduced the partially labelled scenario. In this chapter, we present our theoretical investigation about hypothesis testing in partially labelled data. Our aim is to make use of all available information, in an entirely *classifier-independent* and *inference-free* fashion. Our general strategy is to use *surrogate variables*, instead of the unobservable labels, and explore statistical consequences in each situation.

To do so, in Section 4.1, we will derive all the fully observable surrogate variables in this setting, while we will define which of them can be classified as valid and informed. Then we will compare the performance, in terms of probability of committing a type I and a type II error, of these surrogate approaches with the performance that we would obtain by using the unobservable fully-supervised variable Y. We will explore the behaviour of the surrogate variables in the three missingness scenarios: MCAR in Section 4.2, MAR-F in Section 4.3 and MAR-C in Section 4.4. Finally, we will extend our theoretical results for testing conditional independence in partially labelled data.

4.1 Deriving Surrogate Variables From Partially Labelled Datasets

Given a fully observed variable X, a partially observed Y, and a fully observed $\widetilde{Y}$, we investigate how $\widetilde{Y}$ could be used as a "surrogate" for Y. We define two properties that $\widetilde{Y}$ may possess – validity and informedness – concerning the false positive rate and the false negative rate of a G-test when $\widetilde{Y}$ is used in place of Y.

Definition 4.1. *Valid surrogate variable:* $\widetilde{Y}$ *is a* valid *surrogate for* Y *iff when* $X \perp\!\!\!\perp Y$*, the test* $G(X;\widetilde{Y})$ *has the same false positive rate as the unobservable* $G(X;Y)$*. In other words,* $X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp \widetilde{Y}$*.*

Definition 4.2. *Informed surrogate variable:* $\widetilde{Y}$ *is an* informed *surrogate for* Y *iff:*
1. it is a valid *surrogate (i.e. satisfies definition 4.1) and*
2. the test $G(X;\widetilde{Y})$ *can be corrected to have the same false negative rate as the unobservable* $G(X;Y)$*, simply by increasing the number of samples to* N/κ*, where* κ *is a constant factor calculated using knowledge of the class prior in the domain.*

Each possible surrogate variable provides a potential strategy to use in place of the *unobservable* test $G(X;Y)$. A surrogate variable effectively encodes an assumption that we can make over the unlabelled set – either we ignore it, use a special token, or assume all (unlabelled) examples are all either positive or negative. The differences of the various surrogate variables are illustrated in Table 4.1 and described in Table 4.2.

Surrogate 1: ***"Ignore unlabelled"*** This uses only the labelled set $\mathcal{D}_L$ and ignores the unlabelled. It is also known as *list-wise deletion*.

Surrogate 2: ***"Missingness token"*** This uses $\widetilde{Y}_m$ in place of Y, and assumes that the unlabelled example belong to a new category m. It is also known as *missing indicator method*.

Surrogate 3: ***"Assume negative"*** Denoted by $\widetilde{Y}_0$. It assumes that all unlabelled examples are negative.

Surrogate 4: ***"Assume positive"*** Denoted by $\widetilde{Y}_1$. It assumes that all unlabelled examples are positive.

		$\mathbf{X}$	Y	S	Surrogate 2 ($\widetilde{Y}_m$)	Surrogate 3 ($\widetilde{Y}_0$)	Surrogate 4 ($\widetilde{Y}_1$)
$\mathcal{D}$	$\mathcal{D}_L$	$\mathbf{x}^1$	1	1	1	1	1
		$\mathbf{x}^2$	0	1	0	0	0
		$\mathbf{x}^3$	0	1	0	0	0
		$\mathbf{x}^4$	1	1	1	1	1
		$\mathbf{x}^5$	0	1	0	0	0
	$\mathcal{D}_U$	$\mathbf{x}^6$		0	m	0	1
		$\mathbf{x}^7$		0	m	0	1
		$\mathbf{x}^8$		0	m	0	1
		$\mathbf{x}^9$		0	m	0	1
		$\mathbf{x}^{10}$		0	m	0	1

Table 4.1: *Example of semi-supervised data, and some possible surrogate variables that could be used in place of the unobservable Y. The tilde indicates surrogate variable, and the subscript indicates an assumed value for missing values — e.g. using a missingness token is $\widetilde{Y}_m$.*

4.2 Testing when the Labels are MCAR

In order to use any of the four surrogate approaches, we should first explore if they are *valid* for testing the null hypothesis of independence. In other words, we should check that when the null hypothesis holds for the unobservable test (i.e. $X \perp\!\!\!\perp Y$) then it also holds for the surrogate tests, and vice versa. This proof makes sure that by following a surrogate approach, the probability of committing a type I error will be the same as using the unobservable fully supervised test between X and Y. The following theorem presents our findings when the labels are missing completely at random.

Theorem 4.3 (MCAR: Which surrogate tests are *valid* for testing $X \perp\!\!\!\perp Y$?)**.** *In MCAR we can test independence by using any of the four surrogate approaches:*

$$\begin{aligned}
&\textbf{Surrogate 1 } (\mathcal{D}_L): \; X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp Y | s = 1,\\
&\textbf{Surrogate 2 } (\widetilde{Y}_m): \; X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp \widetilde{Y}_m,\\
&\textbf{Surrogate 3 } (\widetilde{Y}_0): \; X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp \widetilde{Y}_0,\\
&\textbf{Surrogate 4 } (\widetilde{Y}_1): \; X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp \widetilde{Y}_1.
\end{aligned}$$

Proof sketches for each of these situations can be found in Appendix A.

While Theorem 4.3 tells us that the surrogate tests are equivalent to the

Symbol	Short description
$\widetilde{Y}_m$	surrogate variable, fully observed,
	taking value m when Y has missing values.
$\widetilde{y}_m = 1$	a positively labelled example $(y = 1, s = 1)$
$\widetilde{y}_m = 0$	a negatively labelled example $(y = 0, s = 1)$
$\widetilde{y}_m = m$	an unlabelled example $(s = 0)$
$\widetilde{Y}_0$	surrogate variable, fully observed,
	taking value 0 when Y has missing values.
$\widetilde{y}_0 = 1$	a positively labelled example $(y = 1, s = 1)$
$\widetilde{y}_0 = 0$	a negatively labelled or an unlabelled example
$\widetilde{Y}_1$	surrogate variable, fully observed,
	taking value 1 when Y has missing values.
$\widetilde{y}_1 = 0$	a negatively labelled example $(y = 0, s = 1)$
$\widetilde{y}_1 = 1$	a positively labelled or an unlabelled example

Table 4.2: *Notation with short description of the surrogate variables.*

unobservable test for detecting independencies, it says nothing about how well the surrogate approaches perform when the null hypothesis is *false*. To do this we should compare the tests in terms of their power to detect a given effect. The effect size that our work uses is the mutual information $I(X;Y)$ which quantifies the dependency between the random variables, and is the natural effect for the G-test of independence (Section 2.2).

So we will explore the power of the surrogate G-tests of independence in order to detect effects expressed in terms of $I(X;Y)$. To do so, we will re-express the non-centrality parameters of the surrogate tests in terms of the non-centrality parameter of the unobservable test $\lambda_{G(X;Y)} = 2NI(X;Y)$.

Theorem 4.4 (MCAR: Informed surrogate approaches)**.**
In MCAR the non-centrality parameters of the four valid surrogate tests can be written in terms of the non-centrality parameter of the unobservable test as $\lambda_{G(X;\widetilde{Y}_*)} = \kappa\lambda_{G(X;Y)}$ *with the following correction factors:*

$$\begin{aligned}
&\textbf{Surrogate 1 } (\mathcal{D}_L): \ \kappa = p(s=1),\\
&\textbf{Surrogate 2 } (\widetilde{Y}_m): \ \kappa = p(s=1),\\
&\textbf{Surrogate 3 } (\widetilde{Y}_0): \ \kappa = \frac{1-p(y=1)}{1-p(y=1)p(s=1)}p(s=1),\\
&\textbf{Surrogate 4 } (\widetilde{Y}_1): \ \kappa = \frac{1-p(y=0)}{1-p(y=0)p(s=1)}p(s=1).
\end{aligned}$$

Proofs can be found in Appendix A.

A first conclusion from Theorem 4.4 is that all four surrogate tests have smaller non-centrality parameters than the fully-supervised test, and as a result smaller power. Furthermore, the first two tests have the same non-centrality parameter, but they do not share the same power, due to the different degrees of freedom; $\nu_{G(X;Y|s=1)} = (|\mathcal{X}|-1)$ while $\nu_{G(X;\widetilde{Y}_m)} = 2(|\mathcal{X}|-1)$. This happens because $\widetilde{Y}_m$ takes three values, while Y is binary. As a result, $G(X;Y|s=1)$ is more powerful than $G(X;\widetilde{Y}_m)$, because it has fewer degrees of freedom and the same non-centrality parameter (Agresti, 2013). Furthermore it holds; $\lambda_{G(X;Y|s=1)} > \lambda_{G(X;\widetilde{Y}_0)}$ and $\lambda_{G(X;Y|s=1)} > \lambda_{G(X;\widetilde{Y}_1)}$, and since all of these three tests have the same degrees of freedom we can derive the following corollary, which holds under the assumption that the labels are MCAR.

Corollary 4.5 (MCAR: Comparing the power of the surrogate tests)**.**
In MCAR the most powerful of the four surrogate approaches is surrogate 1, that is, to simply ignore the unlabelled data.

To verify experimentally the theoretical results that have been presented so far we will generate synthetic random variables X and Y with different degree of dependency and we plot figures similar to those of Gretton and Györfi (2010). To create the data, firstly we generated the values of Y, by taking N samples from a Bernoulli distribution with parameter $p(y=1)$. Then, we randomly chosen the parameters $p(x|y)$ that guaranteed the desired degree of dependency (expressed in terms of $I(X;Y)$) and we used these parameters to sample the values of X. In the x-axis of the figures we have different effect sizes in terms of mutual information

between X and Y, while in the y-axis we have the acceptance rate of the null hypothesis H_0 (over 1000 independent generations of the data). The y-intercept represents *1 - False Positive Rate*, and should be close to $1 - \alpha$ in order for the tests to be valid, while elsewhere the plots indicate the *False Negative Rate*. Figure 4.1 verifies Theorem 4.3, by showing that all four surrogate tests are valid, since all lines have the same intercept at $1 - \alpha = 0.90$ and as a result the four surrogate tests have the same false positive rate. Furthermore, all the surrogate approaches lead to tests with higher false negative rate, and this verifies Theorem 4.4 – the four tests have less power than the unobservable test. Finally, in that figure we observe that the most powerful among the surrogate approaches is to ignore the unlabelled examples, which verifies Corollary 4.5.

4.3 Testing when the Labels are MAR-F

When the labels are missing at random feature dependent, from the four surrogate approaches only one is valid to test the null hypothesis of independence.

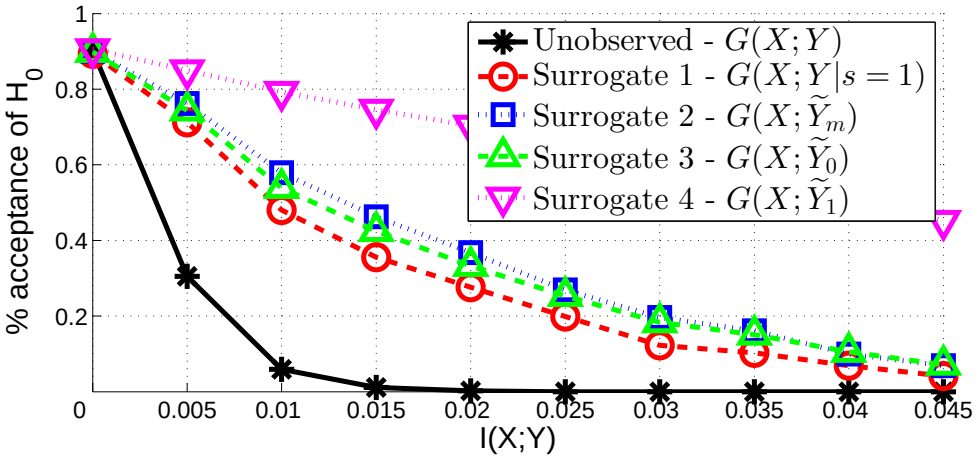
Theorem 4.6 (MAR-F: Which surrogate tests are *valid* for testing $X \perp\!\!\!\perp Y$?)**.** *In MAR-F we can test independence only by:*

$$\textbf{Surrogate 1 } (\mathcal{D}_L) : X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp Y | s = 1.$$

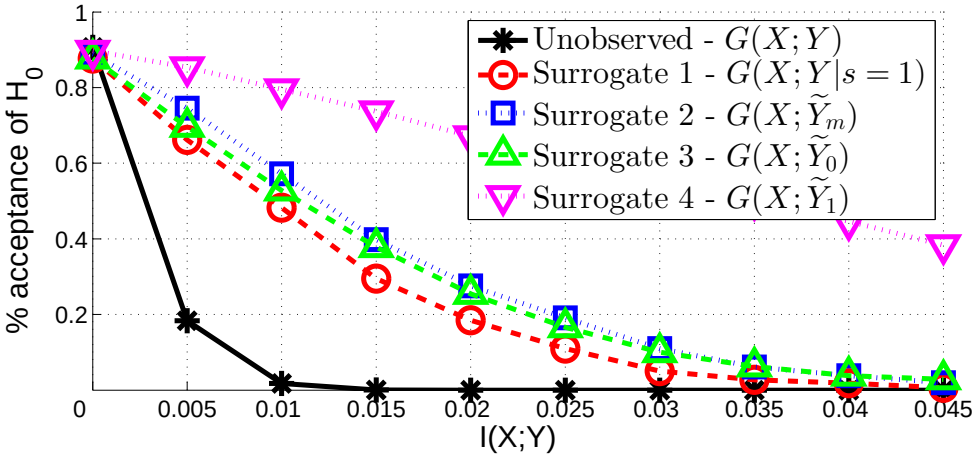
Proof sketch can be found in Appendix A.

Unfortunately, we cannot re-express the non-centrality parameter of the only valid test, $\lambda_{G(X;Y|s=1)}$, in terms of the supervised effect size $I(X;Y)$. So under this scenario, we cannot quantify a-priori the power of the only valid test $G(X;Y|s = 1)$, and as a result no surrogate approach is informed. This happens because the mutual information in the labelled set $I(X;Y|s = 1)$ cannot be re-expressed in terms of the population mutual information $I(X;Y)$.

We verify experimentally our observations in Figure 4.2. We observe that the only valid approach in this missingness scenario is to ignore the unlabelled examples, since the line of the $G(X;Y|s = 1)$-test and of the $G(X;Y)$-test have the same intercept.



(a) $N = 500$ *and* $|\mathcal{X}| = 2$



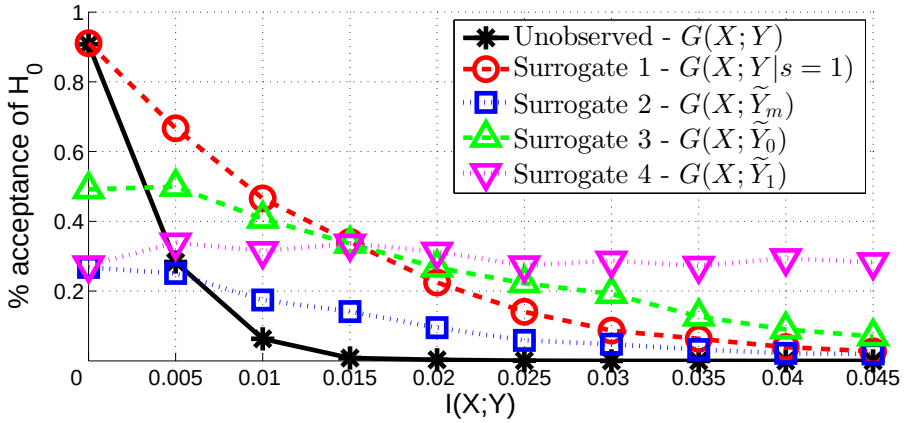
(b) $N = 1000$ *and* $|\mathcal{X}| = 5$

Figure 4.1: *Comparing the Type-I and Type-II error for the tests when the labels are missing completely at random (MCAR). For all figures we have* $\alpha = 0.10$. *In order to generate the semi-supervised dataset we used* $p(s = 1) = 0.25$.

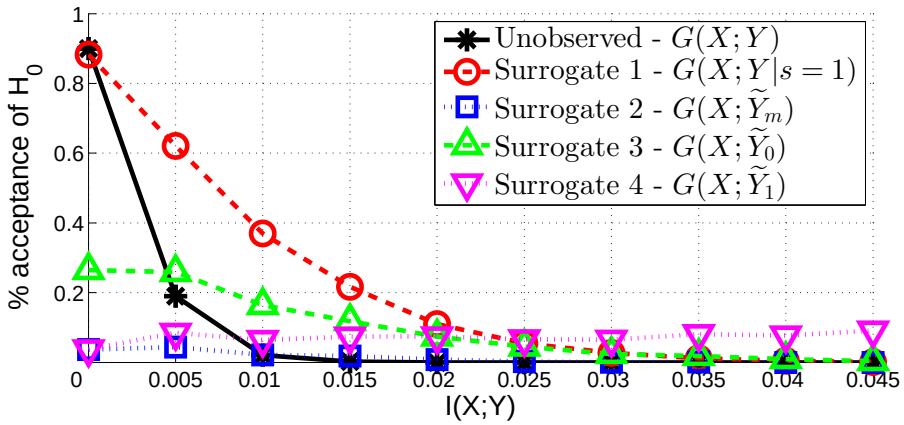
4.4 Testing when the Labels are MAR-C

When the labels are missing at random class dependent we have the same valid tests as in the missing completely at random scenario (Section 4.2).

Theorem 4.7 (MAR-C: Which surrogate tests are *valid* for testing $X \perp\!\!\!\perp Y$?)**.**



(a) $N = 500$ *and* $|\mathcal{X}| = 2$



(b) $N = 1000$ *and* $|\mathcal{X}| = 5$

Figure 4.2: *Comparing the Type-I and Type-II error for the tests when the labels are missing at random feature dependent (MAR-F). For all figures we have* $\alpha = 0.10$. *In order to generate the semi-supervised dataset, we used* $p(s = 1) = 0.25$ *and we label the data such that the marginal distribution of* X *in the labelled set is uniform:* $p(x|s = 1) = \frac{1}{|\mathcal{X}|}, \ \forall \ x \ \in \ \mathcal{X}$

In MAR-C we can test independence by using any of the four surrogate approaches:

$$\textbf{Surrogate 1 } (\mathcal{D}_L): \ X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp Y|s = 1,$$
$$\textbf{Surrogate 2 } (\widetilde{Y}_m): \ X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp \widetilde{Y}_m,$$
$$\textbf{Surrogate 3 } (\widetilde{Y}_0): \ X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp \widetilde{Y}_0,$$
$$\textbf{Surrogate 4 } (\widetilde{Y}_1): \ X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp \widetilde{Y}_1.$$

Proof sketches for each of these situations can be found in Appendix A.

With the following theorem we re-express the non-centrality parameters of the two valid surrogate tests, $G(X;\widetilde{Y}_0)$ and $G(X;\widetilde{Y}_1)$, in terms of the non-centrality parameter for the unobservable fully-supervised test, $G(X;Y)$.

Theorem 4.8 (MAR-C: Informed surrogate approaches)**.**
In MAR-C two of the valid tests are also informed with the following correction factors:

$$\begin{aligned}
\textbf{Surrogate 3 } (\widetilde{Y}_0): \kappa_{\widetilde{Y}_0} &= \frac{1-p(y=1)}{p(y=1)}\frac{p(\widetilde{y}_0=1)}{1-p(\widetilde{y}_0=1)} \\
&= \frac{1-p(y=1)}{p(y=1)}\frac{p(y=1,s=1)}{1-p(y=1,s=1)}, \\
\textbf{Surrogate 4 } (\widetilde{Y}_1): \kappa_{\widetilde{Y}_1} &= \frac{1-p(y=0)}{p(y=0)}\frac{p(\widetilde{y}_1=0)}{1-p(\widetilde{y}_1=0)} \\
&= \frac{1-p(y=0)}{p(y=0)}\frac{p(y=0,s=1)}{1-p(y=0,s=1)}.
\end{aligned}$$

Proofs can be found in Appendix A.

From the above theorem, we observe that by using *"exact" prior knowledge* over $p(y=1)$ and the probabilities of labelling, we can quantify the power of these two surrogate approaches.

Interestingly, to decide which of these two tests is more powerful we do not need exact prior knowledge, but we can do so by using some *"soft" prior knowledge* expressed in terms of inequality. In order to decide which approach is more powerful we need to compare $\kappa_{\widetilde{Y}_1}$ and $\kappa_{\widetilde{Y}_0}$. For example, $G(X;\widetilde{Y}_0)$ is more powerful than $G(X;\widetilde{Y}_1)$ when $\kappa_{\widetilde{Y}_0} > \kappa_{\widetilde{Y}_1}$, which results to the following inequality:

$$\begin{aligned}
p(y=1) &< \frac{1}{1+\sqrt{\frac{(1-p(\widetilde{y}_0=1))p(\widetilde{y}_1=0)}{p(\widetilde{y}_0=1)(1-p(\widetilde{y}_1=0))}}} \Leftrightarrow \\
p(y=1) &< \frac{1}{1+\sqrt{\frac{(1-p(y=1,s=1))p(y=0,s=1)}{p(y=1,s=1)(1-p(y=0,s=1))}}}.
\end{aligned} \tag{4.1}$$

When the opposing inequality holds, the most powerful choice is $G(X;\widetilde{Y}_1)$. When equality holds, both approaches are equivalent. So, by observing $p(y=1,s=1)$ and $p(y=0,s=1)$ from the labelled data, we can use some "soft" prior knowledge over $p(y=1)$ to decide the most powerful option.

Furthermore, these correction factors enable us to use the $G(X;\widetilde{Y}_0)$ and/or $G(X;\widetilde{Y}_1)$ instead of $G(X;Y)$ for power analysis and sample size determination. Taking advantage of the extra degree of freedom in $p(y=1,s=1)$ and/or $p(y=0,s=1)$, we can also determine the *required level of supervision* (i.e. number of labelled examples) needed, following the same procedure as in sample

size determination. Section 7 presents a complete methodology for sample size and labelled set size determination in partially labelled data.

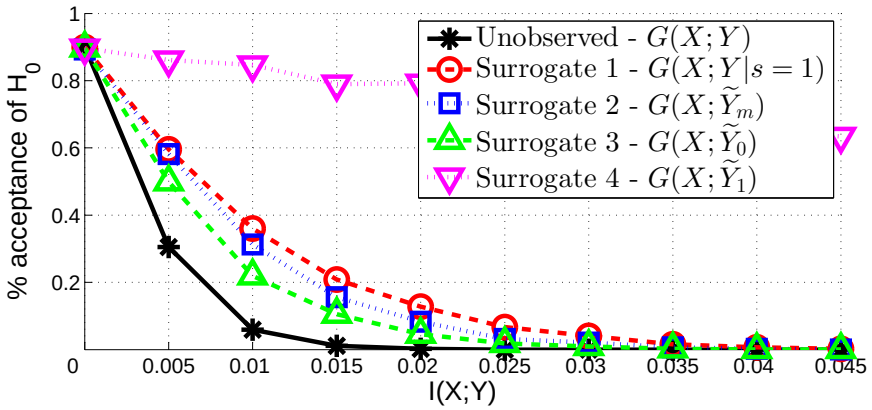
Unfortunately, we cannot derive any similar conclusion for the other two valid approaches, since we cannot express their non-centrality parameters, $\lambda_{G(X;Y|s=1)}$ and $\lambda_{G(X;\widetilde{Y}_m)}$, in terms of the non centrality parameter for the unobservable fully-supervised test, $\lambda_{G(X;Y)}$. But combining our findings on the MAR-C scenario with our findings on the MCAR scenario (Section 4.2), we can consider a useful conjecture. Before that, we should mention that the MCAR scenario can be seen as a restricted version of the MAR-C[1].

Conjecture 4.9 (MAR-C: Comparing the power of the tests)**.**
The closer we are to the MCAR assumption, i.e. $D_{KL}(p(y)||p(y|s=1)) \approx 0$*, then* **Surrogate 1**, $G(X;Y|s=1)$*, will have the highest statistical power. In contrast, the closer we are to extreme MAR-C scenarios, i.e.* $D_{KL}(p(y)||p(y|s=1)) \gg 0$*, then either* **Surrogate 3** *or* **Surrogate 4**, *that is* $G(X;\widetilde{Y}_0)$ *or* $G(X;\widetilde{Y}_1)$*, will have the highest power. In this latter scenario we can identify which of the two will be most powerful using inequality (4.1).*

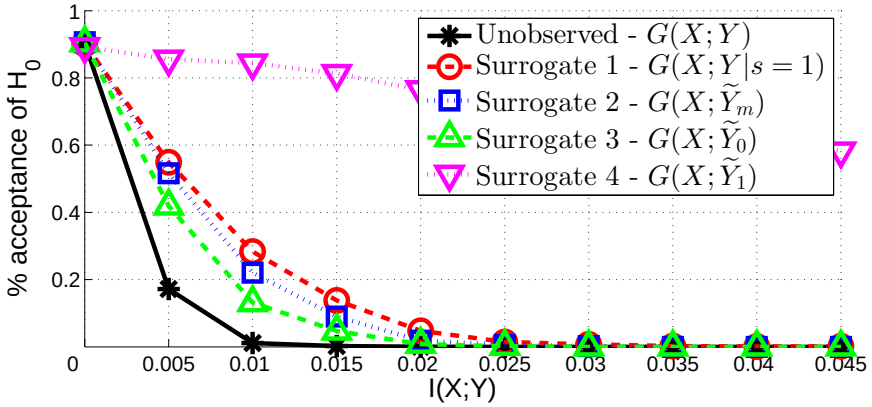
Figures 4.3 and 4.4 verify experimentally our findings and show that using any of the four surrogate tests is a valid approach, since all of the lines have the same intercept (at $1-\alpha$) and, as a result, the tests have the same false positive rate. Furthermore, by incorporating "soft" prior knowledge over $p(y=1)$ and using inequality (4.1), we can decide which of the two tests, $G(X;\widetilde{Y}_0)$ or $G(X;\widetilde{Y}_1)$, is more powerful. For the first setting, showed in Figures 4.3a and 4.3b, we have $p(y=1,s=1) = p(y=0,s=1) = 0.125$, so the *rhs* of inequality (4.1) is equal to 0.50. And by using "soft" knowledge that $p(y=1)$ is less than this value we can conclude that $G(X;\widetilde{Y}_0)$ is more powerful than $G(X;\widetilde{Y}_1)$. Figures 4.3a and 4.3b verify this conclusion. The same also holds for the second setting captured by Figures 4.4a and 4.4b, where we have $p(y=1,s=1) = 0.05$ and $p(y=0,s=1) = 0.15$ and the *rhs* of inequality (4.1) becomes 0.35. Again, by using "soft" knowledge over $p(y=1)$, we can conclude that $G(X;\widetilde{Y}_0)$ is more powerful than $G(X;\widetilde{Y}_1)$.

By comparing the first setting (Figures 4.3a and 4.3b) with the second setting (Figures 4.4a and 4.4b), we can verify our Conjecture 4.9. In the first setting, the MAR-C is more extreme since we have $p(y=1|s=1) = 0.50$ while the population

[1]When the MAR-C holds we have $p(s=1|\mathbf{x},y) = p(s=1|y)$, and we can derive the MCAR if we furthermore assume $p(s=1|y) = p(s)$ for each $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$.



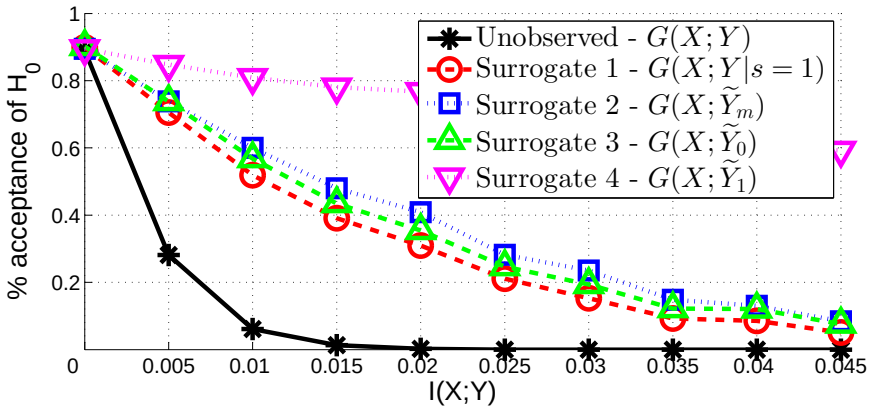
(a) $|\mathcal{X}| = 2$, $N = 500$ *and* $p(s = 1) = 0.25$



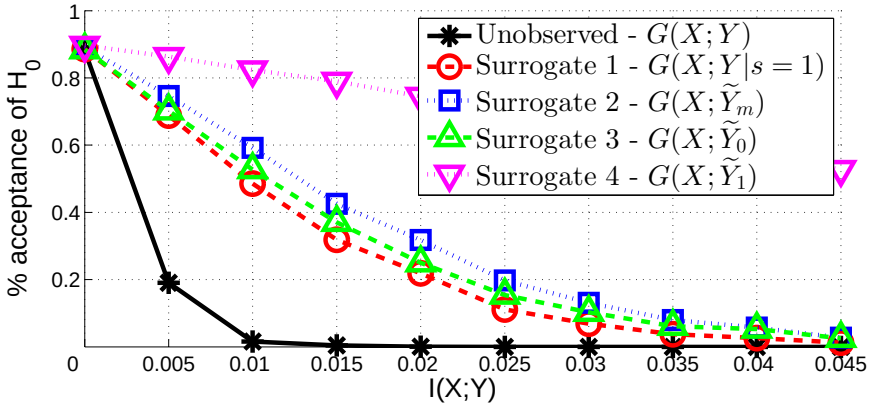
(b) $|\mathcal{X}| = 5$, $N = 1000$ *and* $p(s = 1) = 0.25$

Figure 4.3: *Comparing the Type-I and Type-II error for the tests when the labels are missing at random class dependent, in an extreme MAR-C scenario. For all figures we have* $\alpha = 0.10$ *and* $p(y = 1) = 0.20$. *In order to generate the semi-supervised dataset under the MAR-C assumption, we label the data such that* $p(y = 1, s = 1) = p(y = 0, s = 1) = 0.125$ *or in other words* $p(y = 1|s = 1) = 0.50$ *— an extreme MAR-C scenario since* $D_{KL}(p(y)||p(y|s = 1)) = 0.19$.

prior is much lower $p(y = 1) = 0.20$. So, in this scenario, using the unlabelled examples assuming that they belong to the negative class outperforms the other approaches. In the second setting we are closer to the MCAR assumption, since the probability in the labelled set $p(y = 1|s = 1) = 0.25$ is very close to the population prior $p(y = 1) = 0.20$. As a result, in this scenario we can see that ignoring the unlabelled examples is a more powerful option, as we proved for the MCAR scenario.



(a) $|\mathcal{X}| = 2$, $N = 500$ *and* $p(s = 1) = 0.20$



(b) $|\mathcal{X}| = 5$, $N = 1000$ *and* $p(s = 1) = 0.20$

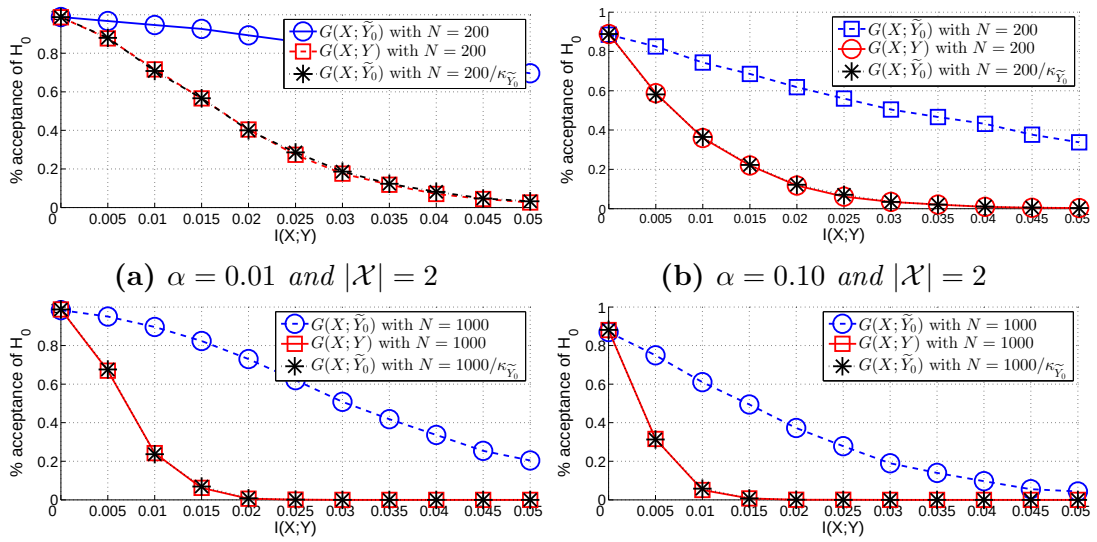
Figure 4.4: *Comparing the Type-I and Type-II error for the tests when the labels are missing at random class dependent, in a MAR-C scenario close to MCAR. For all figures we have* $\alpha = 0.10$ *and* $p(y = 1) = 0.20$. *In order to generate the semi-supervised dataset under the MAR-C assumption we label the data such that* $p(y = 1, s = 1) = 0.05$ *and* $p(y = 0, s = 1) = 0.15$, *or in other words* $p(y = 1|s = 1) = 0.25$ — *a MAR-C scenario close to MCAR since* $D_{KL}(p(y)||p(y|s = 1)) = 0.01$.

An interesting point to mention is that our analysis in this section can be also used when we have labelled examples from one class, such as the positive-unlabelled setting. Under the PU constraint, the surrogate variable of assuming all unlabelled examples being negative ($\widetilde{Y}_0$) is valid and it is also informed by incorporating prior knowledge over $p(y = 1)$. As a result, we can use the $G(X;\widetilde{Y}_0)$-test for experimental design activities, such as sample size determination. This application of our work is presented in Chapter 7.

4.5 Verification of the Correction Factors

An important outcome of our analysis so far is the derivation of the correction factors $\kappa_{\widetilde{Y}_0}, \kappa_{\widetilde{Y}_1}$. For example, in Chapters 7 and 8 we use these correction factors for sample size determination and Markov blanket discovery in partially labelled data. In this section we will verify experimentally the correctness of these factors. We will focus on the $G(X;\widetilde{Y}_0)$-test and its correction factor $\kappa_{\widetilde{Y}_0}$ when the labels are MAR-C, but our results can be extended to any informed test presented so far. We focus on this surrogate test, because it is also observed in the positive-unlabelled scenario.

As a sanity check, in Figure 4.5, we observe that if we increase the sample size of the test between X and $\widetilde{Y}_0$ by a factor $\kappa_{\widetilde{Y}_0}$, the two tests have the same power, and this result verifies Theorem 4.8. No matter what the sample size is, the intercepts are always at the same value (close to the design parameter $1-\alpha$), which again verifies that the surrogate variable $\widetilde{Y}_0$ is valid.



(a) $\alpha = 0.01$ *and* $|\mathcal{X}| = 2$ **(b)** $\alpha = 0.10$ *and* $|\mathcal{X}| = 2$

(c) $\alpha = 0.01$ *and* $|\mathcal{X}| = 10$ **(d)** $\alpha = 0.10$ *and* $|\mathcal{X}| = 10$

Figure 4.5: *Comparing the Type-II error for the unobservable $G(X;Y)$-test and the surrogate $G(X;\widetilde{Y}_0)$-test with and without corrected sample size when the labels are missing at random class dependent (MAR-C). For these figures we have $p(y=1) = 0.20$ and we labelled only 5% of the examples as positives $p(y=1, s=1) = p(\widetilde{y}_0 = 1) = 0.05$.*

4.6 Testing Conditional Independence in Partially Labelled Data

The results that we proved for the testing in MCAR (Section 4.2) and MAR-C (Section 4.4) can be extended to conditional tests. The MCAR extension is straightforward since under this scenario holds the unconditional independence described in Figure 3.1a. Deriving the results in MAR-C is more challenging, and this is the focus of the current section. Firstly we will show that testing conditional independence by assuming the unlabelled examples to be either positive or negative is a valid approach.

Theorem 4.10 (MAR-C: Which surrogate tests are valid for testing $X \perp\!\!\!\perp Y|\mathbf{Z}$?)**.** *In MAR-C we can test conditional independence by the following two surrogate approaches:*

$$\textbf{Surrogate 3 } (\widetilde{Y}_0): \; X \perp\!\!\!\perp Y|\mathbf{Z} \Leftrightarrow X \perp\!\!\!\perp \widetilde{Y}_0|\mathbf{Z},$$
$$\textbf{Surrogate 4 } (\widetilde{Y}_1): \; X \perp\!\!\!\perp Y|\mathbf{Z} \Leftrightarrow X \perp\!\!\!\perp \widetilde{Y}_1|\mathbf{Z}.$$

Proof sketches can be found in Appendix A.

The consequence of this theorem is that the derived conditional tests of independence are valid, but it does not tell us anything about what is happening when the alternative hypothesis holds. To explore that, we will quantify the amount of power that we are loosing by assuming that all unlabelled examples are negative (i.e. using $\widetilde{Y}_0$) or positive (i.e. using $\widetilde{Y}_1$).

Theorem 4.11 (MAR-C: Informed surrogate approaches for conditional testing)**.** *These two valid tests are also informed with the following correction factors:*

$$\begin{aligned}
\textbf{Surrogate 3 } (\widetilde{Y}_0): \kappa_{\widetilde{Y}_0} &= \frac{1-p(y=1)}{p(y=1)} \frac{p(\widetilde{y}_0=1)}{1-p(\widetilde{y}_0=1)} \\
&= \frac{1-p(y=1)}{p(y=1)} \frac{p(y=1,s=1)}{1-p(y=1,s=1)}, \\
\textbf{Surrogate 4 } (\widetilde{Y}_1): \kappa_{\widetilde{Y}_1} &= \frac{1-p(y=0)}{p(y=0)} \frac{p(\widetilde{y}_1=0)}{1-p(\widetilde{y}_1=0)} \\
&= \frac{1-p(y=0)}{p(y=0)} \frac{p(y=0,s=1)}{1-p(y=0,s=1)}.
\end{aligned}$$

Proofs can be found in Appendix A.

A useful observation is that the conditional tests have exactly the same correction factors as the unconditional tests presented in Theorem 4.8. Chapter 8 presents how we can use the results of this section in order to use "exact" or "soft" prior knowledge in feature selection through Markov blanket discovery procedures.

4.7 Chapter Summary

In this section we explored valid and informed surrogate approaches to test independence and conditional independence in partially labelled data. When the labels are MCAR, we can use any of the four surrogate approaches to test independence and achieve a desirable false positive rate, but in order to minimize the false negative rate we should ignore the unlabelled data. When the labels are MAR-F, we have only one way that guarantees that we can achieve a desired level of false positive rate, but unfortunately this way is not informed over the false negatives. Finally, when the labels are MAR-C we can control the false positive rate with all of the four surrogate approaches. Furthermore, by assuming unlabelled examples as positives or negatives and an "exact" knowledge of $p(y)$ we can have an informed decision over the false negatives, while using inequality (4.1) we can decide using "soft" prior knowledge which is the optimal choice between these two surrogate approaches.

When we only have labelled examples from one class, such as positive-unlabelled data, the surrogate approach of assuming the unlabelled examples belong to the other class is valid. It is also informed, since we can control the false positive rate by incorporating prior knowledge over $p(y = 1)$ and using our derived correction factor $\kappa_{\widetilde{Y}_0}$. In Chapter 7 we present how we can use this approach for experimental design activities, such as sample size determination. Chapter 8 presents how we can our correction factors and "exact" or "soft" prior knowledge in feature selection through Markov blanket discovery procedures. In the next chapter we present a theoretical analysis on how to estimate mutual information in the different partially labelled scenarios.

Chapter 5

Theoretical Analysis of Effect Size Estimation in Partially Labelled Data

Hypothesis testing inherently involves a cut-off point, beyond which we consider the result significant or non-significant. In most scenarios we will benefit from knowing a good estimate of the effect size, enabling presentation of confidence intervals, for example. In this section we present a novel set of methodologies for deriving *consistent* estimators of the mutual information in an entirely model-independent/inference-free manner when the labels are MCAR (Section 5.2), MAR-F (Section 5.3) and MAR-C (Section 5.4).

The main challenge is how to derive useful information on the joint probability $p(x, y)$ from the labelled data, while using the unlabelled in order to have a more accurate estimate of the marginal probabilities $p(x)$ and/or $p(y)$. This technique, estimating some parameters from the labelled set while others from the labelled and unlabelled, is known in statistics as *available-case analysis* or *pairwise deletion* (Enders, 2010). Our approach is to suggest consistent estimators by re-expressing mutual information in different ways for the different missingness scenarios, which will enable us to use the labelled and unlabelled information in an efficient manner.

5.1 Re-expressing Mutual Information in Three Different Ways

As we already stated, the main question we try to answer in this section is how we can use the labelled and the unlabelled data in order to suggest consistent estimators. To do so, we will explore the two different ways that the joint probability can be factorized; $\widehat{p}(x,y) = \widehat{p}(y|x)\widehat{p}(x) = \widehat{p}(x|y)\widehat{p}(y)$. By using them we can derive the following equivalent expressions for estimating the mutual information:

$$\widehat{I}(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \widehat{p}(x|y)\widehat{p}(y)\ln\frac{\widehat{p}(x|y)}{\widehat{p}(x)}, \tag{5.1}$$

$$\widehat{I}(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \widehat{p}(y|x)\widehat{p}(x)\ln\frac{\widehat{p}(y|x)}{\widehat{p}(y)} \tag{5.2}$$

$$\widehat{I}(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \widehat{p}(y|x)\widehat{p}(x)\ln\frac{\widehat{p}(x|y)}{\widehat{p}(x)} \tag{5.3}$$

It turns out that each of these *different (but equivalent) expressions for the mutual information are suitable for different partially labelled environments.* For example, when the missingness mechanism is MAR-C, the m-graph in Figure 3.1c shows us clearly that the class conditional probability in the labelled set is the same as the population one, $p(x|y,s=1) = p(x|y)$. As a result, expression (5.1) is more suitable in this scenario since it uses this class-conditional probability.

When the missingness mechanism is MAR-F, the m-graph in Figure 3.1b shows that the conditional probability in the labelled set is the same as the population one $p(y|x,s=1) = p(y|x)$, so expression (5.2) is more suitable for this scenario, since it uses this conditional probability.

Finally, when the missingness mechanism is MCAR (m-graph in Figure 3.1a) both the conditional and the class conditional probabilities in the labelled set are the same with their population values. In this scenario we can take maximal advantage of the unlabelled data by using (5.3), which calculates the marginal probability of the features $p(x)$ two times, and so by using both labelled and unlabelled we can improve the accuracy of the overall estimator. In the next sections we present how we can use these re-expressions to provide consistent estimators despite the partial supervision.

5.2 Estimating Mutual Information when the Labels are MCAR

In this scenario, the estimator derived using only the labelled data is consistent, since it holds that $I(X;Y|s=1) = I(X;Y)$. We can get more insight by re-writing this estimator using equation (5.3).

$$\widehat{I}(X;Y|s=1) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \widehat{p}(x|s=1)\widehat{p}(y|x,s=1)\ln\frac{\widehat{p}(x|y,s=1)}{\widehat{p}(x|s=1)}. \tag{5.4}$$

Although this estimator is consistent, the question remains: "Is there any way of incorporating the unlabelled examples to improve the accuracy?" To estimate (5.4) we need $\widehat{p}(x|s=1)$, but since in MCAR it holds that $p(x|s=1) = p(x)$, we can derive a better estimator by using all data instead of using only the labelled set. With the following theorem we take advantage of these re-writing and we suggest a consistent estimator for $I(X;Y)$ by using both the labelled and the unlabelled information.

Lemma 5.1 (MCAR: Semi-supervised Estimator for $I(X;Y)$)**.**
When the labels are missing completely at random, the following estimator

$$\widehat{I}_{MCAR}(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \widehat{p}(x)\widehat{p}(y|x,s=1)\ln\frac{\widehat{p}(x|y,s=1)}{\widehat{p}(x)}, \tag{5.5}$$

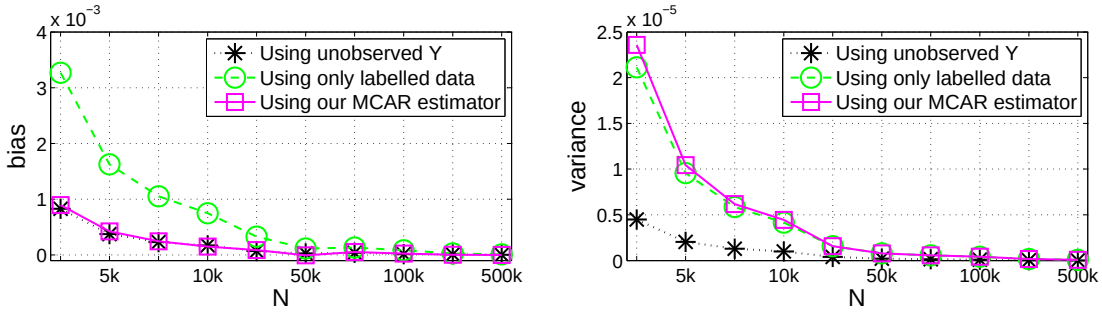
is a consistent estimator for $I(X;Y)$.

The proof is straightforward, since $I_{MCAR}(X;Y) = I(X;Y)$ under the MCAR assumption. We have reached the limit of what a theoretical analysis can tell us about our suggested estimator, but by comparing (5.4) and (5.5), we see that their only difference is that in $\widehat{I}_{MCAR}(X;Y)$ we use all the data to estimate $p(x)$, while in $\widehat{I}(X;Y|s=1)$ only the labelled examples. As a result, we suggest the following conjecture over the accuracy of the two estimators:

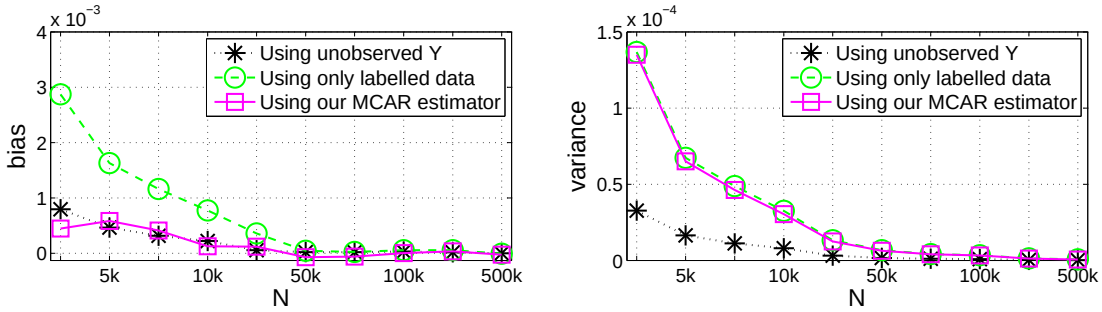
Conjecture 5.2 (MCAR: Most accurate estimator)**.**
To estimate $I(X;Y)$, our suggested estimator $\widehat{I}_{MCAR}(X;Y)$ is more accurate than using only the labelled data $\widehat{I}(X;Y|s=1)$ because it uses all the data for the estimation of the marginal probability $p(x)$.

These results are experimentally demonstrated in Figure 5.1. As we observe, for all the settings, $\widehat{I}_{MCAR}(X;Y)$ has a bias similar to the unobservable estimator

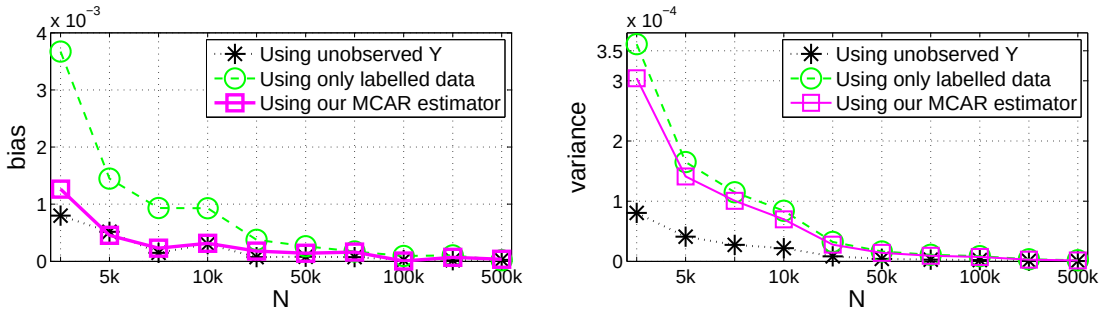
$\widehat{I}(X;Y)$, while its variance is similar to $\widehat{I}(X;Y|s=1)$ – the estimator using only labelled data. So, overall accuracy improves by using the unlabelled examples with our framework.



(a) *Small effect* — $I(X;Y) = 0.005$ *with* $|\mathcal{X}| = 5$.



(b) *Medium effect* — $I(X;Y) = 0.045$ *with* $|\mathcal{X}| = 5$.



(c) *Large effect* — $I(X;Y) = 0.125$ *with* $|\mathcal{X}| = 5$.

Figure 5.1: *Comparing the performance of our suggested MCAR semi-supervised estimator with the estimator using only the labelled data* $\widehat{I}(X;Y|s=1)$ *and the fully supervised unobservable estimator* $\widehat{I}(X;Y)$ *in terms of bias/variance. To estimate bias/variance we average over 2000 runs. The semi-supervised data were generated through MCAR with probability of labelling* $p(s=1) = 0.25$. *Please note different axes for the variance of the different effect levels.*

5.3 Estimating Mutual Information when the Labels are MAR-F

In this scenario, the estimator derived using only the labelled data is not consistent, since it holds $I(X;Y|s=1) \neq I(X;Y)$. So the question we try to answer can be phrased as: "Is there any way of incorporating the unlabelled examples to derive a consistent estimator for the mutual information?"

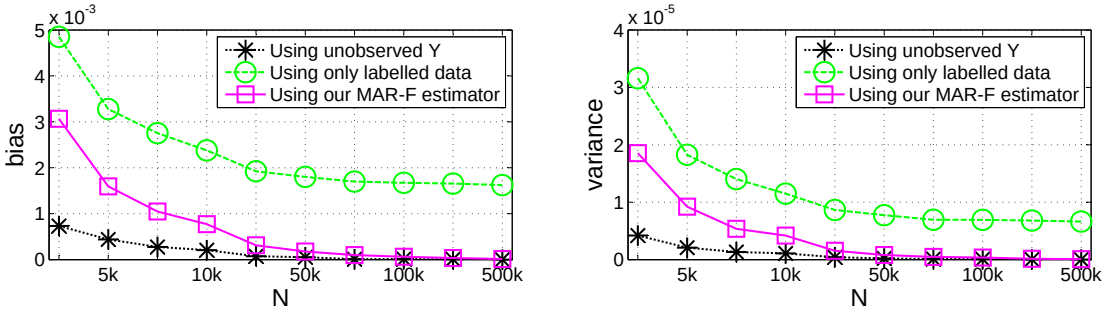
In order to calculate the mutual information, we need the joint distribution $p(x,y)$. Mohan et al. (2013) presented how we can use m-graphs to recover this distribution through a list-wise deletion procedure. With our notation it holds that $p(x,y) = p(y|x)p(x) = p(y|x,s=1)p(x)$. For the first step we used the chain rule; for the second, we used the fact that when the labels are MAR-F it holds that $p(y|x) = p(y|x,s=1)$. So we can calculate the joint distribution through calculating the conditional $p(y|x,s=1)$ from the labelled set, and the marginal $p(x)$ from both the labelled and the unlabelled set. With the following Lemma we show that we can use this methodology to derive a consistent estimator for the $I(X;Y)$ despite the partial supervision and without using any prior knowledge.

Lemma 5.3 (MAR-F: Semi-supervised Estimator for $I(X;Y)$)**.**
When the labels are missing at random feature dependent, the following estimator

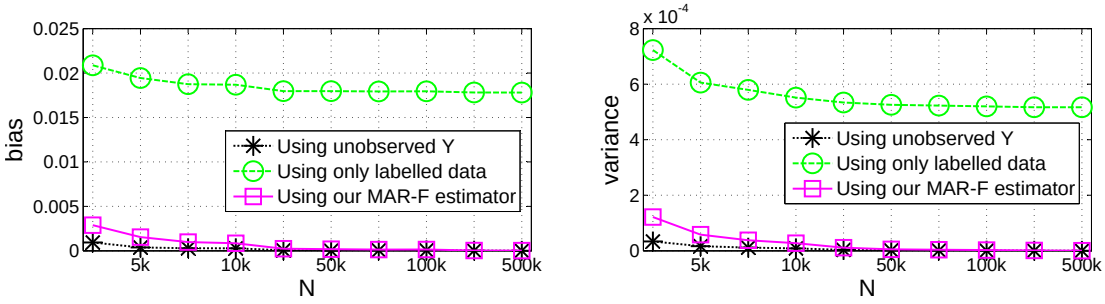
$$\widehat{I}_{MAR-F}(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \widehat{p}(x)\widehat{p}(y|x,s=1)\ln\frac{\widehat{p}(y|x,s=1)}{\sum_{x'\in\mathcal{X}}\widehat{p}(y|x',s=1)\widehat{p}(x')} \tag{5.6}$$

is a consistent estimator for $I(X;Y)$.

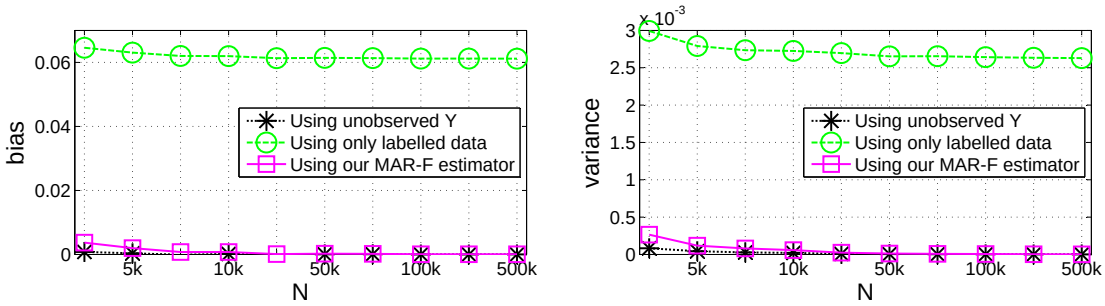
The proof is straightforward, since $I_{MAR-F}(X;Y) = I(X;Y)$ under the MAR-F assumption. Again, we present an experimental demonstration of these results in Figure 5.2. As we observe for all the settings, our estimator that takes into account the unlabelled information is a consistent estimator, since both bias and variance decrease as the sample size gets larger, while by ignoring the unlabelled examples we get an asymptotically biased estimator.



(a) *Small effect* — $I(X;Y) = 0.005$ *with* $|\mathcal{X}| = 5$.



(b) *Medium effect* — $I(X;Y) = 0.045$ *with* $|\mathcal{X}| = 5$.



(c) *Large effect* — $I(X;Y) = 0.125$ *with* $|\mathcal{X}| = 5$.

Figure 5.2: *Comparing the performance of our suggested MAR-F semi-supervised estimator with the estimator using only the labelled data* $\widehat{I}(X;Y|s=1)$ *and the fully supervised unobservable estimator* $\widehat{I}(X;Y)$ *in terms of bias/variance. To estimate bias/variance we average over 2000 runs. The semi-supervised data were generated through MAR-F with probability of labelling an example* $p(s=1) = 0.25$ *and the marginal distribution of* X *in the labelled set is uniform:* $p(x|s=1) = \frac{1}{|\mathcal{X}|}$, $\forall\ x\ \in\ \mathcal{X}$. *Please note different axes for the bias/variance of the different effect levels.*

5.4 Estimating Mutual Information when the Labels are MAR-C

In this scenario the estimator derived using only the labelled data is not consistent, since it holds $I(X;Y|s=1) \neq I(X;Y)$. So the question we try to answer remains the same: "Is there any way of incorporating the unlabelled examples to derive a consistent estimator for the mutual information?"

Again, we will try to explore whether we can calculate the joint distribution from the available information. When the labels are MAR-C, we know that the class conditional distribution in the labelled set is the same with the population class conditional distribution $p(x|y,s=1) = p(x|y)$. By using the chain rule we can re-write the joint distribution as $p(x,y) = p(x|y)p(y) = p(x|y,s=1)p(y)$. Unfortunately, this time we cannot estimate directly the $p(y)$ — therefore, as observed by Mohan and Pearl (2014), the query $p(x,y)$ is *not recoverable* from raw data. However, we will proceed to show that by combining the data with prior knowledge over $p(y)$, the mutual information can in fact still be recovered. One way to incorporate this prior knowledge is through the following estimator:

$$\widehat{I}^{SS}_{MAR-C}(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(y)\widehat{p}(x|y,s=1)\ln\frac{\widehat{p}(x|y,s=1)}{\widehat{p}(x)}. \tag{5.7}$$

This is a consistent estimator for the mutual information, since $I^{SS}_{MAR-C}(X;Y) = I(X;Y)$ under the MAR-C assumption.

Interestingly, in this scenario, we can have consistent estimators by using the labelled information of each class independently. Using prior knowledge over the true probability of an example being positive $p(y=1)$, we can derive with the following theorem a consistent estimator for the $I(X;Y)$ by using only the *positively labelled information.*

Theorem 5.4 (MAR-C: Positive information estimator for $I(X;Y)$)**.** *When the labels are missing at random class dependent, we can use prior knowledge over $p(y=1)$ and derive the following consistent estimator for $I(X;Y)$*

$$\widehat{I}^{Pos}_{MAR-C}(X;Y) = \sum_{x\in\mathcal{X}} p(y=1)\widehat{p}(x|y=1,s=1)\ln\frac{\widehat{p}(x|y=1,s=1)}{\widehat{p}(x)} \tag{5.8}$$
$$+\sum_{x\in\mathcal{X}}(\widehat{p}(x)-p(y=1)\widehat{p}(x|y=1,s=1))\ln\frac{\widehat{p}(x)-p(y=1)\widehat{p}(x|y=1,s=1)}{\widehat{p}(x)\,(1-p(y=1))}.$$

The asymptotic sampling distribution of this estimator is

$$\widehat{I}^{Pos}_{MAR-C}(X;Y) \sim N\left(I(X;Y), \frac{\sigma^2_{Pos}}{N}\right),$$

where

$$\sigma^2_{Pos} = \sum_{x\in\mathcal{X}}\left(p(x,y=1,s=1)\phi^2_{x,y=1,s=1} + (p(x)-p(x,y=1,s=1))\phi^2_{x,\widetilde{y}_0=0}\right)$$
$$-\left(\sum_{x\in\mathcal{X}}(p(x,y=1,s=1)\phi_{x,\widetilde{y}_0=1} + (p(x)-p(x,y=1,s=1))\phi_{x,\widetilde{y}_0=0})\right)^2$$

and

$$\phi_{x,\widetilde{y}_0=1} = \ln\frac{p(x)-p(x|y=1,s=1)p(y=1)}{p(x)}$$
$$+\frac{p(y=1)}{p(y=1,s=1)}\sum_{x'\in\mathcal{X}}(p(x'|y=1,s=1)-\delta_{xx'})\ln\frac{p(x')-p(x'|y=1,s=1)p(y=1)}{p(x'|y=1,s=1)p(y=1)}$$
$$\phi_{x,\widetilde{y}_0=0} = \ln\frac{p(x)-p(x|y=1,s=1)p(y=1)}{p(x)}$$

Proof can be found in Appendix A.

This estimator is very useful, since it can be utilized in the positive-unlabelled scenario under the selected completely at random assumption (more details in Section 3.3.3). In this scenario it holds that $p(x|y=1,s=1)=p(x|y=1)$, but we do not have any available labelled information for the negative class. Despite this labelling constraint, we can still derive a consistent estimator for $I(X;Y)$ using Theorem 5.4.

Interestingly, by interchanging the alphabet of positive and negative classes in the previous theorem, we can derive a consistent estimator for the $I(X;Y)$ by

incorporating prior knowledge in terms of $p(y=0)$ and using only the *negatively labelled information*.

Corollary 5.5 (MAR-C: Negative information estimator for $I(X;Y)$)**.** *When the labels are missing at random class dependent, we can use prior knowledge over $p(y=0)$ and derive the following consistent estimator for $I(X;Y)$*

$$\widehat{I}^{Neg}_{MAR-C}(X;Y) = \sum_{x\in\mathcal{X}} p(y=0)\widehat{p}(x|y=0,s=1)\ln\frac{\widehat{p}(x|y=0,s=1)}{\widehat{p}(x)} \tag{5.9}$$
$$+\sum_{x\in\mathcal{X}}(\widehat{p}(x)-p(y=0)\widehat{p}(x|y=0,s=1))\ln\frac{\widehat{p}(x)-p(y=0)\widehat{p}(x|y=0,s=1)}{\widehat{p}(x)\,(1-p(y=0))}$$

The asymptotic sampling distribution of this estimator is

$$\widehat{I}^{Neg}_{MAR-C}(X;Y) \sim N\left(I(X;Y), \frac{\sigma^2_{Neg}}{N}\right),$$

where

$$\sigma^2_{Neg} = \sum_{x\in\mathcal{X}}\left(p(x,y=0,s=1)\phi^2_{x,\widetilde{y}_1=0} + (p(x)-p(x,y=0,s=1))\phi^2_{x,\widetilde{y}_1=1}\right)$$
$$-\left(\sum_{x\in\mathcal{X}}(p(x,y=0,s=1)\phi_{x,\widetilde{y}_1=0} + (p(x)=p(x,y=0,s=1))\phi_{x,\widetilde{y}_1=1})\right)^2,$$

and

$$\phi_{x,\widetilde{y}_1=0} = \ln\frac{p(x)-p(x|y=0,s=1)p(y=0)}{p(x)},$$
$$+\frac{p(y=0)}{p(y=0,s=1)}\sum_{x'\in\mathcal{X}}(p(x'|y=0,s=1)-\delta_{xx'})\ln\frac{p(x')-p(x'|y=0,s=1)p(y=0)}{p(x'|y=0,s=1)p(y=0)}$$
$$\phi_{x,\widetilde{y}_1=1} = \ln\frac{p(x)-p(x|y=0,s=1)p(y=0)}{p(x)}.$$

Proof can be found in Appendix A.

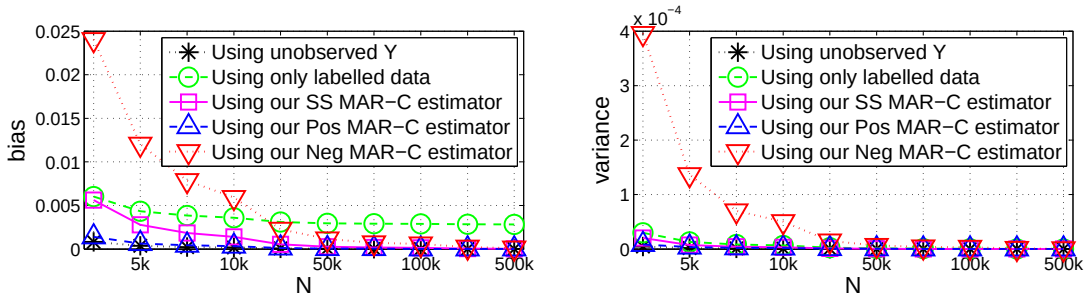
We have reached the limit of what a theoretical analysis can tell us about our suggested estimators, but at this point, an interesting question raises "Which is more trustworthy labelled information?" or in other words "Which of the last two estimators is better?" Because the labelling is class dependent, one class will be over-represented in the labelled set, and the other, under-represented. In order to

decide which of the two estimators is more accurate, we can use our results from testing when the labels are MAR-C (Section 4.4) and decide the most accurate choice through the following conjecture.

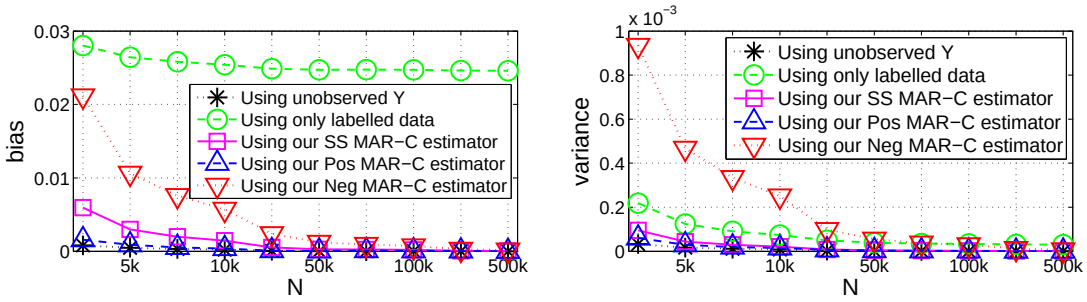
Conjecture 5.6 (MAR-C: Most accurate estimator between Pos and Neg)**.** *In the MAR-C scenario when the inequality (4.1) holds $\widehat{I}^{Pos}_{MAR-C}$ is more accurate than $\widehat{I}^{Neg}_{MAR-C}$. When the opposing inequality holds, the inverse relationship holds. When equality holds, both approaches are equivalent.*

To motivate this conjecture we can make the following observations. The positive information estimator, $\widehat{I}^{Pos}_{MAR-C}(X;Y)$, uses the same probability estimates with the ones used in $\widehat{I}(X;\widetilde{Y}_0)$, but instead of estimating the probability of a positively labelled example, it incorporates prior knowledge over the probability of the positive class. On the other hand, $\widehat{I}^{Neg}_{MAR-C}(X;Y)$ uses the same probability estimates with the ones used in $\widehat{I}(X;\widetilde{Y}_1)$, but instead of estimating the probability of a negatively labelled example, it incorporates prior knowledge over the probability of the negative class. So deciding the most accurate estimator between I^{Pos}_{MAR-C} and I^{Neg}_{MAR-C}, is similar to deciding the most accurate between $\widehat{I}(X;\widetilde{Y}_0)$ and $\widehat{I}(X;\widetilde{Y}_1)$. But we can answer to the latter question by using our findings in hypothesis testing (Section 4.4), since the most powerful test leads to more accurate estimators.

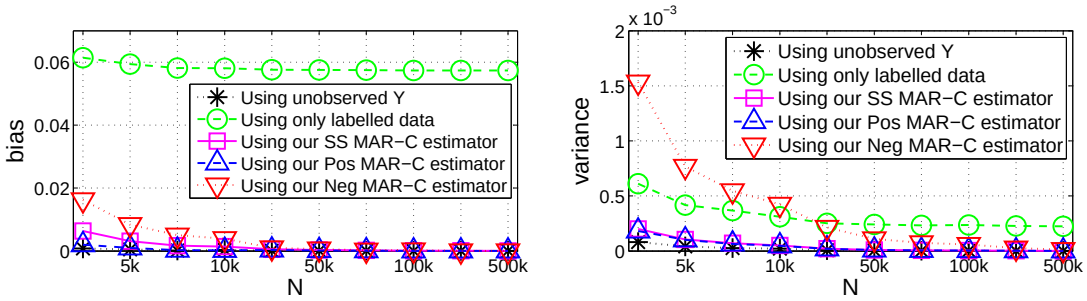
Figure 5.3 shows an experimental demonstration. Our suggested estimators are consistent, shown by both bias and variance decreasing as the sample size gets larger. The alternative, ignoring the unlabelled examples, gives an asymptotically biased estimator. By using soft prior knowledge over $p(y = 1)$ and Conjecture 5.6, we can decide which is the most accurate estimator. For this setting, the *rhs* of inequality (4.1) is equal to 0.50, and since it is smaller than $p(y = 1) = 0.20$ we can conclude that $\widehat{I}^{Pos}_{MAR-C}(X;Y)$ is more accurate than $\widehat{I}^{Neg}_{MAR-C}(X;Y)$, this conclusion is verified both in terms of bias and variance in Figure 5.3.



(a) *Small effect* — $I(X;Y) = 0.005$ *with* $|\mathcal{X}| = 5$.



(b) *Medium effect* — $I(X;Y) = 0.045$ *with* $|\mathcal{X}| = 5$.



(c) *Large effect* — $I(X;Y) = 0.125$ *with* $|\mathcal{X}| = 5$.

Figure 5.3: *Comparing the performance of our suggested Pos/Neg/SS MAR-C semi-supervised estimators with the estimator using only the labelled data* $\widehat{I}(X;Y|s=1)$ *and the fully supervised unobservable estimator* $\widehat{I}(X;Y)$ *in terms of bias/variance. To estimate bias/variance we average over 2000 runs. The class prior is* $p(y=1) = 0.20$ *and the semi-supervised data were generated through MAR-C and with* $p(s=1) = 0.25$ *and uniform distribution for the class in the labelled set:* $p(y=1|s=1) = p(y=0|s=1) = 0.50$. *Please note different axes for the bias/variance of the different effect levels.*

5.4.1 Verifying the Sampling Distribution of the Positive Information Estimator

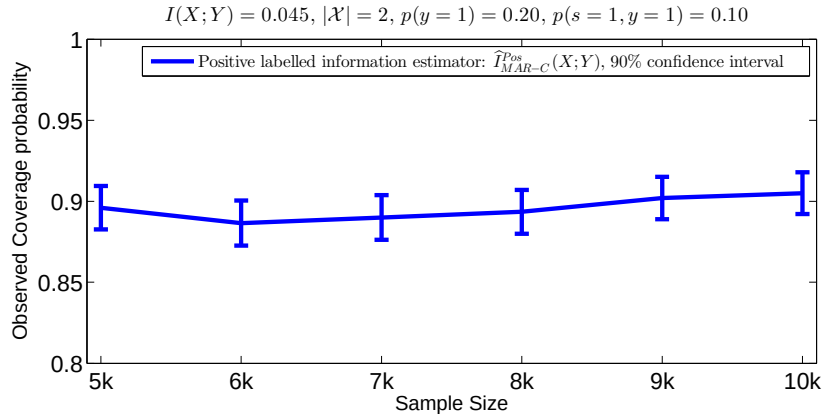
In this section, we will empirically verify the sampling distribution of the estimator suggested in Theorem 5.4. This estimator uses only the positively labelled information to estimate mutual information, and as a result it can be used in positive-unlabelled scenarios.

Figure 5.4 presents the proportion of times (over 2,000 repeats) that the 90% confidence intervals, derived through the sampling distribution of the estimator $\widehat{I}^{Pos}_{MAR-C}(X;Y)$, contains the true value of the mutual information $I(X;Y)$. Since the observed coverage is fluctuating around the nominal value of 90%, we can conclude that the sampling distribution is accurate. So, by labelling only a small fraction of the positive examples we can consistently estimate the mutual information, and furthermore we can provide interval estimates using our estimator despite the partial supervision.

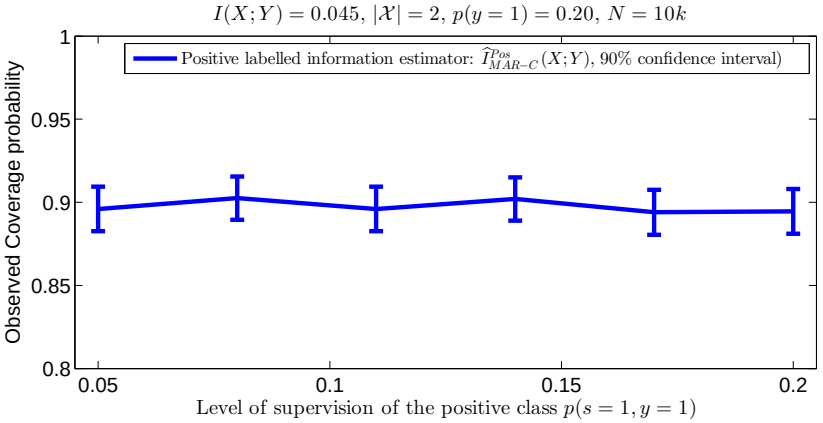
5.5 Chapter Summary

In this chapter we explored the estimation of mutual information in partially labelled data. When the labels are MCAR, using only the labelled set leads to consistent estimators, but we can improve the accuracy (mainly through reducing the bias) by incorporating the unlabelled examples as suggested by equation (5.5). In the MAR-F scenario, the bias in the labelled set leads to inconsistent estimators, but by using the unlabelled examples, we can correct the bias and estimate consistently mutual information by using equation (5.6) without the need of prior knowledge. Finally, when the labels are MAR-C, the labelled set leads to inconsistent estimators, and in order to correct the bias we incorporate prior knowledge over $p(y)$. Using this knowledge we suggested three ways to estimate mutual information (5.7), (5.8) and (5.9). By using inequality (4.1), we can decide the most accurate choice between the last two estimators. Furthermore, expressions (5.8) and (5.9) can be used to estimate mutual information and build confidence intervals.

In the next chapter we will explore how we can use our suggested estimators and the surrogate variables derived in Chapter 4 to rank the features in the different partially labelled scenarios.



(a) *Different sample sizes, for fixed lever of supervision of the positive class* $p(y = 1, s = 1) = 0.10$.



(b) *Different levels of supervision of the positive class, for fixed sample size* $N = 10000$.

Figure 5.4: *Verifying whether the 90% confidence interval suggested from our estimator holds for different sample sizes and different levels of superivsion.*

Chapter 6

Theoretical Analysis of Ranking in Partially Labelled Data

In the previous chapter we explored ways to consistently estimate the mutual information despite the partial supervision. We can use these consistent estimators to rank the features, and recover rankings that are as close as possible to the true population ranking. But, as we will present in this chapter, we can only recover the population ranking by using asymptotically biased estimators, as long as the bias introduced is independent of the characteristics of the features.

This chapter focuses on how to use the unlabelled examples and produce surrogate approaches that rank the features as close as possible to the population ranking. To do so, we will combine the results that we derived in the previous two chapters and we will suggest efficient ways for feature ranking under the three different missingness scenarios: MCAR (Section 6.2), MAR-F (Section 6.3) and MAR-C (Section 6.4). But, before, that we should provide some definitions over the *ranking equivalent* approaches.

6.1 Defining Ranking Equivalent Approaches

For our analysis it is important to define the equivalence between rankings derived using different approaches.

Definition 6.1 (Ranking Equivalence)**.**
Assume that we have a set of features $\mathbf{X} = \{X_1, ..., X_d\}$ *and we use two different approaches to rank them according to their dependency with respect to a target variable, i.e.* $f_1(X)$ *and* $f_2(X)$*. We say that the two approaches* f_1 *and* f_2 *are ranking equivalent if and only if* $\forall\ i,\ j$ *that* $f_1(X_i) < f_1(X_j)$ *it holds that* $f_2(X_i) < f_2(X_j)$*, where* $f_1(X_i)$ *and* $f_1(X_i)$ *represent the score of the feature* X_i *using the first and second approach respectively. When two approaches are ranking equivalent we will use the symbol* $f_1(X) \overset{R}{=} f_2(X)$*.*

For example, assume that we have a set of features $\mathbf{X}$, two random variables Y and W and the following relationship holds: $I(X_i; Y) = aI(X_i; W)\ \forall\ X_i \in \mathbf{X}$, where $a \in \mathbb{R}^+$ is a constant with respect to X_i. Because of this relationship it holds that: $I(X; Y) \overset{R}{=} I(X; W)$. In other words, ranking the features using the variable Y is equivalent to ranking the features using variable W.

One of the main questions we posed in the introduction is how to derive a ranking from *finite samples of data* that would be close to the population ranking. As we saw in Section 2.3, this is related to the bias and variance of the estimator. We can say that the maximum likelihood estimator is *approximately ranking equivalent* to the population ranking, $\widehat{I}(X; Y) \overset{R}{\approx} I(X; Y)$, while the more samples we collect the better the approximation becomes, since this estimator is consistent. In a different wording we can say that the ranking derived using this estimator *converges* to the population valued ranking: $\widehat{I}(X; Y) \underset{N\to\infty}{\overset{R}{=}} I(X; Y)$. Furthermore, using the result that the squared loss and Shannon's mutual information are approximately equal (Section 2.1), we can say that it holds $\widehat{I}_2(X; Y) \overset{R}{\approx} I(X; Y)$, while by using the fact that the G and the X^2-test are asymptotically equivalent (Haberman, 1974, p. 109) we can conclude that $\widehat{I}_2(X; Y) \underset{N\to\infty}{\overset{R}{=}} I(X; Y)$.

In the following sections we will analyse the surrogate approaches to derive rankings in different partially labelled scenarios in terms of *if* and *how fast* they converge to the true population ranking.

6.2 Ranking the Features when the Labels are MCAR

By exploring the population values of the different mutual information quantities compared to the supervised mutual information, $I(X;Y)$, we can derive equivalent rankings. With straightforward calculations, the following relationships hold:

$$\textbf{Surrogate 1 } (\mathcal{D}_L): \ I(X;Y|s=1) = I(X;Y),$$
$$\textbf{Surrogate 2 } (\widetilde{Y}_m): \ I(X;\widetilde{Y}_m) = p(s=1)I(X;Y),$$
$$\textbf{Surrogate 3 } (\widetilde{Y}_0): \ I_2(X;\widetilde{Y}_0) = \frac{p(s=1)-p(s=1)p(y=1)}{1-p(y=1)p(s=1)} I_2(X;Y),$$
$$\textbf{Surrogate 4 } (\widetilde{Y}_1): \ I_2(X;\widetilde{Y}_1) = \frac{p(s=1)-p(s=1)p(y=0)}{1-p(y=0)p(s=1)} I_2(X;Y),$$
$$\textbf{Our MCAR estimator}: \ I_{MCAR}(X;Y) = I(X;Y).$$

We see that all the mutual information quantities of the *lhs* can be written as $aI(X;Y)$, where the factor a is independent of the characteristics of the feature X. So a direct consequence of these relationships is that we can use any mutual information quantity of the *lhs* to rank the features, and the ranking will be the same as if we had used the unobservable $I(X;Y)$. In other words, using a finite dataset to estimate the mutual information quantities, the following approaches are approximately ranking equivalent:

$$\textbf{Surrogate 1 } (\mathcal{D}_L): \ \widehat{I}(X;Y|s=1) \overset{R}{\approx} I(X;Y),$$
$$\textbf{Surrogate 2 } (\widetilde{Y}_m): \ \widehat{I}(X;\widetilde{Y}_m) \overset{R}{\approx} I(X;Y),$$
$$\textbf{Surrogate 3 } (\widetilde{Y}_0): \ \widehat{I}(X;\widetilde{Y}_0) \overset{R}{\approx} I(X;Y),$$
$$\textbf{Surrogate 4 } (\widetilde{Y}_1): \ \widehat{I}(X;\widetilde{Y}_1) \overset{R}{\approx} I(X;Y),$$
$$\textbf{Our MCAR estimator}: \ \widehat{I}_{MCAR}(X;Y) \overset{R}{\approx} I(X;Y).$$

Deciding which of the above approximate rankings is preferable has to do with the accuracy of the estimators. By combining our analysis in testing and estimation we can conclude which is the better option, or in other words, which approximately equivalent ranking will be closer to the population ranking. In Section 4.2, we showed that the most powerful option to test independence is to

ignore the unlabelled examples, and as a result this is the most accurate way to estimate $I(X;Y)$ among the four surrogate approaches presented in Section 4.1. So from the first four approximate rankings, the one that ignores the unlabelled examples will perform better, or in other words, the best option is to use the estimator $\widehat{I}(X;Y|s=1)$. Furthermore, in Section 5.2 we show that $\widehat{I}_{MCAR}(X;Y)$ is a more accurate estimator than $\widehat{I}(X;Y|s=1)$. As a result we can conclude the following Corollary.

Corollary 6.2 (MCAR: Ranking)**.**
In MCAR the rankings derived by the four surrogate approaches and our suggested estimator all converge to the population ranking. However our approach (the MCAR estimator) converges faster.

To verify Corollary 6.2 we will compare the rankings derived by using the different estimators against the population ranking. To check the similarity between the rankings we use Spearman's rank correlation coefficient or Spearman's ρ (Kalousis et al., 2007). The range of values that this coefficient takes is $[-1,1]$, where 1 means that the two rankings are identical, 0 means that there is no correlation between them, and -1 means that they have exactly the inverse order. Since we need to have knowledge over the population ranking, we will use a synthetic dataset. Table 6.1 presents the characteristics of that dataset, which contains effects whose sizes can be classified from small to medium; this is extremely challenging in terms of predicting the population ranking, because the stepwise increase in the population values of the mutual information is 0.0001.

We sample various different dataset sizes (N) from 2500 ($2.5k$) to 500000 ($500k$) examples, to observe the performance when the sample size increases.

# Features	Population values of the effects between the features and the target	Class prior $p(y=1)$
100	$I(X_1;Y)$=0.0351, $I(X_2;Y)$=0.0352,$\cdots$, $I(X_{100};Y)$=0.0450	0.20

Table 6.1: *Characteristics of synthetic dataset used to observe the ranking performance. The arity of features is chosen randomly between the following values* $|\mathcal{X}| = 2, 5, 10$ *and* 20*.*

Figure 6.1 verifies the results of this section. Our suggested semi-supervised estimator $\widehat{I}_{MCAR}(X;Y)$ outperforms the other estimators, especially in small sample sizes, and this is a verification of Corollary 6.2. Furthermore, we see that by

increasing the sample size, all estimators improve their rankings, and they are closer to the population valued ranking. This is a verification of the fact that all approaches converge to the population ranking. In this figure, for ease of visibility, we did not plot surrogate 4, which corresponds to the least powerful option among the four surrogate approaches.

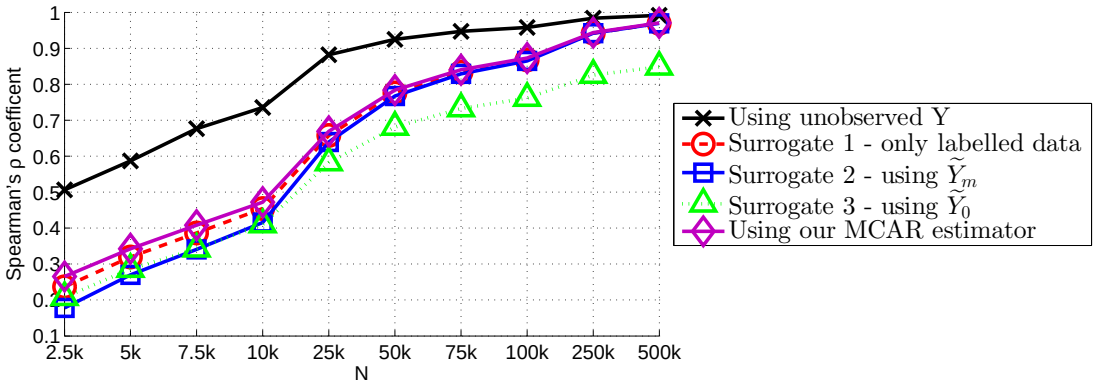


Figure 6.1: *Spearman's ρ coefficient between the population ranking and the ranking derived through different estimators when the labels are MCAR. We plot the average values over* 10 *different sampled datasets for each different sample size N. For each dataset we created* 30 *semi-supervised versions by sampling MCAR with $p(s=1)=0.25$ and we average the values.*

6.3 Ranking the Features when the Labels are MAR-F

When the missingness mechanism is MAR-F, we cannot derive re-expressions similar to the ones of the previous section. As a result we cannot conclude anything about the population ranking and the rankings derived through the surrogate approaches 1-4: $\widehat{I}(X;Y|s=1)$, $\widehat{I}(X;\widetilde{Y}_m)$, $\widehat{I}(X;\widetilde{Y}_0)$ and $\widehat{I}(X;\widetilde{Y}_1)$. On the other hand, we know that $\widehat{I}_{MAR-F}(X;Y)$ is a consistent estimator for $I(X;Y)$, and as a result we can derive the following corollary.

Corollary 6.3 (MAR-F: Ranking)**.**
In MAR-F using the ranking derived through our suggested consistent estimator $I_{MAR-F}(X;Y)$ converges to the population ranking, while the other surrogate approaches do not converge.

Figure 6.2 verifies the result of this section. To make the comparison fair between the features, a different labelled set was generated for each feature X_i in such a way that the distribution of X_i in the labelled set is uniform. So for each feature $X_i \in \mathbf{X} = X_1 ... X_d$ it holds that: $p(x_i|s=1) = \frac{1}{|\mathcal{X}_i|} \; \forall \; x_i \in \mathcal{X}_i$. As we observe, ranking the data through our consistent estimator performs better than ignoring the unlabelled examples, or following one of the partially supervised approaches, and this is the only approach that converges to the population ranking.

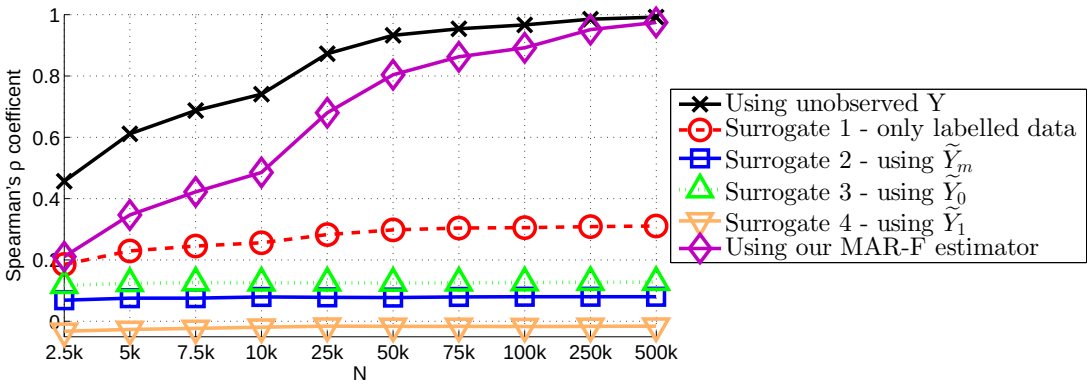


Figure 6.2: *Spearman's ρ coefficient between the population ranking and the ranking derived through different estimators when the labels are MAR-F. We plot the average values over* 10 *different sampled datasets for each different sample size N. For the semi-supervised scenarios we created* 30 *semi-supervised versions by sampling MAR-F with $p(s=1) = 0.25$ and we averaged the values over them.*

6.4 Ranking the Features when the Labels are MAR-C

For this scenario, by exploring the population values of the different mutual information quantities, comparing to the supervised mutual information $I(X;Y)$ we can derive equivalent rankings. Unfortunately, in this scenario we cannot derive relationships for all the approaches presented so far, but only for the following instead:

$$\textbf{Surrogate 3 } (\widetilde{Y}_0): \ I_2(X;\widetilde{Y}_0) = \frac{1-p(y=1)}{p(y=1)}\frac{p(\widetilde{y}_0=1)}{1-p(\widetilde{y}_0=1)}I_2(X;Y),$$
$$\textbf{Surrogate 4 } (\widetilde{Y}_1): \ I_2(X;\widetilde{Y}_1) = \frac{1-p(y=0)}{p(y=0)}\frac{p(\widetilde{y}_1=0)}{1-p(\widetilde{y}_1=0)}I_2(X;Y),$$
$$\textbf{Our SS MAR-C estimator}: \ I^{SS}_{MAR-C}(X;Y) = I(X;Y),$$
$$\textbf{Our Pos MAR-C estimator}: \ I^{Pos}_{MAR-C}(X;Y) = I(X;Y),$$
$$\textbf{Our Neg MAR-C estimator}: \ I^{Neg}_{MAR-C}(X;Y) = I(X;Y).$$

As we observe, the mutual information quantities of the *lhs* can be written as $aI(X;Y)$, where the factor a is independent of the characteristics of the feature X. A consequence of these relationships is that we can use any mutual information quantity of the *lhs* to rank the features, and the ranking will be the same as if we had used the unobservable $I(X;Y)$. So in finite datasets the following approaches are approximately ranking equivalent:

$$\textbf{Surrogate 3 } (\widetilde{Y}_0): \ \widehat{I}(X;\widetilde{Y}_0) \overset{R}{\approx} I(X;Y),$$
$$\textbf{Surrogate 4 } (\widetilde{Y}_1): \ \widehat{I}(X;\widetilde{Y}_1) \overset{R}{\approx} I(X;Y),$$
$$\textbf{Our SS MAR-C estimator}: \widehat{I}^{SS}_{MAR-C}(X;Y) \overset{R}{\approx} I(X;Y),$$
$$\textbf{Our Pos MAR-C estimator}: \widehat{I}^{Pos}_{MAR-C}(X;Y) \overset{R}{\approx} I(X;Y),$$
$$\textbf{Our Neg MAR-C estimator}: \widehat{I}^{Neg}_{MAR-C}(X;Y) \overset{R}{\approx} I(X;Y).$$

An interesting consequence is that we can rank the features without an exact prior knowledge over the $p(y=1)$ by simply using Surrogate 3 or Surrogate 4 approach. By using exact prior knowledge we can use any of our suggested estimators. Deciding which of the estimated rankings converges faster to the population has to do with the accuracy of the estimators. Using the results in Section 5.4, we can suggest the following conjecture.

Conjecture 6.4 (MAR-C: Ranking by using "exact" or "soft" prior knowledge)**.** *In MAR-C Surrogate 3, Surrogate 4 and our three suggested estimators, which incorporate prior knowledge, converge to the population ranking. When we have "exact" prior knowledge we can use (4.1) to decide the optimal choice between* $\widehat{I}^{Pos}_{MAR-C}(X;Y)$ *or* $\widehat{I}^{Neg}_{MAR-C}(X;Y)$*, while when we have "soft" we can use the same inequality to decide the optimal choice between surrogate 3, which uses* $\widehat{I}(X;\widetilde{Y}_0)$

or surrogate 4, which uses $\widehat{I}(X;\widetilde{Y}_1)$.

Figure 6.3 verifies the results of this section. To generate the semi-supervised data, we used the same methodology as in Sections 4.4 and 5.4. When we have "soft" prior knowledge we can decide the optimal choice between $\widehat{I}(X;\widetilde{Y}_0)$ and $\widehat{I}(X;\widetilde{Y}_1)$. In this setting, since $p(\widetilde{y}_0 = 1) = p(\widetilde{y}_1 = 0) = 0.125$, the *rhs* of the inequality (4.1) becomes 0.50, which is larger than 0.20, and as result the ranking derived through $\widehat{I}(X;\widetilde{Y}_0)$ will be closer to the population ranking than the one derived by $\widehat{I}(X;\widetilde{Y}_1)$. When we have "exact" prior knowledge over $p(y = 1)$ we can estimate our three suggested estimators, and using (4.1) we can decide the optimal choice between $\widehat{I}(X;Y)^{Pos}_{MAR-C}$ and $\widehat{I}(X;Y)^{Neg}_{MAR-C}$. By using inequality (4.1) again, we can decide that the optimal choice is $\widehat{I}(X;Y)^{Pos}_{MAR-C}$. In this figure, to help the visibility, we plot only the optimal choices and we see that all of them converge to the population ranking, while using only the labelled data does not.

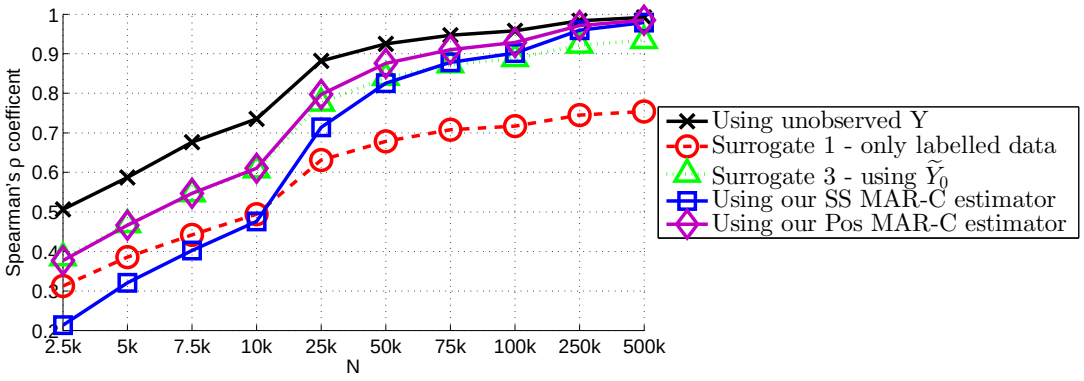


Figure 6.3: *Spearman's* ρ *coefficient and standard deviation over* 10 *different sampled datasets when the labels are MAR-C. For each dataset we created* 30 *semi-supervised versions by sampling MAR-C with* $p(s = 1) = 0.25$ *and we averaged the values over them. For each feature the MAR-C sampling created a unform marginal distribution of* Y *in the labelled set:* $p(y|s = 1) = \frac{1}{|\mathcal{Y}|}\ \forall\ y \in \mathcal{Y}$.

6.5 Chapter Summary

This section presented how to derive rankings of features, in the partially labelled setting, that are as close as possible to the true population ranking.

Firstly, when the labels are MCAR, we saw that our suggested estimator can give the most accurate ranking. Using any of the four surrogate approaches

presented in Chapter 4, will give rankings that converge to the true one. Secondly, when the labels are MAR-F, only our estimator recovers the true ranking. Lastly, when the labels are in the MAR-C setting, we show that our three estimators and the two surrogate approaches (Surrogate 3 and Surrogate 4) produce rankings that converge to the true ranking. To decide which one will converge faster we can use prior knowledge over $p(y = 1)$ and inequality (4.1), as the Conjecture 6.4 describes. It is interesting to mention at this point that we can recover the true ranking, even if we do not have exact knowledge over $p(y = 1)$, by assuming the unlabelled examples as positives or negatives and rank features as usual. By making this assumption, we do not have a consistent estimate of the true mutual information, since $I(X;Y) \neq I(X;\widetilde{Y}_0)$ and $I(X;Y) \neq I(X;\widetilde{Y}_1)$, but the rankings converge to the true population ranking.

The results of our theoretical analysis in testing (Chapter 4), estimation (Chapter 5) and ranking (Chapter 6), naturally lead to three extensions and practical applications of our work in the areas of experimental design, Markov blanket discovery and feature selection — Chapters 7, 8 and 9 respectively.

Chapter 7

Extension 1: Sample size and Labelled set Size Determination in Partially Labelled Data

The first extension of our work is enabled by obtaining *informed* tests (via the correction factors) despite the partial supervision. As we already mentioned, an informed test enables power analysis activities, such as *sample size determination*. Using these tests we can decide the minimum number of examples that we need in order to observe an effect (i.e. expressed in terms of $I(X;Y)$) with a predefined probabilities of committing a type I and type II error.

In partially labelled scenarios, under our analysis, a novel capability naturally arises, which we call *supervision determination*. Using our methodology we can determine a-priori the number of labelled examples that the user must collect before being able to observe a desired statistical effect. In this chapter, firstly we will present our findings in the positive-unlabelled setting (Section 7.1), and then how we can extend our analysis to semi-supervised data (Section 7.2).

7.1 Determining the Sample/Labelled-set Size in Positive-Unlabelled Data

Positive-unlabelled data, under the selected completely at random assumption presented in Section 3.3.3, can be seen as a special case of MAR-C with the further restriction that we do not have any negative labelled examples $p(y = 0, s = 1) = 0$

or equivalently $p(y = 1, s = 1) = p(s = 1)$. By using the surrogate variable $\widetilde{Y}_0$, the last relationship can be written as $p(\widetilde{y}_0 = 1) = p(s = 1)$. Thus, in a scenario under the PU constraint we essentially observe only the surrogate variable $\widetilde{Y}_0$, which is identical to the labelling variable S, i.e. $S \equiv \widetilde{Y}_0$.

From our analysis in Section 4.4, we know that using the surrogate variable $\widetilde{Y}_0$ instead of Y is a valid approach to test independence. Furthermore, it is also an informed test, by incorporating prior knowledge over $p(y = 1)$ and the probability of labelling a positive example $p(y = 1, s = 1) = p(\widetilde{y}_0 = 1)$. In this section we will show the capabilities of the $G(X; \widetilde{Y}_0)$-test when the non-centrality parameter is corrected by the factor $\kappa_{\widetilde{Y}_0}$ presented in Theorem 4.8, including sample size determination under the PU constraint, and a novel capability — determining the minimum number of labelled positive examples necessary to achieve statistical significance. We separate these experiments in two parts: the first one where we have perfect prior knowledge (Section 7.1.1) and the second where we use uncertain prior knowledge (Section 7.1.2).

7.1.1 Using Perfect Prior Knowledge

In this section, firstly we provide some theoretical predictions for sample size and supervision determination, and then we verify them empirically.

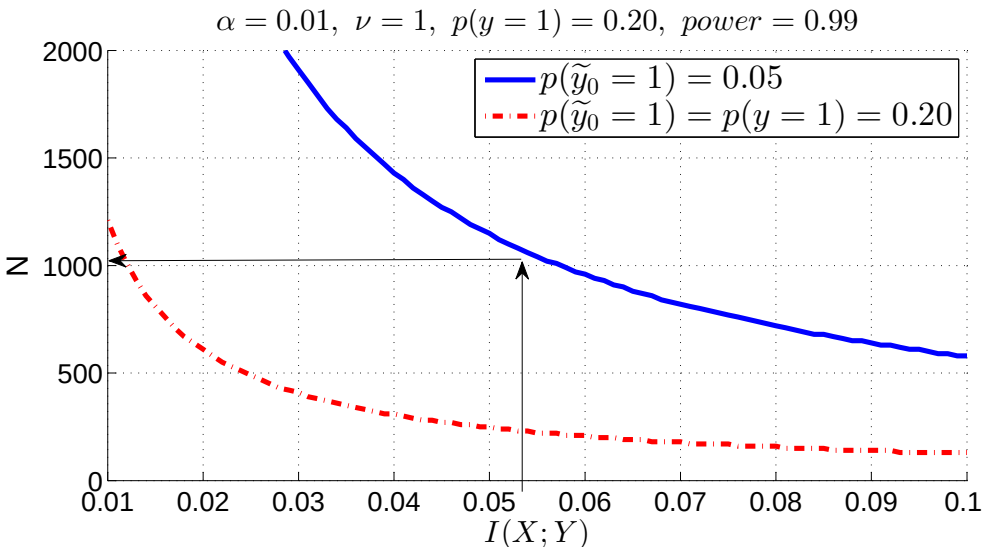
Theoretical predictions for sample size determination

Figure 7.1a shows how classical power analysis changes under the PU constraint. The illustration is for significance level $\alpha = 0.01$, a required power of 0.99, $p(y = 1) = 0.2$ and binary features (degrees of freedom $\nu = 1$).

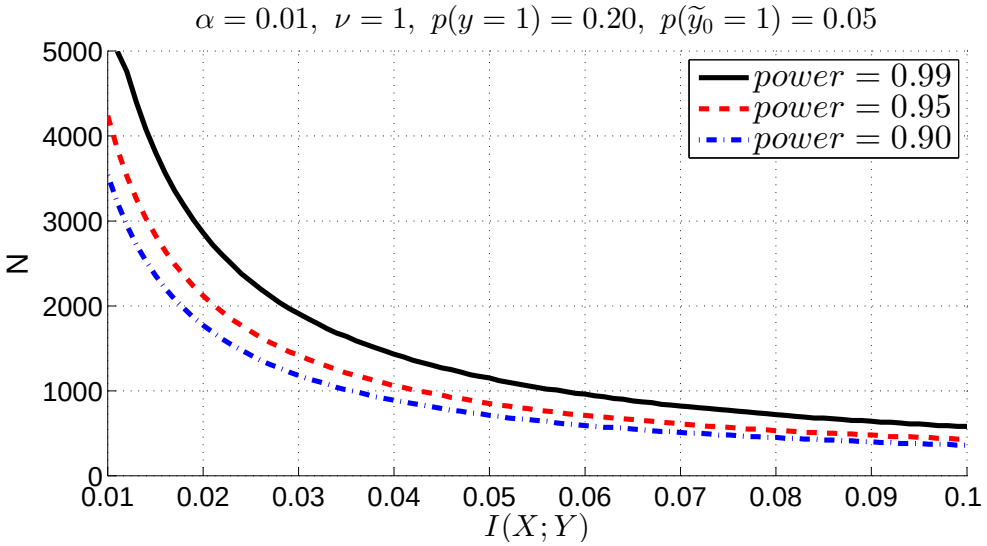
In Figure 7.1a we see the dashed line, which shows classical sample size determination – this is a standard result. The solid line shows the PU case, when we can obtain labels only for 5% of the examples as positives, i.e. $p(\widetilde{y}_0 = 1) = 0.05$.

The figure can be interpreted as follows: if we wish to detect a dependency with mutual information[1] as low as $I(X; Y) = 0.053$ and we want to observe this effect with power 99%, in the fully supervised case (dashed line) we require $N \geq 227$. However in the PU scenario (solid line) with labelling one quarter of the positive examples, $p(\widetilde{y}_0 = 1) = 0.05 \Leftrightarrow p(y = 1|s = 1) = 0.25$, this a-priori

[1] An interpretation of an effect of size 0.053 is that it quantifies the mutual information between Y and X, when X is generated by corrupting 60% of the examples of Y with uniform binary noise. We can calculate analytically that $I(X; Y) = 0.053$ (written to 3 decimal places).



(a)



(b)

Figure 7.1: *Figures for sample size determination. (a) Contrasting classical power analysis, which means we label all the positive examples* $p(\widetilde{y}_0 = 1) = p(y = 1) = 0.20$*, with PU power analysis to determine the minimum sample size. Arrows show that with 5% supervision –* $p(\widetilde{y}_0 = 1) = 0.05$ *– we need* $N \geq 1077$ *examples to achieve the desired power in order to observe a supervised effect* $I(X;Y) = 0.053$*. (b) Sample size determination under the PU constraint. Given a required statistical power, this illustrates the minimum total number of examples needed, assuming we can only label 5% of the instances. For example, if we wish to detect a mutual information as low as* 0.04*, we need* $N \geq 1430$ *to have a power of 99%.*

power analysis indicates we need $N \geq 1077$. Note that the required increase is not a simple multiple of the supervision level: with only 1/4 of the positive examples being labelled one might assume we need a sample $4\times$ larger, which would be 908, however this is insufficient for the required power as shown by the figure. In this case, $\kappa_{\widetilde{Y}_0} = 0.2105$, and the required increase is a multiple of that factor: $227 \times (1/\kappa_{\widetilde{Y}_0}) \approx 1078$. The above results are expanded upon in Figure 7.1b, showing the required N to obtain different power levels.

Theoretical predictions for supervision determination

As we saw, the correction factor $\kappa_{\widetilde{Y}_0}$ enables us to use the $G(X;\widetilde{Y}_0)$ test instead of the $G(X;Y)$ for power analysis activities, such as sample size determination. Taking advantage of the extra degree of freedom in $p(\widetilde{y}_0 = 1)$, we can also *determine the necessary level of supervision* (i.e. number of positively labelled examples), following the same procedure as in sample size determination. This may have implications in active learning (Cohn et al., 1994), where we can request the labels of particular examples. This methodology allows us to predict when we have sufficient labels to have statistically significant results.

Figure 7.2 presents the a-priori PU power analysis, allowing us to determine the minimum level of supervision to achieve certain statistical power. The y-axis is the number of positive examples that have labels $N_L = p(\widetilde{y}_0 = 1)N$. This shows just one scenario, with $\alpha = 0.01$, $N = 1000$, when the true prior is $p(y = 1) = 0.2$. As an illustration, the solid line predicts that to detect a dependency as low as $I(X;Y) = 0.053$, with power greater than 99%, we will need to label at least 54 examples, or in other words, the probability of an example being labeled should be $p(\widetilde{y}_0 = 1) \geq 0.054$.

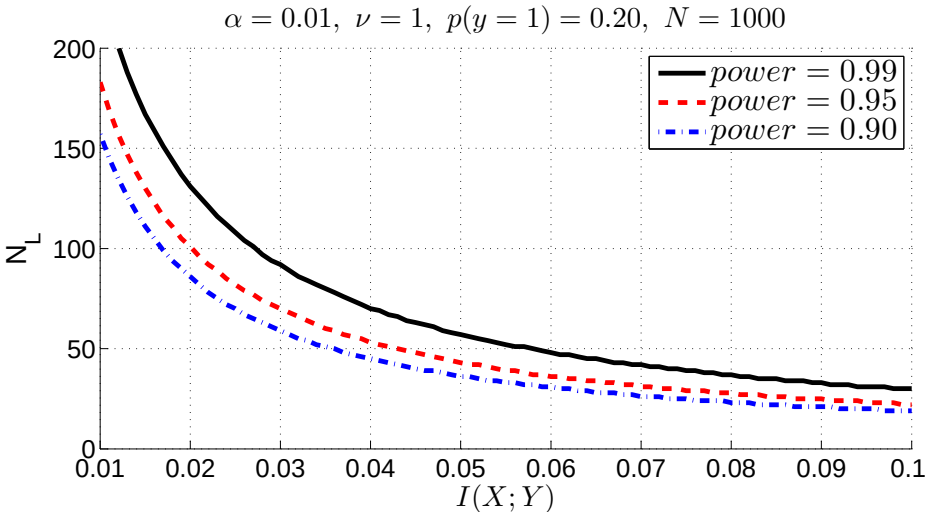


Figure 7.2: *Figure for supervision determination. Determining the required number of labelled examples N_L, assuming $N = 1000$. For example, to detect a mutual information dependency as low as 0.02, with power 95%, we need labels for 101 examples, which means that we need to label at least half of the positive examples.*

Verifying the theoretical predictions

To verify the theoretical predictions of required sample size and supervision level, we generate binary variables with a dependency $I(X;Y) = 0.053$ and observe the ability of a test to reject the null hypothesis, or in other words, the False Negative Rate (type-II error). Since the power is given by $1 - FNR$, any prediction of required N to achieve a particular power will translate directly to a corresponding FNR.

As a sanity check, we first verify the classical sample size determination for the G-test. Figure 7.1a (dashed line) predicts that we will need $N \geq 227$ to detect an underlying effect size of $I(X;Y) = 0.053$, with $\alpha = 0.01$ and power 99%. Figure 7.3a shows the FNR over *10,000* repeats. Note that the FNR crosses below the 1% rate when $N \approx 225$.

The next experiment verifies the PU sample size prediction. Figure 7.1a (solid line) predicted that to detect an effect as small as $I(X;Y) = 0.053$, with $\alpha = 0.01$, we would require $N \geq 1077$ to achieve an FNR below 1%. Again, the FNR over *10,000* repeats is shown in Figure 7.3b, supporting the theory as the FNR crosses 1% when $N \approx 1080$.

Finally we verify the predictions from Figure 7.2. We generate PU data as before, introducing noise so that the true underlying variables have $I(X;Y) =$

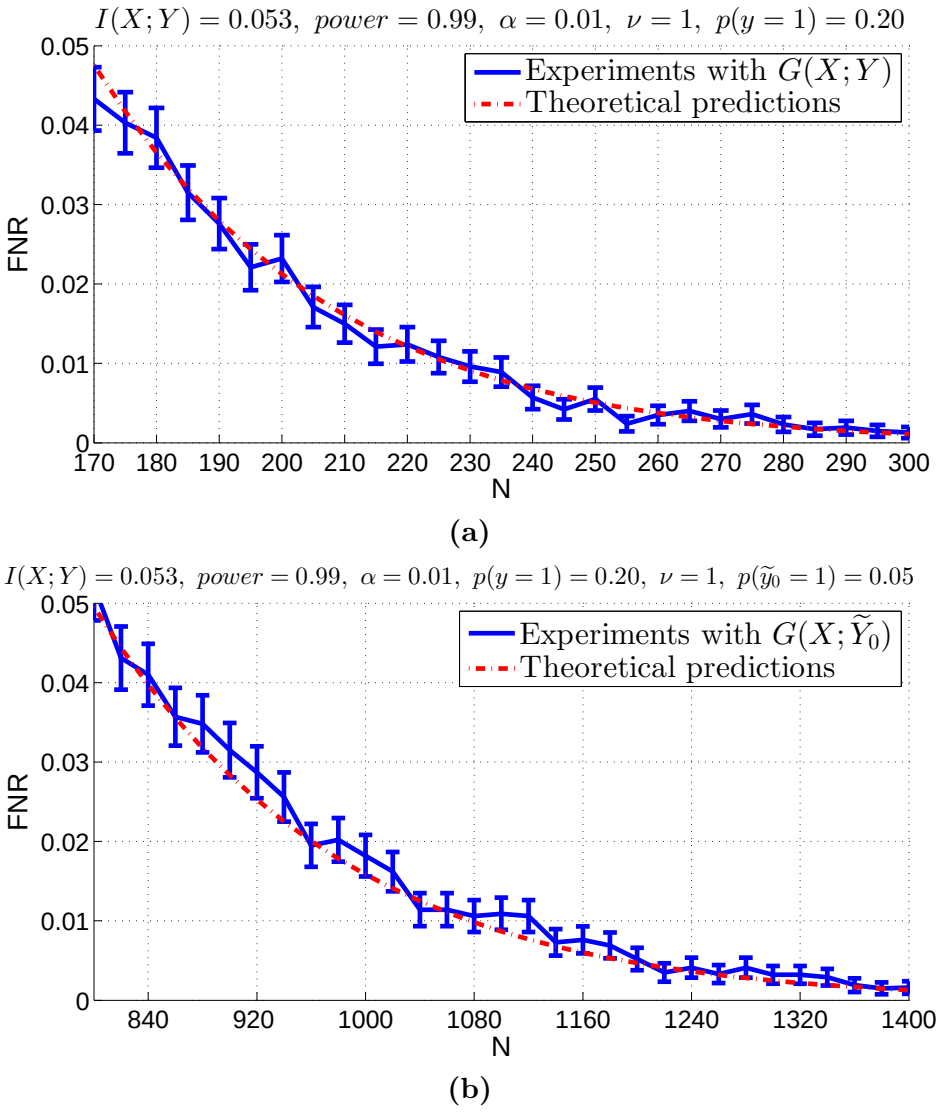


Figure 7.3: *Figures for False Negative Rate. (a) Full supervision, when the true mutual information is* $I(X;Y) = 0.053$. *This verifies the theoretical prediction from Fig. 7.1a. (b) Supervision level* $p(\widetilde{y}_0 = 1) = 0.05$, *supporting the predictions of Fig. 7.1a. The error bars represent 95% confidence intervals.*

0.053. Figure 7.4 shows the FNR, verifying that when we provide labels to the example with probability less than 0.054, or in other words, when we label less than 54 examples the type-II error is greater than 1%, and agrees with Figure 7.2.

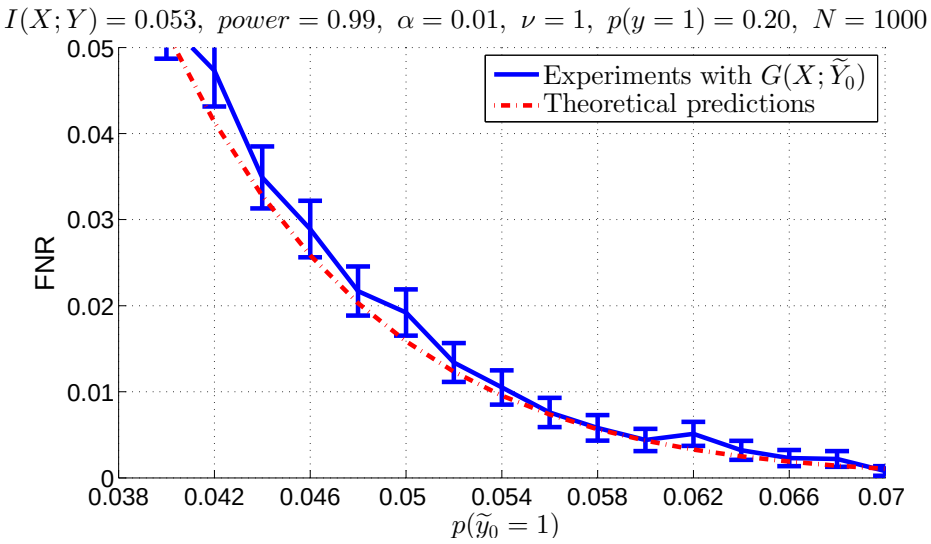


Figure 7.4: *False Negative Rate for varying levels of supervision in the PU constraint, with required power 99%, verifying Figure 7.2 (solid line), which predicted we would need* $p(\widetilde{y}_0 = 1) \geq 0.054 \Leftrightarrow N_L \geq 54$ *to get* $FNR < 0.01$. *The error bars represent 95% confidence intervals.*

7.1.2 Using Uncertain Prior Knowledge

The previous section assumed we somehow knew the exact value of $p(y = 1)$. In a more realistic scenario prior knowledge, $\widetilde{p}$, may be provided as a *distribution* over possible values. We model $\widetilde{p}$ as a generalised Beta distribution, between a minimum and a maximum value (Hahn and Shapiro, 1967), and use Monte-Carlo simulation to explore the resulting uncertainty in the required sample/supervision sizes. Figure 7.5 presents sample size determination when we have uncertain prior knowledge. The dashed vertical line indicates the perfect prior knowledge situation from the previous section. If we use a sample size less than this, we have an increased False Negative Rate. On the other hand, choosing a larger size will achieve at least the desired power, but at the cost of collecting more data.

Figure 7.6 presents how this uncertainty would translate to the required number of labeled examples. The same principle of choosing a value over/under the dashed line applies: here if we select $N_L > 54$ we are unnecessarily increasing our cost of label collection. In Figure 7.7 we observe the behavior when we underestimate (first row) or overestimate (second row) the $p(y = 1)$. A general conclusion is that the uncertainty in the prior translates quite directly to an uncertainty of a similar form over the minimum number of samples and a minimum amount of supervision.

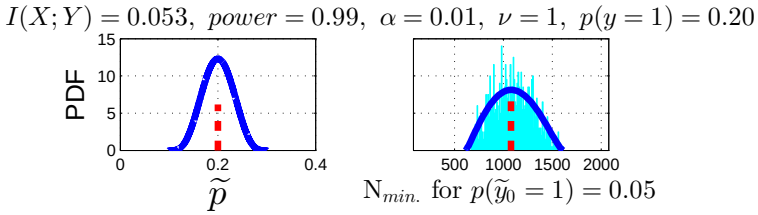


Figure 7.5: *Sample size determination under uncertain prior knowledge. LEFT: The user's prior belief over the value of* $p(y = 1)$*. The dashed line shows the* true *(but unknown) value in the data. RIGHT: The resultant uncertainty in the required sample size when we have only 5% of the examples being labeled, we plot both the histogram of the Monte-Carlo simulation results and a generalized Beta distribution fitted to the data.*

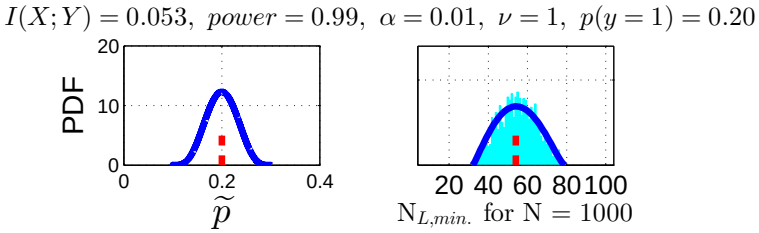


Figure 7.6: *Supervision determination under uncertain prior knowledge. LEFT: The user's prior belief over the value of* $p(y = 1)$*. The dashed line shows the* true *(but unknown) value in the data. RIGHT: The resultant uncertainty in the minimum number of required labeled examples when we have only* $N = 1000$*. The dashed line indicates the the true value with no uncertainty in* $p(y = 1)$*.*

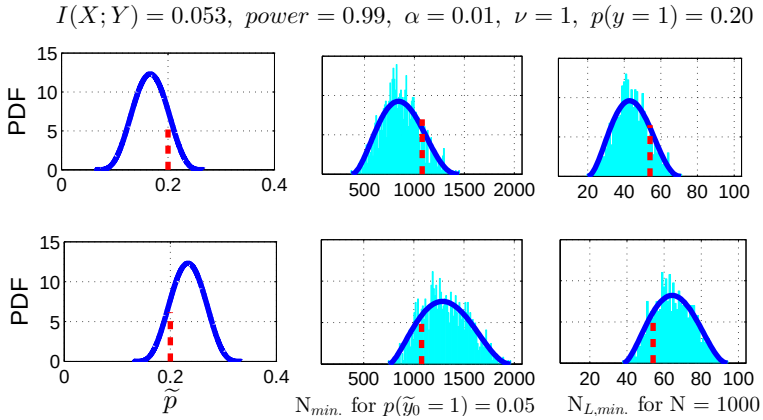


Figure 7.7: *A-priori power analysis under uncertain prior knowledge, when we underestimate (first row) and overestimate (second row) the prior.*

7.1.3 Guidance for Practitioners

For a practitioner the correct course of action depends on the conditions in a given application. To ensure that an effect would not be missed when indeed present, one should *overestimate* the value of $p(y = 1)$, hence leading to a larger number of examples/labels being collected. Conversely, if collection of examples/labels is a costly matter, one can take a more risky, but informed decision, using fewer examples/labels. Furthermore, to achieve a desired statistical power, choosing to fix the amount of supervision or the sample size is application-dependent.

Under our framework we can generate tables for sample size and supervision determination under the PU constraint similar to the ones used in the literature, e.g. Table 7.1a. For a given effect size (ω), degrees of freedom (ν), significance level (α), level of desired power, prior $p(y = 1)$ and fixed supervision level $p(\widetilde{y}_0 = 1)$, in Table 7.1b we can observe the minimum sample size that is needed in order to observe the effect with the given power.

	Effect sizes		
Power	Small	Medium	Large
0.70	962	107	39
0.80	1168	130	47
0.90	1488	166	60
0.95	1782	198	72
0.99	2404	268	97

(a) *Traditional*

	Effect sizes		
Power	Small	Medium	Large
0.70	4566	508	183
0.80	5548	617	222
0.90	7068	786	283
0.95	8462	941	339
0.99	11415	1269	457

(b) *PU with* $p(y = 1) = 0.20$, $p(\widetilde{y}_0 = 1) = 0.05$

Table 7.1: *Sample size required for* $|\mathcal{X}| = 2$ *and* $\alpha = 0.01$.

A new type of table can be generated when we fix the sample size and we want to determine the minimum amount of supervision (or in other words, the minimum number of labelled examples) that we need in order to observe a specific effect with a desired statistical power. Table 7.2 presents the minimum number of labelled positive examples that we need when we have similar conditions as before but now we fix the sample size to be $N = 3000$.

So in practical terms: if we assume we had 3000 examples, and we know that

	Effect sizes		
Power	Small	Medium	Large
0.70	223	27	10
0.80	267	33	12
0.90	331	41	15
0.95	388	49	18
0.99	501	66	24

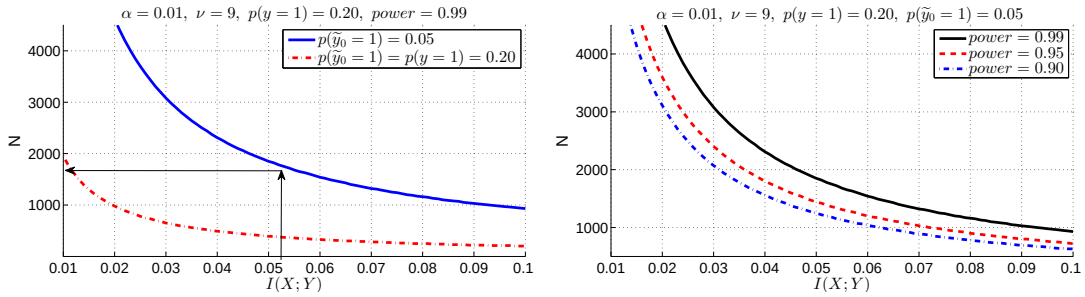
Table 7.2: *Labelled positive examples required for a PU test with* $|\mathcal{X}| = 2$*,* $\alpha = 0.01$*,* $N = 3000$ *and* $p(y = 1) = 0.20$*.*

approximately 600 of them are positive, if we wish to detect a "medium" sized effect (in Cohen's terminology), then, in order to achieve a False Negative Rate of 5% (i.e. power 0.95), we only need to identify correctly 49 of those 600 examples, according to Table 7.2. A different way to read the results is the following: imagine that we want to design an experiment in order to observe a medium effect with a statistical power of 80%, and the prevalence is $p(y = 1) = 0.20$. If we could label both positive and negative cases, we would need 130 examples according to Table 7.1a. So we would need to label 26 positives and 104 negatives. Instead of this we can use the results of Table 7.1b and collect 617 examples out of which we will label only 5%; in other words, we will label only 31 examples as positive and keep the rest as unlabelled. Thus, instead of labelling 104 negative examples, we can label 5 more positive examples and keep 586 as unlabelled. This approach can be useful when it is expensive or difficult to label examples, while it is cheaper to collect unlabelled. Since in the PU context labelling samples is expensive, this methodology can be used to save resources.

Our results can be used in any research involving hypothesis testing in PU data. Our framework has been described in terms of the G-test, and the mutual information as an effect size. We can use the same framework to derive similar expressions for the X^2-test and the squared loss mutual information. The latter one is strongly related to ϕ and Cramer's V coefficients, both widely used as effect sizes for categorical data in statistics. Since both G and X^2 are used extensively in behavioral sciences and biology, our work may have strong relevance in experimental design for partially supervised data (Ellis, 2010). The proposed methods can be used in several machine learning applications. Section 8.1 presents an application to Markov Blanket discovery in PU data, where we use our corrected G-test to decide whether we add an arc or not, since we derived the same correction factor $\kappa_{\widetilde{Y}_0}$ for testing conditional independence (Section 4.6).

7.1.4 Extending our Analysis to Higher Degrees of Freedom

Now, we will reproduce the Figures and the Tables presented in the previous section, but this time having categorical features with $|\mathcal{X}| = 10$ instead of binary.



(a) **(b)**

Figure 7.8: *Figures for sample size determination. (a) Contrasting classical power analysis, which means we label all the positive examples* $p(\widetilde{y}_0 = 1) = p(y = 1) = 0.20$*, with PU power analysis to determine the minimum sample size. Arrows show that with 5% supervision –*$p(\widetilde{y}_0 = 1) = 0.05$*– we need* $N \geq 1743$ *examples to achieve the desired power in order to observe a supervised effect* $I(X;Y) = 0.053$*. (b) Sample size determination under the PU constraint. Given a required statistical power, this illustrates the minimum total number of examples needed, assuming we can only label 5% of the instances. For example, if we wish to detect a mutual information as low as* 0.04*, we need* $N \geq 2310$ *to have a power of 99%.*

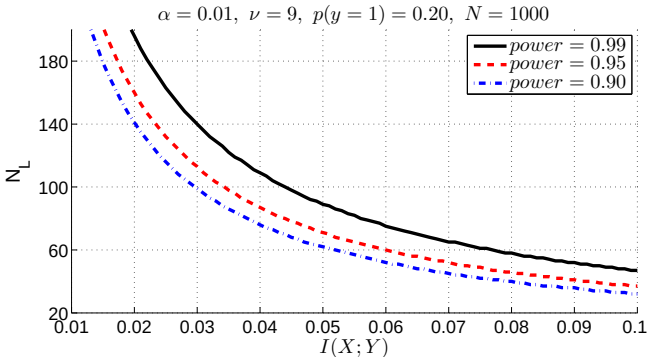
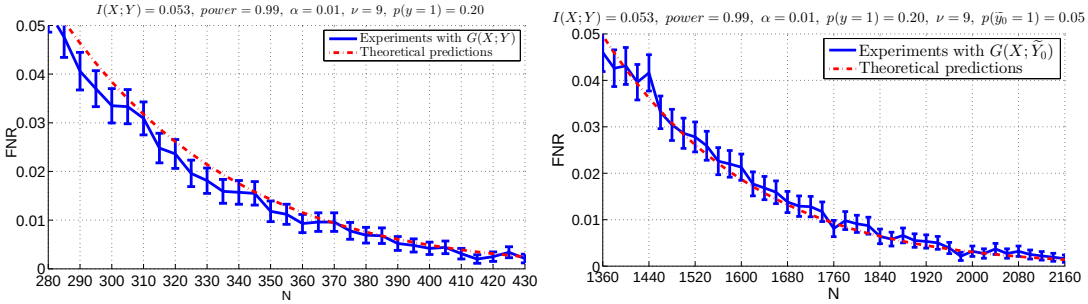


Figure 7.9: *Figure for supervision determination. Determining the required number of labelled examples* N_L*, assuming* $N = 1000$*. For example, to detect a mutual information dependency as low as* 0.02*, in order to have a power of 95%, we need labels for 160 examples, which means that we need to label 80% of the positive examples.*



(a) **(b)**

Figure 7.10: *Figures for False Negative Rate. (a) Full supervision, when the true mutual information is* $I(X;Y) = 0.053$. *This verifies the theoretical prediction from Fig. 7.8a, that the minimum sample size to achieve 99% power is* 367. *(b) Supervision level* $p(\widetilde{y}_0 = 1) = 0.05$, *supporting the predictions of Fig. 7.8a, that the minimum sample size to achieve 99% power is* 1743. *The error bars represent 95% confidence intervals.*

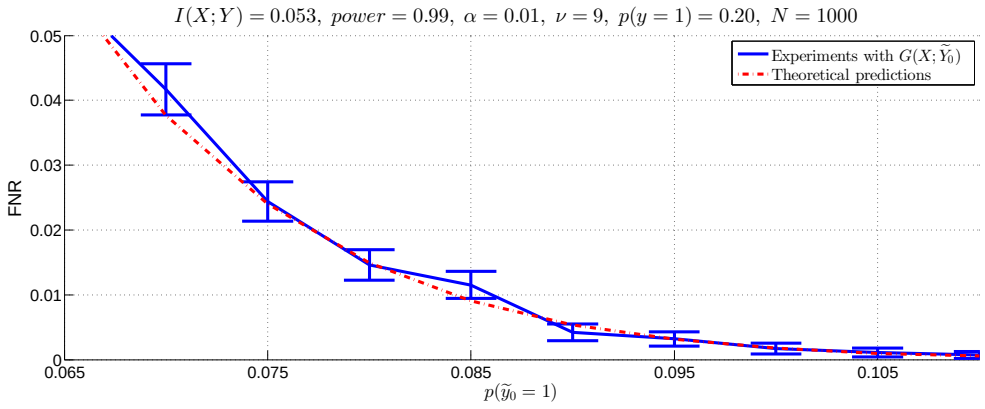


Figure 7.11: *False Negative Rate for varying levels of supervision in the PU constraint, with required power 99%, verifying Figure 7.9 (solid line), which predicted we would need* $p(\widetilde{y}_0 = 1) \geq 0.085 \Leftrightarrow N_L \geq 85$ *to get* $FNR < 0.01$. *The error bars represent 95% confidence intervals.*

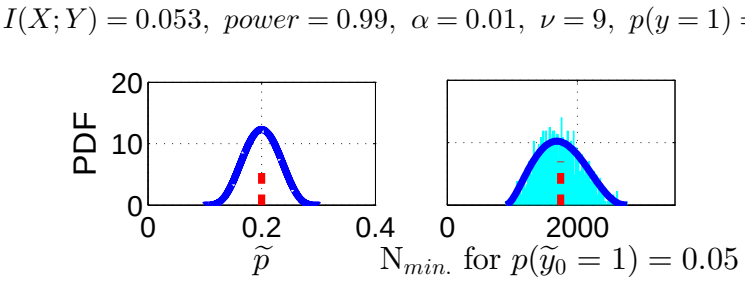


Figure 7.12: *Sample size determination under uncertain prior knowledge. LEFT: The user's prior belief over the value of* $p(y = 1)$*. The dashed line shows the* true *(but unknown) value in the data. RIGHT: The resultant uncertainty in the required sample size when we have only 5% of the examples being labeled, we plot both the histogram of the Monte-Carlo simulation results and a generalized Beta distribution fitted to the data.*

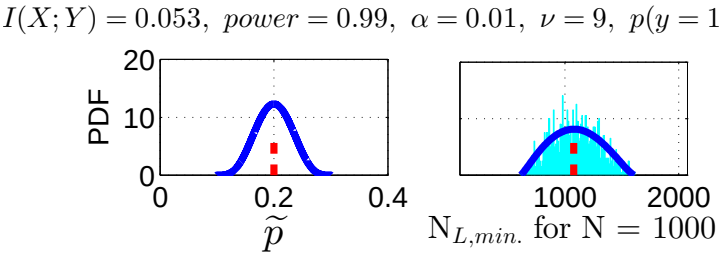


Figure 7.13: *Supervision determination under uncertain prior knowledge. LEFT: The user's prior belief over the value of* $p(y = 1)$*. The dashed line shows the* true *(but unknown) value in the data. RIGHT: The resultant uncertainty in the minimum number of required labeled examples when we have only* $N = 1000$*. The dashed line indicates the the true value with no uncertainty in* $p(y = 1)$*.*

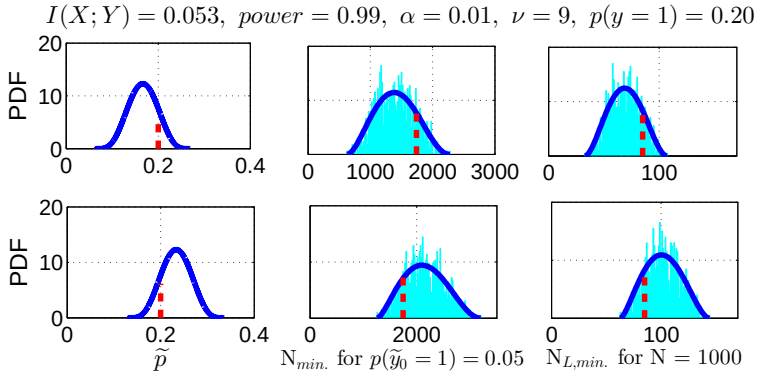


Figure 7.14: *A-priori power analysis under uncertain prior knowledge, when we underestimate (first row) and overestimate (second row) the prior.*

	Effect sizes		
Power	Small	Medium	Large
0.70	1830	204	74
0.80	2143	239	86
0.90	2613	291	105
0.95	3031	337	122
0.99	3890	433	156

(a) *Traditional*

	Effect sizes		
Power	Small	Medium	Large
0.70	8691	966	348
0.80	10179	1131	408
0.90	12409	1379	497
0.95	14395	1600	576
0.99	18474	2053	739

(b) *PU with* $p(y = 1) = 0.20$, $p(\widetilde{y}_0 = 1) = 0.05$

Table 7.3: *Sample size required for* $|\mathcal{X}| = 10$ *and* $\alpha = 0.01$.

	Effect sizes		
Power	Small	Medium	Large
0.70	420	51	19
0.80	484	59	22
0.90	578	72	26
0.95	658	83	31
0.99	814	106	39

Table 7.4: *Labelled positive examples required for a PU test with* $p(y = 1) = 0.20$, $N = 5000$, $|\mathcal{X}| = 10$ *and* $\alpha = 0.01$.

7.2 Determining the Sample/Labelled-set Size in Semi-Supervised Data

The methodology presented in PU scenarios can be extended to any informed test in semi-supervised scenarios. When the labels are MCAR, all four surrogate approaches are informed (Theorem 4.4), so we can use any of them for sample size and supervision determination. The most powerful option, which will lead us to the smallest sample sizes, is to ignore the unlabelled examples. This approach ends up being equivalent to determining the sample size for fully-supervised scenarios.

When the labels are MAR-C, we have two valid tests assuming the unlabelled examples being positives or negatives (Theorem 4.8). The first assumption is similar to solving the positive-unlabelled problem (by using the surrogate test $G(X; \widetilde{Y}_0)$), presented in the previous section, while the second assumption corresponds to the negative-unlabelled problem (by using the surrogate test $G(X; \widetilde{Y}_1)$), which can be solved in a similar way. Following the same methodology as in the previous section and incorporating prior knowledge over $p(y = 1)$, we can use either of these approaches for sample size determination. The sample size for the most powerful option — using inequality (4.1) to decide which it is — will be smaller than the sample size for the less powerful option. Furthermore, we can use these tests for supervision determination, to decide the minimum number of positively labelled and negatively labelled examples. In the following chapter we will present applications of our work in the area of Markov blanket discovery in partially-supervised data.

7.3 Chapter Summary

With our work in Chapter 4, we developed a set of novel methodologies, enabling statistical hypothesis testing activities in PU data. We proved, for example, that in positive-unlabelled data, the very common assumption, of all unlabelled examples being negative, is sufficient for detecting *independence*. However, a G-test using this assumption is less powerful than the fully supervised version, indicating the assumption is invalid for more complex power analysis activities. We solve this problem by deriving a *correction factor* for the test, incorporating prior knowledge from the user.

In this chapter, we presented how we can use these correction factors for determining the sample size in order to take a statistical decision that controls both errors: type I and type II. Furthermore, we explored a novel capability, which we called *supervision determination*: determining *the required minimum number of labelled examples* in order to control both types of error.

In the next chapter, we will use these correction factors for feature selection through Markov blanket discovery.

Chapter 8

Extension 2: Markov Blanket Discovery in Partially Labelled Data

Now we will present how to derive the Markov blanket around partially supervised nodes in an inference-free manner. The importance of Markov blanket discovery algorithms is twofold: they can be used as the main building block in constraint-based structure learning of Bayesian network algorithms, and as a technique to derive the optimal set of features in filter feature selection approaches. Equally, learning from partially labelled data is a crucial and demanding area of Machine Learning, and extending techniques from fully to partially supervised scenarios is a challenging problem. While there are many different algorithms to derive the Markov blanket of fully supervised nodes, the partially labelled problem is far more challenging, and there is a lack of principled approaches to solve it in the literature.

In this chapter, first we will present how to discover the blankets of positive-unlabelled targets and how we can incorporate "perfect" prior knowledge (Section 8.1). Then we will extend our methodologies to semi-supervised targets by presenting a practical algorithm that incorporates "soft" prior knowledge, and we will show that our approach outperforms the state of the art in a real semi-supervised scenario (Section 8.2).

8.1 Markov Blanket Discovery in Positive-Unlabelled Data

In this section we present how we can use our novel methodology for testing conditional independence in PU data to derive the Markov blanket despite the labelling restriction.

In the PU setting, the surrogate variable $\widetilde{Y}_0$ is fully observed despite this labelling restriction, and due to this restriction it is identical to the labelling variable S. Using this surrogate instead of Y is a valid approach to test conditional independence, because of Theorem 4.10. This will result in the same number of *false positive errors* for the two tests, or in Markov Blanket context, using the surrogate variable $\widetilde{Y}_0$ instead of the unobservable Y will result in the same number of nodes that were falsely added to the blanket.

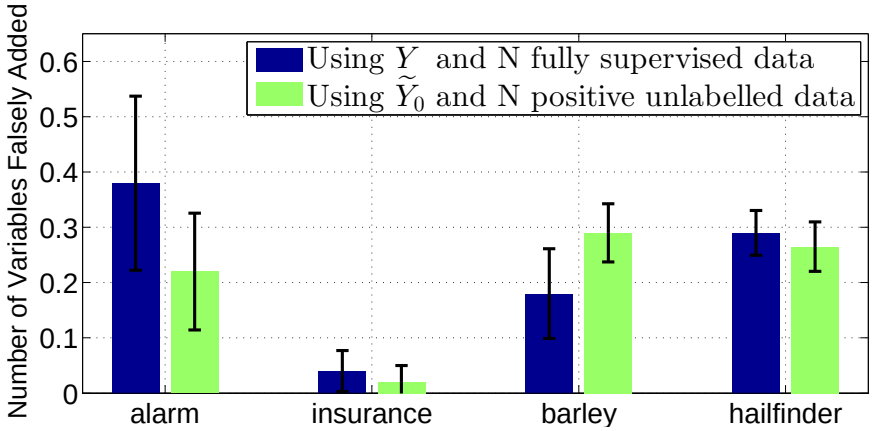
Now we will verify the consequences of this theorem in the context of Markov blanket discovery. We use four widely utilized standard benchmark networks for Markov blanket discovery taken from the Bayesian network repository[1]. Table 8.1 presents the summary of these networks. For target variables we used nodes that have at least one child, one parent and one spouse node in their Markov blanket. Multi-class target nodes transformed to binary by keeping the examples with value 1 as positives and the rest of the examples formed the negative class. Furthermore, we examined the nodes where the positive class is the minority with minimum prior probability of 0.15, to make sure that we will be able to label a minimum number of positive examples. For these networks we knew the true Markov blankets around each target variable and we compared them with the discovered blankets through the IAMB algorithm. For the supervised scenarios (i.e. when we used the variable Y) we performed 10 trials of size $N = 2000$ and 5000. For each trial we sampled 30 different partially labelled datasets, and the overall outcome of the partially labelled approaches was the most frequently derived Markov blanket.

As we observe from Figure 8.1, using $\widetilde{Y}_0$ instead of Y in the IAMB algorithm does not result to a statistically significant difference in the false positive rate. In Markov blanket terminology, the blankets derived from these two approaches are similar in terms of the variables that were *falsely added to the blanket.*

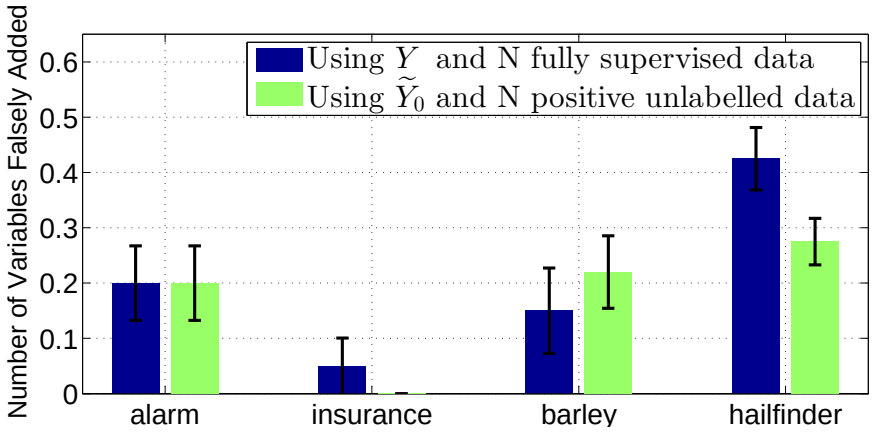
[1]Downloaded from `http://www.bnlearn.com/bnrepository/`

Network	Number of target nodes	Total number of nodes	Average MB length of target nodes	Average prior $p(y = 1)$ of target nodes
alarm	5	37	5.6	0.21
insurance	10	27	6.2	0.32
barley	10	48	5.6	0.31
hailfinder	20	56	4.9	0.31

Table 8.1: *Networks used in Markov blanket discovery experiments.*



(a) $N = 2000,\ p(\widetilde{y}_0 = 1) = 0.05$



(b) $N = 5000,\ p(\widetilde{y}_0 = 1) = 0.05$

Figure 8.1: *Verification of Theorem 4.10. This illustrates the average number of variables falsely added in MB and the 95% confidence intervals over 10 trials when we use IAMB with Y and $\widetilde{Y}_0$. (a) for total sample size $N = 2000$ out of which we label only 100 positive examples and (b) for total sample size $N = 5000$ out of which we label only 250 positives.*

8.1.1 Incorporating "Exact" Prior Knowledge in Sample Size Determination

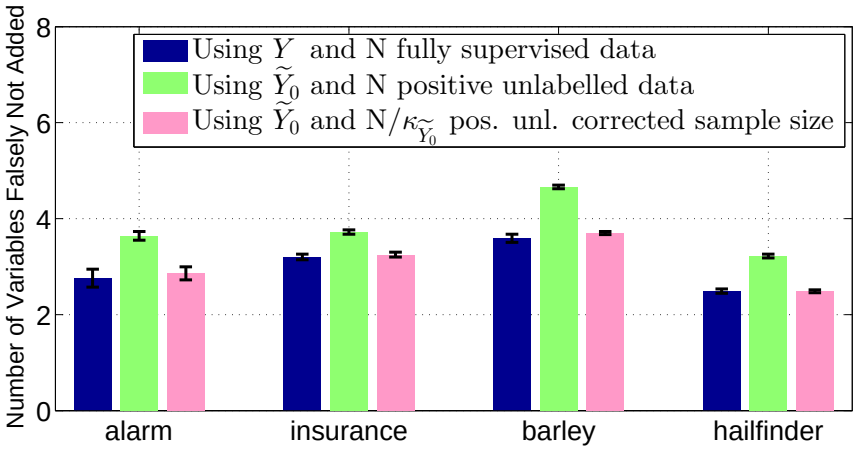
While the surrogate approach guarantees the same number of false positive errors, a direct consequence of Theorem 4.11 is that using $\widetilde{Y}_0$ instead of Y results in a higher number of *false negative* errors. By using the correction factor $\kappa_{\widetilde{Y}_0}$ and "exact" prior knowledge over the $p(y = 1)$ we can use the surrogate test for sample size determination, and decide the amount of data that we need in order to have similar performance with the unobservable fully supervised test in terms of false negatives.

In the MB discovery context this will result in a larger number of variables falsely not added to the predicted blanket, since we assumed that the variables were independent when in fact they were dependent. In order to verify experimentally this conclusion we will compare again the discovered blankets using $\widetilde{Y}_0$ instead of Y. As we see in Figure 8.2, the number of variables that were falsely not added is higher when we are using $\widetilde{Y}_0$. This Figure also verifies Theorem 4.11, where we see that the number of variables falsely removed when using the surrogate test $G(X; \widetilde{Y}_0|\mathbf{Z})$ with increased sample size $N/\kappa_{\widetilde{Y}_0}$ is the same as when using the unobservable test $G(X; Y|\mathbf{Z})$ with N data.

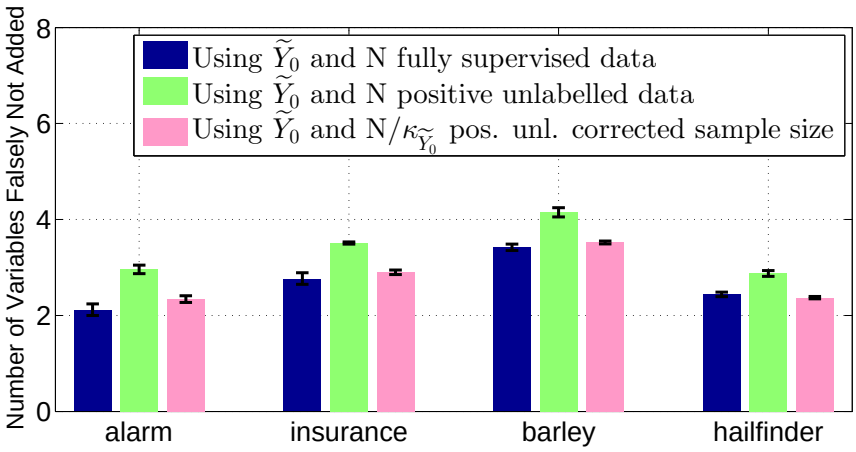
For completeness, at this point we should explore the order in which the features are evaluated in the growing and the shrinkage phases when we are using $\widetilde{Y}_0$ instead of Y. As we already mentioned, in Alg. 1 - Lines 4 and 10, we rank the features according to the p-values of the conditional tests of independence. Using $\widetilde{Y}_0$ instead of Y results to equivalent rankings, and this can be proved by combining the results of Section 6.4 with Theorem 4.11.

8.1.2 Evaluation of Markov Blanket Discovery in PU Data

For an overall evaluation of the derived blankets using $\widetilde{Y}_0$ instead of Y, we will use the F-measure, which is the harmonic mean of precision and recall, against the ground truth (Pocock et al., 2012). In Figure 8.3, we observe that the assumption of all unlabelled examples being negative gives worse results than the fully-supervised scenario, and that the difference between the two approaches gets smaller as we increase the sample size. Furthermore, using the correction factor $\kappa_{\widetilde{Y}_0}$ to increase the sample size of the surrogate approach makes the two techniques perform similarly.

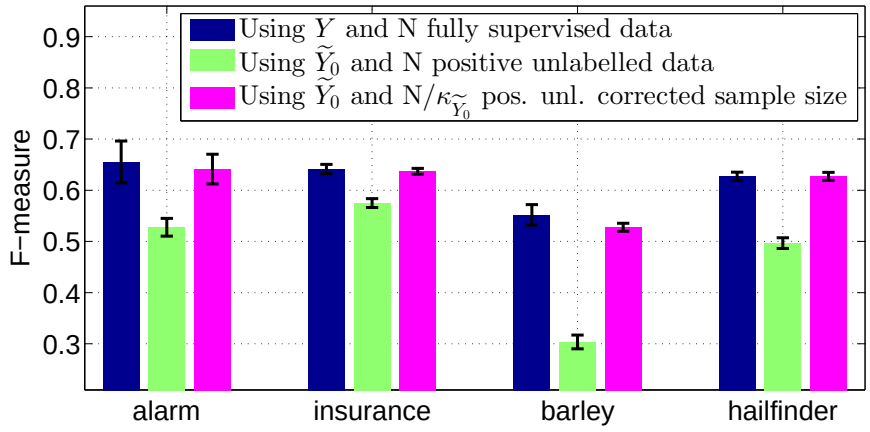


(a) $N = 2000,\ p(\widetilde{y}_0 = 1) = 0.05$

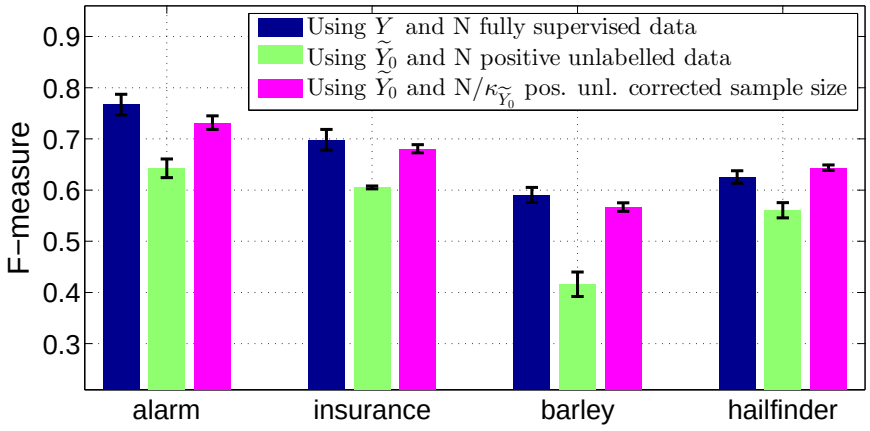


(b) $N = 5000,\ p(\widetilde{y}_0 = 1) = 0.05$

Figure 8.2: *Verification of Theorems 4.11. This illustrates the average number of variables falsely not added to the MB and the 95% confidence intervals over 10 trials when we use IAMB with Y and $\widetilde{Y}_0$. (a) for total sample size $N = 2000$ and (b) for total sample size $N = 5000$. In all the scenarios we label 5% of the total examples as positives.*



(a) $N = 2000,\ p(\widetilde{y}_0 = 1) = 0.05$



(b) $N = 5000,\ p(\widetilde{y}_0 = 1) = 0.05$

Figure 8.3: *Comparing the performance in terms of F-measure when we use IAMB with Y and $\widetilde{Y}_0$. (a) for total sample size $N = 2000$ and (b) for total sample size $N = 5000$. In all the scenarios we label 5% of the total examples as positives.*

8.2 Markov Blanket Discovery in Semi-Supervised Data

In this section we present how to discover the Markov blanket by incorporating "soft" prior knowledge in choosing the optimal way for testing conditional independence. Then we will explore the performance of this approach in a practical semi-supervised scenario.

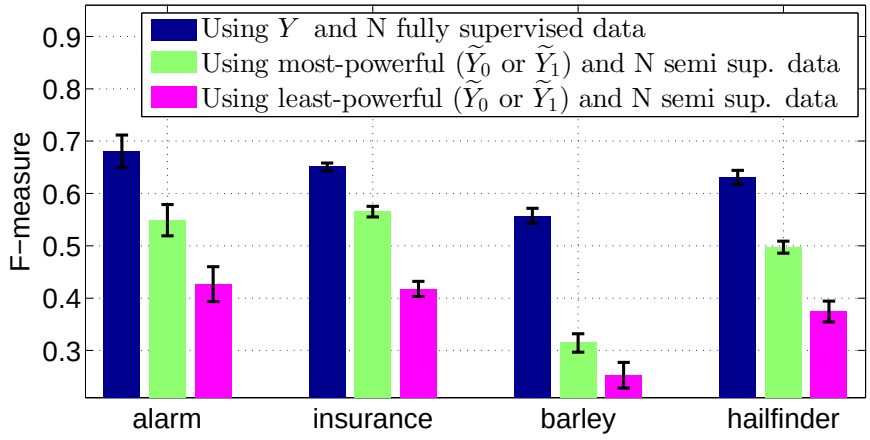
8.2.1 Incorporating "Soft" Prior Knowledge for Optimal Decision

Since the correction factors in the non-centrality parameters of the unconditional tests (Theorem 4.8) are the same with that of the conditional tests (Theorem 4.11), we can use inequality (4.1) and incorporate *"soft" prior knowledge* to decide which surrogate approach is the most powerful. In other words, in order to have the smallest number of falsely missing variables from the Markov blanket we should use $\widetilde{Y}_0$ instead of $\widetilde{Y}_1$, when inequality (4.1) holds. When the opposing inequality holds the most powerful choice is $\widetilde{Y}_1$. When equality holds, both approaches are equivalent.

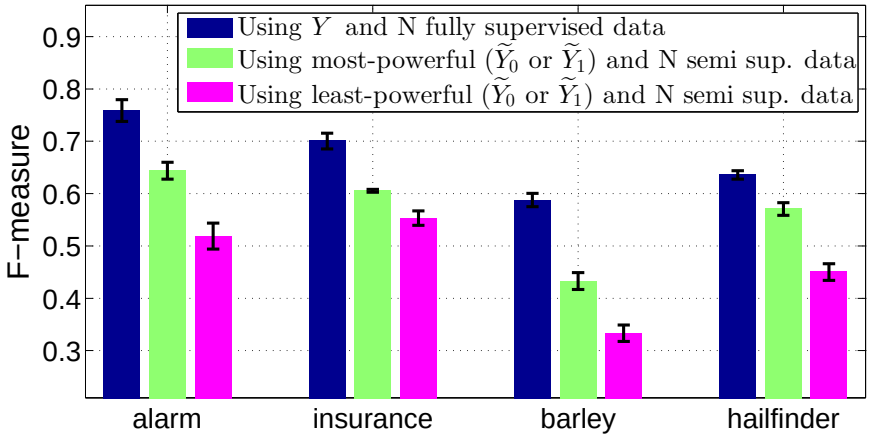
In Figure 8.4 we compare in terms of F-measure the derived Markov blankets when we use the most powerful and the least powerful choice. As we observe, by incorporating "soft" prior knowledge, and using inequality (4.1) to choose the most powerful option, get remarkably better performance than with the least powerful option.

8.2.2 Exploring our Framework Under Class Prior Change: When and How the Unlabelled Data Help

In this section, we will present how our approach performs in a real world problem where the class balance in the labelled set does not reflect the balance over the overall population; such situation is known as *class-prior-change* (Plessis and Sugiyama, 2012), Section 3.3.3 gives more details about the assumptions behind this scenario. We compare our framework with the following two approaches: ignoring the unlabelled examples (known as listwise deletion), or using the unlabelled data to have more reliable estimates for the marginal counts of the features



(a) $N = 2000,\ N_{\widetilde{y}_0=1} = 100,\ N_{\widetilde{y}_1=0} = 100$



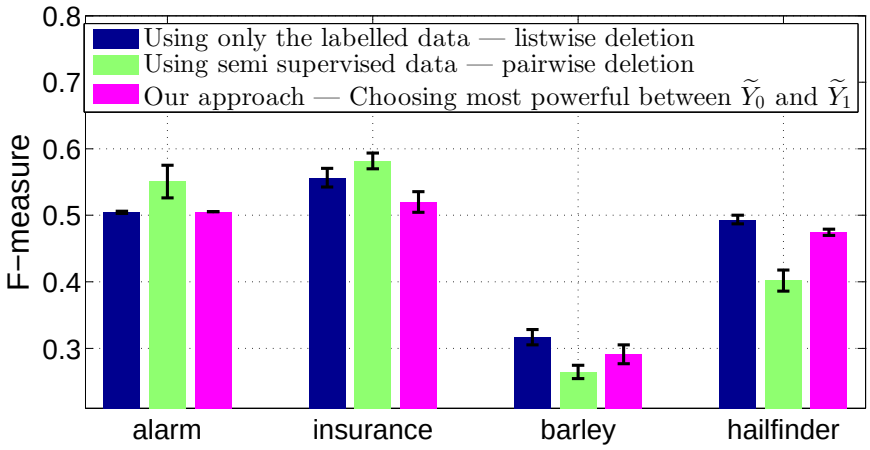
(b) $N = 5000,\ N_{\widetilde{y}_0=1} = 250,\ N_{\widetilde{y}_1=0} = 250$

Figure 8.4: *Comparing the performance in terms of F-measure when we use the unobservable variable Y and the most and least powerful choice between $\widetilde{Y}_0$ and $\widetilde{Y}_1$. (a) for sample size $N = 2000$ out of which we label only 100 positive and 100 negative examples and (b) for sample size $N = 5000$ out of which we label only 250 positive and 250 negative examples.*

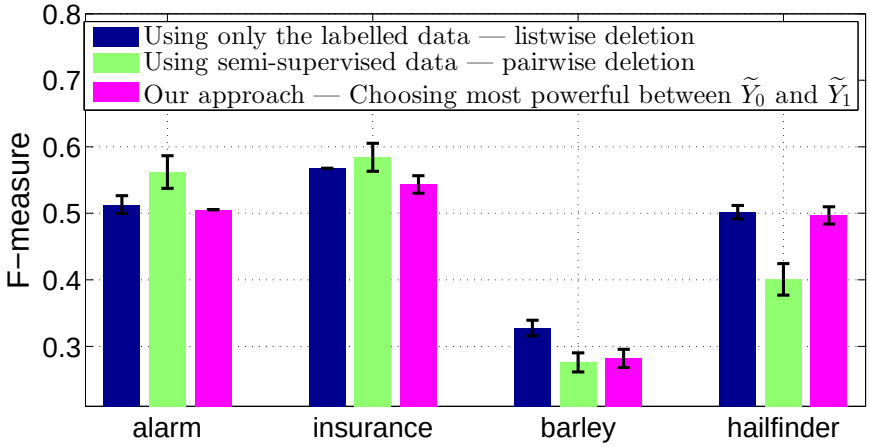
(known as pairwise deletion). The latter approach is followed in BASSUM (Cai et al., 2011); Section 2.4.3 gives more details about this approach and its limitations.

Firstly, let us assume that the semi-supervised data are generated under the "traditional semi-supervised" scenario, where the labelled set is an unbiased sample from the overall population, or in other words the labels are MCAR. As a result, the class-ratio in the labelled set is the same as the population class-ratio: $\frac{p(y=1|s=1)}{p(y=0|s=1)} = \frac{p(y=1)}{p(y=0)}$, where the *lhs* is the class-ratio in the labelled set and in *rhs* the population class-ratio. As we observe in Figure 8.5, choosing the most

powerful option between $\widetilde{Y}_0$ and $\widetilde{Y}_1$ performs similarly to ignoring completely the unlabelled examples. As it was expected, using the semi-supervised data with pairwise deletion has an unpredictable performance and often performs much worse than using only the labelled examples.



(a) $N = 2000,\ N_{s=1} = 200$

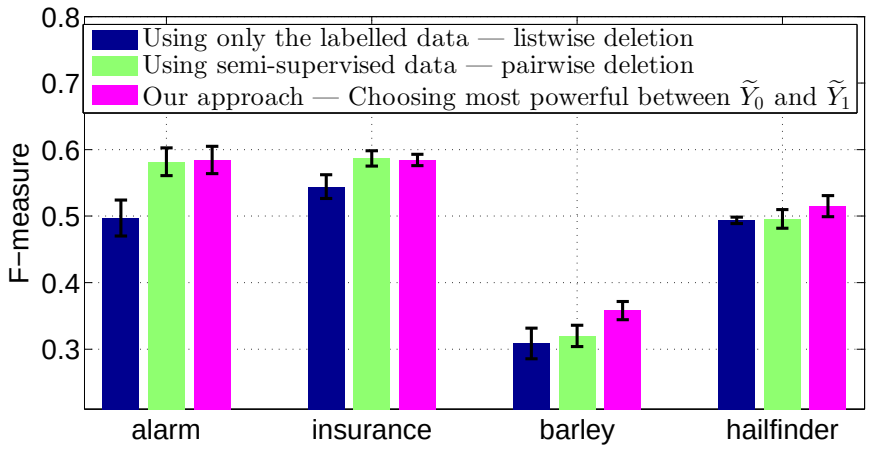


(b) $N = 5000,\ N_{s=1} = 250$

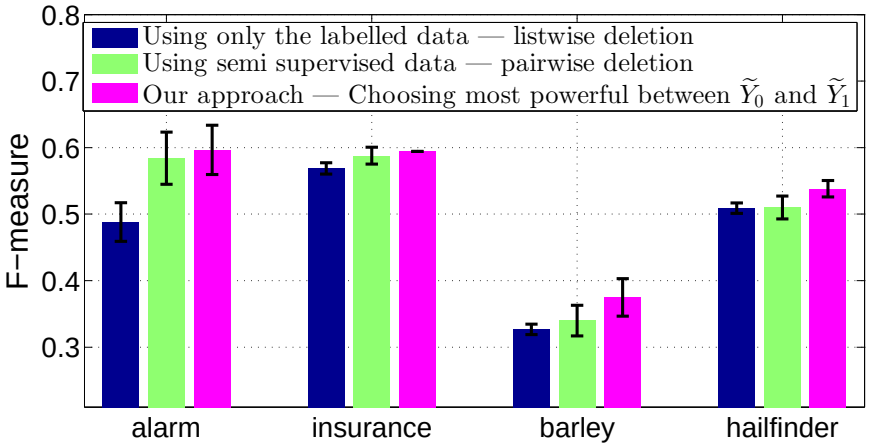
Figure 8.5: *Traditional semi-supervised scenario. Comparing the performance in terms of F-measure when we have the same class-ratio in the labelled-set as in the overall population. (a) for sample size* $N = 2000$ *out of which we label only 200 examples and (b)* $N = 5000$ *out of which we label only 250 examples.*

Now, let us assume we have semi-supervised data under the class-prior-change scenario (for more details see Section 3.3.3), or in other words the labels are MAR-C. In our simulation we sample the labelled data in order to have a class ratio in the labelled set inverse than the population ratio: $\frac{p(y=1|s=1)}{p(y=0|s=1)} = \left(\frac{p(y=1)}{p(y=0)}\right)^{-1}$, where the *lhs* is the class-ratio in the labelled set and in *rhs* the inverse of the

population class-ratio. As we observe in Figure 8.6, choosing the most powerful option between $\widetilde{Y}_0$ and $\widetilde{Y}_1$ performs statistically better than ignoring the unlabelled examples. Our approach performs better on average than the pairwise deletion, while the latter one performs comparably to the listwise deletion in many settings.



(a) $N = 2000,\ N_{s=1} = 200$



(b) $N = 5000,\ N_{s=1} = 250$

Figure 8.6: *Class prior change semi-supervised scenario. Comparing the performance in terms of F-measure when we have inverse class-ratio in the labelled-set than in the overall population. (a) for sample size* $N = 2000$ *out of which we label only 200 examples and (b)* $N = 5000$ *out of which we label only 250 examples.*

Furthermore, our approach can be applied in scenarios where we have labelled examples only from one class, which cannot be handled with the other two approaches. Also, with our approach, we can control the power of our tests, which

is not the case in the pairwise deletion procedure. To sum up, in class-prior-change scenarios we can use inequality (4.1) and some "soft" prior knowledge over $p(y = 1)$ in order to decide which of the following two assumptions is better: to assume all unlabelled examples are negative (i.e. use $\widetilde{Y}_0$) or to assume all unlabelled examples are positive (i.e. use $\widetilde{Y}_1$).

A future research direction could be to explore how we can use our methodology for structure learning of Bayesian networks. Since our techniques are informative in terms of power, they can be used in structure learning approaches that have control over the false negative rate to prevent over-constrained structures; for example, our framework generalises the work presented by Bacciu et al. (2013) for partially labelled data. Furthermore, our work for structure learning in partially labelled data can be used in combination with recently suggested methods for parameter learning from incomplete data by Van den Broeck et al. (2015).

8.3 Chapter Summary

With our work in Chapter 4, we derived a generalization of conditional tests of independence for partially labelled data, while in this chapter we presented a framework on how we use these tests for discovering Markov blankets around partially labelled target nodes.

In positive-unlabelled data, we proved that assuming all unlabelled examples to be negative is sufficient for testing conditional independence but it will increase the number of the variables that are falsely missing from the predicted blanket. Furthermore, with a correction factor, we quantified the amount of power we are losing by this assumption, and we presented how we can take this into account for adjusting the sample size in order to achieve similar performance to the fully-supervised scenarios.

Then, we extended our methodology to semi-supervised data, where we can make two valid assumptions over the unlabelled examples: to assume them either positive or negative. We explored the consequences of these two assumptions again in terms of possible errors in Markov blanket discovery procedures, and we suggested a way to use some "soft" prior knowledge to take the optimal decision. Finally, we presented a practical semi-supervised scenario in which the use of unlabelled examples under our framework proved to be more beneficial compared to other suggested approaches.

While in this chapter we used testing activities for feature selection, in the following one we will explore how we can use our suggested estimators to derive new heuristics for information theoretic feature selection in partially labelled scenarios.

Chapter 9

Extension 3: Filter Feature Selection in Partially Labelled Data

In the era of the Big Data, many datasets have a common characteristic, the large number of features. High dimensional feature spaces are associated with a number of problems, such as over-fitting to irrelevant features and high computational complexity. As a result, selecting the relevant features and ignoring the irrelevant and redundant is a key challenge.

In this chapter we explore how feature selection techniques can be extended from a fully to a partially labelled setting (Section 9.1), by using our analysis on estimation and ranking in partially labelled data (Chapters 5 and 6 respectively). Then, we explore the performance of our suggested techniques in a real world semi-supervised scenario: the class-prior-change scenario. Firstly, in Section 9.2, we focus on how similar are the selected feature subsets under our partially labelled approaches, in comparison to the fully supervised techniques. Then, Section 9.3 presents how the accuracy of a classifier can be improved by using the unlabelled data in order to select informative features.

9.1 Filter Feature Selection: From Fully to Partially Labelled Data

As we mentioned in Section 2.3, in a recent work, Brown et al. (2012) unified many heuristics for information theoretic feature selection under a single framework: maximising the conditional likelihood. This procedure results in a greedy optimization process, where at each step we select the feature X_k that maximizes the conditional mutual information:

$$X_k = \underset{X_k \in \mathbf{X}_{\tilde{\theta}}}{\arg\max}\, J_{cmi}(X_k) = \underset{X_k \in \mathbf{X}_{\tilde{\theta}}}{\arg\max}\, I(X_k; Y|X_\theta) = \underset{X_k \in \mathbf{X}_{\tilde{\theta}}}{\arg\max}\, I(\mathbf{X}_\theta X_k; Y),$$

where $\mathbf{X}_\theta$ are the features already selected, while $X_{\tilde{\theta}}$ the unselected. In the last step, we used the chain rule for mutual information (Cover and Thomas, 2006).

As the number of selected features grows, the dimension of $\mathbf{X}_\theta$ also grows, and this makes our estimates less reliable. Low order criteria have been derived to overcome this issue. Our work focuses on two low order criteria: the *Mutual Information Maximization* (MIM) (Lewis, 1992) and the Joint Mutual Information (JMI) (Yang and Moody, 1999), but can be naturally extended to any other criterion. The MIM criterion ranks the features according to their relevance with the target variable, and at each step k it selects the feature X_k with the highest score: $J_{mim}(X_k) = I(X_k; Y)$. JMI takes into account both the relevancy and the redundancy and selects the feature X_k with the highest score: $J_{jmi}(X_k) = \sum_{X_i \in \mathbf{X}_\theta} I(X_i X_k; Y)$. Brown et al. (2012) present a set of nice properties that JMI satisfies.

In the partially labelled scenario, we cannot estimate directly the mutual information between the features and the target. Instead we can use surrogate variables or our suggested estimators to derive ranking equivalent approaches (more details in Chapters 5 and 6). In the rest of this section we will explore the behaviour of the following partially labelled feature selection techniques:

- **Using only the labelled data $\mathcal{D}_L$**
  MIM version: $J^{\mathcal{D}_L}_{mim}(X_k) = \widehat{I}(X_k; Y|s=1)$
  JMI version: $J^{\mathcal{D}_L}_{jmi}(X_k) = \sum_{X_i \in \mathbf{X}_\theta} \widehat{I}(X_i X_k; Y|s=1)$.
- **Using our $MCAR$ estimator**
  MIM version: $J^{MCAR}_{mim}(X_k) = \widehat{I}_{MCAR}(X_k; Y)$
  JMI version: $J^{MCAR}_{jmi}(X_k) = \sum_{X_i \in \mathbf{X}_\theta} \widehat{I}_{MCAR}(X_i X_k; Y)$.
- **Using our $MAR - F$ estimator**
  MIM version: $J^{MAR-F}_{mim}(X_k) = \widehat{I}_{MAR-F}(X_k; Y)$
  JMI version: $J^{MAR-F}_{jmi}(X_k) = \sum_{X_i \in \mathbf{X}_\theta} \widehat{I}_{MAR-F}(X_i X_k; Y)$.
- **Using our Pos/Neg $MAR - C$ estimator and "exact" prior knowledge**
  MIM version: $J^{Pos/Neg}_{mim}(X_k) = \widehat{I}^{Pos}_{MAR-C}(X_k; Y)$ or $\widehat{I}^{Neg}_{MAR-C}(X_k; Y)$
  JMI version: $J^{Pos/Neg}_{jmi}(X_k) = \sum_{X_i \in \mathbf{X}_\theta} \widehat{I}^{Pos}_{MAR-C}(X_i X_k; Y)$ or
  $\sum_{X_i \in \mathbf{X}_\theta} \widehat{I}^{Neg}_{MAR-C}(X_i X_k; Y)$.
  These estimators require "exact" prior knowledge.
  To decide between Pos and Neg we use the prior as Conjecture 6.4 describes.
- **Using $\widetilde{Y}_0$ or $\widetilde{Y}_1$ instead of Y and "soft" prior knowledge**
  MIM version: $J^{\widetilde{Y}_0/\widetilde{Y}_1}_{mim}(X_k) = \widehat{I}(X_k; \widetilde{Y}_0)$ or $\widehat{I}(X_k; \widetilde{Y}_1)$
  JMI version: $J^{\widetilde{Y}_0/\widetilde{Y}_1}_{jmi}(X_k) = \sum_{X_i \in \mathbf{X}_\theta} \widehat{I}(X_i X_k; \widetilde{Y}_0)$ or $\sum_{X_i \in \mathbf{X}_\theta} \widehat{I}(X_i X_k; \widetilde{Y}_1)$.
  To decide between $\widetilde{Y}_0$ and $\widetilde{Y}_1$ we use "soft" prior as Conjecture 6.4 describes.

The last two approaches can also be used in partially labelled scenarios where we have labelled information only from one class (i.e. positive-unlabelled). Now, we will explore how the above ways for selecting features in partially labelled environments perform under the class-prior-change scenario.

9.2 Exploring the Consistency of the Selected Subsets Under Class Prior Change

An interesting question to explore is "how do the selected feature subsets obtained through the partially labelled approaches differ from the ones that we would have using the unobservable target variable Y?" To evaluate the performance of the different approaches, we will measure the similarity between the feature subset that is returned by them and the feature subset that we would have if we had full supervision over the target variable. For our experiments we are always selecting the top-10 features, while the similarity will be measured by Kuncheva's

consistency index (Kuncheva, 2007).

Definition 9.1. *The consistency between two subsets of features* $\mathbf{X_i}, \mathbf{X_j} \subset \mathbf{X}$ *of the same size* $|\mathbf{X_i}| = |\mathbf{X_j}| = k$*, where* $0 < k < |\mathbf{X}|$*, is*

$$C(\mathbf{X_i}, \mathbf{X_j}) = \frac{rn - k^2}{k(n-k)},$$

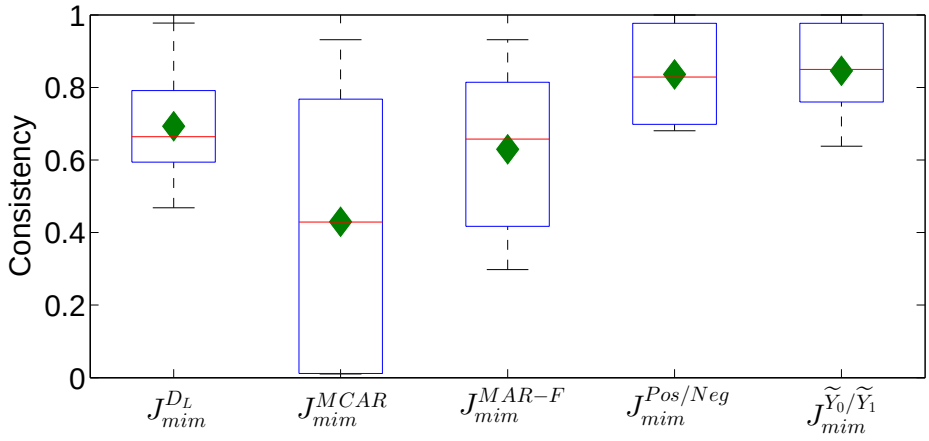
where $r = |\mathbf{X_i} \cap \mathbf{X_j}|$.

Table 9.1 gives details over the six datasets that we use in our experiments. Multi-class datasets were transformed to binary by using the minority class as positive. Continuous features, using Sturges' rule (Sturges, 1926), were discretized to: $\lceil \log_2 N + 1 \rceil$ bins. These datasets are fully-supervised, and we sample them in order to generate semi-supervised versions under class prior change (the labels are MAR-C). We label only 10% of the examples, $p(s = 1) = 0.10$, in such a way that the ratio between positives and negatives in the labelled set to be $2 : 1$. This sampling generates class prior change, since in the population situation the positive class is the minority, while in the labelled set it becomes majority.

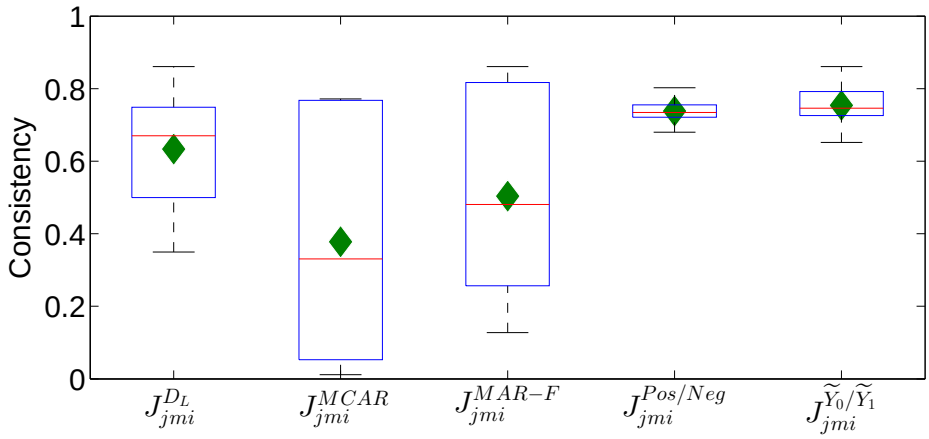
Dataset	# examples	# features	$\widehat{p}(y = 1)$
krvskp	3196	36	0.48
landsat	6435	36	0.10
musk2	6598	166	0.16
semeion	1593	256	0.10
splice	3175	60	0.24
waveform	5000	40	0.33

Table 9.1: *Datasets used in the feature selection experiments.*

Figure 9.1 shows that the approaches that use the $Pos/Neg\ MAR - C$ estimator and incorporate perfect prior knowledge and the approach that uses $\widetilde{Y}_0/\widetilde{Y}_1$ and "soft" prior knowledge outperform the other approaches. This trend is more obvious in the JMI case, where we estimate the mutual information more times than in MIM. Another interesting point is that the approach that uses "soft" prior knowledge, and assumes the unlabelled examples are all positive or all negative, performs very similarly to the approach that uses "perfect" prior knowledge.



(a) *MIM*



(b) *JMI*

Figure 9.1: *Kuncheva's Consistency index between the feature subsets returned through fully supervised MIM/JMI and the ones returned using the partially labelled approaches. In this graph we present box plots and expected values (diamonds) across the 6 datasets. We average the index over 30 semi-supervised datasets with labels MAR-C and* $p(s = 1) = 0.10$*, such that the probability of a positive example in the labelled set to be* $p(y = 1|s = 1) = 2/3$ *and* $p(y = 0|s = 1) = 1/3$*.*

9.3 Exploring the Misclassification Error Under Class Prior Change

In this section we will explore the performance of the semi-supervised criteria in terms of misclassification error. We used 10 train/test splits with 50% of the data used for training and 50% for testing. In order to explore the performance of the partially labelled feature selection approaches we average the misclassification

error over 30 semi-supervised datasets. To generate the semi-supervised data we labelled 10% of the training examples $p(s = 1) = 0.10$ in the same way as in the previous section, to generate a ratio of positive to negative examples $2 : 1$ in the labelled set. We selected the top 10 features by using different criteria, and then we used a k-nearest neighbor classifier ($k = 3$), since this classifier does not make any probabilistic assumption

Table 9.2 presents the misclassification error (average values and standard deviations of the 10 different splits). As we observe, by taking into account the unlabelled examples and by using some "soft" prior knowledge to decide between $\widetilde{Y}_0/\widetilde{Y}_1$, we perform better than using only the labelled data. Furthermore, our suggested JMI version of this technique, $J_{jmi}^{\widetilde{Y}_0/\widetilde{Y}_1}$, which takes into account both the relevancy and the redundancy, outperforms the MIM version, $J_{mim}^{\widetilde{Y}_0/\widetilde{Y}_1}$.

Dataset	Partially labelled			
	Ignore unlabelled		Use unlabelled	
	$J_{mim}^{\mathcal{D}_L}$	$J_{jmi}^{\mathcal{D}_L}$	$J_{mim}^{\widetilde{Y}_0/\widetilde{Y}_1}$	$J_{jmi}^{\widetilde{Y}_0/\widetilde{Y}_1}$
krvskp	0.0715 ± 0.0009	0.0674 ± 0.0026	0.0831 ± 0.0012	**0.0671** ± 0.0030
landsat	0.1071 ± 0.0028	0.0981 ± 0.0006	0.1040 ± 0.0016	**0.0928** ± 0.0006
musk2	0.1013 ± 0.0008	0.0879 ± 0.0022	0.1030 ± 0.0006	**0.0802** ± 0.0043
semeion	0.0821 ± 0.0009	0.0769 ± 0.0004	0.0660 ± 0.0029	**0.0648** ± 0.0010
splice	0.1100 ± 0.0001	0.1089 ± 0.0010	0.1078 ± 0.0006	**0.1065** ± 0.0009
waveform	0.1554 ± 0.0012	0.1455 ± 0.0019	0.1583 ± 0.0001	**0.1451** ± 0.0010

Table 9.2: *Comparison of the misclassification error using features derived from different criteria. Bold indicated the best performance.*

9.4 Chapter Summary

In this chapter, we showed how popular feature selection techniques can be extended from a fully to a partially labelled setting. By extending our theoretical results, we suggested different criteria by using different estimators for the mutual information. Furthermore, we presented, in the class-prior-change semi-supervised scenario, how to use the unlabelled examples and "soft" prior knowledge, in order to improve the performance that we would have by using only the labelled information. We leave exploring the performance of the suggested estimators under different missingness scenarios, such as MAR-F, as a future work.

Chapter 10

Conclusions and future directions

By focusing on the generalised test of independence and on the mutual information, we have presented a comprehensive study on testing, estimation, and ranking, in partially labelled data in an entirely classifier-independent/inference-free manner. This encompasses semi-supervised and positive-unlabelled scenarios. Furthermore, we showed how we can use these results for feature selection via testing and estimation. In the present chapter we provide a summary of the contributions of this thesis, provide guidance for practitioners and present several interesting areas for feature work.

10.1 What we Have Learnt About ...

10.1.1 Testing, Estimation and Ranking in Partially Labelled Data

In the beginning of this work (Chapter 1) we posed three tangled questions on testing, estimation and ranking of features. To give sensible answers in the partially labelled scenarios, we modelled the underlying mechanism of the missing labels using the formalism of m-graphs (Mohan et al., 2013). We connected this formalism with the three main assumptions used in partially labelled scenarios: MCAR, MAR-F and MAR-C, and we explored how to answer our three main questions in each one of them. Here we summarize the answers to these three questions:

Question 1 – Testing: ***"Is feature X significantly correlated to the partially observed label Y?"***

To answer this question, in our first theoretical result, we showed how *surrogate* variables could be used to conduct valid and informed hypothesis testing activities, in place of the partially observed class variable (Chapter 4). By incorporating the user's belief over the prior probability of the class we can use these surrogate valid tests for a-priori power analysis activities, such as sample size determination. Furthermore, we can decide which is the optimal surrogate approach in terms of minimising the false negative rate.

Question 2 – Estimation: ***"How strong is the dependency between X and the partially observed Y?"***

To answer this question, we derived consistent estimators for the mutual information, *despite the partial labelling*, and we showed how to incorporate the user's prior belief over the class distribution in order to overcome the biases introduced due to partial labelling (Chapter 5).

Question 3 – Ranking: ***"Using a finite sample of data, can we recover a ranking of features that will be close to what we would obtain if we had access to the full data distribution?"***

In our last theoretical result, we presented how to use the surrogate variables and our suggested estimators in order to rank the features as if we had fully supervised data (Chapter 6).

We answer these questions for both semi-supervised and positive-unlabelled data. For example, in positive-unlabelled scenario our contributions can be summarised as follows. (1) We proved that assuming all unlabelled examples are negative cases is sufficient for independence testing, but not for power analysis activities. (2) We suggested a new methodology that compensates for this and enables power analysis, allowing sample size determination for observing an effect with a desired power by incorporating users prior knowledge over the prevalence of positive examples. (3) We provide a new capability, supervision determination, which allows us to determine a-priori the number of labelled examples the user must collect before being able to observe a desired statistical effect. (4) We derived an estimator of the mutual information in positive-unlabelled data, and its asymptotic distribution. (5) Finally, we showed how to rank features with or without prior knowledge. Table 10.1 summarizes our theoretical findings, and gives advice to practitioners on how they can test, estimate and rank under

different partially labelled scenarios.

Building upon our theoretical results on testing, estimation and ranking, we proceeded to show three useful extensions of them on the area of experimental design, Markov blanket discovery and information theoretic feature selection.

10.1.2 Extension 1: Experimental Design in Partially Labelled Data

In Chapter 7, we explored how our theoretical results can be used for experimental design activities, such as sample size determination, in partially labelled scenarios. Furthermore, we presented a new capability, not previously demonstrated — which we call "supervision" determination — determining the number of *labelled* examples that are needed to achieve a given false positive and false negative rate. One particularly interesting example (Table 7.2) demonstrates that given $N = 3000$ examples, with knowledge of a class prior at $p(y = 1) = 0.2$, we need to manually label only 49 positive examples of the 3000, to exactly recover the dynamics of the original (unobservable) independence test, leaving the remaining 2951 entirely unlabelled.

10.1.3 Extension 2: Markov Blanket Discovery in Partially Labelled Data

With our work we derived a generalization of conditional tests of independence for partially labelled data and we presented a framework on how we can use unlabelled data for discovering Markov blankets around partially labelled target nodes (Chapter 8). In positive-unlabelled data, we proved that assuming all unlabelled examples are negative is sufficient for testing conditional independence but it will increase the number of variables that are falsely missing from the predicted blanket. Furthermore, with a correction factor, we quantified the amount of power we are losing with this assumption, and we presented how we can take this into account for adjusting the sample size in order to achieve similar performance to the fully-supervised scenarios. Then, we extended our methodology to semi-supervised data, where we can make two valid assumptions over the unlabelled examples: assume them either all positive or all negative. We explored the consequences of these two assumptions again in terms of possible errors in Markov blanket discovery procedures, and we suggested a way to use some "soft"

Data type	Task	Recommendation
MCAR	Testing	Ignore unlabelled data, compute G-test as usual. This enables you to control both FPR and FNR.
	Estimation	Use equation (5.5), i.e. compute $p(x)$ from $\mathcal{D}_L \cup \mathcal{D}_U$, but conditionals only from $\mathcal{D}_L$.
	Ranking	As above, use equation (5.5).
MAR-F	Testing	Ignore unlabelled data, compute G-test as usual. This is the only case where the FPR can be controlled.
	Estimation	Use equation (5.6), i.e. compute $p(x)$ and $p(y)$ from $\mathcal{D}_L \cup \mathcal{D}_U$, but conditionals only from $\mathcal{D}_L$.
	Ranking	As above, use equation (5.6).
MAR-C	Testing	Assume missing values are either all positive, or all negative, then compute G-test as usual. This enables you to control both FPR and FNR, the latter by having "exact" prior knowledge. Using "soft" prior knowledge and inequality (4.1) we can decide which approach minimizes FNR.
	Estimation	Use equation (5.8) or (5.9) , i.e. compute $p(x)$ from $\mathcal{D}_L \cup \mathcal{D}_U$, conditionals only from $\mathcal{D}_L$ and "exact" prior knowledge for $p(y)$. Use inequality (4.1) to decide between the two equations, if it holds then use equation (5.8), otherwise use equation (5.9). With this way you can derive point and interval estimates.
	Ranking	With "exact" prior knowledge estimate as above. With "soft" prior knowledge use inequality (4.1) do decide whether to assume unlabelled examples as positives or negatives and rank features as usual.
Positive unlabelled	Testing	Assume missing values are all negative, then compute G-test as usual. This enables you to control both FPR and FNR, the latter by having "exact" prior knowledge.
	Estimation	Use equation (5.8), i.e. compute $p(x)$ from $\mathcal{D}_L \cup \mathcal{D}_U$, conditionals only from $\mathcal{D}_L$ and use "exact" prior knowledge for $p(y = 1)$. This approach enables you to derive point and interval estimates.
	Ranking	With "exact" prior knowledge estimate as above. With no prior knowledge assume unlabelled examples as negatives and rank features as usual.

Table 10.1: *Guidance to practitioners. When referring to "prior knowledge" we mean of the class probability $p(y)$.*

prior knowledge to take the optimal decision. Finally, we presented a practical semi-supervised scenario in which the usage of unlabelled examples under our framework proved to be more beneficial compared to other suggested approaches.

10.1.4 Extension 3: Information Theoretic Feature Selection in Partially Labelled Data

Finally, in Chapter 9, we explored different ways to derive information theoretic feature selection criteria in partially labelled data. We suggested criteria that capture both the relevancy and the redundancy, while we showed how to incorporate prior knowledge and how to use the unlabelled data in an efficient way.

10.2 Future Work

This thesis reveals many interesting areas for future work.

Can we extend the framework to handle continuous features?
Our work explored how to test, estimate and rank categorical features when the binary target variable is partially observed. An interesting extension will be to explore these activities when we have continuous features. In the categorical setting, our work built upon the relationship of the G-test of independence and the mutual information as a measure of the effect size. In statistics, one of the main ways to measure the correlation between a continuous and a binary (dichotomous) variable, is to use the point-biserial correlation coefficient as measure of the effect size. There is a natural relationship between this measure and the unpaired two sample t-test (Kornbrot, 2005; Rosenthal et al., 2000). Building upon this observation, we can explore the unpaired two sample t-test in semi-supervised and positive-unlabelled data, under the different missingness scenarios, and how we can suggest consistent estimates for the point-biserial coefficient in these scenarios.

Can we use our surrogate variables and our estimators for Markov blanket discovery and feature selection when the labels are MAR-F?
Sections 8.2.2, 9.2 and 9.3 explored how we can use our results for Markov blanket discovery and feature selection, when we have semi-supervised data

with class-prior change under the MAR-C scenario. It will be interesting to explore practical algorithms to deal with semi-supervised data where the labels are MAR-F. In this scenario, we do not have any informed approach to test independence, but, as we presented, we can have consistent estimates of the mutual information. In the MAR-F scenario the missingness variable S will depend on a subset of these features. Because of that we cannot treat all the features in the same way. Firstly, we should identify the features that are parents of the variable S. This can be done by using all the available information. Then, we can use our MAR-F estimator to decide the strength of the effect between these features and the partially observed target. Finally, in order to decide the strength of the effect between other features which are not parents of S and the target variable Y, we can use our MCAR estimators if we condition over the parents of S. Combining the ideas above we can suggest practical algorithms for feature selection when the labels are MAR-F, and we can see how we can extend them even in scenarios where the labels are MNAR.

How can we explore test, estimation and ranking when the labels are MNAR?

Another possible direction is to explore under which assumptions over the model and what type of prior knowledge do we need in order to perform testing, estimation and ranking when the labels are MNAR — one potential strategy can be to decompose the problem into MAR-C and MAR-F subproblems.

Can we extend the framework to Bayesian testing and estimation?

In our work we incorporate prior knowledge over $p(y)$ in many different ways, but always in a frequentist setting. On the other hand, Bayesian methods provide us with a natural way to incorporate prior knowledge over the parameters. As a result, an interesting extension of our work is to explore Bayesian testing and estimation in partially labelled data. In Bayesian statistics, the main way to test hypothesis is through the Bayes factors, which require the specification of a prior distribution on both null and alternative hypothesis models. This raises many problems, since the values of the factors depend critically on these priors, while they raise many computational issues, especially in high dimensional settings. To overcome

these problems, Johnson (2005) presented a different approach to define Bayes factors based on modelling directly the test statistic. There is a natural connection between the frequentist local alternative assumption, a main building block of our analysis, and the modelling followed by Johnson (2005). Building upon this, we can suggest Bayes factors using the G-test statistic in a fully-supervised setting, and we can further explore how this approach can be extended in dealing with partially labelled data. Furthermore, by modelling directly the test statistic, we can derive credible intervals for the mutual information.

Can we apply our results to different kinds of uncertain data, such as when the target variable is under-reported?

Underreported variables are commonplace in many medical scenarios — for example 10% of pregnant mothers may admit they smoke, but anecdotal reporting suggests this percentage may be significantly higher in reality. This non-disclosure bias creates a problem for comparisons of effect sizes between studies. Other examples of under-reported variables include childhood abuse or HIV infection, due to perceived stigma or fear of reprisal. By using our suggested estimator in Section 5.4, under some assumptions, we can correct the mutual information in this scenario to account for such under-reported variables. Unlike previous approaches (e.g. Yang et al. (2010)) our solution to these problems constitutes an entirely *inference-free* approach, i.e. does not involve imputation of values.

Can we adapt the MRMR feature selection criterion for different missingness scenarios?

Brown et al. (2012) re-wrote the CMI criterion as a sum of three terms describing the relevance, the redundancy and the conditional redundancy:

$$J'_{cmi}(X_k) = I(X_k;Y) - \beta \sum_{X_j \in S} I(X_k;X_j) + \gamma \sum_{X_j \in S} I(X_k;X_j|Y). \qquad (10.1)$$

By giving different values to the parameters β and γ we can derive different criteria. For example, by setting $\beta = \gamma = 0$ we derive the MIM, by setting $\beta = \gamma = \frac{1}{|S|}$ we derive the JMI, while by setting $\beta = \frac{1}{|S|}$ and $\gamma = 0$ we derive the Max-Relevance Min-Redundancy (MRMR) criterion (Peng et al., 2005), a popular feature selection technique. In the semi-supervised

data, by using the re-writing of equation (10.1), we can use the labelled and unlabelled data to estimate the redundancy term, $I(X_k; X_j)$. On the other hand, we can use only the labelled set to estimate the relevance, $I(X_k; Y)$, and the conditional redundancy term, $I(X_k; X_j|Y)$. A recent work by He et al. (2015) follows this idea but focuses only on the MRMR criterion. The authors suggest a transductive version that uses both labelled and unlabelled examples for estimating the redundancy under the MCAR assumption. With our work, we can derive semi-supervised generalisations for the MRMR criterion under any missingness scenario (MCAR, MAR-F and MAR-C).

Appendix A

Proofs and sketches of proofs

Theorem 2.1

Before giving the proof, we will formally introduce the delta method (Agresti, 2013, Section 16.1.4).

Lemma A.1 (Delta method)**.**
Suppose that cell counts $\boldsymbol{n} = \{n_{x,y}\}$ *have a multinomial distribution with cell probabilities* $\boldsymbol{p} = \{p(x,y)\}$, $\forall\ x \in \mathcal{X},\ y \in \mathcal{Y}$. *Let* $N = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} n_{x,y}$, *and let* $\hat{\boldsymbol{p}}$ *denote the sample proportions:* $\hat{p}(x,y) = n_{x,y}/N$. *Let* $g(\boldsymbol{p}) \in \mathbb{R}$ *be a differentiable function, and let* $\phi_{x,y} = \frac{\partial g}{\partial p(x,y)}(\boldsymbol{p})$, $\forall\ x \in \mathcal{X},\ y \in \mathcal{Y}$. *Assume that at least one* $\phi_{x,y}$ *is nonzero and then the distribution* $\sqrt{N}\left[g(\hat{\boldsymbol{p}}) - g(\boldsymbol{p})\right]$ *converges to the normal distribution* $\mathcal{N}(0, \sigma^2)$ *when* $N \to \infty$, *where* $\sigma^2 = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y)\phi_{x,y}^2 - \left(\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y)\phi_{x,y}\right)^2$.

In order to compute the partial derivatives $\frac{\partial g(\boldsymbol{p})}{\partial p(x,y)} = \frac{\partial I(X;Y)}{\partial p(x,y)}, \forall\ x \in \mathcal{X},\ y \in \mathcal{Y}$ we first need to calculate the following partial derivatives:

- $\frac{\partial p(x',y')}{\partial p(x,y)} = \delta_{xx'}\delta_{yy'}$, where $\delta_{xx'}\delta_{yy'}$ is the Kronecker delta: $\delta_{xx'}\delta_{yy'} = 1$ if $x = x'$ and $y = y'$, otherwise it equals 0
- $\frac{\partial p(x')}{\partial p(x,y)} = \delta_{xx'}$, where $\delta_{xx'}$ is the Kronecker delta: $\delta_{xx'} = 1$ if $x = x'$, otherwise it equals 0
- $\frac{\partial p(y')}{\partial p(x,y)} = \delta_{yy'}$
- $\frac{\partial (p(x')p(y'))}{\partial p(x,y)} = \delta_{xx'}p(y') + \delta_{yy'}p(x')$

- $\frac{\partial \ln \frac{p(x',y')}{p(x')p(y')}}{\partial p(x,y)} = \frac{1}{p(x',y')} \frac{\frac{\partial p(x',y')}{\partial p(x,y)} p(x')p(y') - p(x',y') \frac{\partial (p(x')p(y'))}{\partial p(x,y)}}{p(x')p(y')}$
$= \frac{1}{p(x',y')} \left(\frac{\partial p(x',y')}{\partial p(x,y)} - \frac{p(x',y')}{p(x')p(y')} \frac{\partial (p(x')p(y'))}{\partial p(x,y)} \right) =$
$= \frac{1}{p(x',y')} \left(\delta_{xx'}\delta_{yy'} - \frac{p(x',y')}{p(x')p(y')} \left(\delta_{xx'} p(y') + \delta_{yy'} p(x') \right) \right) =$
$= \frac{1}{p(x',y')} \left(\delta_{xx'}\delta_{yy'} - \frac{p(x,y')}{p(x)} - \frac{p(x',y)}{p(y)} \right)$

We are now ready to compute the partial derivatives of the mutual information estimator:

$$
\begin{aligned}
\phi_{x,y} = \frac{\partial g(\boldsymbol{p})}{\partial p(x,y)} = \frac{\partial I(X;Y)}{\partial p(x,y)} &= \frac{\partial \sum_{x' \in \mathcal{X}} \sum_{y' \in \mathcal{Y}} p(x',y') \ln \frac{p(x',y')}{p(x')p(y')}}{\partial p(x,y)} = \\
&= \sum_{x' \in \mathcal{X}} \sum_{y' \in \mathcal{Y}} \left(\frac{\partial p(x',y')}{\partial p(x,y)} \ln \frac{p(x',y')}{p(x')p(y')} + p(x',y') \frac{\partial \ln \frac{p(x',y')}{p(x')p(y')}}{\partial p(x,y)} \right) = \\
&= \sum_{x' \in \mathcal{X}} \sum_{y' \in \mathcal{Y}} \left(\delta_{xx'}\delta_{yy'} \ln \frac{p(x',y')}{p(x')p(y')} + \delta_{xx'}\delta_{yy'} - \frac{p(x,y')}{p(x)} - \frac{p(x',y)}{p(y)} \right) = \\
&= \ln \frac{p(x,y)}{p(x)p(y)} + 1 - \sum_{y' \in \mathcal{Y}} \frac{p(x,y')}{p(x)} - \sum_{x' \in \mathcal{X}} \frac{p(x',y)}{p(y)} = \\
&= \ln \frac{p(x,y)}{p(x)p(y)} + 1 - \frac{p(x)}{p(x)} - \frac{p(y)}{p(y)} = \\
&= \ln \frac{p(x,y)}{p(x)p(y)} - 1
\end{aligned}
$$

so the sample variance become

$$
\begin{aligned}
\sigma^2 &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \phi_{x,y}^2 - \left(\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \phi_{x,y} \right)^2 \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \left(\ln \frac{p(x,y)}{p(x)p(y)} - 1 \right)^2 \\
&\quad - \left(\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \left(\ln \frac{p(x,y)}{p(x)p(y)} - 1 \right) \right)^2 \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \left(p(x,y) \left(\ln \frac{p(x,y)}{p(x)p(y)} \right)^2 + p(x,y) - 2p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \right) \\
&\quad - \left(\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} - 1 \right)^2
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\left(\ln\frac{p(x,y)}{p(x)p(y)}\right)^2 + 1 - 2\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\ln\frac{p(x,y)}{p(x)p(y)} \\
&\quad - \left(\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}\right)^2 - 1 + 2\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\ln\frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\left(\ln\frac{p(x,y)}{p(x)p(y)}\right)^2 - \left(\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\ln\frac{p(x,y)}{p(x)p(y)}\right)^2 \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\left(\ln\frac{p(x,y)}{p(x)p(y)}\right)^2 - I(X;Y)^2
\end{aligned}
$$

□

Theorem 2.2

This sampling distribution can be derived by the fact that the G-statistic, under the local alternative assumption, follows a non-central χ^2 distribution with:

$$
G(X;Y) = 2N\hat{I}(X;Y) \sim \chi^2(\nu,\lambda) \text{ with } \begin{cases} \nu = (|\mathcal{X}|-1)(|\mathcal{Y}|-1) \\ \lambda = 2NI(X;Y) \end{cases}.
$$

For more details check the references of Section 2.2. □

Theorem 4.3

When the labels are MCAR, the following equations hold, which are useful for our proofs:

$$
p(x,y|s=1) = p(x,y), \tag{A.1}
$$

$$
p(x|s=1) = p(x)), \tag{A.2}
$$

$$
p(y|s=1) = p(y). \tag{A.3}
$$

Surrogate 1 ($\mathcal{D}_L$): To prove that $X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp Y|s=1$, we need to prove:

$$
p(x,y|s=1) = p(x|s=1)p(y|s=1) \Leftrightarrow p(x,y) = p(x)p(y)\ \forall\ x\ \in\ \mathcal{X},\ y\ \in\ \mathcal{Y}.
$$

The proof is straightforward:

$$p(x, y|s = 1) = p(x|s = 1)p(y|s = 1) \overset{(A.1),(A.2),(A.3)}{\Leftrightarrow} p(x, y) = p(x)p(y)$$

Surrogate 2 ($\widetilde{Y}_m$), **Surrogate 3** ($\widetilde{Y}_0$) and **Surrogate 4** ($\widetilde{Y}_1$): We can prove that these surrogate approaches are valid by following the same methodology, or we can check that the independence relationships hold from the m-graph of MCAR in Figure 3.1a. □

Theorem 4.4

Surrogate 1 ($\mathcal{D}_L$): The non-centrality parameter of this surrogate test is equal to $\lambda_{G(X;Y|s=1)} = 2N_{s=1}I(X;Y|s = 1)$, where $N_{s=1}$ represents the size of the labelled set. With straightforward calculations, and using eq. (A.1) - (A.3), we can show that when the labels are MCAR it holds $I(X;Y|s = 1) = I(X;Y)$. So the non-centrality parameter of the surrogate approach can be written as:

$$\lambda_{G(X;Y|s=1)} = 2N_{s=1}I(X;Y) \Leftrightarrow \lambda_{G(X;Y|s=1)} = \frac{N_{s=1}}{N}2NI(X;Y)$$

The fraction $\frac{N_{s=1}}{N}$ represents the probability of labelling an example $p(s = 1)$, while $2NI(X;Y)$ is the non-centrality parameter of the unobservable test. Thus, $\lambda_{G(X;Y|s=1)} = p(s = 1)\lambda_{G(X;Y)}$, and the correction factor is $\kappa = p(s = 1)$.

Surrogate 2 ($\widetilde{Y}_m$): The non-centrality parameter of this surrogate test is equal to $\lambda_{G(X;\widetilde{Y}_m)} = 2NI(X;\widetilde{Y}_m)$. With straightforward calculations, and using eq. (A.1) - (A.3), we can show that when the labels are MCAR it holds $I(X;\widetilde{Y}_m) = p(s = 1)I(X;Y)$. So the non-centrality parameter of the surrogate approach can be written as:

$$\lambda_{G(X;\widetilde{Y}_m)} = p(s = 1)2NI(X;Y) \Leftrightarrow \lambda_{G(X;\widetilde{Y}_m)} = p(s = 1)\lambda_{G(X;Y)}$$

Thus the correction factor is $\kappa = p(s = 1)$.

Surrogate 3 ($\widetilde{Y}_0$): In order to prove that relationship we will use the result of Haberman (1974) which states that when we assume local alternatives the

X^2 and the G-test have the same asymptotic power (Haberman, 1974, p. 109). In other words, the non-centrality parameters of the distributions of the two statistics converge to a common value as $N \to \infty$ (Agresti, 2013, Section 16.3.5). So instead of exploring the relationship of the non-centrality parameters for the G-tests between $X, \widetilde{Y}_0$ and X, Y, we can explore the relationship between the non-centrality parameters of the X^2-tests between $X, \widetilde{Y}_0$ and X, Y. The non-centrality parameter of this X^2 surrogate test is equal to $\lambda_{X^2(X;\widetilde{Y}_0)} = 2NI_2(X;\widetilde{Y}_0)$. With straightforward calculations and using eq. (A.1) - (A.3), we can show that when the labels are MCAR it holds $I_2(X;\widetilde{Y}_0) = \frac{1-p(y=1)}{1-p(y=1)p(s=1)}p(s=1)I_2(X;Y)$. So the non-centrality parameter of the surrogate approach can be written as:

$$\begin{aligned}\lambda_{X^2(X;\widetilde{Y}_0)} &= \frac{1-p(y=1)}{1-p(y=1)p(s=1)}p(s=1)2NI_2(X;Y) \Leftrightarrow \\ \lambda_{X^2(X;\widetilde{Y}_0)} &= \frac{1-p(y=1)}{1-p(y=1)p(s=1)}p(s=1)\lambda_{X^2(X;Y)}.\end{aligned}$$

By using the result that the non-centrality parameters for the X^2 and the G-test converge to a common value as $N \to \infty$, we can re-write the above relationship using the non-centrality parameter of the G-test:

$$\lambda_{G(X;\widetilde{Y}_0)} = \frac{1-p(y=1)}{1-p(y=1)p(s=1)}p(s=1)\lambda_{G(X;Y)}$$

Thus, the correction factor is $\kappa = \frac{1-p(y=1)}{1-p(y=1)p(s=1)}p(s=1)$.

Surrogate 4 ($\widetilde{Y}_1$): We can prove this correction factor by following the same methodology as for $\widetilde{Y}_0$. This time it holds $I_2(X;\widetilde{Y}_1) = \frac{1-p(y=0)}{1-p(y=0)p(s=1)}p(s=1)I_2(X;Y)$, and as a result the correction factor is $\kappa = \frac{1-p(y=0)}{1-p(y=0)p(s=1)}p(s=1)$. □

Theorem 4.6

When the labels are MAR-F, the following equation hold, which are useful for our proofs.

$$p(y|x, s=1) = p(y|x), \tag{A.4}$$

Surrogate 1 ($\mathcal{D}_L$): To prove $X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp Y|s = 1$, we need to prove:

$$p(x, y|s = 1) = p(x|s = 1)p(y|s = 1) \Leftrightarrow p(x, y) = p(x)p(y) \; \forall \; x \; \in \; \mathcal{X}, \; y \; \in \; \mathcal{Y}.$$

The proof is straightforward by using equation (A.4) and Dawid's (1979) definition of independence, eq. (IIb) in Dawid (1979). A similar proof is given for the conditional independence by Didelez et al. (2010, Theorem 6). □

Theorem 4.7

When the labels are MAR-C, the following equations hold, which are useful for our proofs:

$$p(x|y, s = 1) = p(x|y), \tag{A.5}$$

$$p(x|\widetilde{y}_0 = 1) = p(x|y = 1), \tag{A.6}$$

$$p(x|\widetilde{y}_1 = 0) = p(x|y = 0). \tag{A.7}$$

Surrogate 1 ($\mathcal{D}_L$): To prove $X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp Y|s = 1$, we need to prove:

$$p(x, y|s = 1) = p(x|s = 1)p(y|s = 1) \Leftrightarrow p(x, y) = p(x)p(y) \; \forall \; x \; \in \; \mathcal{X}, \; y \; \in \; \mathcal{Y}.$$

The proof is straightforward by using equation (A.5) and Dawid's (1979) definition of independence, eq. (IIb) in Dawid (1979). A similar proof is given for the conditional independence by Didelez et al. (2010, Theorem 6).

Surrogate 2 ($\widetilde{Y}_m$), **Surrogate 3** ($\widetilde{Y}_0$) and **Surrogate 4** ($\widetilde{Y}_1$): We can prove that these surrogates are valid by following the same methodology, or we can check that the independence relationships hold from the m-graph of MARC in Figure 3.1c. For completeness we will give the analytical proof for one scenario, i.e. Surrogate 3. To prove $X \perp\!\!\!\perp Y \Leftrightarrow X \perp\!\!\!\perp \widetilde{Y}_0$, we need to prove that:

$$p(x, \widetilde{y}_0) = p(x)p(\widetilde{y}_0) \Leftrightarrow p(x, y) = p(x)p(y) \; \forall \; x \; \in \; \mathcal{X}, \; y \; \in \; \mathcal{Y} \; and \; \widetilde{y}_0 \; \in \; \widetilde{\mathcal{Y}}_0.$$

Since the random variable Y is binary it is sufficient to prove this for the two

classes. So for the first class we have:

$$\begin{aligned} p(x,\widetilde{y}_0=1)=p(x)p(\widetilde{y}_0=1) &\Leftrightarrow p(x|\widetilde{y}_0=1)=p(x) \overset{A.6}{\Leftrightarrow} \\ p(x|y=1)=p(x) &\Leftrightarrow p(x,y=1)=p(x)p(y=1). \end{aligned}$$

Using the above result for the first class, we will prove it also for the second class:

$$\begin{aligned} &p(x,\widetilde{y}_0=0)=p(x)p(\widetilde{y}_0=0) \Leftrightarrow p(x)-p(x,\widetilde{y}_0=1)=p(x)(1-p(\widetilde{y}_0=1)) \Leftrightarrow \\ &p(x,\widetilde{y}_0=1)=p(x)p(\widetilde{y}_0=1) \Leftrightarrow p(x,y=1)=p(x)p(y=1) \Leftrightarrow \\ &p(x)-p(x,y=0)=p(x)(1-p(y=0)) \Leftrightarrow p(x,y=0)=p(x)p(y=0). \end{aligned}$$

□

Theorem 4.8

Surrogate 3 ($\widetilde{Y}_0$): In order to prove that relationship, we will use again the result that we obtain when we assume local alternatives that the X^2 and the G-test have the same asymptotic power. So instead of exploring the relationship of the non-centrality parameters for the G-tests between $X, \widetilde{Y}_0$ and X, Y, we can explore the relationship between the non-centrality parameters of the X^2-tests between $X, \widetilde{Y}_0$ and X, Y. The non-centrality parameter of this X^2 surrogate test is equal to $\lambda_{X^2(X;\widetilde{Y}_0)} = 2NI_2(X;\widetilde{Y}_0)$. With straightforward calculations, and using eq. (A.6), we can show that when the labels are MAR-C it holds $I_2(X;\widetilde{Y}_0) = \frac{1-p(y=1)}{p(y=1)}\frac{p(\widetilde{y}_0=1)}{1-p(\widetilde{y}_0=1)}I_2(X;Y)$. So the non-centrality parameter of the surrogate approach can be written as:

$$\begin{aligned} \lambda_{X^2(X;\widetilde{Y}_0)} &= \frac{1-p(y=1)}{p(y=1)}\frac{p(\widetilde{y}_0=1)}{1-p(\widetilde{y}_0=1)}2NI_2(X;Y) \Leftrightarrow \\ \lambda_{X^2(X;\widetilde{Y}_0)} &= \frac{1-p(y=1)}{p(y=1)}\frac{p(\widetilde{y}_0=1)}{1-p(\widetilde{y}_0=1)}\lambda_{X^2(X;Y)}. \end{aligned}$$

By using the result that the non-centrality parameters for the X^2 and G-test converge to a common value, we can re-write the above relationship using the non-centrality parameter of the G-test:

$$\lambda_{G(X;\widetilde{Y}_0)} = \frac{1-p(y=1)}{p(y=1)}\frac{p(\widetilde{y}_0=1)}{1-p(\widetilde{y}_0=1)}p(s=1)\lambda_{G(X;Y)}$$

Thus the correction factor is $\kappa_{\widetilde{Y}_0} = \frac{1-p(y=1)}{p(y=1)} \frac{p(\widetilde{y}_0=1)}{1-p(\widetilde{y}_0=1)}$.

Surrogate 4 ($\widetilde{Y}_1$): We can prove this correction factor by following the same methodology as for $\widetilde{Y}_0$. This time, by eq. (A.7), we can prove that $I_2(X;\widetilde{Y}_1) = \frac{1-p(y=0)}{p(y=0)} \frac{p(\widetilde{y}_1=0)}{1-p(\widetilde{y}_1=0)} I_2(X;Y)$, and as a result the correction factor is $\kappa_{\widetilde{Y}_1} = \frac{1-p(y=0)}{p(y=0)} \frac{p(\widetilde{y}_1=0)}{1-p(\widetilde{y}_1=0)}$. □

Theorem 5.4

We can use the surrogate variable $\widetilde{Y}_0$ to represent the event of having a positive and labelled example $(y=1, s=1)$, as $(\widetilde{y}_0 = 1)$ and rewrite the estimator:

$$\hat{I}^{Pos}_{MAR-C}(X;Y) = \sum_{x\in\mathcal{X}} p(y=1)\hat{p}(x|\widetilde{y}_0=1) \ln \frac{\hat{p}(x|\widetilde{y}_0=1)}{\hat{p}(x)} + \sum_{x\in\mathcal{X}} (\hat{p}(x) - p(y=1)\hat{p}(x|\widetilde{y}_0=1)) \ln \frac{\hat{p}(x) - p(y=1)\hat{p}(x|\widetilde{y}_0=1)}{\hat{p}(x)\,(1-p(y=1))}.$$

To derive the asymptotic distribution of $\hat{I}^{Pos}_{MAR-C}(X;Y)$ we will use the delta method; the notation that we follow can be found in (Agresti, 2013, Section 16.1.4). Since in the expression of $\hat{I}^{Pos}_{MAR-C}(X;Y)$ we have the maximum likelihood estimates for the probabilities $p(x), p(x|\widetilde{y}_0=1)$ the first step is to calculate the partial derivatives of these quantities with respect to the parameters of this model $p(x,\widetilde{y}_0=1)$ and $p(x,\widetilde{y}_0=0)$:

$$\frac{\partial p(x')}{\partial p(x,\widetilde{y}_0=1)} = \delta_{xx'}, \qquad \frac{\partial p(x')}{\partial p(x,\widetilde{y}_0=0)} = \delta_{xx'},$$
$$\frac{\partial p(x'|\widetilde{y}_0=1)}{\partial p(x,\widetilde{y}_0=1)} = \frac{\delta xx' - p(x'|\widetilde{y}_0=1)}{p(\widetilde{y}_0=1)}, \qquad \frac{\partial p(x'|\widetilde{y}_0=1)}{\partial p(x,\widetilde{y}_0=0)} = 0,$$

where $\delta_{xx'}$ is the Kronecker delta, which takes the value of 1 if $x = x'$ and 0 otherwise. By using the above partial derivatives we have the following results:

$$\phi_{x,\widetilde{y}_0=1} = \frac{\partial \hat{I}^{Pos}_{MAR-C}(X;Y)}{\partial p(x,\widetilde{y}_0=1)} = \ln \frac{p(x) - p(x|\widetilde{y}_0=1)p(y=1)}{p(x)} + \frac{p(y=1)}{p(\widetilde{y}_0=1)} \sum_{x'\in\mathcal{X}} (p(x'|\widetilde{y}_0=1) - \delta_{xx'}) \ln \frac{p(x') - p(x'|\widetilde{y}_0=1)p(y=1)}{p(x'|\widetilde{y}_0=1)p(y=1)}$$

$$\phi_{x,\widetilde{y}_0=0} = \frac{\partial \hat{I}^{Pos}_{MAR-C}(X;Y)}{\partial p(x,\widetilde{y}_0=0)} = \ln \frac{p(x) - p(x|\widetilde{y}_0=1)p(y=1)}{p(x)}.$$

So by using delta method the asymptotic variance of the estimator equals

$$\begin{aligned}\sigma^2_{Pos} = &\sum_{x\in\mathcal{X}} \left(p(x,\widetilde{y}_0=1)\phi^2_{x,\widetilde{y}_0=1} + p(x,\widetilde{y}_0=0)\phi^2_{x,\widetilde{y}_0=0}\right) \\ &- \left(\sum_{x\in\mathcal{X}} \left(p(x,\widetilde{y}_0=1)\phi_{x,\widetilde{y}_0=1} + p(x,\widetilde{y}_0=0)\phi_{x,\widetilde{y}_0=0}\right)\right)^2,\end{aligned}$$

where $\phi_{x,\widetilde{y}_0=1}$ and $\phi_{x,\widetilde{y}_0=0}$ are calculated earlier and are functions of $p(y=1)$. Furthermore, the estimator $\hat{I}^{Pos}_{MAR-C}(X;Y)$ is asymptotically normally distributed around $I(X;Y)$, since $I^{Pos}_{MAR-C}(X;Y) = I(X;Y)$ under the MAR-C assumption and because of equation (A.6). So the $\hat{I}^{Pos}_{MAR-C}(X;Y)$ estimator follows the ditribution

$$\hat{I}^{Pos}_{MAR-C}(X;Y) \sim \mathcal{N}\left(I(X;Y), \frac{\sigma^2_{Pos}}{N}\right).$$

□

Corollary 5.5

We can use the surrogate variable $\widetilde{Y}_1$ to represent the event of having a negative and labelled example $(y=0, s=1)$ as $(\widetilde{y}_1=0)$ and rewrite the estimator:

$$\begin{aligned}\hat{I}^{Neg}_{MAR-C}(X;Y) = &\sum_{x\in\mathcal{X}} p(y=0)\hat{p}(x|\widetilde{y}_1=0)\ln\frac{\hat{p}(x|\widetilde{y}_1=0)}{\hat{p}(x)} \\ &+ \sum_{x\in\mathcal{X}} \left(\hat{p}(x) - p(y=0)\hat{p}(x|\widetilde{y}_1=0)\right)\ln\frac{\hat{p}(x) - p(y=0)\hat{p}(x|\widetilde{y}_1=0)}{\hat{p}(x)\left(1-p(y=0)\right)}.\end{aligned}$$

Following the same methodology as in the previous Theorem and using the delta method, we can prove the following sampling distribution:

$$\hat{I}^{Neg}_{MAR-C}(X;Y) \sim \mathcal{N}\left(I(X;Y), \frac{\sigma^2_{Neg}}{N}\right),$$

$$\text{with} \quad \sigma^2_{Neg} = \sum_{x\in\mathcal{X}} \left(p(x,\widetilde{y}_1=0)\phi^2_{x,\widetilde{y}_1=0} + p(x,\widetilde{y}_1=1)\phi^2_{x,\widetilde{y}_1=1}\right)$$

$$- \left(\sum_{x \in \mathcal{X}} (p(x, \widetilde{y}_1 = 0)\phi_{x,\widetilde{y}_1=0} + p(x, \widetilde{y}_1 = 1)\phi_{x,\widetilde{y}_1=1}) \right)^2,$$

$$\phi_{x,\widetilde{y}_1=0} = \frac{\partial \hat{I}^{Neg}_{MAR-C}(X;Y)}{\partial p(x, \widetilde{y}_1 = 0)} = \ln \frac{p(x) - p(x|\widetilde{y}_1 = 0)p(y = 0)}{p(x)},$$

$$+ \frac{p(y = 0)}{p(\widetilde{y}_1 = 0)} \sum_{x' \in \mathcal{X}} (p(x'|\widetilde{y}_1 = 0) - \delta_{xx'}) \ln \frac{p(x') - p(x'|\widetilde{y}_1 = 0)p(y = 0)}{p(x'|\widetilde{y}_1 = 0)p(y = 0)}$$

$$\phi_{x,\widetilde{y}_1=1} = \frac{\partial \hat{I}^{Neg}_{MAR-C}(X;Y)}{\partial p(x, \widetilde{y}_1 = 1)} = \ln \frac{p(x) - p(x|\widetilde{y}_1 = 0)p(y = 0)}{p(x)}.$$

□

Theorem 4.10

Surrogate 3 ($\widetilde{Y}_0$): In order to prove the theorem we will use the following useful lemma:

Lemma A.2.
When the labels are MAR-C, the following equations hold, for any subset of features $\mathbf{z} \in \boldsymbol{\mathcal{Z}}$*:*

$$p(x|y = 1, \mathbf{z}) = p(x|\widetilde{y}_0 = 1, \mathbf{z}) \quad \forall \ \mathbf{z} \in \boldsymbol{\mathcal{Z}},$$

Proof. To prove this Lemma we will start from the *rhs* of the desired equation:

$$p(x|\widetilde{y}_0 = 1, \mathbf{z}) \overset{\text{when } \widetilde{y}_0=1 \text{ then } y=1}{=\!=\!=\!=\!=\!=\!=\!=\!=\!=\!=\!=\!=} p(x|\widetilde{y}_0 = 1, y = 1, \mathbf{z}).$$

Then by using the Bayes theorem and the chain rule we get:

$$p(x|\widetilde{y}_0 = 1, y = 1, \mathbf{z}) \overset{\text{Bayes theorem}}{=\!=\!=\!=\!=\!=\!=\!=\!=\!=} \frac{p(x, \widetilde{y}_0 = 1|y = 1, \mathbf{z})}{p(\widetilde{y}_0 = 1|y = 1, \mathbf{z})} \overset{\text{Chain rule}}{=\!=\!=\!=\!=\!=\!=\!=\!=}$$
$$= \frac{p(\widetilde{y}_0 = 1|x, y = 1, \mathbf{z})p(x|y = 1, \mathbf{z})}{p(\widetilde{y}_0 = 1|y = 1, \mathbf{z})}.$$

Because of the MAR-C assumption:

$$p(\widetilde{y}_0 = 1|y = 1, x, \mathbf{z}) = p(\widetilde{y}_0 = 1|y = 1, \mathbf{z}). \tag{A.8}$$

As a result the last expression becomes:

$$\frac{p(\widetilde{y}_0 = 1|x, y = 1, \mathbf{z})p(x|y = 1, \mathbf{z})}{p(\widetilde{y}_0 = 1|y = 1, \mathbf{z})} \overset{\text{eq. (A.8)}}{=\!=\!=} p(x|y = 1, \mathbf{z})$$

This finishes the proof of this lemma, since we derived the *lhs* of the desired equation. An interesting point to clarify is that equation (A.8) holds for any subset of features. To show that, without loss of generality, let us assume that the entire set of features $\mathbf{x}$ consists of the variables $x, \mathbf{z}$ and $\mathbf{w}$, where x is a single variable and $\mathbf{z}$, $\mathbf{w}$ sets of variables. The $x, \mathbf{z}$ and $\mathbf{w}$ can be created by any feature combination as long as their intersection is the empty set and their union is the entire feature space. Now we can re-write the MAR-C assumption as:

$$\begin{aligned}
p(\widetilde{y}_0 = 1|y = 1, \mathbf{x}) &= p(\widetilde{y}_0 = 1|y = 1) \Leftrightarrow \\
p(\widetilde{y}_0 = 1|y = 1, x, \mathbf{z}, \mathbf{w}) &= p(\widetilde{y}_0 = 1|y = 1) \Leftrightarrow \\
p(\widetilde{y}_0 = 1, x, \mathbf{z}, \mathbf{w}|y = 1) &= p(\widetilde{y}_0 = 1|y = 1)p(x, \mathbf{z}, \mathbf{w}|y = 1).
\end{aligned}$$

Now marginalising out the variable $\mathbf{w}$ we get:

$$\begin{aligned}
\sum_{\mathbf{w}\in\mathcal{W}} p(\widetilde{y}_0 = 1, x, \mathbf{z}, \mathbf{w}|y = 1) = p(\widetilde{y}_0 = 1|y = 1) \sum_{\mathbf{w}\in\mathcal{W}} p(x, \mathbf{z}, \mathbf{w}|y = 1) \Leftrightarrow \\
p(\widetilde{y}_0 = 1, x, \mathbf{z}|y = 1) = p(\widetilde{y}_0 = 1|y = 1)p(x, \mathbf{z}|y = 1) \Leftrightarrow \quad \text{(A.9)} \\
p(\widetilde{y}_0 = 1|y = 1, x, \mathbf{z}) = p(\widetilde{y}_0 = 1|y = 1). \quad \text{(A.10)}
\end{aligned}$$

Furthermore, in equation (A.9) by marginalising out the variable x we get:

$$\begin{aligned}
\sum_{x\in\mathcal{X}} p(\widetilde{y}_0 = 1, x, \mathbf{z}|y = 1) = p(\widetilde{y}_0 = 1|y = 1) \sum_{x\in\mathcal{X}} p(x, \mathbf{z}|y = 1) \Leftrightarrow \\
p(\widetilde{y}_0 = 1, \mathbf{z}|y = 1) = p(\widetilde{y}_0 = 1|y = 1)p(\mathbf{z}|y = 1) \Leftrightarrow \\
p(\widetilde{y}_0 = 1|y = 1, \mathbf{z}) = p(\widetilde{y}_0 = 1|y = 1). \quad \text{(A.11)}
\end{aligned}$$

Thus, from equations (A.10) and (A.11) we can derive equation (A.8). □

To prove $X \perp\!\!\!\perp Y|\mathbf{Z} \Leftrightarrow X \perp\!\!\!\perp \widetilde{Y}_0|\mathbf{Z}$, we need to prove that

$$p(x, \widetilde{y}_0|\mathbf{z}) = p(x|\mathbf{z})p(\widetilde{y}_0|\mathbf{z}) \Leftrightarrow p(x, y|\mathbf{z}) = p(x|\mathbf{z})p(y|\mathbf{z}) \forall x \in \mathcal{X}, y \in \mathcal{Y}, \widetilde{y}_0 \in \widetilde{\mathcal{Y}}_0, \mathbf{z} \in \boldsymbol{\mathcal{Z}}.$$

Since the random variables $\widetilde{Y}_0$ and Y are binary it is sufficient to prove this for

the two classes. For the first class we have:

$$p(x, \widetilde{y}_0 = 1|\mathbf{z}) = p(x|\mathbf{z})p(\widetilde{y}_0 = 1|\mathbf{z}) \Leftrightarrow p(x|\widetilde{y}_0 = 1, \mathbf{z}) = p(x|\mathbf{z}) \overset{\text{Lemma } A.2}{\Leftrightarrow}$$
$$p(x|\widetilde{y}_0 = 1, \mathbf{z}) = p(x|\mathbf{z}) \Leftrightarrow p(x, y = 1|\mathbf{z}) = p(x|\mathbf{z})p(y = 1|\mathbf{z}).$$

Using the above result for the first class, we will prove it for the second:

$$\begin{aligned}
&p(x, \widetilde{y}_0 = 0|\mathbf{z}) = p(x|\mathbf{z})p(\widetilde{y}_0 = 0|\mathbf{z}) \Leftrightarrow \\
&p(x|\mathbf{z}) - p(x, \widetilde{y}_0 = 1|\mathbf{z}) = p(x|\mathbf{z})(1 - p(\widetilde{y}_0 = 1|\mathbf{z})) \Leftrightarrow \\
&p(x, \widetilde{y}_0 = 1|\mathbf{z}) = p(x|\mathbf{z})p(\widetilde{y}_0 = 1|\mathbf{z}) \Leftrightarrow \\
&p(x, y = 1|\mathbf{z}) = p(x|\mathbf{z})p(y = 1|\mathbf{z}) \Leftrightarrow \\
&p(x|\mathbf{z}) - p(x, y = 0|\mathbf{z}) = p(x|\mathbf{z})(1 - p(y = 0|\mathbf{z})) \Leftrightarrow \\
&p(x, y = 0|\mathbf{z}) = p(x|\mathbf{z})p(y = 0|\mathbf{z}).
\end{aligned}$$

Surrogate 4 ($\widetilde{Y}_1$): The proof can be derived by following the same methodology as we did for the case of Surrogate 3. □

Theorem 4.11

Surrogate 3 ($\widetilde{Y}_0$): By using the chain rule of the mutual information (Cover and Thomas, 2006) the non-centrality parameter can be written as:

$$\lambda_{G(X;\widetilde{Y}_0|\mathbf{Z})} = 2NI(X; \widetilde{Y}_0|\mathbf{Z}) = 2NI(X\mathbf{Z}; \widetilde{Y}_0) - 2NI(\mathbf{Z}; \widetilde{Y}_0) = \lambda_{G(X\mathbf{Z};\widetilde{Y}_0)} - \lambda_{G(\mathbf{Z};\widetilde{Y}_0)}.$$

Using Theorem 4.8, we can associate the non-centrality parameters of the G-tests $X, \widetilde{Y}_0$ and X, Y, so we have:

$$\begin{aligned}
\lambda_{G(X;\widetilde{Y}_0|\mathbf{Z})} &= \kappa_{\widetilde{Y}_0}\lambda_{G(X\mathbf{Z};Y)} - \kappa_{\widetilde{Y}_0}\lambda_{G(\mathbf{Z};Y)} = \\
&= \kappa_{\widetilde{Y}_0}2NI(X\mathbf{Z}; Y) - \kappa_{\widetilde{Y}_0}2NI(\mathbf{Z}; Y) = \kappa_{\widetilde{Y}_0}2N(I(X\mathbf{Z}; Y) - I(\mathbf{Z}; Y)).
\end{aligned}$$

And, by using again the chain rule, the last expression can be written as:

$$\lambda_{G(X;\widetilde{Y}_0|\mathbf{Z})} = \kappa_{\widetilde{Y}_0}2NI(X; Y|\mathbf{Z}) = \kappa_{\widetilde{Y}_0}\lambda_{G(X;Y|\mathbf{Z})}.$$

Surrogate 4 ($\widetilde{Y}_1$): The proof can be derived by following the same methodology as we did for the case of Surrogate 3. □

Bibliography

Alan Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 3rd edition, 2013.

Constantin F. Aliferis, Alexander Statnikov, Ioannis Tsamardinos, Subramani Mani, and Xenofon D Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *The Journal of Machine Learning Research (JMLR)*, 11:171–234, 2010.

Paul D. Allison. *Missing Data*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136, 2001.

Davide Bacciu, Terence A Etchells, Paulo JG Lisboa, and Joe Whittaker. Efficient identification of independence networks using mutual information. *Computational Statistics*, 28(2):621–646, 2013.

Fazia Bellal, Haytham Elghazel, and Alex Aussem. A semi-supervised feature ranking method with ensemble learning. *Pattern Recognition Letters*, 33(10): 1426–1433, 2012.

James O. Berger. Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, 18(1):1–32, 2003.

G. Blanchard, G. Lee, and C. Scott. Semi-Supervised Novelty Detection. *The Journal of Machine Learning Research (JMLR)*, 11, March 2010.

Veronica Bolón-Canedo, Konstantinos Sechidis, Noelia Sánchez-Maroño, Amparo Alonso-Betanzos, and Gavin Brown. Some guidelines for distributed feature ranking. *Under review*, 2015.

David R. Brillinger. Some data analyses using mutual information. *Brazilian Journal of Probability and Statistics*, 18(6):163–183, 2004.

Gavin Brown, Adam Pocock, Ming-Jie. Zhao, and Mikel Lujan. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research (JMLR)*, 13:27–66, 2012.

Ruichu Cai, Zhenjie Zhang, and Zhifeng Hao. BASSUM: A Bayesian semi-supervised method for classification feature selection. *Pattern Recognition*, 44 (4):811–820, 2011.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-supervised learning*, volume 2. MIT press Cambridge, 2006.

Nitesh V. Chawla and Grigoris I. Karakoulas. Learning from labeled and unlabeled data: An empirical study across techniques and domains. *Journal of Artificial Intelligence Research (JAIR)*, 23:331–366, 2005.

So Yeon Chun and Alexander Shapiro. Normal versus noncentral chi-square asymptotics of misspecified models. *Multivariate Behavioral Research*, 44(6): 803–827, 2009.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Routledge Academic, 1988.

David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine Learning*, 15(2):201–221, 1994.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006.

H. Cramér. *Mathematical Methods of Statistics*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1999.

Noel Cressie and Timothy R.C. Read. Pearson's X^2 and the loglikelihood ratio statistic G^2: a comparative review. *International Statistical Review/Revue Internationale de Statistique*, pages 19–43, 1989.

Philip A. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–31, 1979.

Vanessa Didelez, Svend Kreiner, and Niels Keiding. Graphical models for inference under outcome-dependent sampling. *Statistical Science*, 25(3):368–387, 2010.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data mining (KDD)*, pages 213–220, 2008.

Paul D. Ellis. *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press, 2010.

Craig K. Enders. *Applied Missing Data Analysis*. Methodology in the social sciences. Guilford Press, 2010.

Wei Fan and Ian Davidson. On sample selection bias and its efficient correction via model averaging and unlabeled examples. In *SIAM International Conference on Data Mining (SDM)*, 2007.

Robert M. Fano. *Transmission of Information: A Statistical Theory of Communications*. MIT Press Classics. Massachusetts Institute of Technology Press, 1961.

François Fleuret. Fast binary feature selection with conditional mutual information. *The Journal of Machine Learning Research (JMLR)*, 5:1531–1555, 2004.

Patrick Fox-Roberts and Edward Rosten. Unbiased generative semi-supervised learning. *Journal of Machine Learning Research (JMLR)*, 15:367–443, 2014.

Bernhard Goebel, Zaher Dawy, Joachim Hagenauer, and Jakob C Mueller. An approximation to the distribution of finite sample size mutual information estimates. In *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, volume 2, pages 1102–1106. IEEE, 2005.

Arthur Gretton and László Györfi. Consistent nonparametric tests of independence. *The Journal of Machine Learning Research (JMLR)*, 99:1391–1423, 2010.

Dato N. M. de Gruijter and Leo J. T. van der Kamp. *Statistical Test Theory for the Behavioral Sciences*. Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences. CRC Press, 2007.

Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A. Zadeh. *Feature Extraction: Foundations and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

Shelby J. Haberman. *The Analysis of Frequency Data*. Midway reprints. University of Chicago Press, 1974.

Gerald J. Hahn and Samuel S. Shapiro. *Statistical Models in Engineering*. Wiley Series on Systems Engineering and Analysis Series. John Wiley & Sons, 1967.

Dan He, Irina Rish, David Haws, and Laxmi Parida. MINT: Mutual information based transductive feature selection for genetic trait prediction. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 2015.

James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

Matthias Hein. Binary classification under sample selection bias. In Joaquin Quiñonero Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors, *Dataset Shift in Machine Learning*, chapter 3, pages 41–64. The MIT Press Cambridge, 2009.

Thibault Helleputte and Pierre Dupont. Partially supervised feature selection with regularized linear models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 409–416. ACM, 2009.

Martin E. Hellman and Josef Raviv. Probability of error, equivocation, and the Chernoff bound. *Information Theory, IEEE Transactions on*, 16(4):368–372, 1970.

Valen E. Johnson. Bayes factors based on test statistics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):689–701, 2005.

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1):95–116, 2007.

Ivan Kojadinovic. On the use of mutual information in data analysis: an overview. In *Proceedings of 11th International Symposium on Applied Stochastic Models and Data Analysis (ASMDA'05)*, pages 738–747, 2005.

Daphne Koller and Mehran Sahami. Toward optimal feature selection. In *International Conference of Machine Learning (ICML)*, pages 284–292, 1996.

Diana Kornbrot. *Point Biserial Correlation*, volume 3, page 15521553. John Wiley & Sons, Ltd, 2005.

Ludmila I. Kuncheva. A stability index for feature selection. In *Artificial intelligence and applications*, pages 421–427, 2007.

John Lafferty and Larry Wasserman. Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems (NIPS) 21*, 2007.

Neil D. Lawrence and Michael I. Jordan. Gaussian processes and the null-category noise model. In Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors, *Semi-Supervised Learning*, chapter 8, pages 137–150. The MIT Press Cambridge, 2006.

Erich Leo Lehmann. Some concepts of dependence. *The Annals of Mathematical Statistics*, pages 1137–1153, 1966.

David D. Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of the workshop on Speech and Natural Language*, pages 212–217. Association for Computational Linguistics, 1992.

Wenkai Li, Qinghua Guo, and Charles Elkan. Can we model the probability of presence of species without absence data? *Ecography*, 34(6):1096–1105, 2011.

Roderick J. A. Little and Donald B. Rubin. *Statistical analysis with missing data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, 2 edition, 2002.

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. In *Advances in Neural Information Processing Systems (NIPS) 12*, pages 505–511. MIT Press, 1999.

Douglas A. McManus. Who invented local power analysis? *Econometric Theory*, 7(2):pp. 265–268, 1991.

George A. Miller. Note on the bias of information estimates. In Henry Quastler, editor, *Information Theory in Psychology: Problems and Methods II-B*, pages 95–100. Glencoe, IL: Free Press, 1955.

Karthika Mohan and Judea Pearl. Graphical models for recovering probabilistic and causal queries from missing data. In *Advances in Neural Information Processing Systems (NIPS) 27*, pages 1520–1528, 2014.

Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems (NIPS) 26*, pages 1277–1285, 2013.

Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521 – 530, 2012.

Jerzy Neyman, Herman Chernoff, and D. G. Chapman. Discussion of hoeffding's paper. *The Annals of Mathematical Statistics*, 36(2):pp. 401–408, 1965.

Regina Nuzzo. Scientific method: Statistical errors. *Nature*, 506(7487):150–152, February 2014.

Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, June 2003.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

Karl Pearson. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.

Jean-Philippe Pellet and André Elisseeff. Using markov blankets for causal structure learning. *The Journal of Machine Learning Research (JMLR)*, 9:1295–1342, 2008.

Hanchuan Peng, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

Edwin J. G. Pitman. *Some basic theory for statistical inference*. Monographs on applied probability and statistics. Chapman and Hall, 1979.

Marthinus Christoffel du Plessis and Masashi Sugiyama. Semi-supervised learning of class balance under class-prior change by distribution matching. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

Adam Pocock, Mikel Luján, and Gavin Brown. Informative priors for markov blanket discovery. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

Richard F. Potthoff, Gail E. Tudor, Karen S. Pieper, and Vic Hasselblad. Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research*, 15:213–34, 2006.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

Matthew Reimherr and Dan L. Nicolae. On quantifying dependence: A framework for developing interpretable measures. *Statistical Science*, 28(1):116–130, 2013.

Jiangtao Ren, Zhengyuan Qiu, Wei Fan, Hong Cheng, and S Yu Philip. Forward semi-supervised feature selection. In *Advances in Knowledge Discovery and Data Mining*, pages 970–976. Springer, 2008.

Robert Rosenthal. Parametric measures of effect size. In H.M. Cooper and L.V. Hedges, editors, *The Handbook of Research Synthesis*, chapter 16, pages 231–244. Russell Sage Foundation, 1994.

Robert Rosenthal, Ralph L. Rosnow, and Donald B. Rubin. *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach*. McGraw-Hill series in psychology. Cambridge University Press, 2000.

Saharon Rosset, Ji Zhu, Hui Zou, and Trevor J Hastie. A method for inferring label sampling mechanisms in semi-supervised learning. In *Advances in Neural Information Processing Systems (NIPS) 18*, 2004.

Myra L. Samuels, Jeffrey A. Witmer, and Andrew Schaffner. *Statistics for the Life Sciences*. Prentice Hall, 4 edition, 2012.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. Semi-supervised learning in causal and anticausal settings. In *Empirical Inference*, pages 129–141. Springer, 2013.

Konstantinos Sechidis and Gavin Brown. Markov blanket discovery in positive-unlabelled and semi-supervised data. In *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 351–366. Springer Berlin Heidelberg, 2015a.

Konstantinos Sechidis and Gavin Brown. Hypothesis testing and feature selection in semi-supervised data. *Under review*, 2015b.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 145–158. Springer Berlin Heidelberg, 2011.

Konstantinos Sechidis, Borja Calvo, and Gavin Brown. Statistical hypothesis testing in positive unlabelled data. In *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 66–81. Springer Berlin Heidelberg, 2014a.

Konstantinos Sechidis, Nikolaos Nikolaou, and Gavin Brown. Information theoretic feature selection in multi-label data through composite likelihood. In *Structural, Syntactic, and Statistical Pattern Recognition (SSPR)*, pages 143–152. Springer Berlin Heidelberg, 2014b.

Matthias Seeger. Learning with labeled and unlabeled data. Technical report, Technical report, University of Edinburgh, 2002.

Aarti Singh, Robert Nowak, and Xiaojin Zhu. Unlabeled data: Now it helps, now it doesn't. In *Advances in Neural Information Processing Systems (NIPS) 22*, pages 1513–1520, 2009.

Andrew Smith and Charles Elkan. A Bayesian network framework for reject inference. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery and Data mining (KDD)*, pages 286–295, 2004.

Andrew T. Smith and Charles Elkan. Making generative classifiers robust to selection bias. In *Proceedings of the 13th ACM SIGKDD International conference on Knowledge Discovery and Data mining (KDD)*, pages 657–666, 2007.

Robert R. Sokal and F. James Rohlf. *Biometry: The principles and practice of Statistics in Biological data*. W. H. Freeman & Co., third edition, 1995.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press Cambridge, second edition, 2001.

Eleftherios Spyromitros-Xioufis, Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. MLKD's participation at the CLEF 2011 photo annotation and concept-based retrieval tasks. In *ImageClef Lab of CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation*, 2011.

Amos Storkey. When training and test sets are different: Characterizing learning transfer. In Joaquin Quiñonero Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors, *Dataset Shift in Machine Learning*, chapter 1, pages 3–28. The MIT Press Cambridge, 2009.

Herbert A. Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66, 1926.

Masashi Sugiyama. Machine learning with squared-loss mutual information. *Entropy*, 15(1):80–112, 2012.

Felix Thoemmes and Karthika Mohan. Graphical representation of missing data problems. *Structural Equation Modeling: A Multidisciplinary Journal*, pages 1–13, 2015.

Jin Tian. Missing at random in graphical models. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

Ioannis Tsamardinos and Constantin F. Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In *Proceedings of the Ninth International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2003.

Ioannis Tsamardinos and Giorgos Borboudakis. Permutation testing improves Bayesian network learning. In *Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 322–337. Springer Berlin Heidelberg, 2010.

Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *ACM SIGKDD*, 2003.

Guy Van den Broeck, Karthika Mohan, Arthur Choi, and Judea Pearl. Efficient algorithms for Bayesian network parameter learning from incomplete data. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015.

Howard Wainer and Daniel H Robinson. Shaping up the practice of null hypothesis significance testing. *Educational Researcher*, 32(7):22–30, 2003.

Barnet Woolf. The log likelihood ratio test (the G-test). *Annals of Human Genetics*, 21(4):397–409, 1957.

Shaomin Wu and Peter A. Flach. Feature selection with labelled and unlabelled data. In M. Bohanec, B. Kasek, N. Lavrac, and D. Mladenic, editors, *ECML/P-KDD'02 workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning*, pages 156–167. University of Helsinki, August 2002.

Zenglin Xu, Irwin King, Michael Rung-Tsong Lyu, and Rong Jin. Discriminative semi-supervised feature selection via manifold regularization. *Neural Networks, IEEE Transactions on*, 21(7):1033–1047, 2010.

Howard Hua Yang and John Moody. Data visualization and feature selection: New algorithms for nongaussian data. In S.A. Solla, T.K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems (NIPS) 12*, pages 687–693. MIT Press, 1999.

Sha Yang, Yi Zhao, and Ravi Dhar. Modeling the underreporting bias in panel survey data. *Marketing Science*, 29(3):525–539, 2010.

Sandeep Yaramakala and Dimitris Margaritis. Speculative markov blanket discovery for optimal feature selection. In *5th ICDM*. IEEE, 2005.

Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, 2004.

Yiteng Zhai, Yew-Soon Ong, and Ivor W. Tsang. The emerging "Big Dimensionality". *Computational Intelligence Magazine, IEEE*, 9(3):14–26, Aug 2014.

Jidong Zhao, Ke Lu, and Xiaofei He. Locality sensitive semi-supervised feature selection. *Neurocomputing*, 71(10):1842–1849, 2008.

Ming-Jie Zhao, Narayanan Edakunni, Adam Pocock, and Gavin Brown. Beyond Fano's inequality: bounds on the optimal F-score, BER, and cost-sensitive risk and their implications. *Journal of Machine Learning Research (JMLR)*, 14(1): 1033–1090, 2013.

Hui Zou, Ji Zhu, and Trevor Hastie. Automatic Bayes carpentry using unlabeled data in semi-supervised classification. In *Advances in Neural Information Processing Systems (NIPS) 18*, 2004.