BAYESIAN NETWORKS: MODEL SELECTION AND APPLICATIONS

DMITRY RUSAKOV

BAYESIAN NETWORKS: MODEL SELECTION AND APPLICATIONS

Research Thesis

Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

DMITRY RUSAKOV

Submitted to the Senate of the Technion — Israel Institute of Technology

ADAR, 5764 HAIFA MARCH, 2004

The Research Thesis Was Done Under The Supervision of Assoc. Prof. Dan Geiger in the Department of Computer Science

ACKNOWLEDGMENT

THE GENEROUS FINANCIAL HELP OF TECHNION IS GRATEFULLY ACKNOWLEDGED

Contents

Abstract 1

List of Symbols 3

1 Introduction 5

2 Background 11

2.1 Bayesian Networks 11

2.2 Exponential Families of Distributions 13

2.2.1 An Example of Exponential Family - Normal Distribution 14

2.2.2 An Example of Exponential Family - Multinomial Distribution 15

2.2.3 Graphical Models as Exponential Families of Distributions 17

2.3 Bayesian Model Selection Procedure 18

2.4 Asymptotic Approximation of Integrals 20

2.4.1 A Surface of Stationary Points 24

2.4.2 A Self Crossing Stationary Curve in Two Dimensional Space 27

2.4.3 The General Approximation Method by Watanabe 33

3 Asymptotics for Naive Bayesian Networks 39

3.1 Introduction 39

3.2 Naive Bayesian Models 41

3.3 Main Claims 42

3.4 Proof Outline of the Main Result 48

3.4.1 Useful Transformations 48

3.4.2 Preliminary Lemmas . . . 51
3.4.3 Analysis of Type 2 Singularity. . . . 52
3.5 Full Proof of the Main Theorem . . . 56
3.5.1 Proof of Lemma 11 . . . 56
3.5.2 Regular Statistics Case . . . 59
3.5.3 Type 1 Singularity . . . 60
3.5.4 Type 2 Singularity . . . 64
3.5.5 Proof of Claims (d,e) of Theorem 8 (Case $n = 2$) . . . 68
3.5.6 Proof of Theorem 8f (Case $n = 1$) . . . 70
3.6 Proof of the Second Theorem . . . 71
3.7 Discussion . . . 73

4 Automated Evaluation of Marginal Likelihood 75
4.1 Introduction . . . 75
4.2 Technical Background . . . 77
4.3 Automatic Effective Dimensionality Computations . . . 80
4.3.1 Evaluation: Naive Bayesian Models . . . 81
4.3.2 Evaluation: Other Models . . . 83
4.4 Marginal Likelihood for Singular Statistics . . . 84
4.4.1 Evaluation . . . 85
4.5 Discussion . . . 86

5 Parameter Priors 89
5.1 Introduction . . . 89
5.2 Assessment of Parameter Priors for DAG Models . . . 91
5.3 Globally Independent Priors for Two Node Networks . . . 93
5.3.1 The Approach to the Proof of Theorem 15 . . . 94
5.4 Multiple Node Networks . . . 96
5.4.1 Binary-Valued Networks . . . 98
5.4.2 The General Network . . . 99
5.5 Dirichlet Priors: The Minimal Set of Assumptions . . . 100

5.6 Proofs and Derivations . . . 102
5.6.1 Lemmas . . . 102
5.6.2 The Proof of Theorem 15 . . . 105
5.6.3 The Proof of Lemma 16 . . . 111
5.6.4 The Proof of Lemma 17 . . . 112
5.6.5 The Proof of Theorem 18 . . . 113
5.6.6 The Proof of Theorem 21 for Two Node Networks . . . 116
5.7 Matlab Code . . . 118
5.8 Discussion . . . 120

6 Model Averaging for Complex Disease Analysis 121

6.1 Introduction . . . 121
6.2 Genetic Linkage Analysis . . . 123
6.3 Linkage Analysis for Complex Diseases . . . 125
6.4 Averaging of Penetrances . . . 126
6.5 Inheritance Models . . . 127
6.6 Experimental Results . . . 128
6.7 Discussion . . . 131

List of Figures

2.1 An example of Bayesian network. . . . 13

2.2 Representation of normal distribution as an exponential distribution. . . . 15

2.3 Representation of trinomial distribution as an exponential distribution. . . . 16

2.4 A linear exponential model represented by a simple Bayesian network. . . . 18

2.5 Laplace approximation of integrals. . . . 22

2.6 Two dimensional $\mathbf{M}$ transformation. . . . 25

2.7 Two dimensional E^+ and E^- boxes. . . . 26

2.8 Maclaurin Trisectrix as an example of self-crossing stationary curve. . . . 28

2.9 The form and isolines of the function $f(x, y) = (x^2 - y^2)^2$. . . . 29

2.10 Non-classical Laplace type integrals. . . . 36

3.1 A naive Bayesian model. Class variable C is latent. . . . 42

3.2 An example of incorrect Bayesian model selection by the standard BIC score. . . . 48

3.3 Critical surface for type 2 singularity. . . . 54

3.4 Critical surface of type 1 singularity. . . . 61

3.5 The integration domain U. . . . 68

4.1 Effective dimensionality computation algorithm. . . . 81

4.2 W-structure. . . . 83

4.3 Algorithm for asymptotic approximation of the marginal likelihood. . . . 85

5.1 Two node complete DAG model for random variables X and Y. . . . 94

5.2 An example of parameter independence correspondence. . . . 97

6.1 Penetrance tables for the two-locus inheritance models. 129
6.2 Power graph for the **R+R** model as a function of the number of nuclear families segregating the disease. 130
6.3 Power curves to achieve a given Z value. 134

List of Tables

2.1 Summary of asymptotic approximations of $I[N]$ (Eq. 2.12) under various conditions. 23

4.1 Degenerate naive Bayesian models found by the effective dimensionality algorithm. . 82

4.2 Asymptotic approximations to the marginal likelihood found by Algorithm 4.3. . . . 86

6.1 Average Z scores for different generating models, for one-marker analysis. 131

6.2 Power to achieve a given Z value, for one-marker analysis on nuclear families data. . 132

6.3 Average Z scores for multi-marker and large families analysis. 132

6.4 Power to achieve a given Z value, for five-marker analysis on nuclear families data. . 132

Abstract

A Bayesian network is a representation of joint probability distribution for a collection of random variables via a Directed Acyclic Graph (DAG) and a set of associated parameters. In particular, each node in the DAG corresponds to a random variable, and the lack of an edge between two nodes represents a conditional independence assumption. The Bayesian network formally encodes the joint probability distribution for its domain, yet includes a human-oriented qualitative structure that facilitates communication between a user and a system incorporating the probabilistic model.

This thesis focuses on learning a structure of a Bayesian network based on data. A critical step in learning a structure of a Bayesian network is model comparison. We are taking a Bayesian approach to model comparison and selection which requires the evaluation of the marginal likelihood of data given a network structure. Such evaluation requires specification of an appropriate conjugate prior distribution for network parameters or, alternatively, the development of asymptotic (large-sample) approximation formulas for the marginal likelihood integrals under study. This work addresses the problems of asymptotic approximation of marginal likelihood integrals, specification of appropriate conjugate prior distributions for Bayesian networks and application of Bayesian network model selection for genetic linkage analysis.

It was shown before that for Bayesian networks without hidden variables the marginal likelihood of data can be asymptotically approximated by the Bayesian Information Criterion (BIC). The standard BIC score is equal to the maximum likelihood of data in the model penalized by half the number of model parameters multiplied by the logarithm of number of training samples. The first part of this thesis investigates the applicability of the BIC score for approximating the marginal likelihood of data given a Bayesian network with hidden variables. We develop a closed form asymptotic formula for the marginal likelihood of data given a naive Bayesian model with two hidden states and binary features. This formula deviates from the standard BIC score. Therefore, this work provides a concrete example that the standard BIC score is generally incorrect for model selection among Bayesian networks with hidden variables.

The second part presents the implementation of an algorithmic approach used in the approximation of the marginal likelihood given a naive Bayesian model. The approach is generalized and the underlying algorithms are implemented in the Matlab and Maple computer systems.

The third part of the thesis investigates the minimal set of general conditions, known as *global* and *local* parameter independence, that are required in order to ensure the Dirichlet prior on network parameters and, consequently, to allow a closed form evaluation of the marginal likelihood. It is shown that the class of admissible distributions arising from the global independence assumption for parameters of discrete Bayesian network is strictly larger than the class of Dirichlet distributions. In addition, the minimal set of global and local parameter independence assumptions required in order to ensure Dirichlet prior is explicated.

The final part of this thesis presents an application of Bayesian model selection to the problem of finding the location of disease genes in the context of genetic linkage analysis. Two models are being compared in this application, one describing the dominant model of disease penetrance and another describing the recessive model of disease penetrance. Contrary to the problem of asymptotic model selection, to which the major part of this thesis is devoted, the Bayesian model selection in this particular case of genetic linkage analysis reduces to the problem of numerical evaluation of the marginal likelihood integral using a small number of evaluations of the likelihood function. The search for disease gene using the penetrance model selection has been incorporated into the SUPERLINK system for genetic linkage analysis and its success in detecting the disease gene locations was demonstrated by a series of experiments.

List of Symbols

Symbol	Definition
$X = \{X_1, \ldots, X_n\}$	A set of random variables.
$\mathbf{D} = \{D_1, \ldots, D_n\}$	Domain of random variables X.
M, $M(s, \mathcal{F}_s)$	Bayesian network model that consist of structure s and a set of local distributions $\mathcal{F}_s$.
$\theta \in \Theta$, $\omega \in \Omega$	Model parameters.
d	The dimensionality of model M. I.e., the number of independent parameters of M.
$\eta \in \Omega_\eta$	Canonical parameters of exponential family of distributions.
$x \in D$	A sampled data from a distribution on X.
N	Number of samples in D.
$y(x)$	A sufficient statistics of sample x.
$Y_D = \frac{1}{N} \sum_{x \in D} y(x)$	The averaged sufficient statistics of sample D.
$S(N, Y_D, j) = \ln P(D \vert M_j)$	BIC score of model M_j, i.e., the logarithm of the marginal likelihood of data D given model M_j. This quantity is used in Bayesian model selection procedures.
$\mathbb{I}[N, Y_D] = \int_\Omega e^{-Nf(\omega, Y_D)} \mu(\omega) d\omega$	The marginal likelihood integral of data D given model M.
$f(\omega, Y_D)$	The minus log-likelihood function of a single sample with statistics Y_D given model M.

Chapter 1

Introduction

A Bayesian network model is a representation of a family of joint probability distributions for a collection of random variables via a Directed Acyclic Graph (DAG). In particular, each node in the DAG corresponds to a random variable, and the lack of an edge between two nodes represents a conditional independence assumption. A specific joint probability distribution is represented by a given directed acyclic graph together with specific values for the set of associated parameters. Bayesian networks have been extensively studied in AI, Statistics, Machine learning, and in many application areas (Pearl, 1988; Lauritzen, 1996; Dawid & Lauritzen, 1993; Heckerman, Geiger & Chickering, 1995; Friedman, Geiger & Goldszmidt, 1997).

Bayesian networks encode a probability distribution with a manageable number of parameters, due to the factorization introduced by an underlying graph, therefore reducing the complexity of the representation and reducing the complexity of decision making based on this distribution. Bayesian networks are also useful when constructed directly from expert knowledge because they introduce cause-effect relationships that are intuitive to human experts. These features made Bayesian networks a premier tool for representing probabilistic knowledge and reasoning under uncertainty.

This thesis focuses on learning - the process of updating both the parameters and the structure of a Bayesian network based on data (Buntine, 1994; Heckerman, Geiger & Chickering, 1995). A critical step in learning a structure of a Bayesian network is model comparison and selection. We are taking a Bayesian approach to model comparison and selection which requires the evaluation of the marginal likelihood of data given a network structure. Such evaluation requires specification of an appropriate conjugate prior distribution for network parameters or, alternatively, the development of asymptotic (large-sample) approximation formulas for the marginal likelihood integrals under study.

In the Bayesian approach to model selection, a model M is chosen according to the maximum posteriori probability of M given the observed data D:

$$P(M|D) \propto P(M,D) = P(M)P(D|M) = P(M)\int_{\Omega} P(D|M,\omega)P(\omega|M)d\omega$$

where ω denotes the model parameters and Ω denotes the domain of the model parameters. In particular, we focus on model selection using large sample approximation for $P(M,D)$, called *BIC - Bayesian Information Criterion*.

The critical computational part in using this criterion is evaluating the marginal likelihood integral $P(D|M) = \int_{\Omega} P(D|M,\omega)P(\omega|M)d\omega$. To compute marginal likelihood of data given a network structure in a closed form, researchers have made a number of assumptions. In particular, for learning Bayesian networks for a set of discrete random variables, the assumptions of global and local parameter independence for all network structures, Dirichlet distribution on network parameters, and some other technical assumptions were made (Spiegelhalter & Lauritzen, 1990; Cooper & Herskovits, 1992; Dawid & Lauritzen, 1993). It was later shown that the assumption of global and local parameter independence for all nodes in every complete network structure dictates that the only possible prior parameter distribution for discrete DAG models is a Dirichlet prior (Heckerman, Geiger & Chickering, 1995; Geiger & Heckerman, 1997).

In the general settings, when one is not willing to commit himself to a particular prior, or when the statistical model is too complex, as in the case of Bayesian networks with hidden variables, the marginal likelihood can not be computed in closed form and an asymptotic approximation to $P(D|M)$ is seeked.

Given an exponential model M we write $P(D|M)$ as a function of the averaged sufficient statistics Y_D of the data D, and the number N of data points in D:

$$\mathbb{I}[N, Y_D, M] = \int_{\Omega} e^{loglikelihood(Y_D, N|\omega, M)} \mu(\omega|M)d\omega$$

where $\mu(\omega|M)$ is the prior parameter density for model M. Recall that the sufficient statistics for multinomial samples of n binary variables $(X_1, \ldots, X_n)$ is simply the counts $N \cdot Y_D$ for each of the possible 2^n joint states. Often the prior $P(M)$ is assumed to be equal for all models, in which case

Bayesian model selection is performed by maximizing $\mathbb{I}[N, Y_D, M]$. The quantity represented by $S(N, Y_D, M) \equiv \ln \mathbb{I}[N, Y_D, M]$ is called the *BIC score* of model M.

For many types of models the asymptotic evaluation of $\mathbb{I}[N, Y_D, M]$, as $N \to \infty$, uses a classical Laplace procedure. This evaluation was first performed for Linear Exponential (LE) models (Schwarz, 1978) and then for Curved Exponential (CE) models under some additional technical assumptions (Haughton, 1988). It was shown that

$$S(N, Y_D, M) = N \cdot \ln P(Y_D|\omega_{ML}) - \frac{d}{2} \ln N + R,$$

where $\ln P(Y_D|\omega_{ML})$ is the log-likelihood of Y_D given the maximum likelihood parameters of the model and d is the model dimension, i.e., the number of parameters. The error term $R(N, Y_D, M)$ was shown to be bounded for a fixed Y_D (Schwarz, 1978) and uniformly bounded for all $Y_D \to Y$ in CE models (Haughton, 1988) as $N \to \infty$. For convenience, the dependence on M is often suppressed from our notation.

The use of BIC score for Bayesian model selection for Graphical Models is valid for Undirected Graphical Models without hidden variables because these are LE models (Lauritzen, 1996). The justification of this score for Directed Graphical Models (called Bayesian Networks) is somewhat more complicated. On one hand discrete and Gaussian DAG models are CE models (Spirtes, Richardson & Meek, 1997; Geiger, Heckerman, King & Meek, 2001). On the other hand, the theoretical justification of the BIC score for CE models has been established under the assumption that the model contains the true distribution - the one that has generated the observed data. This assumption limits the applicability of the proof of BIC score's validity for Bayesian networks in practical setups.

Haughton (1988) proves that if at least one of several models contains the true distribution, then the BIC score is the correct approximation to $\mathbb{I}[N, Y_D]$ (shorthand notation for $\mathbb{I}[N, Y_D, M]$) and the correct model will be chosen by BIC score with probability 1 as $N \to \infty$. However, this claim does not guarantee correctness of the asymptotic expansion of $\mathbb{I}[N, Y_D]$ for models that do not contain the true distribution, nor does it guarantee correctness of model selection for finite N. The last problem is common to all asymptotic methods, but having a correct asymptotic approximation for $\mathbb{I}[N, Y_D]$ provides some confidence in this choice.

The evaluation of the marginal likelihood $\mathbb{I}[N, Y_D]$ for Bayesian networks with hidden variables is a wide open problem because the class of distributions represented by Bayesian networks with hidden variables is significantly richer than curved exponential models and it falls into the class of Stratified Exponential (SE) models (Geiger, Heckerman, King & Meek, 2001). The evaluation of the marginal likelihood for this class is complicated by two factors. First, some of the parameters of the model may be redundant, and should not be accounted in the BIC score (Geiger, Heckerman & Meek, 1996; Settimi & Smith, 1998). Second, the set of maximum likelihood points is sometimes a complex self-intersecting surface rather than a single maximum likelihood point as in the proven cases for linear and curved exponential models. Recently, major progress has been achieved in analyzing and evaluating this type of integrals (Watanabe, 2001). Herein, we apply these techniques to model selection among Bayesian networks with hidden variables.

Chapter 3 focuses on the asymptotic evaluation of $\mathbb{I}[N, Y_D]$ for a binary naive Bayesian model with binary features. This model, described fully in Section 3.2, is useful in classification of binary vectors into two classes (Friedman, Geiger & Goldszmidt, 1997). Our results are derived under similar assumptions to the ones made by Schwarz (1978) and Haughton (1988). In this sense, this work generalizes the mentioned works, providing valid asymptotic formulas for a new type of marginal likelihood integrals. The resulting asymptotic approximations, presented in Theorem 8, deviate from the standard BIC score. Hence the standard BIC score is not justified for Bayesian model selection among Bayesian networks with hidden variables. Moreover, no uniform score formula exists for such models; our *adjusted BIC score* changes depending on the different types of singularities of the sufficient statistics, namely, the coefficient of the $\ln N$ term in approximation of $\ln \mathbb{I}[N, Y_D]$ is no longer $-\frac{d}{2}$ but rather a function of the sufficient statistics Y_D. An additional result presented in Theorem 9 describes the asymptotic marginal likelihood given a degenerate (missing links) naive Bayesian model; it complements the main result presented by Theorem 8.

Chapter 4 presents algorithms that address two major difficulties in analytic asymptotic approximation of marginal likelihood integrals. First, we implement the method for effective dimensionality computation presented in (Geiger, Heckerman & Meek, 1996) and optimize it by decomposing the input network into independent components. The algorithm is implemented in Matlab and is capable of evaluating effective dimensionality of large Bayesian networks with hidden variables. Second, we fill in the details and implement the algorithmic approach suggested in (Watanabe, 2001) for analytic asymptotic approximation of "hard" integrals. Our algorithm combines state-of-the-art algorithms of algebraic geometry (Bodnar & Schicho, 2000; Bravo, Encinas & Villamayor, 2002) with specific analytic methods for marginal likelihood evaluation suitable for Bayesian networks, which were developed in Chapter 3. The latest algorithm, implemented in Maple , is capable of computing the approximation of the marginal likelihood not only for Bayesian networks with hidden variables, but for a larger set of probabilistic models, for which the log-likelihood function can be represented (or bounded) by a polynomial. We demonstrate the usage of our algorithms in evaluating marginal likelihood formulas on a number of Bayesian networks with hidden variables and on other models.

Chapter 5 analyses the general conditions that dictate conjugate parameter priors and therefore allow closed form evaluation of marginal likelihood integral. It shows that, while global independence dictates a Normal-Wishart prior for Gaussian DAG models with more than 3 nodes (Geiger & Heckerman, 2002), global independence alone does not dictate a Dirichlet prior for discrete DAG models with more than 3 nodes. We provide a minimal set of assumptions needed to dictate a Dirichlet prior for discrete Bayesian network and, in addition, we specify the class of discrete probability distributions, which is larger than the Dirichlet family, that arise under the global independence assumption alone via a solution of a new set of functional equations.

Chapter 6 of this thesis presents an application of Bayesian model selection to the problem of finding the location of a disease gene in the context of genetic linkage analysis. Two models are being compared in this application, one describing the dominant model of disease penetrance and another describing the recessive model of disease penetrance. The likelihood of data given an inheritance model, either recessive or dominant, is computed by averaging the likelihood of data given this model over different penetrance values, using a flat prior. Contrary to the problem of asymptotic model selection, to which the major part of this thesis is devoted, the Bayesian model selection in this particular case of genetic linkage analysis reduces to the problem of numerical evaluation of marginal likelihood integral. The search for a disease gene using the penetrance model selection, which we call Maximizing Bayesian LOD score ($MBLOD$), has been implemented within SUPERLINK (Fishelson & Geiger, 2002), a genetic linkage analysis program based on Bayesian networks, and we have demonstrated its advantages via simulation.

In summary, the rest of the thesis is organized as follows. Chapter 2 gives the background on

Bayesian networks, exponential families of distributions and asymptotic approximation of Laplace-type integrals. Chapter 3 presents one of the main results of this thesis, namely, the asymptotic approximations of the marginal likelihood of data given a naive Bayesian network model. Chapter 4 continues this presentation describing algorithms for automatic analytic approximation of complex marginal likelihood integrals. Chapter 5 describes a minimal set of conditions that ensure a Dirichlet prior for discrete Bayesian networks and, therefore, allow closed form computation of the marginal likelihood. Finally, an application of Bayesian model selection in the context of genetic linkage analysis is described in Chapter 6.

Some background material is repeated in every chapter in order to facilitate its independent reading. Notations slightly change between chapters to underline the specific focus of each chapter. Each chapter ends with a summary and discussion of future directions.

Chapter 2

Background

This chapter introduces the concepts of Bayesian networks, exponential families of distributions and Bayesian model selection procedures. It also provides an introduction to the asymptotic approximation of integrals that arise in large sample Bayesian model selection.

2.1 Bayesian Networks

Let $X = \{X_1, \ldots, X_n\}$ be a collection of random variables each associated with a set of possible values D_i. A graphical model $M = M(s, \mathcal{F}_s)$ for X is a set of joint probability distributions with sample space $\mathbf{D} = D_1 \times \ldots \times D_n$ that is specified via two components: a structure s and a set of local distribution families $\mathcal{F}_s$. When s is an undirected graph, the model m is called an *undirected graphical model*. If s is a directed acyclic graph (DAG) then the model m is called a *DAG model*, or a *Bayesian network*. While the last two terms are often used interchangeably, the term Bayesian network usually refers to the network structure coupled with particular set of parameters and representing some specific distribution.

Given a directed acyclic graph s on nodes X, we denote the parents of X_i by $\mathbf{Pa}_i^s$. The graph s represents the set of conditional independence assertions in model $M(s, \mathcal{F}_s)$, and only these conditional independence assertions, which are implied by a factorization of a joint distribution for X given by $p(\mathbf{x}) = \prod_{i=1}^n p(x_i|\mathbf{pa}_i^s)$, where $\mathbf{x}$ is a value for X (an n-tuple), x_i is a value for X_i and $\mathbf{pa}_i^s$

is a value for $\mathbf{Pa}_i^s$.[1] When x_i has no incoming arcs in s (no parents), $p(x_i|\mathbf{pa}_i^s)$ stands for $p(x_i)$. A DAG model is *complete* if it has no missing arcs. Note that any two complete DAG models for X encode the same set of conditional independence assertion, namely none.

The local distributions are the n conditional and marginal distributions that constitute the factorization of $p(\mathbf{x})$. Each such distribution belongs to the specified family of allowable probability distributions $\mathcal{F}_s$, which depends on a finite set of numerical parameters $\theta_m \in \Theta_m \subseteq \mathbb{R}^k$ (a parametric family). The parameters θ_m^i for a local distribution is a set of real numbers that completely determine the functional form of $p(x_i|\mathbf{pa}_i^s)$.

Herein we are dealing with discrete Bayesian networks, which describe the distributions on discrete variables $\{X_1, \ldots, X_n\}$. Each such network consists the directed acyclic graph s and the local distributions $p(x_i|\mathbf{pa}_i^s)$ specified by multinomial parameters $\theta_m^i = \{\theta_{x_i|\mathbf{pa}_i^s} | x_i \in D_i, \mathbf{pa}_i^s \in \mathbf{D}_{\downarrow \mathbf{Pa}_i^s}\}$, where $\mathbf{D}_{\downarrow \mathbf{Pa}_i^s}$ is the set of possible values of $\mathbf{Pa}_i^s$.

A simple discrete Bayesian network is presented in Figure 2.1. This network describes a situation where a person receives a phone call from the neighbor about the working alarm and should decide if the alarm went of because of a burglary or an earthquake. This Bayesian network consist of a directed graph on nodes E, B, R, A and C, which represent the 5 binary variables of the problem, and 10 network parameters that specify the dependence of each one of the nodes on its' parents. For example, the parameters of node E (earthquake) consist of one parameter θ^E, that describes an a priori earthquake probability, and the parameters of node A (alarm) consist of four parameters describing the conditional probability of alarm going off in the case of earthquake and burglary, earthquake and no burglary, no earthquake and burglary, and none of these events. Formally, the Bayesian network in Figure 2.1 can represent any distribution on $\{E, B, R, A, C\}$ of the form $P(E, B, R, A, C) = P(E)P(B)P(R|E)P(A|E, B)P(C|A)$.

An important class of Bayesian network models are *Bayesian networks with hidden variables*. In such networks the structure s, in addition to the observable variables X, includes latent variables $H = \{H_1, \ldots, H_m\}$ that are never observed. Such situation may happen when, for example, variable A (alarm) of Figure 2.1 is never directly observed (i.e., $H = \{A\}$), but still, we are interested in learning or working with this network. In such networks, the distribution on X is the marginal

[1]Throughout the thesis we denote random variables by capital letters and their values by the corresponding small letters. Collections (sets or vectors) of variables or their values are typesetted in bold and $\downarrow$ sign denotes the restriction of the collection to a particular subset. E.g., $\mathbf{D}_{\downarrow \mathbf{Pa}_i^s}$ denotes the restriction of $\mathbf{D}$ to variables $\mathbf{Pa}_i^s$, which are the parents of X_i in s.

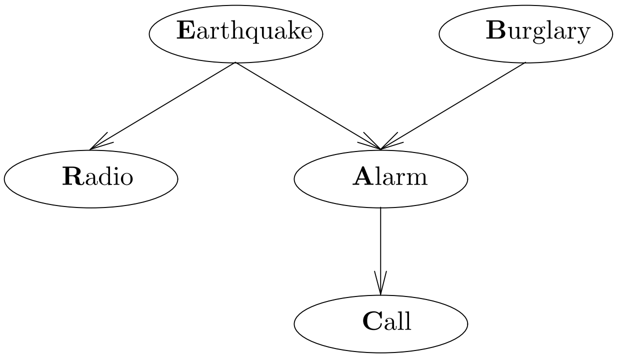


Figure 2.1: An example of simple Bayesian network that describes the probabilistic dependences between events of earthquake, burglary, alarm, phone-call from neighbors, and earthquake announcement on radio. Such network, for example, may be used in order to evaluate the probability of burglary given the values of 'call' and 'radio' variables. This example is borrowed from (Pearl, 1988).

distribution summed over all hidden variables H.

2.2 Exponential Families of Distributions

Let x denote the vector of random variables. The family of probability distributions having the form

$$f(x|\omega) = h(x)a(\omega)e^{c(\omega)\cdot y(x)}, \tag{2.1}$$

where $\omega \in \Omega$ are the distribution parameters, $c(\omega)$ and $y(x)$ are vectors of the same dimension and "$\cdot$" denotes the scalar product, is called the *exponential* family or *Koopman-Darmois* family of distributions. It follows that $y(x)$ is a sufficient statistics for the given exponential family (DeGroot, 1970; page 156).

Since $\int_\Omega f(x|\omega)dx = 1$, the function $a(\omega)$ is completely determined by $c(\omega)$. Thus we can, without loss-of-generality, use $Image(c)$ instead of Ω as a parameter space. In this way we get the following representation of exponential family, which is very common,

$$f(x|\eta) = h(x)e^{\eta\cdot y(x) - b(\eta)}, \tag{2.2}$$

where $\eta = c(\omega)$ and $b(\eta) = \ln a(\omega(\eta))$. The parameters η are called the *natural* or *canonical* exponential parameters and sufficient statistics $y(x)$ is called the *canonical* statistics of x (Barndorff-Nielsen & Cox, 1994).

Statisticians are often faced with the problem of estimation of *maximum likelihood parameters* for a given sample $x = \{x_1, x_2, \ldots, x_N\}$. Recall that the maximum likelihood parameters are the parameters η that maximize $f(x|\eta)$. A useful property of the exponential distributions is the possibility to use the *averaged* canonical statistics of a number of independent samples in order to evaluate the maximum likelihood parameters for this sample. I.e.,

$$f(x_1, x_2, \ldots, x_N|\eta) \propto \prod_{i=1}^{N} e^{\eta \cdot y(x_i) - b(\eta)} = e^{\eta \cdot Y(x) - b(\eta)},$$

where $x = \{x_1, \ldots, x_N\}$ is set of independent samples and $Y(x) = \frac{1}{N}\sum_{i=1}^{N} y(x_i)$ is the averaged canonical statistics.

Provided the canonical parameter space Ω_η consists of all points η such that $\int e^{\eta \cdot y(x)} h(x)dx < 0$ we refer to the family of exponential distributions with $\eta \in \Omega_\eta$ as a *full exponential model*. When η is restricted to a linear subspace $\mathcal{L} \in \Omega_\eta$ then the set of exponential distributions specified by $\eta \in \mathcal{L}$ is called a *linear exponential model*. Similarly, when η is restricted to a smooth surface $\mathcal{C} \in \Omega_\eta$ then the set of exponential distributions specified by $\eta \in \mathcal{C}$ is called a *curved exponential model* (Barndorff-Nielsen & Cox, 1994).

A detailed analysis of properties of exponential distributions is offered in (Kass & Vos, 1997; Murray & Rice, 1993; Barndorff-Nielsen & Cox, 1994; Barndorff-Nielsen, 1978). In the following subsections we present a number of examples of exponential families by rewriting well-known distributions in exponential form, and we demonstrate the connection between graphical models and exponential families of distributions.

2.2.1 An Example of Exponential Family - Normal Distribution

Consider a one-dimensional normal distribution (DeGroot, 1970; page 37),

$$f(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma}\right]. \tag{2.3}$$

We can rewrite this distribution in exponential form:

$$\begin{aligned}
&f(x|\mu, \sigma^2) = e^{-b(\eta_1,\eta_2) + \sum_{i=1,2} \eta_i y_i(x)} \\
&y_1(x) = x, \quad y_2(x) = x^2, \\
&\eta_1 = \tfrac{\mu}{\sigma^2}, \quad \eta_2 = -\tfrac{1}{2\sigma^2} \\
&b(\eta_1, \eta_2) = -\tfrac{\eta_1^2}{4\eta_2} + \tfrac{1}{2}\ln\left[-\tfrac{\pi}{\eta_2}\right]
\end{aligned} \tag{2.4}$$

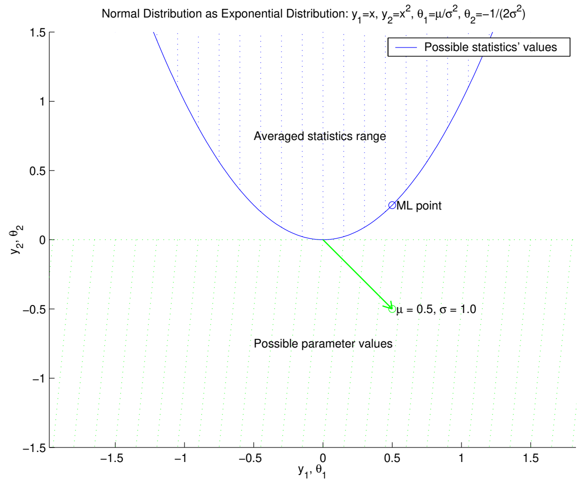


Figure 2.2: Representation of normal distribution family as a subfamily of exponential family of distributions. The graph shows the range of natural parameters $(-\infty, +\infty) \times (-\infty, 0)$; the range of possible values of canonical statistics (y_1, y_2) of each sample (parabola $y_2 = y_1^2$); the range of averaged statistics for sample from normal distribution (parabola interior); and specific parameter vector and most probable data point of for this parameter vector.

In this parameterization, the natural parameter space is $(-\infty, +\infty) \times (-\infty, 0)$ and the sufficient statistics y_1 and y_2 satisfy $y_2 = y_1^2$. This representation is illustrated on Figure 2.2.

This figure illustrates the set of possible values of canonical statistics (y_1, y_2) for a sample (parabola $y_2 = y_1^2$), range of averaged canonical statistics (parabola interior) and, overlapped on the same graph, is the range of natural parameters η_1, η_2. Note that the most probable data value for the normal distribution $x = \mu$ (marked as "ML point" in the figure) corresponds to the point on the canonical statistics parabola where its normal is parallel to the parameter vector.

2.2.2 An Example of Exponential Family - Multinomial Distribution

Consider a multinomial distribution for the distribution of outcomes of N independent trials of discrete random variable with k states (DeGroot, 1970; page 49). This distribution is defined for $\tilde{\mathbf{x}} = (x_1, \ldots, x_k) \in \mathbb{N}^k$ s.t. $\sum_{i=1}^k x_i = N$ with parameters $\mathbf{p} = (p_1, \ldots, p_k)$, such that $\sum_{i=1}^k p_i = 1$ and $\forall i, p_i > 0$:

$$f(\tilde{\mathbf{x}}|N, \mathbf{p}) = \frac{N!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

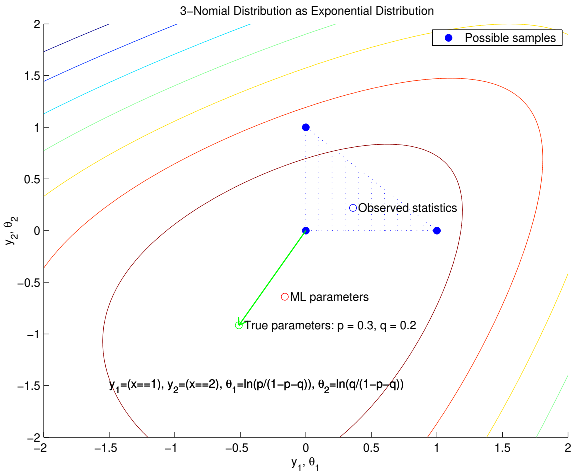


Figure 2.3: Representation of the trinomial distribution family as a subfamily of exponential family of distributions. The range of natural parameters is $\mathbb{R}^2$. The graph shows the possible sample points (black circles), the range of averaged statistics (triangle interior), specific parameter vector and averaged statistics from sampling 100 points from distribution defined by given parameter vector. It also shows the parameter likelihood isolines and maximum likelihood point given the sampled data.

This distribution can be rewritten into the exponential form as

$$\begin{aligned}
&f_N(\mathbf{x} = (x_1, x_2, \ldots, x_{k-1})|\eta) = h(\mathbf{x})e^{\eta\cdot\mathbf{x}-Nb(\eta)},\\
&h(\mathbf{x}) = \tfrac{N!}{x_1!\cdots x_k!}, \quad x_k = N - \textstyle\sum_{i=1}^{k-1} x_i,\\
&\eta = (\eta_1, \ldots, \eta_{k-1}), \quad \eta_i = \ln \tfrac{p_i}{p_k},\\
&b(\eta) = -\ln p_k = \ln(1 + e^{\eta_1} + e^{\eta_2} + \ldots + e^{\eta_{k-1}}).
\end{aligned} \tag{2.5}$$

For $N = 1$ it becomes

$$f_1(y) = e^{\eta\cdot y - b(\eta)} \tag{2.6}$$

where y is the binary vector with $y_i = 1$ if the experiment outcome was i and $y_j = 0$ for all $j \neq i$, and for an outcome of series of N experiments, $\mathbf{x} = y_1 + y_2 + \ldots + y_N$, we have:

$$f_N(\mathbf{x}) = h(\mathbf{x}) \prod_{i=1}^{N} f_1(y_i) = h(\mathbf{x})e^{\eta\cdot\mathbf{x}-Nb(\eta)} = h(\mathbf{x})e^{N(\eta\cdot Y - b(\eta))} \tag{2.7}$$

where $Y = \frac{1}{N}\mathbf{x}$ is the averaged statistics. Figure 2.3 shows the possible values of canonical statistics for the single trial ($N = 1$) and the range of possible averaged statistics for a sequence of trials. In addition this figure shows the approximation of the maximum likelihood parameters after sampling one hundred points from the specified distribution.

2.2.3 Graphical Models as Exponential Families of Distributions

As shown in the previous section, any multinomial distribution on discrete variables $X = \{X_1, \ldots, X_n\}$ is an exponential distribution. Moreover, since a discrete Bayesian network describes a subset of distributions on X it corresponds to some sub-family of an exponential family of distributions on X.

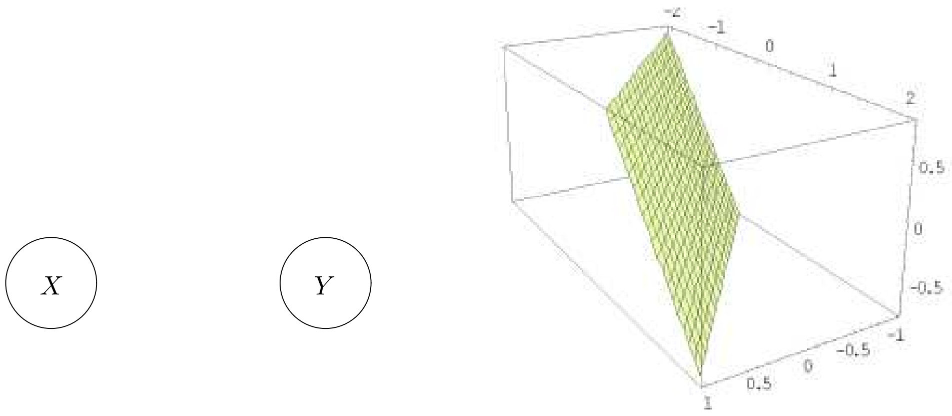
It turns out that the set of distributions that can be represented by an undirected graphical model is a *linear* exponential model (Lauritzen, 1996), namely, the set of distributions that are representable by an undirected graphical model is a linear subspace (hyperplane) in the space of natural parameters of exponential distribution on the network nodes. Furthermore, the set of distributions that can be represented by a Bayesian network without hidden variables is a *curved* exponential model, (Spirtes, Richardson & Meek, 1997; Geiger, Heckerman, King & Meek, 2001). I.e., the set of distributions that are representable by a particular Bayesian network structure is a smooth surface in the space of natural parameters of exponential distribution on the network nodes.

We demonstrate this connection on a simple example. Consider the two nodes Bayesian network depicted on the Figure 2.4a. In this network two binary nodes are independent. Let θ_x and θ_y be the two network parameters. The multinomial parameters are $p_{00} = \theta_x\theta_y$, $p_{01} = \theta_x(1-\theta_y)$, $p_{10} = (1-\theta_x)\theta_y$ and $p_{11} = (1-\theta_x)(1-\theta_y)$ for the pair $\langle X, Y\rangle$ taking the values of 00, 01, 10 and 11 accordingly. Using Eq. 2.5, the natural exponential parameters are

$$\eta_1 = \ln \tfrac{p_1}{p_0} = \ln \tfrac{1-\theta_y}{\theta_y}, \eta_2 = \ln \tfrac{p_2}{p_0} = \ln \tfrac{1-\theta_x}{\theta_x}, \eta_3 = \ln \tfrac{p_3}{p_0} = \ln \tfrac{(1-\theta_x)(1-\theta_y)}{\theta_x\theta_y},$$

yielding the linear constraint $\eta_3 = \eta_1 + \eta_2$. This demonstrates that the set of distributions representable by this two binary nodes network is a linear subspace in the space of natural parameters of joint multinomial distributions on these two binary nodes (Figure 2.4b). Note that in this particular example the empty graph imposed on X and Y is both an undirected graph and a Bayesian network, hence we get a linear exponential model for this set of distributions.

The correspondence between graphical models and linear and curved exponential families allows us to apply the results developed for model selection among exponential models (Schwarz, 1978; Haughton, 1988) for model selection among Bayesian models as well. Unfortunately, the class of families of distributions that are representable by Bayesian network models with *hidden* variables is strictly larger than the class of curved exponential families (Geiger, Heckerman, King & Meek,



(a) (b)

Figure 2.4: An example of a linear exponential model represented by Bayesian network. (a) A simple Bayesian network that consists of two independent binary nodes. (b) A linear subspace of natural exponential parameters of multinomial distributions on two binary nodes that are representable by the given Bayesian network. The entire cube represents the natural parameter space of multinomial distributions on two binary nodes.

2001). Therefore, a number of results for curved exponential models are not valid for Bayesian networks with hidden variables. This work attempts to close this gap.

The next section presents the principle of Bayesian model selection and recites the result of (Schwarz, 1978), which has developed the approximation formula for large-sample Bayesian model selection among linear exponential models.

2.3 Bayesian Model Selection Procedure

Statisticians are often faced with the problem of choosing the appropriate model that best fits a given set of observations. In our case, such problem is the choice of structure in learning of Bayesian networks (Heckerman, Geiger & Chickering, 1995; Cooper & Herskovits, 1992). In model selection the maximum likelihood principle would tend to select the model of highest possible dimension, contrary to the intuitive notion of choosing the right model. Penalized likelihood approaches such as AIC have been proposed to remedy this deficiency (Akaike, 1974).

We focus on the Bayesian approach to model selection by which a model M is chosen according to the maximum a posteriori probability given the observed data D:

$$P(M|D) \propto P(M,D) = P(M)P(D|M) = P(M)\int_{\Omega} P(D|M,\omega)P(\omega|M)d\omega$$

where ω denotes the model parameters and Ω denotes the domain of the model parameters. In particular, we focus on model selection using large sample approximation for $P(M, D)$, called *BIC - Bayesian Information Criterion.*

The critical computational part in using this criterion is evaluating the marginal likelihood integral $P(D|M) = \int_\Omega P(D|M,\omega)P(\omega|M)d\omega$. The factor $P(M)$ is ignored, since it introduces only constant errors in the approximations of $\ln P(D, M) = \ln P(D|M) + \ln P(M)$. Maximizing the logarithm of $P(D|M)$ yields the Bayesian model selection for exponential models of selecting the model M_j that maximizes

$$S(N, Y_D, j) = \ln P(D|M_j) = \ln \int_{\Omega_j} e^{N\cdot[Y_D\cdot\eta(\omega)-b(\eta(\omega))]}\mu_j(\omega)d\omega, \tag{2.8}$$

where Y_D is the averaged sufficient statistics of D and $\omega \in \Omega_j \subseteq \mathbb{R}^{d_j}$ are the model parameters.

Schwarz (1978) proves the following theorem.

Theorem 1 (Schwarz' main result) *If Ω_j represents a linear d_j dimensional exponential model and the prior probability $\mu_j(\omega)$ of ω given model j is bounded and bounded away from zero on Ω_j then for fixed Y and j, as N tends to ∞*

$$S(Y, N, j) = N \ln P(Y|\omega_{ML}) - \frac{1}{2}d_j \ln N + R, \tag{2.9}$$

where ω_{ML} are the maximum likelihood parameters and the remainder $R = R(Y, N, j)$ is bounded in N for a fixed Y and j.

Note that in case we have a finite number of models, $R(N, Y, j)$ is bounded in N for fixed Y, and the bound is independent of j. The score $S(Y, N, j)$ is referred to as the *standard BIC score.*

Later, it was shown that under some additional assumptions, Eq. 2.9 holds for curved exponential models as well (Haughton, 1988). These results allow the use of standard BIC score for undirected and directed graphical models without hidden variables, since these models correspond to linear and curved exponential families (Section 2.2.3). A major part of this thesis is devoted to development of the correct Bayesian scores for model selection among Bayesian models with hidden variables, which fall outside the class of curved exponential models (Geiger, Heckerman, King & Meek, 2001). In the next section we present a number of general results in asymptotic approximation of integrals, which are required for this task.

2.4 Asymptotic Approximation of Laplace Type Integrals

Exact analytical formulas are not available for many integrals arising in practice. In such cases approximate or asymptotic solutions are of interest. Asymptotic analysis is a branch of analysis that is concerned with obtaining approximate analytical solutions to problems where a parameter or some variable in an equation or integral becomes either very large or very small. In this section we review basic definitions and results of asymptotic analysis in relation to the integrals under study.

Let z represent a large parameter. We say that $f(z)$ is *asymptotically equal* to $g(z)$ for $z \to \infty$ if $\lim_{z\to\infty} f/g = 1$, and write

$$f(z) \sim g(z), \text{ as } z \to \infty.$$

Equivalently, $f(z)$ is asymptotically equal to $g(z)$ if $\lim_{z\to\infty} r/g = 0$, denoted $r = o(g)$, where $r(z) = f(z) - g(z)$ is the absolute error of approximation. Note that the error of approximation $r(z) = f(z) - g(z)$ need not be bounded according to this definition, but $r(z)$ is required to be negligible versus $g(z)$, i.e., $\lim_{z\to\infty} r/g = 0$.

We often approximate $f(z)$ by several terms via an iterative approximation of the error terms. An asymptotic approximation by m terms has the form $f(z) = \sum_{n=1}^{m} a_n g_n(z) + o(g_m(z))$, as $z \to \infty$, where $\{g_n\}$ is an *asymptotic sequence* which means that $g_{n+1}(z) = o(g_n(z))$ as $z \to \infty$. An equivalent definition is

$$f(z) = \sum_{n=1}^{m-1} a_n g_n(z) + O(g_m(z)), \text{ as } z \to \infty,$$

where the big 'O' symbol means that the error term is bounded by a constant multiply of $g_m(z)$. The latter definition of asymptotic approximation is often more convenient and we use it herein, mostly for $m = 3$. A good introduction to asymptotic analysis can be found in (Murray, 1984).

One of the objectives of this thesis is deriving asymptotic approximation of marginal likelihood $P(D|M)$ for exponential families, which have the form

$$\mathbb{I}[N, Y_D] = \int_\Omega e^{-Nf(\omega, Y_D)} \mu(\omega) d\omega \tag{2.10}$$

where $f(\omega, Y_D)$ is the minus log-likelihood function.[2] We focus on exponential models, for which the log-likelihood of sampled data is equal to N times the log-likelihood of the averaged sufficient

[2]Throughout this paper we use $\mathbb{I}$ to denote this particular marginal likelihood integral rather than standard 'I' symbol that denote general integrals appearing in theorems, examples and auxiliary derivations.

statistics. Note that as explained in Section 2.2, Bayesian network models discussed in this paper are indeed exponential.

The integral 2.10 is called a Laplace type of integral, since integrals of this form also arise from the Laplace transform of function $f(\omega)$. In the previous section, we have derived the asymptotic approximation of integral $\mathbb{I}[N, Y_D]$ under the assumption of linearity of the log-likelihood function f on Y and ω. Here we present some general results regarding the approximation of Eq. 2.10.

Consider Eq. 2.10 for some fixed Y_D. For large N, the main contribution to the integral comes from the neighborhood of the minimum of f, i.e., the maximum of $-Nf(\omega, Y_D)$. See illustration on Figure 2.5(a,b). Thus, intuitively, the approximation of $\mathbb{I}[N, Y_D]$ is determined by the form of f near its minimum on Ω. In the simplest case $f(\omega)$ achieves a single minimum at ω_{ML} in the interior of Ω and this minimum is non-degenerate, i.e., the Hessian matrix $\mathcal{H}f(\omega_{ML})$ of f at ω_{ML} is of full rank. In this case the isosurfaces of the integrand function near the minimum f are ellipsoids (see Figure 2.5b,c) and the approximation of $\mathbb{I}[N, Y_D]$ for $N \to \infty$ is the classical Laplace approximation (see, e.g., Wong, 1989; page 495), as follows:

Lemma 2 (Laplace Approximation) *Let*

$$I(N) = \int_U e^{-Nf(u)} \mu(u) du$$

where $U \subset \mathbb{R}^d$. Suppose that f is twice differentiable and convex (i.e., $\mathcal{H}f(u)$ is positive definite), the minimum of f on U is achieved on a single internal point u_0, μ is continuous and $\mu(u_0) \neq 0$. If $I(N)$ absolutely converges, then

$$I(N) \sim Ce^{-Nf(u_0)} N^{-d/2} \tag{2.11}$$

where $C = (2\pi)^{d/2} \mu(u_0) [\det \mathcal{H}f(u_0)]^{-\frac{1}{2}}$ is a constant.

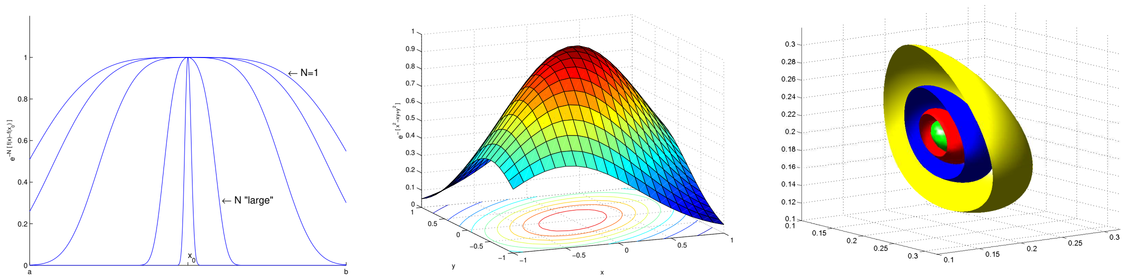
Note that the logarithm of Eq. 2.11 yields the BIC score as presented by Eq. 2.9.

In order to abstract the statistical problem under study, we first consider the evaluation of the integral

$$I[N] = \int_{\mathcal{D}} g(x) e^{-Nf(x)} dx \tag{2.12}$$

where, in our case, $\mathcal{D} = \Omega$, $g(x) = \mu(x)$ and $f(x) = -Y \cdot \theta(x) + b(\theta(x))$. We make a number of technical assumptions:

Assumption 1 [Convergence] *The integral $I[N]$ converges for all $N > 0$.*



(a) (b) (c)

Figure 2.5: The classical Laplace procedure for approximation of integrals of the form $\int e^{-Nf(x)}\mu(x)dx$, where f achieves single minimum in the range of integration. (a) The exponential integrand functions in one dimension, for different N. The larger N becomes, the more mass of the function is concentrated in the small neighborhood of the extremum. (b) The two dimensional integrand function $e^{-(x^2-xy+y^2)}$ for $N=1$. The isosurfaces are ellipses. (c) Ellipsoid-like isosurfaces of the three dimensional log-likelihood multinomial function $f = -[0.2\ln\theta_1 + 0.2\ln\theta_2 + 0.2\ln\theta_3 + 0.4\ln(1-\theta_1-\theta_2-\theta_3)]$.

Assumption 2 [Differentiability] *The functions $f(x)$ and $g(x)$ are sufficiently differentiable for all the operations that follow.*

Assumption 3 [Bounded density] *The density $g(x)$ is bounded and bounded away from zero on $\mathcal{D}$, i.e., there exist constants m and M such that $0 < m < g(x) < M$ for all $x \in \mathcal{D}$.*

Note that for the statistical application (Eq. 2.10), Assumption 1 always holds. To verify this claim, let $f_0 = \inf_{\mathcal{D}} f = \min_{\bar{\mathcal{D}}} f$ and since $f_0 = -\ln P(D|M,\theta_{ML}) > -\infty$ and $\int_{\mathcal{D}} g(x)dx = 1$, observe that

$$I[N] = \int_{\mathcal{D}} g(x)e^{-Nf(x)}dx \leq \int_{\mathcal{D}} g(x)e^{-Nf_0}dx = e^{-Nf_0} < \infty. \tag{2.13}$$

Assumption 3 is introduced for technical convenience. The approximation results can be extended to cases where $g(x) \to 0$ or $g(x) \to \infty$ as $x \to x_0$ and the functional form of $g(x)$ is known. The reader is referred to (Olver, 1974) for discussions on the one-dimensional case with unbounded $g(x)$.

In the rest of this chapter we derive approximations for cases where the maximum likelihood parameters form a surface in the parameter space, rather than a single point (Sections 2.4.1 and 2.4.2). In Section 2.4.3 we present a general approach, due to (Watanabe, 2001), which allows asymptotic approximation of complex integrals of Laplace type. These technical sections are the basis of Chapter 3 that evaluates the marginal likelihood integrals for Bayesian networks with hidden variables. The results of these sections are summarized in Table 2.1.

Table 2.1. Summary of asymptotic approximations of $I[N]$ (Eq. 2.12) under various conditions.

Critical Point(s) ($\nabla f(x_0) = 0$).	**Conditions**	**Asymptotic Approximation**
No critical points. Maximum x_0 on the boundary Γ of $\mathcal{D}$. ($\nabla f(x_0) \neq 0$)		$\ln I[N] = -nf(x_0) - \frac{d+1}{2} \ln n + O(1)$ (Bleistein & Hendelsman, 1975), Section 8.3.
Single critical point, x_0 ($\nabla f(x_0) = 0$).	x_0 is an internal point of $\mathcal{D}$. Hessian $H = \mathcal{H}f(x_0)$ is of full rank.	$I[N] = \left(\frac{2\pi}{N}\right)^{d/2} \frac{g(x_0)}{(\det H)^{1/2}} e^{-Nf(x_0)} \left(1 + O(N^{-1})\right)$ $\ln I[N] = -Nf(x_0) - \frac{d}{2} \ln N + C + O(N^{-1})$ where $C = \frac{d}{2} \ln 2\pi + \ln g(x_0) - \frac{1}{2} \sum_{i=1}^{d} \ln \lambda_i$ with $\{\lambda_i\}_{i=1}^{d}$ being the eigenvalues of H. (Wong, 1989), Section 9.5, (Bleistein & Hendelsman, 1975), Section 8.3.
	x_0 lies on the boundary of $\mathcal{D}$. Hessian $\mathcal{H}f(x_0)$ is of full rank.	$I[N] = \frac{1}{2} \cdot \left(\frac{2\pi}{N}\right)^{d/2} \frac{g(x_0)}{(\det H)^{1/2}} e^{-Nf(x_0)} \left(1 + O(N^{-1})\right)$ (Wong, 1989), Section 9.5, (Bleistein & Hendelsman, 1975), Section 8.3.
Finite number of critical points: x_1, x_2, ..., x_m.	$f(x_1) = f(x_2) = \ldots = f(x_m)$	$I[N] = \sum_{x_i}$ "approximation of $I[N]$ around x_i" $+e^{-Nf(x_1)} r(N)$. $r(N)$ is exponentially small, i.e., $r(N) = O(N^{-\lambda})$ for all $\lambda > 0$.
Critical points form a k-dimensional surface γ. The surface is C^∞ and simple, i.e., no self-intersections (loops).	The critical surface γ is not tangent to the boundary of $\mathcal{D}$. The Hessian of f is of rank $d-k$ for all critical points, i.e. on γ.	$\ln I[N] = -Nf(x_0) - \frac{d-k}{2} \ln N + O(1)$ Section 2.4.1. See (Wong, 1989), Section 8.9 for a similar two dimensional case ($d = 2$, $k = 1$).
Critical points form a k-dimensional surface γ. The surface is C^∞, but can have self-intersections (loops).	$d = 2$, $k = 1$, (x_0, y_0) is a single self-intersection of γ. (x_0, y_0) lies inside $\mathcal{D}$. The critical curve γ is not tangent to the boundary of $\mathcal{D}$ (at the endpoints of γ). The Hessian $\mathcal{H}f$ is of rank 1 for all critical points, except (x_0, y_0).	$\ln I[N] = -Nf(x_0) - \frac{d-k}{2} \ln N + \ln \ln N + O(1)$ Section 2.4.2.
	The general case.	$\ln I[N] = -Nf(x_0) - \lambda \ln N + (m-1) \ln \ln N + O(1)$ where λ and m are the positive rational and natural number which are determined by the singularities in the parametric space. See Section 2.4.3 and (Watanabe, 2001). Application in Chapter 3.

2.4.1 A Surface of Stationary Points

It is known that for certain classes of Bayesian networks the maximum of the log-likelihood function $-f$ is achieved on a sub-surface γ of domain $\mathcal{D}$ (Geiger, Heckerman & Meek, 1996; Chor, Hendy, Holland & Penny, 2000). On the surface γ, $\nabla f = 0$ and such surface is called a *stationary* surface. Our asymptotic evaluation of Integral 2.12 allows stationary surfaces under the following conditions:

(C_1) [No self-crossing] The domain $\mathcal{D}$ contains a C^∞, k-dimensional surface γ such that $\nabla f = 0$ on γ and $\nabla f \neq 0$ on $\mathcal{D} \setminus \gamma$; furthermore γ is simple, i.e., it contains no self-crossings, and if f_0 is a value of f on γ then $f(x) > f_0$ for every x close to, but not on γ.

(C_2) [Non-degenerate Hessian] Let $f_A(t)$ be the function f restricted to the normal subspace N_s to γ at point $A \in \gamma$. Then the Hessian matrix $\frac{\partial^2 f_A}{\partial t_i \partial t_j}$ is of full rank.

(C_3) [Not touching the boundary of $\mathcal{D}$] Let γ be parameterized by $s = \langle s_1, s_2, \ldots, s_m \rangle$, let $\bar{\mathcal{D}}_\gamma$ denote the closed domain of s, and let $x_1 = \xi_1(s)$, $x_2 = \xi_2(s)$, …, $x_k = \xi_k(s)$ for $s \in \bar{\mathcal{D}}_\gamma$. If Γ and Γ_γ are the boundaries of $\mathcal{D}$ and $\mathcal{D}_\gamma$, respectively, then $\xi(\Gamma_\gamma) \subset \Gamma$ and $\xi(s) \notin \Gamma$ for $s \in \text{int } \mathcal{D}_\gamma$.

We state and prove the following

Theorem 3 *Under Assumptions 1-3 and conditions $C1$-$C3$,*

$$I[N] = Ce^{-Nf_0} N^{-\frac{d-k}{2}} \left[1 + O(N^{-1})\right], \quad \text{as } N \to \infty \tag{2.14}$$

where $C = \int_{\mathcal{D}_\gamma} \frac{G(s,0)}{(\det H[s])^{1/2}} ds$ is a constant independent of N.

The immediate consequence of this theorem is that the asymptotic approximation of $\ln I[N]$ is $-Nf_0 - \frac{d-k}{2} \ln N + O(1)$. This claim follows as a special case of a more general result presented in Section 2.4.3. The direct proof is presented here.

Proof: Under condition C_1, $f(x) = f_0 = const$ for $x \in \gamma$. Without the loss of generality, we assume that $f_0 = 0$. Since f_0 is a minimum of f in $\mathcal{D}$, we get that $f(x) > 0$ for every x close to, but not on γ.

We start by noticing that since ϕ and μ are infinitely differentiable (Assumption 2) in $\bar{\mathcal{D}}$ and γ is a C^∞ surface, we can assume that f and g are extended to C^∞ functions in some open neighborhood of $\bar{\mathcal{D}}$, and that the parameterizing function ξ on γ is extended to C^∞ function in some open neighborhood of $\bar{\mathcal{D}}_\gamma$, see (Malgrange, 1968; page 10).

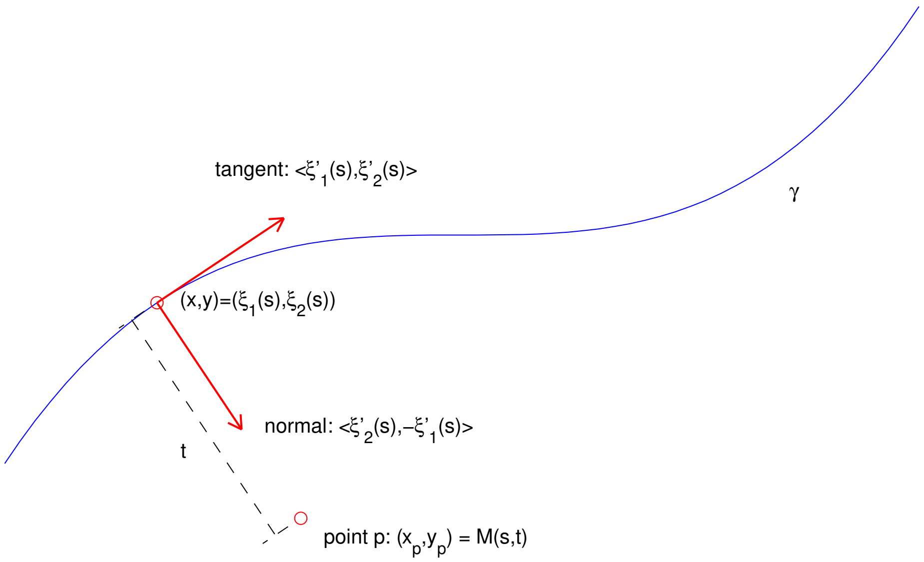


Figure 2.6. Two dimensional $\mathbf{M}$ transformation.

We now define a transformation $\mathbf{M} : (s_1, s_2, \ldots, s_k, t_1, \ldots, t_{d-k}) \to (x_1, \ldots, x_d)$ by

$$x = \xi(s) + A(s) \cdot t \tag{2.15}$$

where $t = \langle t_1, \ldots, t_{k-m} \rangle$ and $A(s)$ is a $k \times (k-m)$ matrix with columns that represent a basis of a vector subspace of $\mathbb{R}^k$ which is orthogonal to a tangent subspace of γ at $\xi(s)$. I.e., $span(A)$ is orthogonal to $span(\frac{\partial \xi(s)}{\partial s_1}, \ldots, \frac{\partial \xi(s)}{\partial s_m})$. The motivation behind the mapping $\mathbf{M}$ is to "straighten" the stationary surface into the Euclidean k-dimensional subspace of $\mathbb{R}^d$.

For an example, in a case of one-dimensional stationary curve in the two dimensional space: $d = 2$, $k = 1$, $s = \langle s \rangle$, $t = \langle t \rangle$, the tangent line to γ at $\xi(s)$ is given $\alpha \langle \xi_1'(s), \xi_2'(s) \rangle$ and $A(s)$ consists of one vector $\langle \xi_2'(s), -\xi_1'(s) \rangle$. Note that $|A| = 1$ is s is an arc length of γ and thus $\xi_1'(s)^2 + \xi_2'(s)^2 = 1$. This transformation is illustrated by Figure 2.6.

As another example, in the case of one-dimensional stationary curve in the three dimensional space: $d = 3$, $k = 1$, $s = \langle s \rangle$, $t = \langle t_1, t_2 \rangle$, the tangent line to γ at $\xi(s)$ is given by $\alpha \cdot \langle \xi_1'(s), \xi_2'(s), \xi_3'(s) \rangle$ and $A(s)$ consists of two linearly independent vectors $a_1(s) \propto \langle \xi_2' + \xi_3', -\xi_1' + \xi_3', -\xi_1' - \xi_2' \rangle$ and $a_2(s) \propto a_1(s) \times \xi'(s) = \langle -\xi_1'\xi_2' - \xi_2'^2 + \xi_1'\xi_3' - \xi_3'^2, \xi_1'^2 + \xi_1'\xi_2' + \xi_2'\xi_3' + \xi_3'^2, -\xi_1'^2 + \xi_1'\xi_3' - \xi_2'^2 - \xi_2'\xi_3' \rangle$. Note that there are many possible matrices $A(s)$ and columns $a_i(s)$ can always be made some C^∞ functions of $\xi'(s)$.

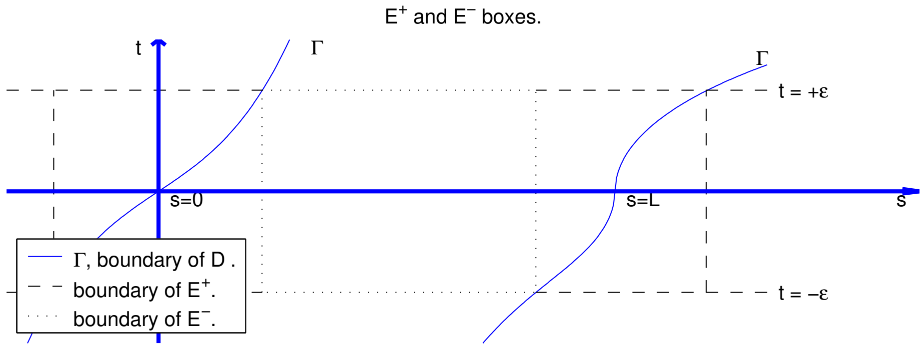


Figure 2.7. Two dimensional E^+ and E^- boxes.

The determinant of the Jacobian matrix (in short, *Jacobian determinant*) of the transformation $\mathbf{M}$ is

$$\begin{aligned} J_{\mathbf{M}}(s_1,\ldots,s_k,t_1,\ldots,t_{d-k}) &= \det \tfrac{\partial(x_1,x_2,\ldots,x_d)}{\partial(s_1,\ldots,s_k,t_1,\ldots,t_{d-k})} \\ &= \det\left(\tfrac{\partial\xi}{\partial s_1}+\tfrac{\partial A}{\partial s_1}\cdot t;\ldots;\tfrac{\partial\xi}{\partial s_k}+\tfrac{\partial A}{\partial s_k}\cdot t;a_1;\ldots;a_{d-k}\right) \end{aligned} \tag{2.16}$$

where $a_1,\ldots,a_{d-k}$ are column vectors of A. For all points $(s_1,\ldots,s_k,0,\ldots,0)$ on γ we have

$$J_{\mathbf{M}}(s_1,\ldots,s_k,0,\ldots,0) = \det\left(\frac{\partial\xi}{\partial s_1};\ldots;\frac{\partial\xi}{\partial s_k};a_1(s);\ldots;a_{d-k}(s)\right). \tag{2.17}$$

Note that column vectors of $A(s)$: $a_1(s),\ldots,a_{d-k}(s)$ are linearly independent and orthogonal to the tangent space of γ at s. Thus, the matrix $\left(\frac{\partial\xi}{\partial s_1};\ldots;\frac{\partial\xi}{\partial s_k};a_1(s);\ldots;a_{d-k}(s)\right)$ is of full rank, and so Jacobian determinant $J_{\mathbf{M}}(s_1,\ldots,s_k,0,\ldots,0)\neq 0$ for $s\in\mathcal{D}_\gamma$.

Since, by definition, $A(s)$ is a C^∞ function of $\xi'(s)$ and therefore it is a C^∞ function of s, we get that $J_{\mathbf{M}}(s_1,\ldots,s_k,t_1,\ldots,t_{d-k})$ is continuous and there exists ϵ such that $J_{\mathbf{M}}\neq 0$ on $\mathcal{D}_\gamma\times[-\epsilon,\epsilon]^{d-k}$. Thus, $\mathbf{M}$ is one-to-one from $E^+=\mathcal{D}_\gamma^+\times[-\epsilon,\epsilon]^{d-k}$ to some neighborhood of γ^+, where $\mathcal{D}_\gamma^+$ is some open neighborhood of $\bar{\mathcal{D}}_\gamma$ and $\gamma^+=\xi(\mathcal{D}_\gamma^+)$.

Let $\mathcal{D}_\gamma^-$ be an open subset of $\mathcal{D}_\gamma$ such that transformation $\mathbf{M}$ maps $E^-=\mathcal{D}_\gamma^-\times[-\epsilon,\epsilon]^{d-k}$ inside $\mathcal{D}$. Changing the variables in Eq. 2.12 from x to (s,t) we obtain (up to exponentially small errors)

$$\int_{E^-} G(s,t)e^{-NF(s,t)}dtds = I^-[N] \leq I[N] \leq I^+[N] = \int_{E^+} G(s,t)e^{-NF(s,t)}dtds \tag{2.18}$$

where $F(s,t)=f(x)$ and $G(s,t)=g(x)$. This idea for the two-dimensional case is illustrated by Figure 2.7.

Let

$$J_s[N] = \int_{[-\epsilon,\epsilon]^{d-k}} G(s,t)e^{-NF(s,t)}dt. \tag{2.19}$$

The function $f_s(t) = F(s,t)$ has a single minimum on $R_\epsilon = [-\epsilon,\epsilon]^{d-k}$ achieved at $t = 0$. The asymptotic approximation to $J_s[N]$ for $N \to \infty$ is (Wong, 1989)

$$J_s[N] = \left(\frac{2\pi}{N}\right)^{(d-k)/2} G(s,0)(\det H[s])^{-1/2} \exp[-NF(s,0)]\left[1 + O(N^{-1})\right] \tag{2.20}$$

where $H[s] = \frac{\partial^2 f_s}{\partial t_i \partial t_j}|_{t=0}$ denotes a Hessian matrix of $f_s(t)$ at $t = 0$. We assume that $H[s]$ is non-degenerate (C_2). Denote by $r(s,N)$ the relative error of approximation, $|r(s,N)| = O(N^{-1})$. From Eq. 2.18 we get

$$\begin{aligned} I^+[N] &= \int_{\mathcal{D}_\gamma^+} J_s[N]ds = \int_{\mathcal{D}_\gamma^+} \left(\tfrac{2\pi}{N}\right)^{(d-k)/2} \tfrac{G(s,0)}{(\det H[s])^{1/2}} e^{-NF(s,0)}(1 + r(s,N))ds \\ &= \left(\tfrac{2\pi}{N}\right)^{(d-k)/2} e^{-Nf_0}\left[1 + O(N^{-1})\right] \cdot C_\epsilon^+ \end{aligned} \tag{2.21}$$

where $C_\epsilon^+ = \int_{\mathcal{D}_\gamma^+} \frac{G(s,0)}{(\det H[s])^{1/2}} ds$. Note that a similar approximation exists for $I^-[N]$ with $C_\epsilon^- = \int_{\mathcal{D}_\gamma^-} \frac{G(s,0)}{(\det H[s])^{1/2}} ds$. Therefore, we get

$$\left(\frac{2\pi}{N}\right)^{(d-k)/2} e^{-Nf_0}\left[1 + O(N^{-1})\right] \cdot C_\epsilon^- \le I[N] \le \left(\frac{2\pi}{N}\right)^{(d-k)/2} e^{-Nf_0}\left[1 + O(N^{-1})\right] \cdot C_\epsilon^+ \tag{2.22}$$

with exponential errors subsumed by the $O(N^{-1})$ error. By dividing all parts by $\left(\frac{2\pi}{N}\right)^{(d-k)/2} e^{-Nf_0}$ and setting $N \to \infty$ we get

$$I[N] \sim C_0 \left(\frac{2\pi}{N}\right)^{(d-k)/2} e^{-Nf_0} \tag{2.23}$$

where $C_0 = \int_{\mathcal{D}_\gamma} \frac{G(s,0)}{(\det H[s])^{1/2}} ds$. Similar analysis can be performed for the next term of the asymptotic expansions of $I^-[N]$ and $I^+[N]$ and the order of the error term can be established to be $O(N^{-1})$. ■

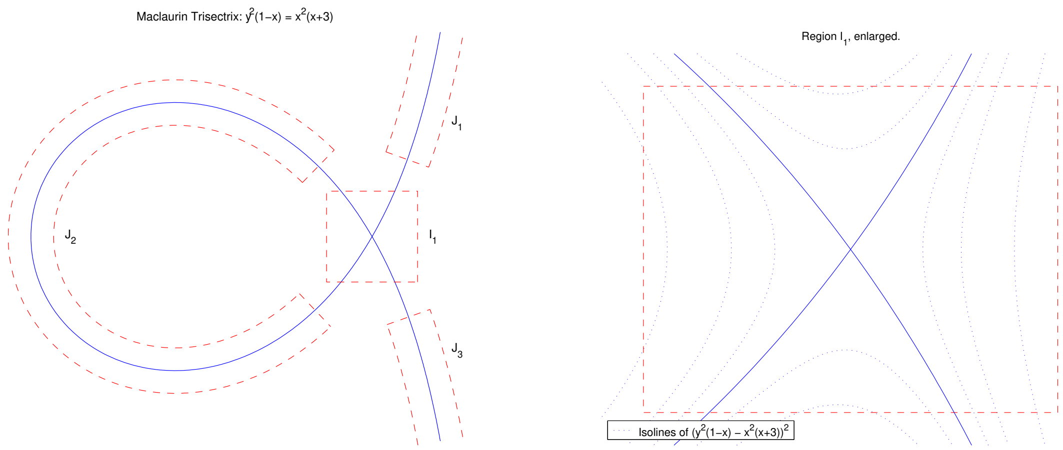
2.4.2 A Self Crossing Stationary Curve in Two Dimensional Space

We now consider the two dimensional version of Integral 2.12, namely,

$$I[N] = \int\int_{\mathcal{D}} g(x,y)e^{-Nf(x,y)}dxdy \tag{2.24}$$

We consider the case where f achieves its minimum on some one-dimensional critical curve γ inside $\mathcal{D}$ and that this curve has a self-intersection. An example of such critical curve is shown on Figure 2.8.

Note that we need to concentrate only on the asymptotic contribution to $I[N]$ from the region I_1 in Figure 2.8, since contributions from regions J_1, J_2 and J_3 are evaluated by the methods described in Section 2.4.1.



(a) (b)

Figure 2.8: Maclaurin Trisectrix as an example of self-crossing stationary curve. Such a critical curve can be generated by the function $f(x,y) = (y^2(1-x) - x^2(x+3))^2$.

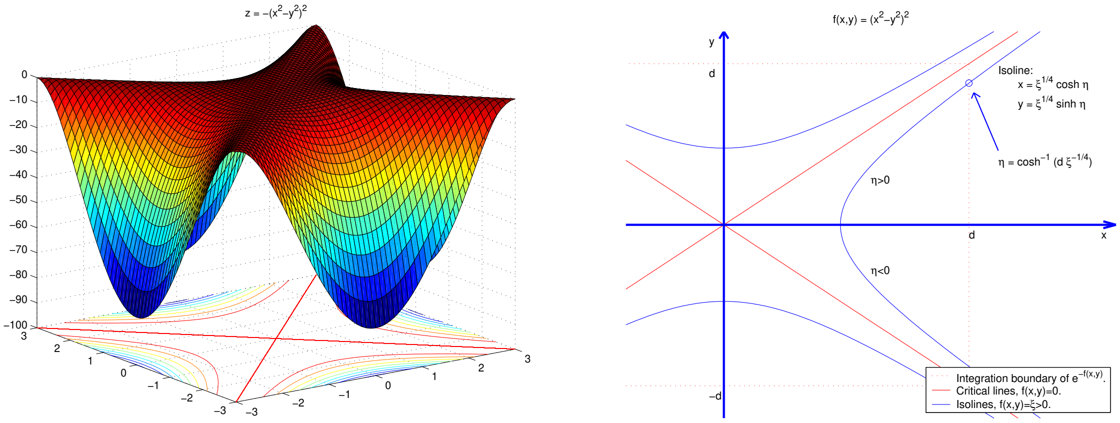
Without loss of generality we assume that a self crossing occurs at the point $(0,0)$. We evaluate $I[N]$ under the following conditions:

(C_4) [Two crossing curves] $\mathcal{D}$ contains two C^∞, one-dimensional curves γ_1 and γ_2 such that $\nabla f = 0$ on γ_1 and γ_2 and $\nabla f \neq 0$ on $\mathcal{D} \setminus (\gamma_1 \cup \gamma_2)$; furthermore, γ_1 intersect γ_2 at point $(0,0)$ and γ_1 and γ_2 are not tangent to each other at the intersection point.

(C_5) [Location of crossing and curves in $\mathcal{D}$] Let γ_1 be parameterized by s, the arc length of γ_1 and let γ_2 be parameterized by t, the arc length of γ_2. We write $(x,y) = (\xi_1(s), \xi_2(s))$ for γ_1 where $-a_1 \leq s \leq a_2$. Similarly, function η_1 and η_2 define a parameterization of γ_2, $t \in [-b_1, b_2]$. The intersection point is $(\xi_1(0), \xi_2(0)) = (\eta_1(0), \eta_2(0)) = (0,0)$. If $A_1 = (\xi_1(-a_1), \xi_2(-a_1))$, $A_2 = (\xi_1(a_2), \xi_2(a_2))$ are the endpoints of γ_1 and B_1, B_2 are the endpoints of γ_2, then $A_1, A_2, B_1, B_2 \in \Gamma$ and these are the only points of intersection of γ_1 and γ_2 with Γ, where Γ is the boundary of $\mathcal{D}$.

(C_6) [Non-degenerate forth derivative of f] Let $x = \xi_1(s) + \eta_1(t)$ and $y = \xi_2(s) + \eta_2(t)$ and let $F(s,t) = f(x,y)$ in some neighborhood of $(0,0)$ then $\frac{\partial^4 F}{\partial^2 s \partial^2 t}(0,0) \neq 0$.

We claim the following.



(a) (b)

Figure 2.9. The form and isolines of the function $f(x, y) = (x^2 - y^2)^2$.

Theorem 4 *Under Assumptions 1-3 and conditions C_4-C_6,*

$$I[N] = \int\int_{\mathcal{D}} g(x,y)e^{-Nf(x,y)}dxdy \sim Ce^{-Nf_0}N^{-\frac{1}{2}}\ln N + e^{-Nf_0}O\left(N^{-\frac{1}{2}}\right) \tag{2.25}$$

where $C = 4\sqrt{\pi}g(0,0)\kappa$ is a constant independent of N, with

$$\begin{aligned} \kappa &= 2\left|\frac{\partial(x,y)}{\partial(s,t)}\right|_{(0,0)}\left[\frac{\partial^4 F}{\partial^2 s\partial^2 t}(0,0)\right]^{-1/2} \\ &= \frac{ad-bc}{\sqrt{c^2(6d^2f_{04}+3bdf_{13}+b^2f_{22})+ac(3d^2f_{13}+4bdf_{22}+3b^2f_{31})+a^2(d^2f_{22}+3bdf_{31}+6b^2f_{40})}} \end{aligned} \tag{2.26}$$

and $a = \xi_1'(0)$, $b = \eta_1'(0)$, $c = \xi_2'(0)$, $d = \eta_2'(0)$.

A special case of Theorem 4, when $f(x,y) = (x^2 - y^2)^2$ (Figure 2.9), is proved first.

Theorem 5 *Let*

$$I[N] = \int_{-a}^{a}\int_{-a}^{a} g(x,y)e^{-N(x^2-y^2)^2}dxdy. \tag{2.27}$$

Then under Assumption 3

$$I[N] = CN^{-\frac{1}{2}}\ln N + O(N^{-1/2}) \tag{2.28}$$

where $C = 2\sqrt{\pi}g(0,0)$ is a constant independent of N.

Proof: We apply the theorem of resolution of multiple integrals (Wong, 1989, Chapter 8), and get

$$I[N] = \int_0^K h(t)e^{-Nt}dt \tag{2.29}$$

with

$$h(t) = \int_{f(x,y)=t} \frac{g(x,y)}{|\nabla f|} d\sigma \tag{2.30}$$

where $|\nabla f| = \sqrt{f_x^2 + f_y^2}$ and $d\sigma$ is a length element of a curve $f(x,y) = t$.

Let

$$x = \xi^{1/4} \cosh \eta, \quad y = \xi^{1/4} \sinh \eta \tag{2.31}$$

with $\eta \in \mathbb{R}$ and $\xi > 0$. This transformation works in the right quadrant ($y < |x|$). For other quadrants, similar transformations can be introduced, but since $f(x) = (x^2 - y^2)^2$ is symmetric the contributions from other quadrants are the same. We have $\xi = (x^2 - y^2)^2 = f(x,y)$ and

$$\frac{d\sigma}{d\eta} = \sqrt{\frac{dx}{d\eta}^2 + \frac{dy}{d\eta}^2} = \xi^{1/4} \sqrt{\cosh 2\eta}. \tag{2.32}$$

Furthermore,

$$|\nabla f| = \sqrt{f_x^2 + f_y^2} = 4|(x^2 - y^2)| \sqrt{x^2 + y^2} = 4\xi^{1/2} \xi^{1/4} \sqrt{\cosh 2\eta}. \tag{2.33}$$

Thus, the contribution of the right quadrant to Integral 2.30 is

$$h_r(t) = \int_{-\delta(t)}^{\delta(t)} g(t^{1/4} \cosh \eta, t^{1/4} \sinh \eta) \frac{1}{4} t^{-1/2} d\eta. \tag{2.34}$$

where $\delta(t) = \cosh^{-1}(a \cdot t^{-1/4})$.

Note that since $g(x,y)$ is bounded and bounded away from zero on $\mathcal{D} = (-a,a)^2$ (Assumption 3), we have

$$2c_1 t^{-1/2} \cosh^{-1}(dt^{-1/4}) < h(t) < 2c_2 t^{-1/2} \cosh^{-1}(dt^{-1/4}), \qquad \forall t \in (0, K) \tag{2.35}$$

where $c_1 = \inf_{\mathcal{D}} g(x,y) \geq m$, $c_2 = \sup_{\mathcal{D}} g(x,y) \leq M$ and m and M are bounds provided by Assumption 3.

Recall that

$$\cosh^{-1}(z) = \ln(z + \sqrt{z^2 - 1}). \tag{2.36}$$

Hence

$$\cosh^{-1}(at^{-1/4}) = \ln(a + \sqrt{a^2 - t^{1/2}}) - \frac{1}{4} \ln t \sim -\frac{1}{4} \ln t + \ln 2a, \qquad \text{as } t \to \infty \tag{2.37}$$

Eqs. 2.29 and 2.34 yield the following upper bound on $I[N]$:

$$\begin{aligned}
I[N] = 4\int_0^K h_r(t)e^{-Nt}dt &< -2\int_0^K c_2 t^{-1/2}\ln t e^{-Nt}dt + 8\int_0^K c_2\ln(2a)t^{-1/2}e^{-Nt}dt \\
&= 2c_2\int_0^\infty \lambda^{-1/2}N^{1/2}(\ln N - \ln\lambda)e^{-\lambda}N^{-1}d\lambda \\
&\quad +8c_2\ln(2a)\int_0^\infty \lambda^{-1/2}N^{1/2}e^{-\lambda}N^{-1}d\lambda \\
&= 2c_2N^{-1/2}\ln N\int_0^\infty \lambda^{-1/2}e^{-\lambda}d\lambda - 2c_2N^{-1/2}\int_0^\infty \lambda^{-1/2}\ln\lambda e^{-\lambda}d\lambda \\
&\quad +8c_2\ln(2a)N^{-1/2}\int_0^\infty \lambda^{-1/2}e^{-\lambda}d\lambda \\
&\sim 2c_2\sqrt{\pi}N^{-1/2}\ln N + N^{-1/2}\left[8c_2\sqrt{\pi}\ln(2a) - 2c_2J\right]
\end{aligned} \tag{2.38}$$

where $J = \int_0^\infty \lambda^{-1/2}\ln\lambda e^{-\lambda}d\lambda$. Since $-\lambda^{-1} < \ln\lambda < \lambda$, we have that integral J converges and that $\Gamma(-1/2) < J < \Gamma(3/2)$.[3]

Since a can be made arbitrarily small and the "branches" of $f(x, y)$ contribute only $O(N^{-1/2})$ (by Theorem 3), we obtain via an argument similar to that before Eq. 2.23 that

$$I[N] = 2\sqrt{\pi}g(0,0)N^{-1/2}\ln N + O(N^{-1/2}). \blacksquare \tag{2.39}$$

Proof of Theorem 4: The main idea of the proof is to make a coordinate change that transforms Integral 2.24 to the form given by Eq. 2.27 which is resolved by Theorem 5.

First we show that only the forth derivatives of f can be non-zero. We apply the Maclaurin expansion, i.e., the Taylor expansion of $f(x, y)$ around $(0, 0)$. Consider the initial terms up to the third order.

$$f(x,y) = f_{00} + f_{10}x + f_{01}y + \frac{1}{2}f_{20}x^2 + f_{11}xy + \frac{1}{2}f_{02}y^2 + O(x^3 + x^2y + xy^2 + y^3) \tag{2.40}$$

where $f_{ij} = \frac{\partial^{i+j}f}{\partial^i x\partial^j y}(0, 0)$. The constant term $f_{00} = f(0, 0) = f_0$ contributes the factor e^{-Nf_0} to the final expansion and we can assume $f_{00} = 0$ without loss of generality. The function f achieves a minimum at $(0, 0)$, so f_{10} and f_{01} are zero. Suppose that f_{20} is not zero. Let $x = z - \frac{f_{11}}{f_{20}}w$ and $y = w$. The Jacobian determinant of this transformation is 1. We have

$$f(x,y) \approx \frac{1}{2}f_{20}x^2 + f_{11}xy + \frac{1}{2}f_{02}y^2 = \lambda z^2 + \mu w^2 \tag{2.41}$$

where $\lambda = \frac{f_{20}}{2}$ and $\mu = \frac{f_{20}f_{02} - f_{11}^2}{2f_{20}}$. On the one hand λ and μ have to be non-negative, because $f(0, 0)$ is minimum. On the other hand, λ and μ can not be positive, since then there will be no two

[3] Calculations using Mathematica show that $J = -\sqrt{\pi}(\mathrm{Eu} + \ln 4) \approx -3.48$, where Eu is the Euler constant approximately equal to 0.577216.

crossing curves with $f = 0$. Thus $f_{20} = 0$ and, by a similar argument, $f_{02} = 0$. Now f_{11} has to be 0 to ensure that $f(0,0)$ is minimum when approached by the lines $y = \pm x$.

We have just shown that first two derivatives of f are zero at $(0,0)$. Consider now the next term of the expansion

$$f(x,y) = \frac{1}{6}f_{30}x^3 + \frac{1}{2}f_{21}x^2y + \frac{1}{2}f_{12}xy^2 + \frac{1}{6}f_{03}y^3 + O(x^4 + x^3y + x^2y^2 + xy^3 + y^4). \tag{2.42}$$

The terms f_{30} and f_{03} are zero because $f(0,0)$ is minimum. Now, $f(x,y) \approx \frac{1}{2}xy(f_{21}x + f_{12}y)$, and let us analyze the behavior of f on the line $y = ax$ for $x \to 0$. We get $f(x, ax) \approx ax^3(f_{21} + af_{12})$, and for small enough a the sign of $f(x, ax)$ is determined by the signs of a and f_{21} and thus $f_{21} = 0$. Similarly, $f_{12} = 0$.

We have

$$f(x,y) = \frac{1}{24}f_{40}x^4 + \frac{1}{6}f_{31}x^3y + \frac{1}{4}f_{22}x^2y^2 + \frac{1}{6}f_{13}xy^3 + \frac{1}{24}f_{04}y^4 + O\left(\sum_{i+j=5} x^iy^j\right) \tag{2.43}$$

We define a transformation $\mathbf{M}$ in the neighborhood of $(0,0)$, $x(s,t) = \xi_1(s) + \eta_1(t)$ and $y(s,t) = \xi_2(s) + \eta_2(t)$. The Jacobian determinant of this transformation is

$$J_{\mathbf{M}}(s,t) = \left|\frac{\partial(x,y)}{\partial(s,t)}\right| = \left|\begin{bmatrix} \xi_1'(s) & \eta_1'(t) \\ \xi_2'(s) & \eta_2'(t) \end{bmatrix}\right| \tag{2.44}$$

and it is non-zero at $(0,0)$ by condition C_4. Thus, since the Jacobian determinant function is continuous, the transformation $\mathbf{M}$ is one-to-one in some neighborhood N_ϵ of $(0,0)$. Let $F(s,t) = f(x(s,t), y(s,t))$. We have

$$I[N] = \int_{N_\epsilon} G(s,t)e^{-NF(s,t)}dsdt + O(N^{-1/2}) \tag{2.45}$$

where $G(s,t) = g(x(s,t), y(s,t))J_{\mathbf{M}}(s,t)$. Note that the $O(N^{-1/2})$ contribution comes from the branches of f that lie in $\mathcal{D} \setminus N_\epsilon$.

By our definitions, F achieves its minimum on lines $s = 0$ and $t = 0$. Hence $F_{j0}(0,0)$ and $F_{0j}(0,0)$ are zero for $j \geq 1$, $F_{01}(s,0) = 0 = \mathit{const}$ and $F_{10}(0,t) = 0 = \mathit{const}$. Thus, $F_{j1}(0,0) = 0$ and $F_{1j} = 0$ for $j \geq 1$. We have

$$F(s,t) = \frac{1}{4}F_{22}s^2t^2 + O\left(\sum_{i+j=5;\ i,j\geq 2} s^it^j\right) \tag{2.46}$$

and in terms of the original variables

$$F_{22} = 4\left[c^2(6d^2 f_{04} + 3bdf_{13} + b^2 f_{22}) + ac(3d^2 f_{13} + 4bdf_{22} + 3b^2 f_{31}) + a^2(d^2 f_{22} + 3bdf_{31} + 6b^2 f_{40})\right] \tag{2.47}$$

where $a = \xi_1'(0)$, $b = \eta_1'(0)$, $c = \xi_2'(0)$ and $d = \eta_2'(0)$.

Eq. 2.46 can be written as

$$F(s,t) = \frac{1}{4}F_{22}s^2t^2\left(1 + P(s,t)\right) \tag{2.48}$$

where $P(s,t)$ is a power series in s and t such that $P(0,0) = 0$. We now make the change of variables

$$u = \frac{1}{2}\sqrt{F_{22}}s \quad \text{and} \quad v = t\left[1 + P(s,t)\right]^{1/2} \tag{2.49}$$

and

$$z = \frac{u+v}{2} \quad \text{and} \quad w = \frac{u-v}{2} \tag{2.50}$$

with Jacobian determinant at $(0,0)$ equal $\left|\frac{\partial(s,t)}{\partial(z,w)}\right|_{(0,0)} = \left|\frac{\partial(s,t)}{\partial(u,v)}\right| \cdot \left|\frac{\partial(u,v)}{\partial(z,w)}\right|_{(0,0)} = (\frac{1}{2}\sqrt{F_{22}})^{-1} \cdot 2 = 4F_{22}^{-1/2} > 0$. So there exists $\epsilon > 0$ such that

$$I[N] = \int_{-\epsilon}^{\epsilon}\int_{-\epsilon}^{\epsilon} H(z,w)e^{-N(z^2-w^2)^2}dzdw + O(N^{-1/2}) \tag{2.51}$$

where

$$H(z,w) = \left|\frac{\partial(s,t)}{\partial(z,w)}\right| G(s(z,w),t(z,w)) \tag{2.52}$$

and H and G are bounded in the neighborhood of zero, because the Jacobian determinants of these transformations are continuous and non-zero. We now apply Theorem 5 and obtain, including the term e^{-Nf_0},

$$I[N] = Ce^{-Nf_0}N^{-\frac{1}{2}}\ln N + e^{-Nf_0}O\left(N^{-\frac{1}{2}}\right) \tag{2.53}$$

where, from Theorem 5 and Eqs. 2.44 and 2.47,

$$C = 2\sqrt{\pi}H(0,0) = 2\sqrt{\pi}4F_{22}^{-1/2}J_{\mathbf{M}}(0,0)g(0,0) = 4\sqrt{\pi}g(0,0)\kappa \tag{2.54}$$

and κ is specified by Eq. 2.26. ■

2.4.3 The General Approximation Method by Watanabe

In many cases, and, in particular, in the case of naive Bayesian networks to be defined in the next chapter, the minimum of f (Eq. 2.10) is achieved on a variety $W_0 \subset \Omega$. Sometimes, this variety may be d'-dimensional surface (smooth manifold) in Ω in which case the computation of the integral is locally equivalent to the $d - d'$ dimensional classical case. The hardest cases to evaluate happen when the variety W_0 contains self-intersections. Section 2.4.2 dealt with the simplest case of such a variety, namely, a self-crossing line in the plane.

Recently, an advanced mathematical method for approximating this type of integrals has been introduced to the machine learning community by Watanabe (2001). Below we briefly describe this method and state the main results. First, we introduce the main theorem that enables us to evaluate the asymptotic form of $\mathbb{I}[N, Y_D]$ (Eq. 2.10) as $N \to \infty$ computed in a neighborhood of a maximum likelihood point.

Theorem 6 (based on Watanabe, 2001) *Let*

$$I(N) = \int_{W_\varepsilon} e^{-Nf(w)} \mu(w) dw$$

where W_ε is some closed ε-box around w_0, which is a minimum point of f in W_ε, and $f(w_0) = 0$. Assume that f and μ are analytic functions, $\mu(w_0) \neq 0$. Then,

$$\ln I(N) = \lambda_1 \ln N + (m_1 - 1) \ln \ln N + O(1)$$

where the rational number $\lambda_1 < 0$ and the natural number m_1 are the largest pole and its multiplicity of the meromorphic (analytic + poles) function that is analytically continued from

$$J(\lambda) = \int_{f(w)<\epsilon} f(w)^\lambda \mu(w) dw \qquad (Re(\lambda) > 0) \tag{2.55}$$

where $\epsilon > 0$ is a sufficiently small constant. [4]

This theorem states the main claim of the proof of Theorem 1 in (Watanabe, 2001). Consequently, the approximation of the marginal likelihood integral $\mathbb{I}[N, Y_D]$ (Eq. 2.10) can be determined by the poles of

$$J_{w_0}(\lambda) = \int_{W_\epsilon} \left[f(w) - f(w_0)\right]^\lambda \mu(w) dw$$

evaluated in the neighborhoods W_ϵ of points w_0 on which f attains its minimum. This claim, which is further developed in Section 3.4, holds because the minimum of $f(w) - f(w_0)$ is zero and the main contribution to $\mathbb{I}[N, Y_D]$ comes from the neighborhoods around the minimums of f.

[4]Recall that the pole of the complex function $f(z)$ is the point where it has a finite number of negative terms in its Laurent expansion, i.e., $f(z) = a_{-m}/(z - z_0)^m + \ldots + a_0 + a_1(z - z_0) + \ldots$. In this case it is said that $f(z)$ has a pole of order (or multiplicity) m at z_0, e.g., (Lang, 1993).

Often, however, it is not easy to find the largest pole and multiplicity of $J(\lambda)$ defined by Eq. 2.55. Here, another fundamental mathematical theory is helpful. The *resolution of singularities* in algebraic geometry transforms the integral $J(\lambda)$ into a direct product of integrals of a single variable.

Theorem 7 (Atiyah, 1970**, Resolution Theorem)** *Let $f(w)$ be a real analytic function defined in a neighborhood of $0 \in \mathbb{R}^d$. Then there exists an open set W that includes 0, a real analytic manifold U, and a proper analytic map $g : U \to W$ such that:*

1. *$g : U \setminus U_0 \to W \setminus W_0$ is an isomorphism, where $W_0 = f^{-1}(0)$ and $U_0 = g^{-1}(W_0)$.*

2. *For each point $p \in U$ there are local analytic coordinates $(u_1, \ldots, u_d)$ centered at p so that, locally near p, we have*

$$f(g(u_1, \ldots, u_d)) = a(u_1, \ldots, u_d)u_1^{k_1} \ldots u_d^{k_d},$$

where $k_i \geq 0$ and $a(u)$ is an analytic function with analytic inverse $1/a(u)$.

This theorem is based on the fundamental results of Hironaka (1964) and the process of changing to u-coordinates is known as resolution of singularities.

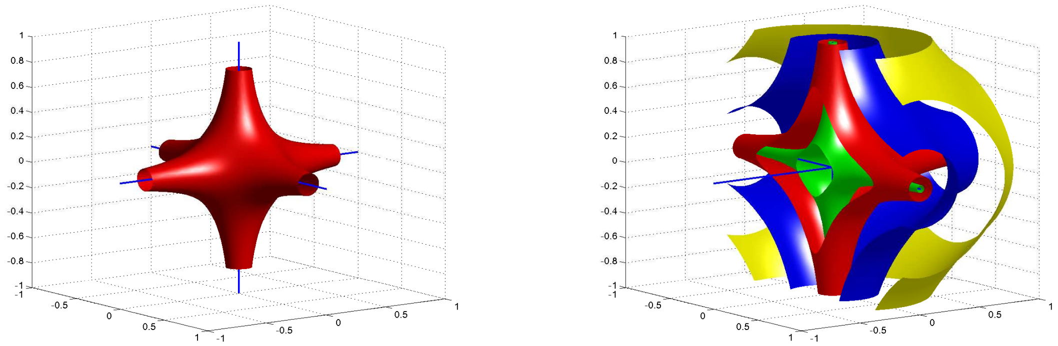
Theorems 6 and 7 provide an approach for computing the leading terms in the asymptotic expansion of $\ln \mathbb{I}[N, Y_D]$:

1. Cover the integration domain Ω by a finite union of open neighborhoods W_α. This is possible under the assumption that Ω is compact.

2. Find a resolution map g_α and manifold U_α for each neighborhood W_α by resolution of singularities. Note that in the process of resolution of singularities U_α may be further divided into subregions $U_{\alpha\beta}$ by neighborhoods of different points $p \in U_\alpha$, as specified by Theorem 7. Select a finite cover of U_α by $U_{\alpha\beta}$, which is possible because the closure of each U_α is also compact.

3. Compute the integral $J(\lambda)$ (Eq. 2.55) in each region $W_{\alpha\beta} = g_\alpha(U_{\alpha\beta})$ and find its poles and their multiplicity. This integral, denoted by $J_{\alpha\beta}$, becomes

$$\begin{aligned} J_{\alpha\beta}(\lambda) &= \int_{W_{\alpha\beta}} f(w)^\lambda \mu(w) dw \\ &= \int_{U_{\alpha\beta}} f(g_\alpha(w))^\lambda \mu(g_\alpha(u)) |g'_\alpha(u)| du \\ &= \int_{U_{\alpha\beta}} a(u)^\lambda \, u_1^{\lambda k_1} u_2^{\lambda k_2} \ldots u_d^{\lambda k_d} \, \mu(g_\alpha(u)) \, |g'_\alpha(u)| \, du. \end{aligned} \tag{2.56}$$



(a) (b)

Figure 2.10: Part (a) depicts an isosurface of $e^{-N(u_1^2u_2^2+u_1^2u_3^2+u_2^2u_3^2)}$ (or alternatively of $u_1^2u_2^2+u_1^2u_3^2+u_2^2u_3^2$) and its set of maximum (minimum) points which coincide with the three axis. Part (b) depicts four isosurfaces of the same function for its different values. The isosurfaces are not ellipsoids as in the classical Laplace case of a single maximum (see Figure 2.5c).

where $|g'_a(u)|$ is the Jacobian determinant. The last integration (up to a constant) is done by bounding $a(u)$ and $\mu(g_\alpha(u))$, using the Taylor expansion for $|g'_\alpha|$, and integrating each variable u_i separately. The largest pole $\lambda_{\alpha\beta}$ of $J_{\alpha\beta}$ and its multiplicity $m_{\alpha\beta}$ are now found.

4. The largest pole and multiplicity of $J(\lambda)$ are $\lambda_{(\alpha\beta)^*} = \max_{(\alpha\beta)} \lambda_{\alpha\beta}$ and the corresponding multiplicity $m_{(\alpha\beta)^*}$. If the $(\alpha\beta)^*$ values that maximize $\lambda_{\alpha\beta}$ are not unique, then the $(\alpha\beta)^*$ value that maximizes the corresponding multiplicity $m_{(\alpha\beta)^*}$ is chosen.

In order to demonstrate this method, we approximate the integral

$$I[N] = \int_{-\epsilon}^{+\epsilon}\int_{-\epsilon}^{+\epsilon}\int_{-\epsilon}^{+\epsilon} e^{-N(u_1^2u_2^2+u_1^2u_3^2+u_2^2u_3^2)}\, du_1du_2du_3 \tag{2.57}$$

as N tends to infinity. This approximation of $I[N]$ is an important component in establishing of main results of Chapter 3. The key properties of the integrand function in Eq. 2.57 are illustrated in Figure 2.10.

Watanabe's method calls for the analysis of the poles of the following function

$$J(\lambda) = \int_{-\epsilon}^{+\epsilon}\int_{-\epsilon}^{+\epsilon}\int_{-\epsilon}^{+\epsilon} (u_1^2u_2^2+u_1^2u_3^2+u_2^2u_3^2)^\lambda\, du_1du_2du_3. \tag{2.58}$$

To find the poles of $J(\lambda)$ we transform the integrand function into a more convenient form by changing to new coordinates via the process of resolution of singularities. To obtain the needed transformations for the integral under study, we apply a technique called *blowing-up* which consists of a series of *quadratic transformations*. For an introduction to these techniques see (Abhyankar, 1990).

Rescaling the integration range to $(-1,1)$ and then taking only the positive octant yields

$$\begin{aligned} J(\lambda) &= 8\epsilon^{4\lambda+3}\int_{(0,1)^3}(u_1^2u_2^2+u_1^2u_3^2+u_2^2u_3^2)^\lambda du \\ &= 8\epsilon^{4\lambda+3}\left(\int_{0<u_2,u_3<u_1<1}+\int_{0<u_1,u_3<u_2<1}+\int_{0<u_1,u_2<u_3<1}\right)(u_1^2u_2^2+u_1^2u_3^2+u_2^2u_3^2)^\lambda du. \end{aligned}$$

The three integrals are symmetric, so we evaluate only the first. Using the quadratic transformation $u_2=u_1u_2$, $u_3=u_1u_3$, which modifies the integration range $0<u_2,u_3<u_1<1$ to be $(0,1)^3$, yields

$$J_1(\lambda)=\int_{0<u_2,u_3<u_1<1}(u_1^2u_2^2+u_1^2u_3^2+u_2^2u_3^2)^\lambda du=\int_{(0,1)^3}u_1^{4\lambda+2}(u_2^2+u_3^2+u_2^2u_3^2)^\lambda du.$$

We now divide the range $(0,1)^3$ to the regions $0<u_3<u_2<1$ and $0<u_2<u_3<1$. Again these cases are symmetric and so we continue to evaluate only the first using the transformation $u_3=u_2u_3$,

$$J_{11}(\lambda)=\int_{0<u_3<u_2<1}u_1^{4\lambda+2}(u_2^2+u_3^2+u_2^2u_3^2)^\lambda du=\int_{(0,1)^3}u_1^{4\lambda+2}u_2^{2\lambda+1}(1+u_3^2+u_2^2u_3^2)du.$$

Since the function $(1+u_3^2+u_2^2u_3^2)$ is bounded on the region of integration, namely $1\le 1+u_3^2+u_2^2u_3^2\le 3$ for all $0\le u_2,u_3\le 1$, it follows that

$$8\epsilon^{4\lambda+3}\int_{(0,1)^2}u_1^{4\lambda+2}u_2^{2\lambda+1}du_1du_2\ \le\ J(\lambda)\ \le\ 24\epsilon^{4\lambda+3}\int_{(0,1)^2}u_1^{4\lambda+2}u_2^{2\lambda+1}du_1du_2, \tag{2.59}$$

yielding

$$8\epsilon^{4\lambda+3}\frac{1}{(4\lambda+3)(2\lambda+2)}\ \le\ J(\lambda)\ \le\ 24\epsilon^{4\lambda+3}\frac{1}{(4\lambda+3)(2\lambda+2)}.$$

Thus $J(\lambda)$ has poles at $\lambda=-3/4$ and $\lambda=-1$ with multiplicity $m=1$. The largest pole is $\lambda=-3/4$ with multiplicity $m=1$. We conclude, using Theorem 6, that $I[N]$ defined by Eq. 2.57 is asymptotically equal to $cN^{-\frac{3}{4}}$.

We note that in this process of resolution of singularities we have implicitly computed the terms k_1, k_2, k_3, the function $a(u)$ and the Jacobian determinant $|g'(u)|$ (in Eq. 2.56). In particular, we have established that $k_1=4$, $k_2=2$, $k_3=0$, $a(u)=1+u_3^2+u_2^2u_3^2$ and $|g'(u)|=u_1^2u_2$ for the appropriate range under study. The mapping g (of Theorem 7) is the composition of the two transformations we used and is defined via $u_1=u_1$, $u_2=u_1u_2$ and $u_3=u_1u_2u_3$. However, this explicit form is not needed for the evaluation of the target integral, as long as the values of k_i and $|g'(u)|$ are derived.

In the proof of the theorems in Chapter 3 we perform a similar process of resolution of singularities producing implicitly the mapping g which is guaranteed to exist according to Theorem 7, and which determines the values of k_i and $|g'(u)|$ needed for evaluation of poles of function $J(\lambda)$ as required by Theorem 6.

Chapter 3

Asymptotic Model Selection for Naive Bayesian Networks

We develop a closed form asymptotic formula to compute the marginal likelihood of data given a naive Bayesian network model with two hidden states and binary features. This formula deviates from the standard BIC score. It provides a concrete example that the BIC score is generally incorrect for statistical models that belong to stratified exponential families. This claim stands in contrast to linear and curved exponential families, where the BIC score has been proven to provide a correct asymptotic approximation for the marginal likelihood. A version of this chapter has been published as (Rusakov & Geiger, 2004).

3.1 Introduction

We focus on the Bayesian approach to model selection by which a model M is chosen according to the maximum posteriori probability given the observed data D, $P(M|D) \propto P(M,D) = P(M)P(D|M) = P(M)\int_\Omega P(D|M,\omega)P(\omega|M)d\omega$, where ω denotes the model parameters and Ω denotes the domain of the model parameters. Given an exponential model M we write $P(D|M)$ as a function of the averaged sufficient statistics Y_D of the data D, and the number N of data points in D:

$$\mathbb{I}[N, Y_D, M] = \int_{\Omega} e^{loglikelihood(Y_D, N|\omega, M)} \mu(\omega|M) d\omega \tag{3.1}$$

where $\mu(\omega|M)$ is the prior parameter density for model M. Recall that the sufficient statistics for multinomial samples of n binary variables $(X_1, \ldots, X_n)$ is simply the counts $N \cdot Y_D$ for each of the possible 2^n joint states. The model selection principle that uses a large sample approximations to Eq. 3.1 is called *BIC - Bayesian Information Criterion.*

For many types of models the asymptotic evaluation of Eq. 3.1, as $N \to \infty$, uses a classical Laplace procedure (Section 2.4). This evaluation was first performed for Linear Exponential (LE) models (Schwarz, 1978) and then for Curved Exponential (CE) models under some additional technical assumptions (Haughton, 1988). It was shown that

$$\ln \mathbb{I}[N, Y_D, M] = N \cdot \ln P(Y_D|\omega_{ML}) - \frac{d}{2} \ln N + O(1), \tag{3.2}$$

where $\ln P(Y_D|\omega_{ML})$ is the log-likelihood of Y_D given the maximum likelihood parameters of the model M and d is the model dimension, i.e., the number of parameters. We call the above approximation the *standard BIC score.*

As explained in Sections 2.2 and 2.3, the use of BIC score for Bayesian model selection for Graphical Models is valid for Undirected Graphical Models without hidden variables because these are linear exponential models (Lauritzen, 1996), and for Bayesian networks without hidden variables because these are curved exponential models (Geiger, Heckerman, King & Meek, 2001; Spirtes, Richardson & Meek, 1997).

The evaluation of the marginal likelihood $\mathbb{I}[N, Y_D]$ for Bayesian networks with hidden variables is more complicated because the class of distributions represented by Bayesian networks with hidden variables is significantly richer than curved exponential models and it falls into the class of stratified exponential models (Geiger, Heckerman, King & Meek, 2001). The evaluation of the marginal likelihood for this class is complicated by two factors. First, some of the parameters of the model may be redundant, and should not be accounted in the BIC score (Geiger, Heckerman & Meek, 1996; Settimi & Smith, 1998). Second, the set of maximum likelihood points is sometimes a complex self-intersecting surface rather than a single maximum likelihood point as in the proven cases for linear and curved exponential models. Recently, major progress has been achieved in analyzing and evaluating this type of integrals (Watanabe, 2001). Herein, we apply these techniques to model selection among Bayesian networks with hidden variables.

The focus of this chapter is the asymptotic evaluation of $\mathbb{I}[N, Y_D]$ for a binary naive Bayesian model with binary features. This model, described fully in Section 3.2, is useful in classification of binary vectors into two classes (Friedman, Geiger & Goldszmidt, 1997). Our results are derived under similar assumptions to the ones made by Schwarz (1978) and Haughton (1988). In this sense, this chapter generalizes the mentioned works, providing valid asymptotic formulas for a new type of marginal likelihood integrals. The resulting asymptotic approximations, presented in Theorem 8, deviate from the standard BIC score. Hence the standard BIC score is not justified for Bayesian model selection among Bayesian networks with hidden variables. Moreover, no uniform score formula exists for such models; our *adjusted BIC score* changes depending on the different types of singularities of the sufficient statistics, namely, the coefficient of the $\ln N$ term (Eq. 3.2) is no longer $-\frac{d}{2}$ but rather a function of the sufficient statistics. An additional result presented in Theorem 9 describes the asymptotic marginal likelihood given a degenerate (missing links) naive Bayesian model; it complements the main result presented by Theorem 8.

The rest of this chapter is organized as follows. Section 3.2 reviews naive Bayesian models and explicates the relevant marginal likelihood integrals for these models. Section 3.3 states and explains our main results and Section 3.4 gives a proof outline of Theorem 8 that demonstrates the mathematical techniques used herein. The full proof of our theorems is deferred to Sections 3.5 and 3.6. Section 3.7 discusses our contributions and outlines future research directions.

3.2 Naive Bayesian Models

A naive Bayesian model M for discrete variables $X = \{X_1, \ldots, X_n\}$ is a set of joint distributions for X that factor according to the tree structure depicted on Figure 3.1, where the class variable C is never observed. Formally, a probability distribution $P(X = x)$ belongs to a naive Bayesian model if and only if

$$P(X = x) = \sum_{j=1}^{r} P(C = c_j) \prod_{i=1}^{n} P(X_i = x_i | C = c_j),$$

where $x = (x_1, \ldots, x_n)$ is the n-dimensional binary vector of values of X, r is the number of hidden states and c_j denotes a particular unobserved state (class). Intuitively, this model describes the generation of data x that comes from r sources $c_1, \ldots, c_r$. Naive Bayesian models are a subclass of

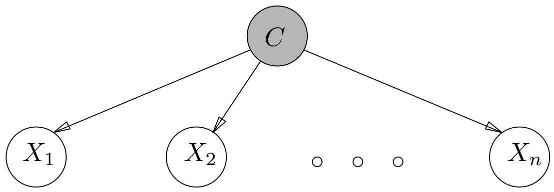


Figure 3.1. A naive Bayesian model. Class variable C is latent.

Bayesian networks (Pearl, 1988) and they are widely used in clustering (Cheeseman & Stutz, 1995).

In this work we focus on naive Bayesian networks that have two hidden states ($r = 2$) and n binary feature variables $X_1, \ldots, X_n$. We denote the parameters defining $p(x_i = 1|c_1)$ by a_i, the parameters defining $p(x_i = 1|c_2)$ by b_i, and the parameters defining $p(c_1 = 1)$ by t. These parameters are called the *model parameters.* We denote the *joint space parameters* $P(X = x)$ by θ_x. The following mapping, denoted T, relates these two sets of parameters:

$$\theta_x = t \prod_{i=1}^{n} a_i^{x_i}(1 - a_i)^{1-x_i} + (1 - t) \prod_{i=1}^{n} b_i^{x_i}(1 - b_i)^{1-x_i}, \tag{3.3}$$

and the marginal likelihood integral (Eq. 3.1) for these models becomes

$$\mathbb{I}[N, Y_D] = \int_{(0,1)^{2n+1}} e^{N \sum_x Y_x \ln \theta_x(\omega)} \mu(\omega) d\omega \tag{3.4}$$

where $\omega = (a_1, \ldots, a_n, b_1, \ldots, b_n, t)$ are the model parameters, N is the sample size, and the averaged sufficient statistics Y_x is the number of samples for which $X = x$ divided by the sample size N.

3.3 Main Claims

This section presents an asymptotic approximation of the integral $\mathbb{I}[N, Y_D]$ (Eq. 3.4) for naive Bayesian networks consisting of binary variables $X_1, \ldots, X_n$ and two hidden states. It is based on two results. First, the classification of singular points for these types of models (Geiger, Heckerman, King & Meek, 2001). Second, Watanabe's approach as explained in Section 2.4, which provides a method to obtain the correct asymptotic formula of $\mathbb{I}[N, Y_D]$ for the singular points not covered by the classical Laplace approximation scheme.

Let $\Upsilon = \{(y_1, \ldots, y_{2^n})|y_i \geq 0, \sum y_i = 1\}$ be the set of possible values of the averaged sufficient statistics $Y_D = (Y_1, \ldots, Y_{2^n})$ for data $D = \{(x_{i,1}, \ldots, x_{i,n})\}_{i=1}^{N}$. In our asymptotic analysis we let the sample size N grow to infinity.

Let $\Upsilon_0 \subset \Upsilon$ be the points $(y_1, \ldots, y_{2^n})$ that correspond to the joint space parameters of the distributions that can be represented by binary naive Bayesian models with n binary variables. In other words, assuming the indices of y_i are written as vectors $(\delta_1, \ldots, \delta_n)$ of n zeros and ones, points in Υ_0 are those that can be parameterized via

$$y_{(\delta_1,\ldots,\delta_n)} = t\prod_{i=1}^{n} a_i^{\delta_i}(1-a_i)^{1-\delta_i} + (1-t)\prod_{i=1}^{n} b_i^{\delta_i}(1-b_i)^{1-\delta_i} \tag{3.5}$$

where t, $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$ are the $2n+1$ model parameters, as defined in Section 3.2.

Geiger et al. (2001) classify the singular points of the algebraic variety of the parameters of binary naive Bayesian networks into two classes S and S'. This classification is used here to classify the possible statistics arising from binary naive Bayesian networks with different parameters; The set S is the set of points $(y_1, \ldots, y_{2^n})$ such that Eq. 3.5 holds and all $a_i = b_i$ except for at most two indices in $\{1, \ldots, n\}$. Intuitively, each such point represents a probability distribution that can be defined by a naive Bayesian model (Figure 3.1) with all links removed except at most two.

The set $S' \subset S$ is the set of points represented by a naive Bayesian model, just as the set S does, but with all links removed; namely, a distribution where all variables are mutually independent and independent of the class variable as well. These statistics are parameterized via $y_{(\delta_1,\ldots,\delta_n)} = \prod_{i=1}^{n} a_i^{\delta_i}(1-a_i)^{1-\delta_i}$.

Clearly $S' \subset S \subset \Upsilon_0 \subset \Upsilon$. We call points in $\Upsilon_0 \setminus S$ *regular points*, and points in sets $S \setminus S'$ and S' *type* 1 and *type* 2 *singularities*, respectively. We now present our main result.

Theorem 8 (Asymptotic Marginal Likelihood Formula) *Let $\mathbb{I}[N, Y_D]$ (Eqs. 3.3 and 3.4) be the marginal likelihood of data with averaged sufficient statistics Y_D given the naive Bayesian model with binary variables and two hidden states with parameters $\omega = (a, b, t)$. Namely,*

$$\begin{aligned} \mathbb{I}[N, Y_D] &= \textstyle\int_{(0,1)^{2n+1}} e^{N\sum_x Y_x \ln \theta_x(\omega)} \mu(\omega) d\omega, \\ \theta_{(x_1,\ldots,x_n)} &= t\textstyle\prod_{i=1}^{n} a_i^{x_i}(1-a_i)^{1-x_i} + (1-t)\prod_{i=1}^{n} b_i^{x_i}(1-b_i)^{1-x_i}, \end{aligned} \tag{3.6}$$

where $x = (x_1, \ldots, x_n)$ denotes the binary vector of length n and the vectors Y_D and θ of length 2^n are indexed by x. Let Y_D and μ satisfy the following assumptions:

A1 Bounded density. *The density $\mu(\omega)$ is bounded and bounded away from zero on $\Omega = (0,1)^{2n+1}$.*

A2 Positive statistics. *The statistics $Y_D = (Y_1, \ldots, Y_{2^n})$ are such that $Y_i > 0$ for $i = 1, \ldots, 2^n$.*

A3 Statistics stability. *There exists a sample size N_0 such that the averaged sufficient statistics Y_D is equal to a limiting statistics Y for all sample sizes $N \geq N_0$.*

Then, for $n \geq 3$ as $N \to \infty$:

(a) If $Y \in \Upsilon_0 \setminus S$ (regular point)

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{2n+1}{2} \ln N + O(1), \tag{3.7}$$

(b) If $Y \in S \setminus S'$ (type 1 singularity)

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{2n-1}{2} \ln N + O(1), \tag{3.8}$$

(c) If $Y \in S'$ (type 2 singularity)

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{n+1}{2} \ln N + O(1), \tag{3.9}$$

where ω_{ML} are the maximum likelihood parameters for the averaged sufficient statistic Y. Moreover, for $n = 2$, $S = \Upsilon_0 = \Upsilon$ and

(d) If $Y \notin S'$ (namely, $Y \in S \setminus S'$),

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{3}{2} \ln N + O(1), \tag{3.10}$$

(e) If $Y \in S'$,

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{3}{2} \ln N + 2 \ln \ln N + O(1), \tag{3.11}$$

and for $n = 1$,

(f)

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{1}{2} \ln N + O(1), \tag{3.12}$$

as $N \to \infty$.

The first assumption that the prior density μ is bounded has been made by all earlier works; in some applications it holds and in some it does not. The proof and results, however, can be easily modified to apply to any particular kind of singularity of μ, as long as the form of singularity is specified. The second and third assumptions are made to ease the proof; the third assumption was also made by (Schwarz, 1978). Removing these assumptions is beyond the scope of this work.

Note that Eq. 3.8 corresponds to selecting $\lambda_1 = -\frac{2n-1}{2}$ and $m_1 = 1$ in Watanabe's method and Eq. 3.9 corresponds to selecting $\lambda_1 = -\frac{n+1}{2}$ and $m_1 = 1$. Both formulas are different from the standard BIC score, given by Eq. 3.7, which only applies to regular points, namely, the points in $\Upsilon_0 \setminus S$. In contrast to the standard BIC score, which is uniform for all points Y_D, the asymptotic approximation given by our *adjusted BIC score* depends on the value of $Y = Y_D$ through the coefficient of $\ln N$.

One might be tempted to think that the coefficient of the $-\ln N$ term can be guessed by various intuitive considerations. We now discuss three such erroneous attempts. First, the number of parameters of the model that generates a singular point Y_D is $n+1$ for case (c) because there are $n+1$ independent binary variables (the class variable and n feature variables). This may seem to explain the coefficient of $\ln N$ in case (c). However, using the same reasoning for case (b) yields the coefficient $(n+3)/2$ which differs from the correct coefficient. Another attempt is to claim that the coefficient of $-\ln N$ is half the number of parameters in the naive Bayesian model minus the number of redundant parameters in the model that generates Y_D. In particular, for case (b), the number of redundant parameters in the generative model is $(n+3)-(n+1) = 2$ and so the speculated coefficient should be $(2n+1-2)/2 = (2n-1)/2$ which is the correct coefficient. However, using the same reasoning for case (c) yields the coefficient $2n/2$ which is wrong. Finally, computing the maximum rank of the Jacobian of the map from the model parameters to the joint space parameters (defined by Eq. 22) at the maximum likelihood parameters w_{ML} for singular statistics Y_D yields the correct coefficient for case (b) but the wrong coefficient $(2n-1)/2$ for case (c).

The next theorem specifies the asymptotic behavior of marginal likelihood integrals for degenerate naive Bayesian models, namely, when some of the links are missing. This theorem complements Theorem 8 and its proof is explicated in Section 3.6.

Theorem 9 *Let M be the degenerate naive Bayesian model with two hidden states and n binary feature variables of which m are independent of the hidden state and let $\omega = (a_1, \ldots, a_{n-m}, b_1, \ldots, b_{n-m}, t, c_{n-m+1}, \ldots, c_n)$ be the $2n-m+1$ model parameters of M. Let $\mathbb{I}[N, Y_D]$ be the marginal likelihood of data D with averaged sufficient statistics Y_D given model M. Namely,*

$$\begin{aligned} &\mathbb{I}[N, Y_D] = \int_{(0,1)^{2n+1}} e^{N \sum_x Y_x \ln \theta_x(\omega)} \mu(\omega) d\omega, \\ &\theta_x = \left(t \prod_{i=1}^{n-m} a_i^{x_i}(1-a_i)^{1-x_i} + (1-t) \prod_{i=1}^{n-m} b_i^{x_i}(1-b_i)^{1-x_i}\right) \prod_{i=n-m+1}^{n} c_i^{x_i}(1-c_i)^{1-x_i}, \end{aligned} \tag{3.13}$$

where $x = (x_1, \ldots, x_n)$ denotes the binary vector of length n and the vectors Y_D and θ of length 2^n are indexed by x. Let Y_D and μ satisfy the following assumptions:

A1 Bounded density. *The density $\mu(\omega)$ is bounded and bounded away from zero on $\Omega = (0,1)^{2n+1}$.*

A2 Positive statistics. *The statistics $Y_D = (Y_1, \ldots, Y_{2^n})$ are such that $Y_i > 0$ for $i = 1, \ldots, 2^n$.*

A3 Statistics stability. *There exists a sample size N_0 such that the averaged sufficient statistics Y_D is equal to a limiting statistics Y for all sample sizes $N \geq N_0$.*

Assume also that $Y \in \Upsilon_0$ and that the parameterization of Y (as is Eq. 3.5) corresponds to a binary naive Bayesian model M', which shares k links with model M. Then, for $m \leq n-3$ as $N \to \infty$:

(a) If $k \geq 3$ (regular point)

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{2n+1-m}{2} \ln N + O(1),$$

(b) If $k = 2$ (type 1 singularity)

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{2n-1-m}{2} \ln N + O(1),$$

(c) If $k = 0$ or $k = 1$ (type 2 singularity)

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{n+1}{2} \ln N + O(1),$$

where ω_{ML} are the maximum likelihood parameters of statistics Y.
Furthermore, for $m = n-2$

(d) If $k = 2$ (type 1 singularity)

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{n+1}{2} \ln N + O(1).$$

Note that here $n+1 = 2n-m-1$, since $m = n-2$.

(e) If $k = 0$ or $k = 1$ (type 2 singularity)

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{n+1}{2} \ln N + 2 \ln \ln N + O(1),$$

and for $m = n-1$ or $m = n$,

$$\text{(f)} \qquad \ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{n}{2} \ln N + O(1),$$

regardless of k as $N \to \infty$.

An adversary may argue that evaluating the marginal likelihood on singular points is not needed because one could exclude from the model all singular points which only have measure zero. The remaining set would be a smooth manifold defining a curved exponential model, and so the standard BIC score would be a correct asymptotic expansion as long as the point Y_D has not been excluded. However, this proposed remedy is not perfect because in some situations the data may come from a model that yields singular statistics relative to the models being compared.

As an example of incorrect Bayesian model selection by the standard BIC score, consider the problem of selecting between two naive Bayesian models M_1 and M_2, as depicted on Figure 3.3. Suppose that the data is generated by the third model M_T. Both models M_1 and M_2 can not represent the target distribution (M_T) exactly, therefore, given a large enough sample, the choice of the model depends on the particular distribution represented by M_T and its parameters. Intuitively, if the dependencies of X_1 and X_2 on the hidden node C in model M_T are stronger than the dependency of X_4 on the hidden node, then one should prefer model M_1 over model M_2, and vice versa.

Now, if the maximum likelihoods of the data given model M_1 and given model M_2 happen to be equal, then the standard choice of the model is dictated by the penalty term of the BIC score (Eq. 3.2). The penalty term is smaller for M_1, which contains less parameters than M_2, and, consequently, the model preferred by the standard BIC score is M_1. However, the adjusted BIC approximation formula for the marginal likelihood for models with hidden variables penalizes model M_2 less than model M_1 (Theorem 9). Therefore, the marginal likelihood of the data given model M_2 is asymptotically larger than that of model M_1 and it should be chosen according to a Bayesian model selection procedure, given enough data.

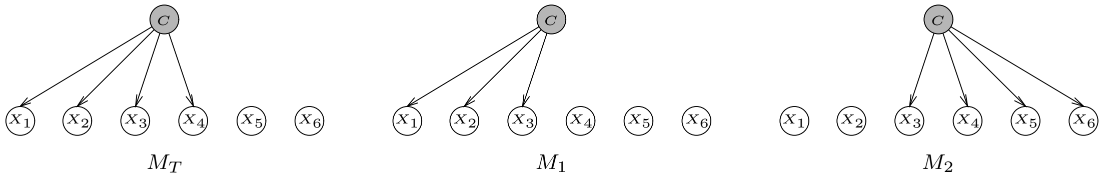


Figure 3.2: An example of incorrect Bayesian model selection by the standard BIC score. M_T represents the generating model, and M_1, M_2 represent models being compared. If the maximum likelihoods of data given M_1 and M_2 happen to be equal, which is possible when X_4 depends strongly on C in M_T, then the model selection procedure based on the standard BIC score will prefer model M_1, as it is less penalized compared to M_2. Using the adjusted BIC formula (Theorem 9), on the other hand, gives an advantage to M_2, reflecting its higher marginal likelihood.

3.4 Proof Outline of Theorem 8

The proof of Theorem 8 consists of two logical parts. The first part is the proof of claim (a) of Theorem 8 that follows from the fact that for regular statistics $Y \in \Upsilon_0 \setminus S$ there are only two (symmetric) maximum likelihood points at each of which the log-likelihood function is properly convex. Hence, the marginal likelihood integral can be approximated by the classical Laplace method (Lemma 2). The proof of Theorem 8a, which reflects standard practice, is provided in Section 3.5.2. The second logical part consists of the proofs of claims (b) and (c) of Theorem 8 and requires the advanced techniques of Watanabe (Section 2.4). First, the integral $\mathbb{I}[N, Y_D]$ is transformed by a series of transformations into a simpler one. Second, the sets of extremum points of the exponent (maximum log-likelihood points) are found, and then the new integral is computed in the neighborhoods of extremum points. Finally, the logarithm of the largest contribution gives the desired asymptotic approximation of the original integral. We first focus on one thread of our proof, which demonstrates this method, deferring the full proof to Section 3.5.

3.4.1 Useful Transformations

Decomposing the transformation T from the model parameters (a, b, t) to the joint space parameters θ_x, as defined by Eq. 3.6, facilitates the evaluation of the integral $\mathbb{I}[N, Y_D]$. We decompose T into a series of three transformations T_1, T_2, T_3 such that $T = T_3 \circ T_2 \circ T_1$. We call the model parameters (a, b, t) - *the source coordinates* and the parameters θ_x - *the target coordinates*. The transformations T_1 and T_3 are diffeomorphisms, namely, one-to-one differentiable mappings with differentiable inverses, that change the source and target coordinates, respectively, and are defined in such a way that the intermediate transformation T_2, which carries all the information about the singularities, is simple to analyze. These transformations are from (Geiger, Heckerman, King & Meek, 2001).

Denote the domain of the model parameters by $\Omega = [0,1]^{2n+1}$ and the domain of the joint space parameters by $\Theta = \bar{\Delta}_{2^n-1}$, where $\bar{\Delta}_{2^n-1} = \{(\alpha_1, \ldots, \alpha_{2^n-1}) | \alpha_i \geq 0, \sum \alpha_i \leq 1\}$ is the closed $2^n - 1$ dimensional unit simplex. Let $U = T_1(\Omega)$ be the image of T_1, $\Lambda = T_3^{-1}(\Theta)$ be the preimage of T_3, and $T_2 : U \to \Lambda$ be the transformation that relates these sets. These transformations are chained as follows:

$$\Omega_{(a,b,t)} \xleftrightarrow{\mathbf{T_1}} U_{(x,u,s)} \xrightarrow{\mathbf{T_2}} \Lambda_{(z)} \xleftrightarrow{\mathbf{T_3}} \Theta_{(\theta)}$$

where the indices denote the names of the coordinates used to describe the corresponding spaces. We now present these three transformations.

Transformation T_1: We define $T_1 : \Omega \to U$ via

$$s = 2t - 1, \quad u_i = \frac{a_i - b_i}{2}, \quad x_i = ta_i + (1-t)b_i, \quad i = 1, \ldots, n. \tag{3.14}$$

The mapping T_1 is a diffeomorphism with $|\det J_{T_1}| = 2^{-n+1}$. The inverse transformation is given by

$$t = (s+1)/2, \quad a_i = x_i + (1-s)u_i, \quad b_i = x_i - (1+s)u_i, \quad i = 1, \ldots, n. \tag{3.15}$$

Furthermore, it can be verified that U is the set of points $(x, u, s) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}$ such that

$$0 \leq x_i \leq 1, \quad -1 \leq s \leq 1, \quad -x_i \leq (1-s)u_i \leq 1 - x_i, \quad x_i - 1 \leq (1+s)u_i \leq x_i. \tag{3.16}$$

Transformation T_3: We define $T_3 : \Lambda \to \Theta$ as the inverse of a composition of two transformations T_{31} and T_{32}. First, consider the nonsingular transformation $T_{31} : \Theta \to \Lambda'$ defined by

$$\nu_{ij\ldots k} = \sum_{(x_1,\ldots,x_n),\ \text{s.t.}\ x_i = x_j = \ldots = x_k = 1} \theta_{(x_1,\ldots,x_n)}$$

where ν_i stands for the probability of the ith feature being true, ν_{ij} stands for the probability that the ith and jth features are both true, etc. Note that $\nu_\emptyset$, if defined, will be always 1. We now express $\nu_{ij\ldots k}$ using the model parameters (a, b, t) via

$$\nu_{ij\ldots k} = ta_i a_j \ldots a_k + (1-t)b_i b_j \ldots b_k. \tag{3.17}$$

Using Eq. 3.15, we rewrite Eq. 3.17 obtaining

$$\begin{aligned}
&\nu_i = x_i, \quad \nu_{ij} = x_i x_j + (1-s^2)u_i u_j, \\
&\nu_{ijk} = x_i x_j x_k + (1-s^2)(x_i u_j u_k + u_i x_j u_k + u_i u_j x_k) - 2s(1-s^2)u_i u_j u_k \\
&\nu_{12\ldots r} = x_1 x_2 \cdots x_r + \textstyle\sum_{i=2}^r p_i(s) \left(\sum \text{“products of } i\ u'\text{s and } r-i\ x'\text{s''}\right)
\end{aligned} \tag{3.18}$$

where $p_i(s) = 1/2\left[(1-s)^i(1+s) + (-1-s)^i(1-s)\right]$, and, in particular, $p_2(s) = 1-s^2$ and $p_3(s) = -2s(1-s^2)$.

Now we subtract products of the first n coordinates to remove the leading terms. So, we do $z_{ij} = \nu_{ij} - \nu_i\nu_j$. Then we subtract products of the first n coordinates with one of the new coordinates to remove the second terms, namely, $z_{ijk} = \nu_{ijk} - \nu_i\nu_j\nu_k - z_{ij}\nu_k - z_{ik}\nu_j - z_{jk}\nu_i$, and so forth. We end up with the transformation $T_{32} : \Lambda' \to \Lambda$ defined by

$$z_i = \nu_i, \quad z_{ij} = \nu_{ij} - \nu_i\nu_j, \quad z_{ijk} = \nu_{ijk} - \nu_i\nu_j\nu_k - z_{ij}\nu_k - z_{ik}\nu_j - z_{jk}\nu_i, \quad \text{etc.} \tag{3.19}$$

where the indices of the z coordinates are non-empty subsets of $\{1,\ldots,n\}$. In particular, the z coordinate corresponding to a set $I \subseteq \{1,\ldots,n\}$ is z_I, the z coordinate corresponding to $\{i\}$ is z_i, and the z coordinate corresponding to $\{i,j,k\} \subseteq \{1,\ldots,n\}$ is z_{ijk}, etc.

The transformations T_{31} and T_{32} are diffeomorphisms with Jacobian determinant 1. The transformation T_3 is defined by $T_3 = T_{31}^{-1} \circ T_{32}^{-1} : \Lambda \to \Theta$. Hence, T_3 is a diffeomorphism with Jacobian determinant equal to 1.

Transformation T_2: We define $T_2 : U \subset \mathbb{R}^{2n+1} \to \Lambda \subset \mathbb{R}^{2^n-1}$ via

$$z_i = x_i, \quad z_{ij} = p_2(s)u_i u_j, \quad \ldots, \quad z_{12\ldots r} = p_r(s)u_1 u_2 \ldots u_r \tag{3.20}$$

obtained by combining Eqs. 3.18 and 3.19. We use the notation $z_I(x,u,s)$ when the dependence of z_I on (x,u,s) needs to be explicated. Note that this transformation is not a diffeomorphism for $n > 3$.

Transformations T_1, T_2 and T_3 are similar to transformations used by (Settimi & Smith, 2000) in the study of the geometry of parametric spaces for Bayesian networks with hidden variables. These transformations can be regarded as reparameterizations of the naive Bayesian models in terms of moments. In particular, if the hidden and observable nodes are assumed to have states -1 and 1, then $s = \mathbb{E}[C]$, $u_i = Cov(X_i, C)/Var(C)$, $p_i(s) = \mathbb{E}[(C-s)^i]$ and $z_{12\ldots r} = \mathbb{E}[\prod_{i=1}^r (X_i - \mathbb{E}[X_i])]$.

3.4.2 Preliminary Lemmas

Based on the transformations T_1, T_2 and T_3, we present two lemmas that facilitate the evaluation of the integral $\mathbb{I}[N, Y_D]$. The first lemma states that under Assumptions $A1$ and $A3$, the integral $\mathbb{I}[N, Y_D]$ can be asymptotically evaluated in the (x, u, s) coordinates for a limiting statistics Y, while dismissing the contribution of the density function μ. The second lemma shows that the resulting integral $\tilde{\mathbb{I}}[N, Y]$ can be evaluated using the quadratic form in the z coordinates.

Lemma 10 *Let $\mathbb{I}[N, Y_D]$ be defined by Eq. 3.6, namely,*

$$\mathbb{I}[N, Y_D] = \int_{(0,1)^{2n+1}} e^{N \sum_x Y_x \ln \theta_x(\omega)} \mu(\omega) d\omega$$

and assume μ is bounded ($A1$) and Y_D is stable ($A3$). Let

$$\tilde{\mathbb{I}}[N, Y] = \int_U e^{-Nf(\theta[x,u,s])} dx\, du\, ds \tag{3.21}$$

where

$$\begin{aligned} &f(x, u, s) = f_Y - \textstyle\sum_{i=1}^{2^n} Y_i \ln \theta_i[x, u, s], \\ &\theta[x, u, s] = (T_3 \circ T_2)[x, u, s], \quad \theta_{2^n}[x, u, s] = 1 - \textstyle\sum_{i=1}^{2^n-1} \theta_i[x, u, s], \end{aligned} \tag{3.22}$$

and where $f_Y = \max_{(x,u,s) \in U} \sum_{i=1}^{2^n} Y_i \ln \theta_i[x, u, s]$ and Y is the limiting statistics of Y_D as specified by Assumption A3, namely, $Y_D = Y$ for $N \geq N_0$.

Then, $f_Y = P(Y|\omega_{ML})$ and

$$\ln \mathbb{I}[N, Y_D] = N f_Y + \ln \tilde{\mathbb{I}}[N, Y] + O(1) \tag{3.23}$$

for all $N > 1$.

Proof: Since T_1 is a diffeomorphism, $f_Y = P(Y|\omega_{ML})$ and the integral $\mathbb{I}[N, Y_D]$ can be evaluated in (x, u, s) coordinates by introducing the constant factor of Jacobian determinant of transformation T_1, $J_{T_1} = 2^{-n+1}$. Moreover $\mu(\omega)$ is bounded and thus the integral evaluated with $\mu(\omega) \equiv 1$ is within a constant factor of $\mathbb{I}[N, Y_D]$ and since Y_D is equal to Y starting from N_0, fixing Y_D to Y introduces finite number of approximation errors for $N < N_0$ that can be bounded. Thus, $\tilde{\mathbb{I}}[N, Y]$ is within a constant factor of the integral $\mathbb{I}[N, Y_D]$ multiplied by e^{Nf_Y} with the constants independent on N and Y_D. Eq. 3.23 expresses this fact in a logarithmic scale. ■

Note that the assumption of boundedness ($A1$) can not be replaced by the assumption of *local* boundedness, because the density function μ may explode on the boundary of Ω and these areas can not be excluded from the integration because in the singular cases the maximum likelihood parameters are located on the boundary of the parameter space (e.g., cases $C3$-$C5$ in Section 3.4.3).

Lemma 11 *Consider $\tilde{\mathbb{I}}[N, Y]$ and $f(x, u, s)$ as defined in Lemma 10 (Eqs. 3.21 and 3.22). Let the zero set $U_0 = \arg\min_{(x,u,s)\in U} f(x, u, s)$ be the set of minimum points of $f(x, u, s)$ in U. Let*

$$\mathbb{J}[N, Y] = \max_{p_0 \in U_0} \mathbb{J}_{p_0}[N, Y] \quad \text{and} \quad \mathbb{J}_{p_0}[N] = \int_{U_\epsilon \cap U} e^{-N \sum_I (z_I(x,u,s) - z'_I)^2} dx\, du\, ds, \tag{3.24}$$

where $z_I(x, u, s)$ is the I-th coordinate of $z = T_2[x, u, s]$, z'_I is the I-th coordinate of $T_2[x', u', s']$ and U_ϵ is an ϵ-box neighborhood of $p_0 = (x', u', s') \in U_0$. (Note that $\mathbb{J}_{p_0}[N]$ does not depend on Y, while $\mathbb{J}[N, Y]$ depends on Y through the form of set U_0.)

If Y is positive ($A2$) and $Y \in \Upsilon_0$, then

$$\ln \tilde{\mathbb{I}}[N, Y] = \ln \mathbb{J}[N, Y] + O(1) \quad \text{for all } N > 1. \tag{3.25}$$

The proof of this lemma uses the facts that T_3 is a diffeomorphism, U is compact, the contributions of non-maximum regions of $-f$ are exponentially small, and the 2^n dimensional point $Y > 0$ corresponds to a maximum likelihood parameters of naive Bayesian network with binary variables and two hidden states. The proof is explicated in Section 3.5.1.

Lemmas 10 and 11 jointly state that the asymptotic forms of $\ln \mathbb{J}[N, Y]$ and $\ln \mathbb{I}[N, Y_D]$ are identical up to an additive term $N f_Y$ and a constant provided that Y is the limiting statistics of Y_D (Assumption $A3$).

3.4.3 Analysis of Type 2 Singularity.

We now focus on the proof of Theorem 8c that deals with the singular points in S'. Let $Y \in S'$. Our starting point in proving Theorem 8c is integral $\mathbb{J}[N, Y]$ (Eq. 3.24), which by Lemmas 10 and 11 specifies the asymptotic form of $\mathbb{I}[N, Y_D]$. We evaluate the contributions $\mathbb{J}_{p_0}[N]$ to $\mathbb{J}[N, Y]$ from the neighborhoods of extremum points $p_0 = (x', u', s') \in U_0$. The largest contribution determines the asymptotic form of integral $\mathbb{I}[N, Y_D]$ as $N \to \infty$ and $Y_D = Y$.

Let $\gamma = (\gamma_1, \ldots, \gamma_n)$ be the model parameters of the n independent variables that define the 2^n dimensional point $Y \in S'$, namely

$$\gamma_j = \sum_{\delta \in \{0,1\}^n, \text{s.t. } \delta_j = 1} Y_{(\delta_1, \ldots, \delta_n)}, \qquad j = 1, \ldots, n. \tag{3.26}$$

Furthermore, $Y \in S'$ if and only if for all $\delta \in \{0,1\}^n$, equality $Y_{(\delta_1,\ldots,\delta_n)} = \prod_{i=1}^n \gamma_i^{\delta_i}(1-\gamma_i)^{1-\delta_i}$ holds for $\gamma = \{\gamma_1, \ldots, \gamma_n\}$ given by Eq. 3.26.

Let $\bar{V}$ denote the closure of a set V. The zero set U_0 can be written as the union of $n+2$ sets

$$U_0 = \bar{U}_{0-} \cup \bar{U}_{0+} \cup \bigcup_{j=1}^{n} \bar{U}_{0j}, \tag{3.27}$$

where

$$\begin{aligned}
&U_{0-} = \left\{(x=\gamma, u, s=-1) \mid u_i \in \left(\tfrac{-\gamma_i}{2}, \tfrac{1-\gamma_i}{2}\right),\ i=1,\ldots,n\right\}, && W_{0-} = \{(a, b=\gamma, t=0) \mid a_i \in (0,1)\},\\
&U_{0+} = \left\{(x=\gamma, u, s=1) \mid u_i \in \left(\tfrac{\gamma_i-1}{2}, \tfrac{\gamma_i}{2}\right),\ i=1,\ldots,n\right\}, && W_{0+} = \{(a=\gamma, b, t=1) \mid b_i \in (0,1)\},\\
&U_{0j} = \left\{ (x=\gamma, u, s) \mid \begin{array}{l} u_i = 0, \forall i \neq j; \\ u_j \in \left(-\frac{1}{2}, \frac{1}{2}\right); s \in (-1,1); \\ -\gamma_j < (1-s)u_j < 1-\gamma_j, \\ \gamma_j - 1 < (1+s)u_j < \gamma_j \end{array} \right\}, && W_{0j} = \left\{ (a,b,t) \mid \begin{array}{l} a_i = b_i = \gamma_i, \forall i \neq j; \\ ta_j + (1-t)b_j = \gamma_j \end{array} \right\},
\end{aligned} \tag{3.28}$$

and where $W_{0-} = T_1^{-1}(U_{0-})$, $W_{0+} = T_1^{-1}(U_{0+})$, and $W_{0j} = T_1^{-1}(U_{0j})$ are the same sets expressed using the model parameters (a, b, t).

The zero set U_0, namely the minimum points of f, is divided into five disjoint sets:

C1: $(x', u', s') \in U_{0j} \setminus \bigcup_{i \neq j} U_{0i}$.

C2: $(x', u', s') \in \bigcap_j U_{0j}$.

C3: $(x', u', s') \in U_{0-} \cup U_{0+} \setminus \bigcup_j \bar{U}_{0j}$.

C4: $(x', u', s') \in \bigcup_j \left[U_{0-} \cup U_{0+} \cap \bar{U}_{0j} \setminus \bigcup_{i \neq j} \bar{U}_{0i}\right]$.

C5: $(x', u', s') \in (U_{0-} \cup U_{0+}) \bigcap_j \bar{U}_{0j}$.

These five disjoint sets and their boundaries cover U_0, because $U_{0+} \cap U_{0-} = \emptyset$ and $U_{0i} \cap U_{0j} = \bigcap_k U_{0k}$. The set U_0 is shown in Figure 3.3 along with a representative point from $C1$ through $C5$. Note that the border point version of $C2$ point is $C5$ point.

Note that U_0 is a union of two n-dimensional planes U_{0-}, U_{0+} and n two-dimensional planes U_{0j}, $j = 1, \ldots, n$. Consequently, one could perhaps guess from the classical Laplace approximation analysis that because the zero subsets U_{0-}, U_{0+} have dimension n, the coefficient of the $\ln N$ term would be at least $-(2n+1-n)/2 = -(n+1)/2$. Indeed this happens, but a formal proof requires to closely examine the form of f near the different minimum points. This evaluation is complicated by

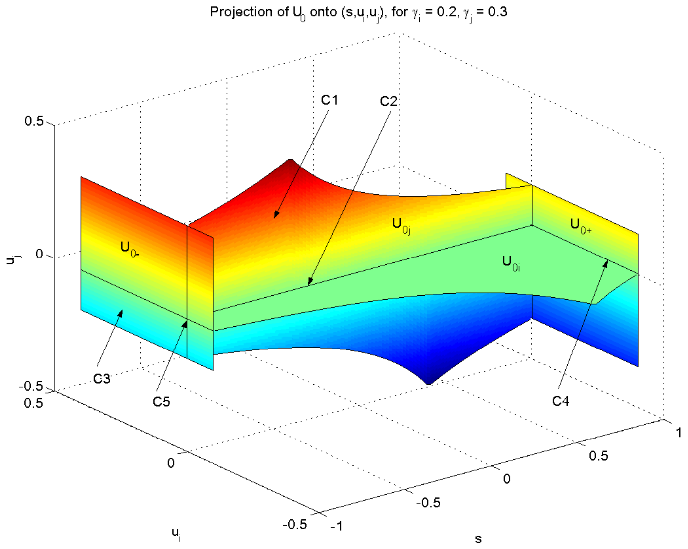


Figure 3.3: The set U_0 projected on (s, u_i, u_j), for $x_i = \gamma_i = 0.2$, $x_j = \gamma_j = 0.3$. Examples of points of types $C1$-$C5$ are marked.

the fact that the zero planes intersect (see Figure 3.3), and such cases ($C2$, $C4$, $C5$) are not covered by the classical Laplace approximation analysis.

The proof proceeds case by case by evaluating the integrals $\mathbb{J}_{p_0}[N]$ (Eq. 3.24) around points $p_0 = (x', u', s')$ from the sets $C1$ through $C5$. Then, the maximal asymptotic value of $\mathbb{J}_{p_0}[N]$ is the approximation of $\mathbb{J}[N, Y]$, as specified by Lemma 11. We now treat case $C2$ which demonstrates the main ideas, deferring the other cases to Section 3.5.

According to case $C2$, $(x', u', s') = \bigcap_j U_{0j}$. Each point of case $C2$ satisfies $u_i' = 0$ and $x_i' = \gamma_i$ for $i = 1, \ldots, n$ and $s' \neq \pm 1$. Furthermore, its z coordinates satisfy $z_i' = x_i'$ for all $i = 1, \ldots, n$ and $z_I' = 0$ for all other indices. Let $\phi(x, u, s) = \sum_I \left[z_I(x' + x, u' + u, s' + s) - z_I'\right]^2$. Note that $\phi(x, u, s)$ is term in the exponent of the integrand of $\mathbb{J}[N, Y]$ centered around the minimum point (x', u', s'). Using transformation T_2 (Eq. 3.20), we obtain

$$\begin{aligned}
\phi(x,u,s) = \quad & \textstyle\sum_I \left[z_I(x'+x, u'+u, s'+s) - z_I'\right]^2 \\
= \quad & \textstyle\sum_i [z_i - z_i']^2 + \sum_{ij, i\neq j}[z_{ij} - z_{ij}']^2 + \sum_{ijk, i\neq j\neq k}[z_{ijk} - z_{ijk}']^2 + \ldots \\
= \quad & \textstyle\sum_i \left[(x_i' + x_i) - x_i'\right]^2 + \sum_{ij, i\neq j}\left[(1-(s'+s)^2)u_iu_j - 0\right]^2 + \text{“higher order terms”} \\
= \quad & \textstyle\sum_i x_i^2 + \sum_{ij, i\neq j}\left[(1-s'^2)u_iu_j - (s+2s')su_iu_j\right]^2 + \text{“higher order terms”}
\end{aligned} \tag{3.29}$$

The higher order terms are multiplication of three, four and more u_i's and their contribution is bounded by the terms explicitly written in Eq. 3.29. For example, third terms are of form $(z_{ijk} - z'_{ijk})^2 = 4(s'+s)^2(1-(s'+s)^2)^2 u_i^2 u_j^2 u_k^2 \leq 5\epsilon^2 u_i^2 u_j^2$ for all $s, u_i, u_j, u_k < \epsilon$ for ϵ small enough. Similar bounds can be obtained for all high order terms in Eq. 3.29. Thus, the principal part of ϕ, that bounds ϕ within the multiplicative constant near zero, is given by

$$\tilde{\phi}(x,u,s) = \sum_i x_i^2 + \sum_{ij, i \neq j} u_i^2 u_j^2. \tag{3.30}$$

and $\tilde{\phi}(x,u,s) \leq \phi(x,u,s) \leq 2\tilde{\phi}(x,u,s)$ for all $s, u_i, u_j < \epsilon$ for ϵ small enough.

Since the multiplicative constants in the exponent can be transferred to the multiplicative constants of integral itself by changing the integration range around zero and rescaling, we only need to evaluate the asymptotic form of integral $\int e^{-N(\sum_i x_i^2 + \sum_{ij,i\neq j} u_i^2 u_j^2)} dxduds$ in order to get the asymptotic form of integral $\mathbb{J}[N, Y \in S']$ (Eq. 3.24) within a constant multiply.

The quadratic form in x_i's contributes an $N^{-n/2}$ factor to the integral $\tilde{\mathbb{J}}[N]$. This can be shown by decomposing the integral and integrating out the x_i's. We are left with the evaluation of the integral

$$\tilde{\mathbb{J}}[N] = \int_{(-\epsilon,+\epsilon)^n} e^{-N\sum_{ij,i\neq j} u_i^2 u_j^2} du.$$

For $n = 3$, this is precisely the integral evaluated as example in Section 2.4 which was found to be asymptotically equal to $cN^{-\frac{3}{4}}$. Generalizing the approach demonstrated in the example in Section 2.4 to $n \geq 3$ we obtain that the largest pole of $J(\lambda)$ is $\lambda_1 = -n/4$ with multiplicity $m = 1$, so $\tilde{\mathbb{J}}[N]$ is asymptotically equal to $cN^{-\frac{n}{4}}$. Thus the contribution of the neighborhood of $(x', u', s') \in \bigcap_j U_{0j}$ to $\mathbb{J}[N, Y \in S']$ is $cN^{-\frac{3n}{4}}$. This contribution does not correspond to the classical BIC formula because it implies that the "effective" dimension is a fraction $\frac{3}{2}n$ rather than $2n - 1$.

In summary, we have analyzed case $C2$, showing that the contribution to $\mathbb{J}[N, Y \in S']$ is $cN^{-\frac{3n}{4}}$. The dominating contributions in the cases $C3$, $C4$, and $C5$, are all equal to $cN^{-\frac{n+1}{2}}$ (the proof of this claim is given in Section 3.5). The dominating contribution in case $C1$ is only $cN^{-\frac{2n-1}{2}}$. Also, the various border points of U_0 do not contribute more than the corresponding internal points. Thus, $\mathbb{J}[N, Y] = cN^{-\frac{n+1}{2}}$ for $Y' \in S$. Consequently, due to Lemmas 10 and 11, $\ln \mathbb{I}[N, Y_D] = N \cdot P(Y|\omega_{ML}) - \frac{n+1}{2} \ln N + O(1)$, as claimed by Theorem 8c. ■

3.5 Full Proof of Theorem 8 (The Main Theorem)

We start with the proof of Lemma 11, which requires two additional lemmas. Then we proceed with a case by case proof of Theorem 8.

3.5.1 Proof of Lemma 11

The proof of Lemma 11 uses Lemmas 12 and 13. In particular, Lemma 12 states that a local version of the claim made by Lemma 11 (Eq. 3.25) holds in the neighborhood of extremum points p_0 under two additional assumptions denoted by $B1$ and $B2$. Lemma 13 shows that $B1$ and $B2$ hold. Finally, the proof of Lemma 11 elevates the local version to the global claim.

Lemma 12 *Let*

$$\begin{aligned} &f(x,u,s) = f_Y - \textstyle\sum_{i=1}^{2^n} Y_i \ln \theta_i[x,u,s], \\ &\theta[x,u,s] = (T_3 \circ T_2)[x,u,s], \quad \theta_{2^n}[x,u,s] = 1 - \textstyle\sum_{i=1}^{2^n-1} \theta_i[x,u,s], \end{aligned} \tag{3.31}$$

where $f_Y = \max_{(x,u,s)\in U} \sum_{i=1}^{2^n} Y_i \ln \theta_i[x,u,s]$ *and* $Y = (Y_1,\ldots,Y_{2^n})$ *is a non-negative vector with sum of elements equal to* 1*. Let the zero set* $U_0 = \arg\min_{(x,u,s)\in U} f(x,u,s)$ *be the set of minimum points of* $f(x,u,s)$ *on* U*, let* $p_0 = (x',u',s')$ *be a point in* U_0 *and let*

$$\tilde{\mathbb{I}}_{p_0}[N,Y] = \int_{U_\epsilon} e^{-Nf(x,u,s)} dx\,du\,ds, \tag{3.32}$$

where U_ϵ *is some small neighborhood of* p_0*. Also, let*

$$\mathbb{J}_{p_0}[N] = \int_{U_\epsilon} e^{-N\sum_I (z_I(x,u,s)-z_I')^2} dx\,du\,ds,$$

where $z_I(x,u,s)$ *is the* I*-th coordinate of* $z(x,u,s) = T_2[x,u,s]$ *and* z_I' *is the* I*-th coordinate of* $T_2[x',u',s']$*.*

Further assume that (x',u',s') *satisfies*

B1. $\theta' = T_3 \circ T_2(x',u',s')$ *is a minimum of* f *as function of* θ*,* $f(\theta') = 0$ *and* $\nabla_\theta f(\theta') = 0$*.*

B2. f*, as a function of* θ*, is strictly convex at* $\theta' = \theta(x',u',s')$*, i.e., the matrix* $\mathcal{H}_\theta f(\theta')$ *is positive definite.*

Then,

$$\ln \tilde{\mathbb{I}}_{p_0}[N,Y] = \ln \mathbb{J}_{p_0}[N] + O(1) \quad \text{for all } N > 1. \tag{3.33}$$

(The right hand side of Eq. 3.33 depends on Y *through the* $O(1)$ *term.)*

Proof: Since $\nabla_\theta f(\theta') = 0$, $\mathcal{H}_\theta f(\theta')$ is positive definite and $T_3 : \Lambda_{(z)} \to \Theta_{(\theta)}$ is a diffeomorphism, it follows that $\nabla_z f(z') = 0$ and $\mathcal{H}_z f(z')$ is positive definite. Also, $f(z') = 0$. Therefore, f as a function of z can be approximated by a quadratic form near $z' = T_2(x', u', s')$ via

$$\eta_1 \sum_I (z_I - z'_I)^2 \; < \; f(z) \; < \; \eta_2 \sum_I (z_I - z'_I)^2, \quad \text{for } z \in \Lambda_\epsilon, \tag{3.34}$$

where Λ_ϵ is some sufficiently small neighborhood of z', and $\eta_1, \eta_2 > 0$ are slightly smaller and larger, respectively, than all eigenvalues of $\mathcal{H}_z f(z')$. Consequently, since $T_2 : U \to \Lambda$ is continuous, there exists neighborhood U_ϵ of p_0 such that $T_2(U_\epsilon) \subseteq \Lambda_\epsilon$ and Inequality 3.34 holds for $z(x, u, s) = T_2(x, u, s)$ for all points (x, u, s) in U_ϵ. Using Inequality 3.34 for evaluating $\tilde{\mathbb{I}}_{p_0}[N, Y]$ (Eq. 3.32) yields

$$\int_{U_\epsilon} e^{-\eta_2 N \sum_I (z_I(x,u,s) - z'_I)^2} dx\, du\, ds \; < \; \tilde{\mathbb{I}}_{p_0}[N, Y] \; < \; \int_{U_\epsilon} e^{-\eta_1 N \sum_I (z_I(x,u,s) - z'_I)^2} dx\, du\, ds.$$

Due to Theorem 6, the bounding integrals are asymptotically equivalent up to a multiplicative constant, because the poles and multiplicities of the corresponding $J(\lambda)$ functions (Eq. 2.55) that determine their asymptotic behavior are the same for any constant multiples of $\sum (z_I(x, u, s) - z'_I)^2$, and in particular, for the multipliers η_1, η_2 and 1. ∎

Lemma 13 *Let $f(u, x, s)$ be as defined by Eq. 3.31, namely,*

$$f(x, u, s) = f_Y - \textstyle\sum_{i=1}^{2^n} Y_i \ln \theta_i[x, u, s],$$
$$\theta[x, u, s] = (T_3 \circ T_2)[x, u, s], \quad \theta_{2^n}[x, u, s] = 1 - \textstyle\sum_{i=1}^{2^n - 1} \theta_i[x, u, s],$$

where $f_Y = \max_{(x,u,s) \in U} \sum_{i=1}^{2^n} Y_i \ln \theta_i[x, u, s]$ and $Y = (Y_1, \ldots, Y_{2^n})$ is a vector in Υ_0 (defined by Eq. 3.5) such that $Y_i > 0$ (A2). Let the zero set $U_0 = \arg\min_{(x,u,s) \in U} f(x, u, s)$ be the set of minimum points of f, and let (x', u', s') be a point in U_0. Then $f(x', u', s') = 0$, and

B1. $\theta' = T_3 \circ T_2(x', u', s')$ is a minimum point of f as function of θ on Θ, $f(\theta') = 0$ and $\nabla_\theta f(\theta') = 0$. Furthermore, $\theta' = (Y_1, \ldots, Y_{2^n - 1})$ and $\nabla f(x', u', s') = 0$.

B2. f as a function of θ is strictly convex at θ', i.e., $\mathcal{H}_\theta f(\theta')$ is positive definite.

B3. If $n \geq 3$ and $Y \in \Upsilon_0 \setminus S$, then $f(x, u, s)$ is strictly convex at (x', u', s'), i.e., the matrix $\mathcal{H}_{(x,u,s)} f(x', u', s')$ is positive definite.

B4. Also, if $n \geq 3$ and $Y \in \Upsilon_0 \setminus S$, then U_0 consists only of two distinct points (x', u', s') and (x'', u'', s''), such that $x' = x''$, $u' = -u''$ and $s' = -s''$.

Proof: The claim $f(x', u', s') = f(\theta') = 0$ follows directly from the definitions of f, θ' and f_Y.

Consider Claim $B1$. The point $\theta_0 = (Y_1, \ldots, Y_{2^n-1})$ is the unique minimum of f, as a function of θ, on Θ, because $f_Y - f(\theta) = \sum_i Y_i \ln \theta_i[x, u, s]$ is the logarithm of a multinomial distribution. Since $Y \in \Upsilon_0$, the distribution specified by θ_0 can be represented by the model parameters, namely, $\theta_0 \in (T_3 \circ T_2)[U_0]$. Consequently, $\theta_0 = (T_3 \circ T_2)[U_0]$ because θ_0 is the unique minimum of f. So, $\theta' = \theta_0 = (Y_1, \ldots, Y_{2^n-1})$. Furthermore, because $Y > 0$, θ' is an internal point of Θ yielding $\nabla_\theta f(\theta') = 0$. Finally $\nabla f(x_0, u_0, s_0) = J^T_{(T_3 \circ T_2)}(x_0, u_0, s_0) \nabla_\theta f(\theta') = 0$ as well.

Claim $B2$ is established by explicit calculations. The Hessian $\mathcal{H}_\theta f(\theta')$ at $\theta' = (Y_1, \ldots, Y_{2^n-1})$ is given by

$$[\mathcal{H}_\theta f(\theta')]_{ij} = \begin{cases} \frac{1}{Y_{2^n}} & \text{for } i \neq j \\ \frac{1}{Y_i} + \frac{1}{Y_{2^n}} & \text{for } i = j \end{cases}$$

Consequently, for any $a \in \mathbb{R}^{2^n-1}$, $a \neq 0$, it follows that

$$a^T \cdot \mathcal{H}_\theta f(\theta') \cdot a = \sum_{i=1}^{2^n-1} \frac{a_i^2}{Y_i} + \frac{1}{Y_{2^n}} \left[\sum_{i=1}^{2^n-1} a_i \right]^2 > 0.$$

Note that Claim $B2$ holds even if $Y_i = 0$ for a single index i.

Claim $B3$ follows from the proof of Theorem 12 of (Geiger, Heckerman, King & Meek, 2001), which shows that the Jacobian of the transformation T_2 is of maximal rank for $n \geq 3$ for points (x', u', s') that satisfy $\theta' = T_2[x', u', s'] \in \Upsilon_0 \setminus S$. The mentioned theorem and claim $B2$ imply that for all $a \in \mathbb{R}^{2n+1}$, $a \neq 0$,

$$\begin{aligned} a^T \cdot \mathcal{H}_{(x,u,s)} f(x_0, u_0, s_0) \cdot a &= a^T \cdot \left[J^T_{(T_3 \circ T_2)}(x_0, u_0, s_0) \cdot \mathcal{H}_\theta f(\theta') \cdot J_{(T_3 \circ T_2)}(x_0, u_0, s_0) \right] \cdot a \\ &= \left[J_{(T_3 \circ T_2)}(x_0, u_0, s_0) \cdot a \right]^T \cdot \mathcal{H}_\theta f(\theta') \cdot \left[J_{(T_3 \circ T_2)}(x_0, u_0, s_0) \cdot a \right] = b^T \cdot \mathcal{H}_\theta f(\theta') \cdot b > 0, \end{aligned}$$

where $b = J_{(T_3 \circ T_2)}(x_0, u_0, s_0) \cdot a$. This proves Claim $B3$ because $\mathcal{H}_\theta f(\theta')$ is positive definite and $b \neq 0$ lest $J_{(T_3 \circ T_2)}$ would not be of maximal rank.

Claim $B4$ follows from claim $B1$ that $\theta' = (Y_1, \ldots, Y_{2^n-1})$ and from Theorem 13 in (Geiger, Heckerman, King & Meek, 2001), which states that for $\theta' \in \Upsilon_0 \setminus S$, there are exactly two source points (x, u, s), precisely the ones specified by Claim $B4$, that satisfy $\theta' = T_2[x, u, s]$. ■

Proof of Lemma 11: Lemma 12 combined with Lemma 13 establish the asymptotic behavior of $\tilde{\mathbb{I}}[N, Y]$ in the ϵ_{p_0} neighborhood of a single minimum p_0 (Eq. 3.33). Now, since U is closed and bounded (Eq. 3.16), it is *compact*. Hence, from an arbitrary infinite set of ϵ-neighborhoods of points in U, there exist a finite subset of disjoint neighborhoods of points in U that cover U. The neighborhoods that do not contain minimum points can be discarded since their contribution to the integral is exponentially small, i.e., a contribution bounded by e^{-Nc_1} versus e^{-Nc_2} where $c_1 > c_2$. Moreover, some neighborhoods may lie partially outside U. Let $U_0' \subseteq U_0$ denote the finite set of points from U_0, the neighborhoods of which are chosen to cover U_0. Also, let $\mathbb{J}[N,Y]$ denote the maximal contribution to $\tilde{\mathbb{I}}[N,Y]$, as in Lemma 11 (Eq. 3.24). We obtain

$$\mathbb{J}[N,Y] \ \leq \ \tilde{\mathbb{I}}[N,Y] \ \leq \ \sum_{p_0 \in U_0'} \mathbb{J}_{p_0}[N] \leq k \cdot \mathbb{J}[N,Y], \tag{3.35}$$

where k is the number of points in U_0'. Taking the logarithm of Eq. 3.35 yields Eq. 3.25 which establishes Lemma 11. ■

Claims $B3$ and $B4$ of Lemma 13 have not been used in the proof of Lemma 11. These claims are needed in the next section.

3.5.2 Proof of Theorem 8a (Regular Statistics Case)

Theorem 8a rephrases standard facts regarding asymptotic expansion of integrals around a single extremum point. Recall that Theorem 8a states that if $Y_D = Y$ for $N \geq N_0$, $Y_i > 0$ for $i = 1, \ldots, 2^n$ and $Y \in \Upsilon_0 \setminus S$, then asymptotic approximation of $\ln \mathbb{I}[N, Y_D]$ (Eq. 3.6) equals $N \ln P(Y|w_{ML}) - \frac{2n+1}{2} \ln N + O(1)$ (Eq. 3.7). To prove this claim we use Lemma 10 which states that $\mathbb{I}[N,Y_D]$ and $\tilde{\mathbb{I}}[N,Y]$ have the same asymptotic approximation up to a multiplicative constant e^{Nf_Y} and compute $\tilde{\mathbb{I}}[N,Y]$ using Lemma 2 (Laplace approximation).

We start by noticing that $\mathbb{I}[N,Y_D]$ absolutely converges for any $N \geq 1$ and $Y_D \geq 0$. That is because the integrand function $e^{N\sum_x Y_x \ln \theta_x(w)} = \prod_x \theta_x(w)^{NY_x}$ satisfies $0 \leq \theta_x(w)^{NY_x} \leq 1$ for all N, Y_D, i and $w = (a,b,t) \in \Omega$ and because $\mu(a,b,t)$ is a probability density function on Ω, thus integral $\mathbb{I}[N,Y_D]$ is finite (and less than 1). Consequently, $\tilde{\mathbb{I}}[N,Y]$ also absolutely converges for any $N \geq 1$ and any $Y \geq 0$, as required in order to use Lemma 2.

Consider now the integral $\tilde{\mathbb{I}}[N,Y] = \int_U e^{-Nf(x,u,s)} dx du ds$. Since the value of $e^{-Nf(x,u,s)}$ outside the small neighborhoods of the minimums f is exponentially small, so the asymptotic behavior of $\tilde{\mathbb{I}}[N,Y]$ on U is actually described by integration of $\tilde{\mathbb{I}}[N,Y]$ in the small neighborhoods of minimums of f (Lemma 11). Since $\tilde{\mathbb{I}}[N,Y]$ converges and Claims $B1$, $B3$ and $B4$ of Lemma 13 hold, it follows that in sufficiently small neighborhoods of the two internal minimum points of f, the integral $\tilde{\mathbb{I}}[N,Y]$ can be computed by Lemma 2 (Laplace Approximation).

Consequently, integrating $\tilde{\mathbb{I}}[N,Y]$ in the full neighborhoods of the maximum likelihood points $(x',u',s') \in U_0$, that lie on the border of U, introduces only a constant multiplicative errors to the approximation. This is shown by considering the integral $\tilde{\mathbb{I}}[N,Y]$ around minimum points of f in the equivalent (since T_1 is a diffeomorphism) coordinates (a,b,t), which have the full integration domain $\Omega = (0,1)^{2n+1}$. In these coordinates, approximating f by a quadratic form (as performed by Laplace approximation) on (a,b,t) and integrating in a full neighborhood of border point results in multiplicative error factor of 2^k where k is the number of border coordinates.

We now apply Lemma 2 to the small neighborhoods of the two minimum points of f and by combining Eq. 3.23 with the logarithm of sum of two approximations described by Eq. 2.11 we obtain the Theorem 8a. ■

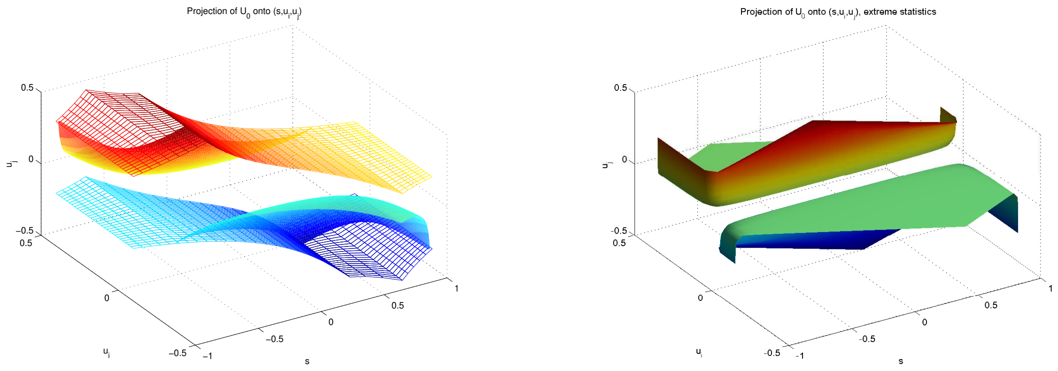
Theorem 8a does not specify the $O(1)$ term. The constant term C is well known in explicit form when the minimum of f is achieved on a single point, as specified by Lemma 2. In our case, the minimum of f is achieved on two points (x',u',s') and (x'',u'',s'') and by taking the integrals $\mathbb{J}_{(x',u',s')}[N]$ and $\mathbb{J}_{(x'',u'',s'')}[N]$ in (a,b,t) coordinates and accounting for the partial integration domains for the border points we obtain

$$C = \frac{2n+1}{2}\ln(2\pi) + \ln\left[\frac{\mu(a',b',t')}{\sqrt{\det \mathcal{H}f(a',b',t')}} + \frac{\mu(a'',b'',t'')}{\sqrt{\det \mathcal{H}f(a'',b'',t'')}}\right] - k\ln 2,$$

where $(a',b',t') = T_1^{-1}(x',u',s')$, $(a'',b'',t'') = T_1^{-1}(x'',u'',s'')$ and k is the number of border coordinates of (a',b',t') (or equivalently of (a'',b'',t'')). Note that $a' = b''$, $b' = a''$ and $t' = 1 - t''$.

3.5.3 Proof of Theorem 8b (Type 1 Singularity)

Theorem 8b states that if $Y_D = Y$ for $N \geq N_0$, $Y_i > 0$ for $i = 1,\ldots,2^n$ and $Y \in S \setminus S'$, then $\ln \mathbb{I}[N,Y_D]$ (Eq. 3.6) is asymptotically equal to $NP(Y|w_{ML}) - \frac{2n-1}{2}\ln N + O(1)$ (Eq. 3.8). To prove this claim we first employ Lemma 10, which relates $\mathbb{I}[N,Y_D]$ with $\tilde{\mathbb{I}}[N,Y]$ (Eqs. 3.21 and 3.23) and Lemma 11, which relates $\tilde{\mathbb{I}}[N,Y]$ with $\mathbb{J}[N,Y]$ (Eqs. 3.24 and 3.25). Consequently, it remains to evaluate $\mathbb{J}[N,Y] = \max_{p_0 \in U_0} \mathbb{J}_{p_0}[N]$. For this task, one needs to examine the neighborhoods of arbitrary minimum points $p_0 \in U_0$ of f. However, for $Y \in S \setminus S'$ (singularity of type 1), the function f can not be approximated by quadratic form and Lemma 2 (Laplace Approximation) no longer



(a) (b)

Figure 3.4: The projection of set U_0 onto (s, u_1, u_2) space. The zero set U_0 is defined by type 1 singularity statistics. (a) Illustration is for $x'_1 = 0.18$, $x'_2 = 0.28$, $z'_{12} = 0.0096$, that correspond to statistics Y generated by true distribution: $a_1 = 0.1$, $a_2 = 0.2$, $b_1 = 0.3$, $b_2 = 0.4$ and $t = 0.6$. Upper and lower bounds on u_1 are shown by mesh-grid. (b) Illustration of set U_0 for extreme (almost type 2) singular statistics of type 1 that is generated by $a_1 = 0.1$, $a_2 = 0.2$, $b_1 = 0.3$, $b_2 = 0.4$ and $t = 0.005$. The zero set is very close to the zero set for type 2 singularity statistics depicted in Figure 3.3(b).

applies. Instead we use Watanabe's method.

Let (a, b, t) be the parameterization of $Y \in S$ as described by the definition of S (Eq. 3.5) with $a_i = b_i$ for all $i \neq l, k$. Also, let $z'_{lk} = T_3^{-1}(Y)_{lk} = (1 - (2t-1)^2) \cdot \frac{a_l - b_l}{2} \cdot \frac{a_k - b_k}{2}$. The zero set U_0 is given by

$$U_0 = \left\{ (x, u, s) \in U \;\middle|\; \begin{array}{l} x_i = a_i, \quad \forall i = 1, \dots, n, \; i \neq l, k, \\ x_l = t a_l + (1-t) b_l, \\ x_k = t a_k + (1-t) b_k; \\ u_i = 0, \quad \forall i = 1, \dots, n, \; i \neq l, k, \\ u_l, u_k, s, \text{ such that } (1 - s^2) u_l u_k = z'_{lk} \end{array} \right\}. \tag{3.36}$$

Note that $z'_{lk} \neq 0$ and $u'_l, u'_k \neq 0$, $s' \neq \pm 1$ for $(x', u', s') \in U_0$, because $Y \notin S'$. The set U_0 is depicted in Figure 3.4.

We now apply the method of Watanabe, as described in Section 2.4, to evaluate the integrals $\mathbb{J}_{p_0}[N, Y]$ for $p_0 \in U_0$. We examine the form of the exponent function in $\mathbb{J}_{p_0}[N, Y]$, $\phi(x, u, s)$ which is equal to $\sum_I [z_I(x, u, s) - z'_I]^2$, in a small neighborhood of $p_0 = (x', u', s') \in U_0$. The coordinates of $z' = T_2(x', u', s')$ are $z'_i = x_i$ for all i, $z'_{lk} = (1 - s'^2) u'_l u'_k$ and all other z'_I's are zero. Substituting z_I as a function of (x, u, s) into ϕ and translating the (x, u, s) coordinates so that (x', u', s') becomes the origin, yields

$$
\begin{aligned}
\phi(x,u,s) = \;& \sum_I \left[z_I(x'+x,u'+u,s'+s) - z'_I\right]^2 \\
= \;& \sum_i [z_i(x'+x,u'+u,s'+s) - z'_i]^2 \\
& + [z_{lk}(x'+x,u'+u,s'+s) - z'_{lk}]^2 \\
& + \sum_{i\neq l,k} \left[z_{il}(x'+x,u'+u,s'+s) - z'_{il}\right]^2 + \left[z_{ik}(x'+x,u'+u,s'+s) - z'_{ik}\right]^2 \\
& + \sum_{i,j\neq l,k} \left[z_{ij}(x'+x,u'+u,s'+s) - z'_{ij}\right]^2 + \ldots \\
= \;& \sum_{i=1}^n [(x'_i + x_i) - x'_i]^2 \\
& + \left[(1-(s'+s)^2)(u'_l+u_l)(u'_k+u_k) - (1-s'^2)u'_l u'_k\right]^2 \\
& + \sum_{i\neq l,k} \left[(1-(s'+s)^2)(u'_l+u_l)u_i - 0\right]^2 + \left[(1-(s'+s)^2)(u'_k+u_k)u_i - 0\right]^2 \\
& + \sum_{i,j\neq l,k} \left[(1-(s'+s)^2)u_i u_j - 0\right]^2 + \ldots \\
= \;& \sum_{i=1}^n x_i^2 \\
& + \left[-2s'u'_l u'_k s + (1-s'^2)u'_k u_l + (1-s'^2)u'_l u_k + \text{“smaller terms''}\right]^2 \\
& + \sum_{i\neq l,k} \left[(1-s'^2)u'_l u_i + \text{“smaller terms''}\right]^2 + \left[(1-s'^2)u'_k u_i + \ldots\right]^2 \\
& + \sum_{i,j\neq l,k} \left[(1-s'^2)u_i u_j + \text{“smaller terms''}\right]^2 + \ldots\,.
\end{aligned}
\tag{3.37}
$$

The phrase “smaller terms” and dots denotes higher order terms that include variables that are present in the explicit terms of the sum and can be discarded for sufficiently small (x,u,s). In particular, the term $z_{lk}(x,u,s) - z'_{lk}$ is rewritten via

$$
\begin{aligned}
z_{lk}(x'+x,u'+u,s'+s) - z'_{lk} = \;& (1-(s+s')^2)(u_l+u'_l)(u_k+u'_k) - (1-s'^2)u'_l u'_k \\
= \;& -(2s'+s)u'_l u'_k s + ((1-s'^2) - 2s's - s^2)(u'_k+u_k)u_l \\
& +((1-s'^2) - 2s's - s^2)u'_l u_k.
\end{aligned}
$$

Consequently for $s' \neq 0$, sufficiently small ϵ and $s, u_l, u_k \in (-\epsilon, \epsilon)$ it follows that $C_1^- < -(2s' + s)u'_l u'_k < C_1^+$, $C_2^- < [(1-s'^2) - 2s's - s^2][u'_k + u_k] < C_2^+$ and $C_3^- < [(1-s'^2) - 2s's - s^2]u'_l < C_3^+$ for C_1^-, C_1^+, C_2^-, C_2^+, C_3^-, C_3^+ slightly smaller and larger than $C_1 = -2s'u'_l u'_k$, $C_2 = (1-s'^2)u'_k$, $C_3 = (1-s'^2)u'_l$.

Consequently, in order to approximate the integral $\mathbb{J}_{p_0}[N]$ (Eq. 3.24) for $p_0 = (x', u', s')$ with $s' \neq 0$, it remains to approximate the integral

$$
\begin{aligned}
& \tilde{\mathbb{J}}_1[N] = \int e^{-N\tilde{\phi}_1(x,u,s)} dx du ds, \\
& \text{where} \quad \tilde{\phi}_1(x,u,s) = \sum_i x_i^2 + \left[\tilde{C}_1 s + \tilde{C}_2 u_l + \tilde{C}_3 u_k\right]^2 + \sum_{i\neq l,k} \tilde{c}_i u_i^2
\end{aligned}
\tag{3.38}
$$

and where $\tilde{C}_1$, $\tilde{C}_2$, $\tilde{C}_3$ and $\tilde{c}_i$ are non-zero constants.

Similar analysis of the principal part of $\phi(x, u, s)$ (Eq. 3.37) function can be applied for the neighborhoods of $p_0 = (x', u', s')$ with $s' = 0$. It reveals that in order to approximate $\mathbb{J}_{p_0}[N]$ for $p_0 = (x', u', s')$ with $s' = 0$ we should approximate the integral

$$\begin{aligned} &\tilde{\mathbb{J}}_2[N] = \int e^{-N\tilde{\phi}_2(x,u,s)} dxduds, \\ &\text{where} \quad \tilde{\phi}_2(x, u, s) = \sum_i x_i^2 + \left[\hat{C}_1 s^2 + \hat{C}_2 u_l + \hat{C}_3 u_k\right]^2 + \sum_{i \neq l,k} \hat{c}_i u_i^2, \end{aligned} \tag{3.39}$$

and where $\hat{C}_1$, $\hat{C}_2$, $\hat{C}_3$ and $\hat{c}_i$ are non-zero constants that are slightly larger or smaller than $u'_l u'_k$, u'_k, u'_l and $u'^2_l + u'^2_k$.

From Eq. 3.38, by changing the coordinates to $v = \tilde{C}_1 s + \tilde{C}_2 u_l + \tilde{C}_3 u_k$, we obtain that in the neighborhoods of the points in U_0 with $s' \neq 0$, that f can be described by quadratic form in $2n-1$ variables, so their contribution to $\mathbb{J}[N, Y]$ is $cN^{\frac{2n-1}{2}}$.

The analysis of neighborhoods of points in U_0 with $s' = 0$ is harder. Integrating out x_i and u_i variables yields $N^{-\frac{2n-2}{2}}$ multiplicative factor to the asymptotic approximation of $\tilde{\mathbb{J}}_2[N]$, leaving us to compute of the contribution of $\int e^{-N[\hat{C}_1 s^2 + \hat{C}_2 u_i + \hat{C}_3 u_j]^2} ds\, du_i\, du_j$. The changes of variables $t = (\tilde{C}_2 u_i + \tilde{C}_3 u_j)/\tilde{C}_1$ transforms the remaining part of $\tilde{\mathbb{J}}_2[N]$ to

$$\tilde{\mathbb{J}}_3[N] = \int_{-\epsilon_1}^{+\epsilon_1} \int_{-\epsilon_2}^{+\epsilon_2} e^{-N(s^2+t)^2} dsdt$$

The zero set of the exponent function is a one-dimensional curve $t = -s^2$, so we expect $\tilde{\mathbb{J}}_3[N]$ be at least $cN^{-\frac{1}{2}}$, as verified below.

Watanabe's method for $\tilde{\mathbb{J}}_3[N]$ calls for the analysis of the poles of the function

$$J(\lambda) = \int_{(-1,1)^2} (s^2 + t)^{2\lambda} dsdt.$$

Here, we transform the original integration range into $(-1, 1)$ by rescaling, introducing only constant multipliers to the integral. The analysis of the poles of $J(\lambda)$ is in the spirit of example shown in Section 2.4. We present this analysis completely to demonstrate a number of important subtle points in the evaluation of integrals by resolution of singularities. E.g., we can not use the binomial formula for expanding $(s^2 + t)^{2\lambda}$, since λ is not necessarily an integer.

The integral $J(\lambda)$ is symmetric relative to s, so we consider only $s > 0$ for the evaluation of its poles. By changing the coordinates via $t = \pm t^2$ we obtain

$$\frac{1}{2} J(\lambda) = \int_{-1}^{1} \int_{0}^{1} (s^2 + t)^{2\lambda} dsdt = \int_{0}^{1} \int_{0}^{1} 2t(s^2 + t^2)^{2\lambda} dsdt + \int_{0}^{1} \int_{0}^{1} 2t(s^2 - t^2)^{2\lambda} dsdt.$$

The first integral is easy to evaluate by standard substitutions $s = ts$ for $0 < s < t < 1$ and $t = st$ for $0 < t < s < 1$. Thus, the first integral contributes a pole at $\lambda = -\frac{3}{4}$ with multiplicity 1. The second integral, however, can not be evaluated in this way, since, the substitution $s = ts$ for $0 < s < t < 1$ yields the integral $\int_0^1 \int_0^1 2t^{4\lambda+2}(s^2-1)^{2\lambda} dsdt$, where the term (s^2-1) is not bounded away from zero on $(0,1)$ and thus can not be ignored when identifying the poles.

To overcome this difficulty let $v = s+t$ and $u = s-t$, yielding

$$\int_0^1 \int_0^1 2t(s^2-t^2)^{2\lambda} dsdt = \frac{1}{2}\int_0^2 \int_{\max(-v,v-2)}^{\min(v,2-v)} (v-u)u^{2\lambda}v^{2\lambda} dudv$$

and

$$\frac{1}{2}\int_0^1 \int_{-v}^{v} (v-u)u^{2\lambda}v^{2\lambda} dudv < \int_0^1 \int_0^1 2t(s^2-t^2)^{2\lambda} dsdt < \frac{1}{2}\int_0^2 \int_{-v}^{v} (v-u)u^{2\lambda}v^{2\lambda} dudv. \quad (3.40)$$

Here is another possible pitfall. Ignoring the correct integration range and just setting it to $(0,1)^2$ gives $\int_0^1 \int_0^1 (v-u)u^{2\lambda}v^{2\lambda} dudv = \frac{1}{(2\lambda+1)(2\lambda+2)} - \frac{1}{(2\lambda+2)(2\lambda+1)} = 0$, which is a wrong result. Computing the lower bound in Eq. 3.40 yields

$$\begin{aligned}\tfrac{1}{2}\int_0^1 \int_{-v}^{v} (v-u)u^{2\lambda}v^{2\lambda} dudv &= \tfrac{1}{2}\int_0^1 \left[v^{2\lambda+1}\tfrac{1}{2\lambda+1}u^{2\lambda+1} - v^{2\lambda}\tfrac{1}{2\lambda+2}u^{2\lambda+2}\Big|_{-v}^{v} \right] dv \\ &= \tfrac{1}{2}\int_0^1 \tfrac{2}{2\lambda+1}v^{4\lambda+2} dv = \tfrac{1}{(2\lambda+1)(4\lambda+3)}.\end{aligned}$$

The upper limit is correspondingly $\frac{2^{4\lambda+3}}{(2\lambda+1)(4\lambda+3)}$. Hence, the largest pole of $J(\lambda)$ is $\lambda = -\frac{1}{2}$, with multiplicity $m = 1$ and the overall contribution of the neighborhoods of points p_0 with $s' = 0$ to $\mathbb{J}[N,Y]$ is again $cN^{-\frac{2n-1}{2}}$, and it is the same as for points p_0 for which $s' \neq 0$. The point (x', u', s') need not be an internal point of U. Such border points have a smaller domain of integration than an internal point, therefore they do not contribute more to $\mathbb{J}[N,Y]$ than internal points. ■

It is interesting to compare Figure 3.4b and Figure 3.3, to see that as a point $Y \in S \setminus S$ approaches $Y' \in S'$, the zero set for Y depicted by Figure 3.4 approaches the zero set for Y' depicted in Figure 3.3.

3.5.4 Proof of Theorem 8c (Type 2 Singularity)

The outline of the proof of Theorem 8c is presented in Section 3.4.3 including the specification of the zero set U_0 and five principal cases $C1$-$C5$ that correspond to different locations of extremum points $(x', u', s') \in U_0$. Recall that we are interested in the evaluation of the contribution of the neighborhood of each of the points of types $C1$-$C5$ to the integral $\mathbb{J}[N,Y]$ (Eq. 3.24). The maximal contribution determines, according to Lemmas 10 and 11, the asymptotic behavior of the integral $\mathbb{I}[N,Y_D]$ (Eq. 3.4) of interest. We now treat these cases one by one.

Case C1: $(x',u',s') \in U_{0j} \setminus \cup_{i\neq j} U_{0i}$ for some j. Each such point (x',u',s') satisfies $u'_i = 0$, for all $i = 1,\ldots,n$, $i \neq j$; $u'_j \neq 0$; $s' \neq \pm 1$; $z'_i = x'_i$; and $z'_{ij..k} = 0$. Using the approach of Watanabe we analyze the form of the exponent function ϕ of integrand of $\mathbb{J}_{p_0}[N]$ near the minimum point $p_0 = (x',u',s')$. Centering (x,u,s) around (x',u',s') yields

$$\begin{aligned}
\phi(x,u,s) = &\ \textstyle\sum_I [z_I(x'+x,x'+u,s'+s) - z'_I]^2 \\
= &\ \textstyle\sum_i [z_i(x'+x,u'+u,s'+s) - z'_i]^2 + \sum_{i\neq j}[z_{ij}(x'+x,u'+u,s'+s) - z'_{ij}]^2 \\
&+ \textstyle\sum_{i,k\neq j}[z_{ik}(x'+x,u'+u,s'+s) - z'_{ik}]^2 + \text{“higher order terms''} \\
= &\ \textstyle\sum_i [(x'_i + x_i) - x'_i]^2 + \sum_{i\neq j}\left[(1-(s'+s)^2)(u'_j+u_j)u_i - 0\right]^2 \\
&+ \textstyle\sum_{i,k\neq j}\left[(1-(s'+s)^2)u_iu_k - 0\right]^2 + \text{“higher order terms''} \\
= &\ \textstyle\sum_i x_i^2 + \sum_{i\neq j}\left[(1-s'^2)u'_ju_i + \text{“smaller terms''}\right]^2 \\
&+ \textstyle\sum_{i,k\neq j}\left[(1-s'^2)u_iu_k - (s+2s')su_iu_k\right]^2 + \text{“higher order terms''}.
\end{aligned}$$

Since, $u'_j \neq 0$ and $s \neq \pm 1$, the principal part of ϕ, that bounds ϕ within a multiplicative constant, is

$$\tilde{\phi}(x,u,s) = \sum_{i=1,\ldots,n} x_i^2 \ + \sum_{i=1,\ldots,n;\, i\neq j} u_i^2.$$

Hence, $\mathbb{J}_{p_0}[N]$ is $cN^{-\frac{2n-1}{2}}$. One should have expected this result because the zero set $U_{0,j}$ is a 2-dimensional surface, yielding a dimensionality drop of 2 due to two locally redundant parameters.

Case C2: $(x',u',s') = \bigcap_j U_{0j}$. This case is analyzed in Section 3.4.3.

Case C3: $(x',u',s') \in U_{0-} \cup U_{0+} \setminus \cup_j \bar{U}_{0j}$. Each such point (x',u',s') satisfies $u'_j \neq 0$ for all $j = 1,\ldots,n$ and $s' = \pm 1$. We have

$$\begin{aligned}
\phi(x,u,s) = &\ \textstyle\sum_I [z_I(x'+x,u'+u,s'+s) - z'_I]^2 \\
= &\ \textstyle\sum_i [z_i(x'+x,u'+u,s'+s) - z'_i]^2 + \sum_{i,j}[z_{ij}(x'+x,u'+u,s'+s) - z'_{ij}]^2 \\
&+ \textstyle\sum_{i,j,k}[z_{ijk}(x'+x,u'+u,s'+s) - z'_{ijk}]^2 + \ldots \\
= &\ \textstyle\sum_i [(x'_i + x_i) - x'_i]^2 + \sum_{i,j}\left[(1-(s'+s)^2)(u'_i+u_i)(u'_j+u_j) - 0\right]^2 \\
&+ \textstyle\sum_{i,j,k}\left[-2(s'+s)(1-(s'+s)^2)(u'_i+u_i)(u'_j+u_j)(u'_k+u_k) - 0\right]^2 + \ldots \\
= &\ \textstyle\sum_i x_i^2 + \sum_{i,j}\left[-2s'u'_iu'_js + \text{“smaller terms''}\right]^2 \\
&+ \textstyle\sum_{i,j,k}\left[4u'_iu'_ju'_ks + \text{“smaller terms''}\right]^2 + \text{“higher order terms''}.
\end{aligned}$$

So, the principal part of ϕ is of the form $\sum_i x_i^2 + s^2$. The fact that integration range for s is one sided, i.e. $s > 0$ (or $s < 0$) changes the integral $\mathbb{J}_{p_0}[N]$ only by a constant multiply (1/2) relatively to the "full" neighborhood. Thus the contribution of this region to $\mathbb{J}[N, Y]$ is $cN^{-\frac{n+1}{2}}$.

Case C4: $(x', u', s') \in \bigcup_j \left[U_{0-} \cup U_{0+} \cap \bar{U}_{0j} \setminus \cap_{i \neq j} \bar{U}_{0i}\right]$, for some j. Each such point (x', u', s') satisfies $u'_j \neq 0$ for some j; $u'_i = 0$ for all $i \neq j$; and $s' = \pm 1$. We have

$$\begin{aligned}
\phi(x,u,s) = \;& \textstyle\sum_I [z_I(x'+x, u'+u, s'+s) - z'_I]^2 \\
= \;& \textstyle\sum_i [z_i(x'+x, u'+u, s'+s) - z'_i]^2 + \sum_{i \neq j} [z_{ij}(x'+x, u'+u, s'+s) - z'_{ij}]^2 \\
& + \textstyle\sum_{i,k \neq j} [z_{ik}(x'+x, u'+u, s'+s) - z'_{ik}]^2 + \text{"higher order terms"} \\
= \;& \textstyle\sum_i [(x'_i + x_i) - x'_i]^2 + \sum_{i \neq j} \left[(1-(s'+s)^2)(u'_j + u_j)u_i - 0\right]^2 \\
& + \textstyle\sum_{i,k \neq j} \left[(1-(s'+s)^2)u_i u_k - 0\right]^2 + \text{"higher order terms"} \\
= \;& \textstyle\sum_i x_i^2 + \sum_{i \neq j} \left[\mp 2 s u'_j u_i \mp 2 s u_j u_i - s^2 u'_j u_i - s^2 u_j u_i\right]^2 \\
& + \textstyle\sum_{i,k \neq j} \left[\mp 2 s u_i u_k - s^2 u_i u_k\right]^2 + \text{"higher order terms"} \\
\approx \;& \textstyle\sum_i x_i^2 + s^2 \sum_{i \neq j} u_i^2.
\end{aligned} \tag{3.41}$$

Integrating out the $\sum_i x_i^2$ terms from $\mathbb{J}_{p_0}[N]$, we see that they contribute factor of $N^{-\frac{n}{2}}$ to $\mathbb{J}_{p_0}[N]$. So, we are left with analysis of the poles of

$$J(\lambda) = \int_{W_\epsilon} s^{2\lambda} \left(\sum_{i=1}^{n-1} u_i^2\right)^{\lambda} ds du.$$

The standard change of variables to $u_i = u_1 u_i$ for $i = 2, \ldots, n-1$ gives

$$J(\lambda) = c \int_{(0,1)^n} s^{2\lambda} u_1^{2\lambda+n-2} (1 + \sum_{i=2}^{n-1} u_i^2)^{\lambda} ds du.$$

Thus the largest pole of $J(\lambda)$ (for $n > 2$) is $\lambda = -\frac{1}{2}$ with multiplicity $m = 1$ and the contribution of the neighborhood of this (x', u', s') is $cN^{-\frac{n+1}{2}}$.

Case C5: $(x', u', s') \in (U_{0-} \cup U_{0+}) \bigcap_j \bar{U}_{0j}$. Each such point (x', u', s') satisfies $u'_i = 0$ for all $i = 1, \ldots, n$ and $s' = \pm 1$. This is the deepest singularity, the crossing of all (except one) zero planes of U_0. We have

$$
\begin{aligned}
\phi(x,u,s) = {} & \sum_I [z_I(x'+x, u'+u, s'+s) - z'_I]^2 \\
= {} & \sum_i [z_i(x'+x, u'+u, s'+s) - z'_i]^2 + \sum_{i,j} [z_{ij}(x'+x, u'+u, s'+s) - z'_{ij}]^2 \\
& + \sum_{i,j,k} [z_{ijk}(x'+x, u'+u, s'+s) - z'_{ijk}]^2 + \dots \\
= {} & \sum_i [(x'_i + x_i) - x'_i]^2 + \sum_{i,j} \left[(1-(s'+s)^2)u_i u_j - 0\right]^2 \\
& + \sum_{i,j,k} \left[-2(s'+s)(1-(s'+s)^2)u_i u_j u_k - 0\right]^2 + \text{“higher order terms''} \\
= {} & \sum_i x_i^2 + \sum_{i,j} \left[\mp 2 s u_i u_j - s^2 u_i u_j\right]^2 \\
& + \sum_{i,j,k} [4 s u_i u_j u_k + \text{“smaller terms''}]^2 + \text{“higher order terms''} \\
\approx {} & \sum_i x_i^2 + s^2 \sum_{i,j} u_i^2 u_j^2 .
\end{aligned}
\tag{3.42}
$$

The higher order terms are bounded by some $s^2 u_i^2 u_j^2$ term, because of the special form of $p_i(s)$ term in $z_{12\dots i}$ (Eq. 3.20). I.e., the function $p_i(s'+s) = 1/2(1-(s'+s)^2)[(1-(s'+s))^{i-1} - (-1)^{i-1}(1+(s'+s))^{i-1}]$ can be rewritten around $s = \pm 1$ as $p_i(s'+s) = s \cdot 1/2(2s'+s)[(1-(s'+s))^{i-1} - (-1)^{i-1}(1+(s'+s))^{i-1}]$. or, in another form

$$
p_i(\pm 1 + s) = \begin{cases} s(\mp 2 - s) \sum_{j=0,2,4,\dots}^{2k-1} \binom{2k-1}{j} (\pm 1 + s)^j & \text{for } i = 2k \\ s(\pm 2 + s) \sum_{j=1,3,5,\dots}^{2k} \binom{2k}{j} (\pm 1 + s)^j & \text{for } i = 2k+1 \end{cases}
$$

Thus, any high-order term $z^2_{ij\dots r}(x'+x, u, \pm 1 + s)$ is of form

$$
z^2_{ij\dots k}(x'+x, u, \pm 1 + s) = s^2 u_i^2 u_j^2 \dots u_k^2 \cdot \tilde{p}(s),
$$

where $\tilde{p}(s) = 1/4(2s'+s)^2[(1-(s'+s))^{r-1} - (-1)^{r-1}(1+(s'+s))^{r-1}]^2$ and where r is the size of index set $\{ij\dots k\}$. Consequently, this term is bounded by $s^2 u_i^2 u_j^2$ for s and u small enough.

The $\sum_i x_i^2$ terms contribute $N^{-\frac{n}{2}}$ multiplicative factor to $\mathbb{J}_{p_0}[N]$, so we should only analyze the poles of

$$
J(\lambda) = \int_{(0,1)^{n+1}} s^{2\lambda} \left(\sum_{l,k} u_l^2 u_k^2 \right)^{\lambda} ds du.
$$

The analysis is similar to the one presented in Section 2.4, but with additional variable s. Thus the largest pole of $J(\lambda)$ for this case is $\lambda = -\frac{1}{2}$ and not $\lambda = -n/4$. The multiplicity of the pole $\lambda = -\frac{1}{2}$ is one and so the contribution of the neighborhoods of $(x', 0, \pm 1)$ is $cN^{-\frac{n+1}{2}}$. This analysis is incorrect for $n = 2$ because then the sum $\sum_{l,k} u_l^2 u_k^2$ contains only one term increasing the multiplicity of the pole to $\lambda = -1/2$.

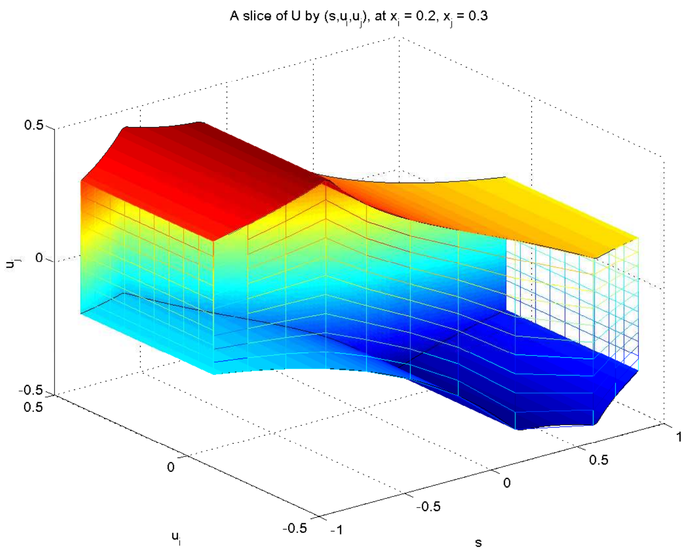


Figure 3.5. The integration domain U of $\tilde{\mathbb{I}}[N,Y]$. Slice of (s,u_i,u_j) for $x_i = 0.2$ and $x_j = 0.3$.

The interesting fact about the last two cases is that in the neighborhood of U_{0-} and U_{0+} the growth of the function ϕ is dominated by s^2 and thus the multiplicity of the maximal pole of $J(\lambda)$ is always one and the $\ln\ln N$ terms do not appear in the approximation of $\ln \mathbb{J}_{p_0}[N]$. This changes in the case $n = 2$, where the dimensionalities of U_{0-} and U_{0+} are the same as of U_{0j}'s, as explicated in the next section.

Summary of Proof of Theorem 8 for type 2 *singularity,* $Y \in S'$: Among the possible cases $C1$-$C5$ the largest contribution to the $\mathbb{J}[N,Y]$ comes from points with $s' = \pm 1$. Note that various border points of U_0 that we do not consider in the above analysis do not contribute more than the corresponding internal points because their domain of integration is smaller. Thus, $\ln \mathbb{J}[N,Y] = -\frac{n+1}{2}\ln N + O(1)$ and due to Lemmas 10 and 11, $\ln \mathbb{I}[N,Y_D] = NP(Y|w_{ML}) - \frac{n+1}{2}\ln N + O(1)$ as claimed. (Figure 3.5 illustrates the integration range, while Figure 3.3 illustrates the zero set U_0.) ■

3.5.5 Proof of Claims (d,e) of Theorem 8 (Case $n = 2$)

Claims (d,e) of Theorem 8 state that if $n = 2$, $Y_D = Y$ for $N \geq N_0$ and $Y_i > 0$ for $i = 1,\ldots,2^n$, $\ln \mathbb{I}[N,Y_D]$ (Eq. 3.6) is asymptotically equal to $NP(Y|w_{ML}) - \frac{3}{2}\ln N + O(1)$ (Eq. 3.10) for $Y \notin S'$ and asymptotically equal to $NP(Y|w_{ML}) - \frac{3}{2}\ln N + 2\ln\ln N + O(1)$ (Eq. 3.11) for $Y \in S'$. Similar to the proofs of Claims (b,c), we first employ Lemma 10, which relates $\mathbb{I}[N,Y_D]$ with $\tilde{\mathbb{I}}[N,Y]$ (Eqs. 3.21

and 3.23) and Lemma 11, which relates $\tilde{\mathbb{I}}[N,Y]$ with $\mathbb{J}[N,Y]$ (Eqs. 3.24 and 3.25). Consequently, it remains to evaluate $\mathbb{J}[N,Y] = \max_{p_0 \in U_0} \mathbb{J}_{p_0}[N]$. For this task, one needs examine the neighborhoods of arbitrary minimum points $p_0 \in U_0$ of the function f. From the definition of Υ, Υ_0 and S (Section 3.3) it follows that $S = \Upsilon_0 = \Upsilon$ for $n = 2$. Note that there is no regular points in this case. We now modify the proofs of type 1 and type 2 singularities to fit to the case $n = 2$.

Type 1 *singularity:* The zero set U_0 is the same set as described by Eq. 3.36 with $l = 1$ and $k = 2$. The analysis of the form of the exponent function ϕ of the integrand of $\mathbb{J}_{p_0}[N]$ gives Eqs. 3.38 and 3.39 without the $\sum_{l \neq i,j} c_l u_l^2$ terms. Thus, by the same analysis, the contribution of these regions to the integral $\mathbb{J}[N,Y]$ is $cN^{-\frac{3}{2}}$ and application of Lemmas 10 and 11 concludes the proof.

Type 2 *singularity:* The zero set $U_0 = \bar{U}_{0-} \cup \bar{U}_{0+} \cup \bar{U}_{01} \cup \bar{U}_{02}$ is the same set as described by Eqs. 3.27 and 3.28. Now, however, $\bar{U}_{0-}$, $\bar{U}_{0+}$, $\bar{U}_{01}$ and $\bar{U}_{02}$ are of the same dimension, namely, two. This fact changes the asymptotic approximation.

Consider the cases $C1$-$C5$ one by one. There is no change in cases $C1$ and $C3$ where the point (x', u', s') lies on the proper two dimensional surfaces U_{01}, U_{02} or U_{0-}, U_{0+}. Here, the function ϕ can be approximated by 3 variables, resulting in the contribution $cN^{-3/2}$ of these regions to $\mathbb{J}[N,Y]$.

The more complex situations are cases $C2$, $C4$ and $C5$, where zero planes of the same dimension meet. Generally, the intersection points of zero surfaces of the same dimension are expected to give rise to a $\ln \ln N$ term. While this is not always a case, e.g., see example in Section 2.4, the $\ln \ln N$ term does appear now. We have:

- C2: The principal part of ϕ is $x_1^2 + x_2^2 + u_1^2 u_2^2$, as specified by Eq. 3.30. Integrating out the x_i^2 terms we obtain through the analysis of the poles of $J(\lambda) = \int u_1^{2\lambda} u_2^{2\lambda} du_1 du_2$ that the largest pole of $J(\lambda)$ is $\lambda = -1/2$ with multiplicity $m = 2$. Thus the contribution of this region to $\mathbb{J}[N,Y]$ is $cN^{-3/2} \ln N$.

- C4: The principal part of ϕ is $x_1^2 + x_2^2 + s^2 u_2^2$ or $x_1^2 + x_2^2 + s^2 u_1^2$ (see Eq. 3.41). Similarly to case $C2$, the contribution of this region to $\mathbb{J}[N,Y]$ is $cN^{-3/2} \ln N$.

- C5: Here, the principal part of ϕ is $x_1^2 + x_2^2 + s^2 u_1^2 u_2^2$ (see Eq. 3.42). Once again, we integrate out the x_i variables and analyze the poles of $J(\lambda) = \int s^{2\lambda} u_1^{2\lambda} u_2^{2\lambda} ds du_1 du_2$. The largest pole is $\lambda = -1/2$ with multiplicity $m = 3$, and thus the contribution of this region to $\mathbb{J}[N,Y]$, including the factors from integrating out the x_i's, is $cN^{-3/2} \ln^2 N$.

Summarizing the contributions of the neighborhoods of various critical points for $Y \in S'$, we see that $\mathbb{J}[N,Y] \sim cN^{-3/2}\ln^2 N$ and, consequently, $\ln \mathbb{I}[N,Y] = Nf_Y - \frac{3}{2}\ln N + 2\ln\ln N + O(1)$. ■

3.5.6 Proof of Theorem 8f (Case $n = 1$)

Theorem 8f states that if $n = 1$, $Y_D = Y$ for $N \geq N_0$ and $Y_1, Y_2 > 0$, then $\ln \mathbb{I}[N, Y_D]$ (Eq. 3.6) is asymptotically equal to $NP(Y|w_{ML}) - \frac{1}{2}\ln N + O(1)$ (Eq. 3.12). Once again, we first employ Lemma 10, which relates $\mathbb{I}[N, Y_D]$ with $\tilde{\mathbb{I}}[N,Y]$ (Eqs. 3.21 and 3.23) and Lemma 11, which relates $\tilde{\mathbb{I}}[N,Y]$ with $\mathbb{J}[N,Y]$ (Eqs. 3.24 and 3.25). Consequently, it remains to evaluate $\mathbb{J}[N,Y] = \max_{p_0 \in U_0} \mathbb{J}_{p_0}[N]$. For this task, one needs examine $\mathbb{J}_{p_0}[N]$ in the neighborhoods of arbitrary minimum points $p_0 \in U_0$ of the function f.

From the definitions of Υ, Υ_0 and S', for $n = 1$, there is no distinction between different type of statistics and $\Upsilon = \Upsilon_0 = S'$. Moreover, according to Theorem 6 the asymptotic form of the integral $J_{p_0}[N] = \int_{U_\epsilon} e^{-N(z_1(x,u,s)-z_1')^2} dx\,du\,ds$ is determined by the poles of $J(\lambda) = \int_{U_\epsilon} (z_1(x,u,s) - z_1')^{2\lambda} dx\,du\,ds$, where, in this case, $z_1(x,u,s) - z_1' = x_1^2$. Once again, contributions of points $p_0 \in U_0$ lying on the boundary of U can be ignored, since their domains of integration are smaller than domains of integration of the corresponding internal points. Thus, the largest pole of $J(\lambda)$ is $\lambda = -1/2$ with multiplicity $m = 1$ and $\ln I[N, Y_D]$ is asymptotically equal to $Nf_Y - \frac{1}{2}\ln N + O(1)$. ■

We can also compute the integral $\mathbb{I}[N, Y_D]$ (Eq. 3.4) directly for $n = 1$ and $Y_D = Y$. It is

$$\mathbb{I}[N,Y] = \int_{(0,1)^3} e^{N(Y_0 \ln[at+b(1-t)] + Y_1 \ln[(1-a)t+(1-b)(1-t)])} \mu(a,b,t)\,da\,db\,dt$$

where $Y_1 = 1 - Y_0$. Ignoring the density $\mu(a,b,t)$ by using the assumption of bounded density $(A1)$ and changing the variables to $x = at + b(1-t)$, we rewrite $\mathbb{I}[N,Y]$ is asymptotically equivalent form

$$\tilde{\mathbb{I}}[N,Y] = \int_0^1 \int_0^1 \frac{1}{b-a} \int_a^b e^{N(Y_0 \ln x + Y_1 \ln[1-x])} dx\,da\,db.$$

Consider now

$$\mathbb{I}_1[N,Y] = \int_a^b e^{N(Y_0 \ln x + Y_1 \ln(1-x))} dx$$

for some $0 \leq a < b \leq 1$ (the case $b > a$ is symmetric). This is the integral of the beta distribution with $\alpha = NY_0 + 1$ and $\beta = NY_1 + 1$ (DeGroot, 1970; page 40). Let $f(x) = Y_0 \ln x + Y_1 \ln(1-x)$. The maximum of the integrand function $f(x)$ on $[0,1]$ is achieved at $x_0 = Y_0$ and it is $e^{Nf(Y_0)}$. There are three cases to consider according to the location of x_0 relative to (a,b).

1. *Internal point,* $x_0 = Y_0 \in (a, b)$. In this case

$$\begin{aligned} f(Y_0+x) = &\quad f(Y_0) + Y_0 \ln\left(1+\tfrac{x}{Y_0}\right) + (1-Y_0)\ln\left(1-\tfrac{x}{1-Y_0}\right) \\ = &\quad f(Y_0) + Y_0\left(\tfrac{x}{Y_0} - \tfrac{x^2}{2Y_0^2} + O(x^3)\right) + (1-Y_0)\left(\tfrac{-x}{1-Y_0} - \tfrac{x^2}{2(1-Y_0)^2} + O(x^3)\right) \\ = &\quad f(Y_0) - \tfrac{1}{2Y_0(1-Y_0)}x^2 + O(x^3). \end{aligned}$$

   Thus, in the small neighborhood of x_0, f can be approximated by quadratic form and the classic Laplace approximation (Lemma 2) can be applied yielding $\mathbb{I}_1[N,Y] \sim c_1 e^{Nf_Y} N^{-1/2}$. Moreover, since $\mathbb{I}_1[N,Y]$ and $e^{Nf(Y_0)}$ are continuous functions of N and $x_0 = Y_0$, uniform asymptotic bounds on $\mathbb{I}_1[N,Y]$ exists for all x_0 in a proper closed subset of (a,b) as $N \to \infty$. I.e., the integral $\mathbb{I}_1[N,Y]$ is bounded within a constant multiplies of $e^{Nf_Y}N^{-\frac{1}{2}}$ and these constants are independent of x_0 and N for all $x_0 \in [a+\epsilon, b-\epsilon]$ and $N \geq 1$. Note that the above approximation of f is only valid for $Y_0 \neq 0, 1$ (Assumption $A2$). Otherwise, the approximation of f is non-quadratic.

2. *Border point,* $x_0 = Y_0 \in \{a, b\}$. The expansion for $f(Y_0 + x)$ is the same, but the integration is performed only on half of the interval, which results in half the constant factor to the final approximation compared with the previous case.

3. *Maximum of f is outside of* $[a, b]$. Let m denote the maximum of $e^{f(x)}$ on $[a, b]$, i.e., $m = \max_{x\in[a,b]} e^{f(x)}$. We have $\mathbb{I}_1[N,Y] \leq (b-a)m^N < c_3 e^{Nf_Y}N^{-1/2}$, for some appropriate constant c_3.

The above analysis shows that $\mathbb{I}_1[N,Y] < c_{upp}e^{Nf_Y}N^{-1/2}$ for some constant c_{upp} for all a and b. Furthermore, $\mathbb{I}[N,Y] > c_{low}e^{Nf_Y}N^{-1/2}$ for some $c_{low} > 0$ for $(a,b) \in \{(a,b)|a < Y_0, b > Y_0, b-a > 2\epsilon > 0\}$. Since the later region has a non-zero Lebesgue measure, it follows that $\mathbb{I}[N,Y] \sim ce^{Nf_Y}N^{-1/2}$ and $\ln \mathbb{I}[N,Y] = Nf_Y - \frac{1}{2}\ln N + O(1)$.

3.6 Proof of Theorem 9

Theorem 9 states the asymptotic approximation for the marginal likelihood given a degenerate binary naive Bayesian model M that has m missing links. In order to prove this theorem we examine the log-likelihood function of the degenerate model and decompose it into a degenerated and a naive

Bayesian parts. These parts define two probability functions that are independent and the marginal likelihood of data is computed relevant to each one of them. Combining the results proves Theorem 9.

Let ψ be the log-likelihood function of the marginal likelihood integral (Eq. 3.13) for the degenerate binary naive Bayesian network described in Theorem 9. We have

$$\begin{aligned}
\tfrac{1}{N}\psi(a,b,t,c) = & \sum_x Y_x \ln \theta_x(\omega) \\
= & \sum_x Y_x \left[\ln \theta_{(x_1,\ldots,x_{n-m})}(a,b,t) + \sum_{i=n-m+1}^n (x_i \ln c_i + (1-x_i)\ln(1-c_i))\right] \\
= & \sum_{(x_1,\ldots,x_{n-m})} \left[\ln \theta_{(x_1,\ldots,x_{n-m})}(a,b,t) \cdot \sum_{(x_{n-m+1},\ldots,x_n)} Y_x\right] \\
& + \sum_{i=n-m+1}^n \left(\sum_x Y_x x_i \ln c_i + \sum_x Y_x (1-x_i)\ln(1-c_i)\right) \\
= & \sum_{(x_1,\ldots,x_{n-m})} Y_{(x_1,\ldots,x_{n-m})} \ln \theta_{(x_1,\ldots,x_{n-m})}(a,b,t) \\
& + \sum_{i=n-m+1}^n (Y_i \ln c_i + (1-Y_i)\ln(1-c_i))
\end{aligned}$$

where $(x_1,\ldots,x_k)$ are binary vectors of length k, $Y_{(x_1,\ldots,x_{n-m})} = \sum_{(x_{n-m+1},\ldots,x_n)} Y_{(x_1,\ldots,x_n)}$ and $Y_i = \sum_{(x_1,\ldots,x_{i-1},1,x_{i+1},\ldots,x_n)} Y_x$. The new statistics $Y_{(x_1,\ldots,x_{n-m})}$ and Y_i's are positive, because Y is positive ($A2$). Using the assumptions of bounded density ($A1$) and stable statistics ($A3$), the marginal likelihood integral $\mathbb{I}[N,Y]$ (Eq. 3.13) can be rewritten as

$$\mathbb{I}[N,Y_D] \sim \hat{\mathbb{I}}[N,Y] = \left[\prod_{i=n-m+1}^n \int_0^1 c_i^{NY_i}(1-c_i)^{N(1-Y_i)} dc_i\right] \int_{(0,1)^{2n-2m+1}} e^{N\sum_{\tilde{x}} Y_{\tilde{x}} \ln \theta_{\tilde{x}}(\omega)} d\omega. \quad (3.43)$$

where $\tilde{x} = (x_1,\ldots,x_{n-m})$. The first m integrals are over the beta distribution (DeGroot, 1970; page 40) and

$$\int_0^1 c_i^{NY_i}(1-c_i)^{N(1-Y_i)} dc_i = \frac{\Gamma(NY_i+1)\Gamma(N(1-Y_i)+1)}{\Gamma(N+2)}$$

The asymptotic behavior of Gamma function is well understood and it is described by Stirling's formula, $\Gamma(z) = e^{-z} z^{z-\frac{1}{2}} \sqrt{2\pi}\left[1+O(z^{-1})\right]$ (Murray, 1984; page 38), and thus $\ln\Gamma(z) = -z + (z-\frac{1}{2})\ln z + O(1)$. Using the equality $\ln(YN+1) = \ln YN + O(1)$, we obtain

$$\begin{aligned}
& \ln \tfrac{\Gamma(NY_i+1)\Gamma(N(1-Y_i)+1)}{\Gamma(N+2)} \\
& = (NY_i + \tfrac{1}{2})\ln(NY_i+1) + (N(1-Y_i)+\tfrac{1}{2})\ln(N(1-Y_i)+1) - (N+\tfrac{3}{2})\ln(N+2) + O(1) \\
& = (NY_i + \tfrac{1}{2})\ln NY_i + (N(1-Y_i)+\tfrac{1}{2})\ln N(1-Y_i) - (N+\tfrac{3}{2})\ln N + O(1) \\
& = -\tfrac{1}{2}\ln N + N\left(Y_i \ln Y_i + (1-Y_i)\ln(1-Y_i)\right) + O(1).
\end{aligned}$$

Hence, the contribution of the first m integrals to $\ln \hat{\mathbb{I}}[N,Y]$ is $N \ln p(Y_{n-m+1},\ldots,Y_n|c_{ML}) - \frac{m}{2}\ln N$. The second integral in Eq. 3.43 is exactly of the type analyzed in Theorem 8, and the theorem follows by summing up the contributions of these two parts. ■

3.7 Discussion

This chapter presents an asymptotic approximation of the marginal likelihood of data given a naive Bayesian model with binary variables (Theorem 8). This Theorem proves that the classical BIC score that penalizes the log-likelihood of a model by $\frac{d}{2}\ln N$ is incorrect for Bayesian networks with hidden variables and suggests an adjusted BIC score. Moreover, no uniform penalty term exists for such models in the sense that the penalty term, i.e., the coefficient of $\ln N$, depends on the averaged sufficient statistics. This result resolves an open problem regarding the validity of the classical BIC score for stratified exponential families, raised in (Geiger, Heckerman, King & Meek, 2001).

The major limitation of Theorem 8 arises from Assumptions $A2$ and $A3$. While Assumption $A1$ (bounded density) is often satisfied in applications, Assumption $A2$ (positive statistics) is only sometimes satisfied and Assumption $A3$ (statistics stability) is never satisfied in practice. Nevertheless, this Theorem is an essential advance towards developing asymptotic Bayesian methods for model selection among naive Bayesian models in particular, and for Bayesian networks with hidden variables in general. We now highlight the steps required for obtaining a valid, practical asymptotic model selection score for arbitrary latent Bayesian networks, namely, for Bayesian networks with hidden variables.

1. Develop a closed form asymptotic formula for marginal likelihood integrals for all types of statistics Y given an arbitrary latent Bayesian model.

2. Extend these solutions by developing *uniform* asymptotic approximations valid for converging statistics $Y_D \to Y$ as $N \to \infty$. A uniform asymptotic approximation is an approximation that has the error term bounded for all Y_D near Y and for all N.

3. Develop an algorithm that, given a Bayesian network with hidden variables and a data set with statistics Y_D, determines the possible singularity types of the limit statistics Y and applies the appropriate asymptotic formula developed in step 2.

Our work provides a first step for naive Bayesian networks and a concrete framework to pursue these tasks.

Theorem 8 shows that when comparing the classical BIC score with our adjusted BIC score (Eq. 3.2 versus Eqs. 3.8, 3.9), one can see that a naive Bayesian network with all links present is somewhat under-evaluated using the classical BIC score for singular statistics Y because the penalty term reduces from $(2n+1)/2$ in the classical score to $(2n-1)/2$ (or $(n+1)/2$) in the adjusted score. We conjecture that such under evaluation occurs for general Bayesian networks with hidden variables. As a result, when the data shows weak dependencies for some links, often resulting in evaluation of the marginal likelihood near singular points of the model, then those models with more links might be under evaluated using BIC, but correctly evaluated with a uniform asymptotic formula that takes the proximity to a singular points into account. An illustrative example of incorrect model choice by the standard BIC score has been presented in Figure 3.3.

We conclude with two remarks. First, we note that the adjusted penalty term (Eqs. 3.8, 3.9) falls within the range of penalty terms, studied by Keribin (2000), that lead to sure consistency estimators in a frequentist's interpretation.

Second, we note that, the sets of singular points S and S' are defined in (Geiger, Heckerman, King & Meek, 2001) as the singular points of the algebraic varieties of distributions represented by binary naive Bayesian networks in the joint space parameters space, while here the same sets are defined as sets of *statistics* points Y which give rise to singular maximum likelihood in the model parameters space. At the singular points of the joint space parameters space, regular local coordinates do not exist and the usual coordinates (i.e., the model parameters) that parameterize the rest of the model variety have a number of coordinates crushed into a single point. This results in complex surfaces of maximum likelihood points in the model parameter space and, consequently, a non-standard behavior of marginal likelihood integrals which we have started to explore in this work. Another ramification of this observation is that a bounded prior density defined on the model parameters may accumulate massively on a single point on the model variety in the joint space parameter space, violating the boundedness assumption of the prior density and thus yielding non-standard approximations to marginal likelihood integrals in the joint space parameters.

Chapter 4

Automated Analytic Asymptotic Evaluation of the Marginal Likelihood for Latent Models

We present two algorithms for analytic asymptotic evaluation of the marginal likelihood of data given a Bayesian network with hidden nodes. As shown in the previous chapter, this evaluation is particularly hard because for these models asymptotic approximation of the marginal likelihood deviates from the standard BIC score. Our algorithms compute regular dimensionality drop for latent models and compute the non-standard approximation formulas for singular statistics for these models. The presented algorithms are implemented in Matlab and Maple and their usage is demonstrated on several examples. A version of this chapter has been published as (Rusakov & Geiger, 2003). Some of the background is repeated herein to facilitate independent reading of this chapter.

4.1 Introduction

In this chapter we address the problem of computing analytic asymptotic approximations of marginal likelihoods and present two computer programs that compute such approximations. Our algorithms are developed in the context of Bayesian networks with hidden variables, where the evaluation of marginal likelihood is particularly hard. Chapter 3 addressed analytically a specific form of such integrals for naive Bayesian networks with two classes and binary feature variables. The algorithms in this chapter deal with more general models with hidden variables.

Recall that under some regularity conditions, the asymptotic form of the logarithm of the marginal likelihood for Bayesian network models without hidden variables is specified by the *standard BIC formula*:

$$\ln P(D|M) = \ln P(D|\omega_{ML}) - \frac{d}{2}\ln N + O(1), \tag{4.1}$$

where N is the number of examples in D, ω_{ML} are the maximum likelihood parameters of M and d is the number of independent parameters.

The evaluation of the marginal likelihood for Bayesian network models with hidden variables is more complex. First, some of the parameters of the model may be redundant, and second, some of the data may give rise to statistics that are *singular* relative to the given model, resulting in non-standard approximation forms (Chapter 3). Various heuristic methods were developed to approximate the marginal likelihood of Bayesian networks with hidden variables (Cheeseman & Stutz, 1995; Chickering & Heckerman, 1997).

In this chapter we address both difficulties. First, we implement the method for effective dimensionality computation presented in (Geiger, Heckerman & Meek, 1996) and optimize it by decomposing the input network into independent components. The algorithm is implemented in Matlab and is capable of evaluating effective dimensionality of large Bayesian networks with hidden variables.

Second, we fill in the details and implement the algorithmic approach suggested in (Watanabe, 2001) for analytic asymptotic approximation of "hard" integrals. Our algorithm combines state-of-the-art algorithms of algebraic geometry (Bodnar & Schicho, 2000; Bravo, Encinas & Villamayor, 2002) with specific analytic methods for marginal likelihood evaluation suitable for Bayesian networks, which were developed in Chapter 3.

The latest algorithm is implemented in Maple and consists of a number of procedures. These procedures compute the approximation of marginal likelihood not only for Bayesian with hidden variables, but for a larger set of probabilistic models, for which the log-likelihood function can be represented (or bounded) by a polynomial. We demonstrate the usage of our algorithm in evaluating marginal likelihood formulas on a number of Bayesian networks with hidden variables and on other models.

The main contribution of this chapter is in the connection and implementation of algorithmic ideas for analytic approximation of marginal likelihood integrals, thus paving the way for practical use of these techniques. The presented procedures are implemented in Matlab and Maple and available at `http://www.technion.ac.il/~rusakov/`, along with commands that replicate all the results reported herein.

The rest of this chapter is organized as follows. Section 4.2 repeats some background on Bayesian networks and asymptotic approximations of marginal likelihood integrals. Sections 4.3 and 4.4 present and evaluate algorithms for asymptotic marginal likelihood evaluation for regular and singular statistics of Bayesian networks with hidden variables. We conclude with a discussion in Section 4.5.

4.2 Technical Background

Recall that a Bayesian network model M for n discrete variables $X = \{X_1, \ldots, X_n\}$, having $r_1, \ldots, r_n$ states respectively, is a set of joint distributions for X defined by a network structure G_M and a set of local distributions $\mathcal{F}_M$. A probability distribution $P(x)$ belongs to a Bayesian network model M if and only if it factors according to the network structure G_M via

$$P(X = x) = \prod_{i=1}^{n} p_i(X_i = x_i | Pa(X_i) = j), \tag{4.2}$$

where x is the n-dimensional vector of values of X, $Pa(X_i)$ denote parents of node X_i in G_M, j denote the values of $Pa(X_i)$ in x and p_i is a conditional distribution from $\mathcal{F}_M$.

We focus on discrete Bayesian networks, for which the local distributions $\mathcal{F}_M$ are multinomial distributions. We denote the *model* parameters defining $p_i(X_i = val_k | Pa(X_i) = j)$ by ω_{ijk} and the *joint space parameters* $P(X = x)$ by θ_x. The mapping that relates these parameters, derived from Eq. 4.2, is

$$\theta_{(x_1,\ldots,x_n)} = \prod_{i=1}^{n} \omega_{ijk}, \tag{4.3}$$

where k and j denote the assignment to X_i and $Pa(X_i)$ as dictated by $(x_1, \ldots, x_n)$. Given a Bayesian network model M, the number of model parameters is given by $ds = \prod_{i=1}^{n}(r_i - 1) \prod_{X_l \in Pa(X_i)} r_l$ and the number of joint space parameters is $dc = \prod_{i=1}^{n} r_i - 1$.

Consider now that in addition to the observable variables X, the model structure G_M includes additional, unobserved variables $H = (H_1, \ldots, H_m)$. In this case, model M consists of the distributions on X that can be parameterized via

$$P(X = x) = \sum_{H=h} \prod_{i=1}^{n} p_i(X_i = x_i|j) \prod_{l=1}^{m} p_l(H_l = h_l|j'),$$

where j and j' denote assignment given by x and h to the parents of X_i and H_l. We denote the nodes of G_M by $\tilde{X} = (\tilde{X}_1 = X_1, \ldots, \tilde{X}_n = X_n, \tilde{X}_{n+1} = H_1, \ldots, \tilde{X}_{n+m} = H_m)$ and we let ω_{ijk} be the model parameters of M defined for all network nodes $\tilde{X}$. Consequently, the joint space parameters on the observable nodes X are described by model parameters via the following formula:

$$\theta_{(x_1,\ldots,x_n)} = \sum_{(h_1,\ldots,h_m)} \prod_{i=1}^{n+m} \omega_{ijk}, \tag{4.4}$$

where k and j denote the assignment to $\tilde{X}_i$ and $Pa(\tilde{X}_i)$ as specified by assignments $(x_1, \ldots, x_n)$ to X and $(h_1, \ldots, h_m)$ to H.

A typical example of Bayesian network with hidden node is the naive Bayesian network, where one hidden (class) node H influences the observable (feature) nodes $X = (X_1, \ldots, X_n)$, as depicted on Figure 3.1. This network is used in clustering and classification (Cheeseman & Stutz, 1995).

As we have seen in Chapter 3 the marginal likelihood $P(D|M)$ can be written as a function of the averaged sufficient statistics $Y_D = (Y_1, \ldots, Y_{dc})$ (i.e., normalized counts) and the number N of samples in D, namely,

$$P(D|M) = \mathbb{I}[N, Y_D] = \int_{\Omega} e^{-Nf(\omega,Y_D)} \mu(\omega) d\omega, \tag{4.5}$$

where ω are the model parameters from the domain $\Omega \subset [0,1]^{ds}$, $\mu(\omega)$ is the prior parameter density and $f(\omega, Y_D) = -loglikelihood(Y_D|\omega) = -\sum_x Y_x \theta_x(\omega)$ is minus the log-likelihood function of model M. Variables $\theta_x(\omega)$ are indexed by vectors in $r_1 \times \ldots \times r_n$ and are specified by Eqs. 4.3 or 4.4.

When the model M is a Bayesian model without hidden variables, $\theta_x(\omega)$ is defined by Eq. 4.3, and the prior parameter density $\mu(\omega)$ is usually assumed to be Dirichlet. Consequently, Integral 4.5 can be evaluated in closed form (Cooper & Herskovits, 1992; Heckerman, Geiger & Chickering, 1995). When the prior parameter density is not known but supposed to be bounded, the approximation of $P(D|M)$ can be performed by bounding its value by two integrals using upper and lower bounds on μ. The two bounding integrals can then be evaluated resulting in $O(1)$ approximation to $\ln P(D|M)$, as specified by Eq. 4.1.

The evaluation of Integral 4.5 for M being a Bayesian network model with hidden variables is more difficult because it is complicated by two factors. First, some of the model parameters may be redundant yielding a surface of maximum likelihood points in the model parameter space Ω, rather than a single point. This can be tested by considering the Jacobian of Transformation 4.4, whose regular rank is the number of independent parameters of the model. Alternatively, this case can be checked directly by considering the set of solutions of polynomial equations for maximum likelihood model parameters. Recall that we call statistics that yield maximum likelihood points surfaces without self-crossings, *regular* statistics. Note that when the model has no redundant parameters all statistics are regular and there are only a finite number maximum likelihood points. (It is not a single point, due to symmetry in the states of the hidden nodes.)

In the neighborhood of each maximum likelihood point on the regular maximum likelihood surface the log-likelihood function can be approximated by a quadratic form on the independent variables. Consequently, using the classical Laplace approximation procedure, one obtains that $\ln P(D|M)$ is approximated by the standard BIC score, Eq. 4.1, with d being the number of independent parameters of M. The number of independent parameters of the model, called the *effective dimensionality*, is computed by evaluating the regular rank of Jacobian of Transformation 4.4 (Geiger, Heckerman & Meek, 1996). In Section 4.3, we present an optimized procedure for evaluating the effective dimensionality of an arbitrary Bayesian model M.

For some *singular* statistics, maximum likelihood surfaces intersect. Similarly, this can be tested by considering the Jacobian of Transformation 4.4, which is not of regular rank at the maximum likelihood surface intersection points (Geiger, Heckerman, King & Meek, 2001). In the neighborhoods of the intersection points the log-likelihood function can not be approximated by a quadratic form and the classical Laplace approximation of integrals fails. It was shown by Watanabe (2001) that an approximation for $\ln P(D|M)$ is given by

$$\ln P(D|M) = \ln P(D|\omega_{ML}) - \lambda \ln N + (m-1) \ln \ln N + O(1), \tag{4.6}$$

where λ is the rational number less then or equal to half the number of independent parameters of M and m is an integer greater or equal to 1. A classical Laplace approximation to $\ln P(D|M)$ corresponds to $\lambda = \frac{d}{2}$ and $m = 1$.

The values of λ and m can be computed in a systematic way by computing the maximum pole and its multiplicity of the function

$$J(\lambda) = \int_{\Omega_\epsilon} [f(\omega, Y_D) - f(\omega_{ML}, Y_D)]^\lambda \, d\omega \tag{4.7}$$

evaluated in small neighborhoods Ω_ϵ of maximum likelihood parameters $\omega_{ML} \in \Omega$. Computation of the integral $J(\lambda)$ can be performed by approximating the integrand function by a polynomial and changing ω to new coordinates in order to transform Integral 4.7 into a product of integrals of independent variables. The required change of coordinates is provided by the process of resolution of singularities from the field of algebraic geometry. Details are available in (Watanabe, 2001) and in Section 2.4.3.

This method has been applied in Chapter 3 to derive *analytic* asymptotic approximation of marginal likelihood for binary naive Bayesian model, showing that non-standard approximations to the marginal likelihood, namely, approximations with $\lambda \neq \frac{d}{2}$ and $m \neq 1$, do happen for Bayesian networks with hidden variables.

Watanabe's method is based on combining the asymptotic theory with the machinery of resolution of singularities in algebraic geometry. Recent advances in computational algebraic geometry (Bodnar & Schicho, 2000; Bravo, Encinas & Villamayor, 2002) made it possible to implement this method. Section 4.4 explicates the algorithmic details of our implementation which automatically produces analytic approximation of "hard" marginal likelihood integrals.

4.3 Automatic Effective Dimensionality Computations

In this section we describe our algorithm to compute the effective dimensionality of a given Bayesian network and evaluate its performance. The algorithm, shown on Figure 4.1, is based on the decomposition of a given Bayesian network into independent components and evaluation of the Jacobian rank of Transformation 4.4 for each sub-network. The actual Jacobian rank need only be evaluated for the Markov neighborhoods of each of the hidden nodes, thus preventing the construction of prohibitively large matrices as would be required by the standard approach (Geiger, Heckerman & Meek, 1996). This idea of decomposing the network has been influenced by the idea of model decomposition for models with isolated hidden nodes (Settimi & Smith, 1999).

Our algorithm is implemented in Matlab and it is available for research use as a Matlab package at `http://www.technion.ac.il/~rusakov/bndim/`. The input to the algorithm is a Bayesian network

Input:
G_M - Bayesian network structure.
$r = (r_1, \ldots, r_n)$ - Number of states of each node.
Algorithm:
1. Find disjoint Markovian neighborhoods of the hidden nodes.
2. For each neighborhood G_i do
   2.1. Set $ds_i = \prod_{\tilde{X}_j \in G_i} (r_j - 1) \prod_{\tilde{X}_k \in Pa_{G_i}(\tilde{X}_j)} r_k$.
   2.2. Construct Jacobian $\mathcal{J}$ of Transformation 4.4.
   2.3. Evaluate regular rank of $\mathcal{J}$ by substitution of 3 random values into $\mathcal{J}$ and evaluation of numerical rank.
   2.4. Effective dimensionality of G_i, de_i, is with probability 1 the maximal numerical rank of $\mathcal{J}$ computed in previous stage at 3 random points.
   2.5. Set $\Delta d_i = ds_i - de_i$.
3. Set $ds = \prod_{i=1}^{n} (r_i - 1) \prod_{\tilde{X}_j \in Pa(\tilde{X}_i)} r_j$.
4. Return: effective dimensionality is $ds - \sum_i \Delta d_i$.

Figure 4.1. Effective dimensionality computation algorithm.

structure, namely, the set of vertices and edges of a Bayesian network, and it does not require any manual coding of derivatives, in contrast to early computer aided computations of effective dimensionality presented in (Geiger, Heckerman & Meek, 1996).

The implemented procedure is useful for evaluating effective dimensionality of various Bayesian networks with hidden variables and as a research tool for testing closed form formulas for effective dimensionality of different types of latent models.

4.3.1 Evaluation: Naive Bayesian Models

We explored the space of naive Bayesian networks to find networks that have effective dimensionality less than the number of model parameters and less than the number of joint space parameters. We call such models *degenerate*. We have chosen naive Bayesian networks as a particular test case for our procedure, since these models, being the most simple of the latent models, have been investigated to some extent by previous works. In particular, we evaluate for each model an upper bound on the effective dimensionality derived by (Kocka & Zhang, 2002) using the results of (Settimi & Smith, 1998). We call this bound the KZ-bound and study its tightness on degenerate models that are found. The results suggest several conjectures regarding the tightness of the KZ-bound and the effective dimensionality of naive Bayesian models.

The space of naive Bayesian models is parameterized by the number of classes, number of features

Table 4.1: Some of the degenerate naive Bayesian models found by the effective dimensionality algorithm. Legend: de - effective dimensionality, *ds* - number of model parameters, *dc* - number of joint space parameters, kz - KZ-bound. The model is described by giving the number of hidden states followed by the number of states of the feature nodes. Dots indicate that all models with the last feature variable having more states than displayed are also degenerate. Models for which the KZ-bound is not tight are in bold.

NB model	de	*ds*	*dc*	kz	Time(sec)
3:2,2,4...	14	17	15	14	0.45
4:2,3,5...	27	31	29	27	1.26
5:2,3,6...	34	44	35	34	1.17
5:2,4,6...	44	49	47	44	1.59
6:2,4,7	53	65	55	53	2.53
6:2,5,7	65	71	69	65	3.33
6:3,3,7	59	65	62	59	2.92
4:3,3,3	**25**	27	26	26	1.23
5:3,4,4	**43**	44	47	47	2.82
7:3,5,5	**73**	76	74	74	4.64
10:3,7,7	**145**	149	146	146	27.69
3:2,2,2,2	**13**	14	15	14	0.65
5:2,2,3,3	**33**	34	35	34	2.92
6:2,2,2,7...	53	59	55	53	6.09

and number of states of each feature. First, we explore naive Bayesian networks with three feature nodes. The results of search for degenerate models among the models with less than or equal to 7 states in all variables are summarized in the upper part of Table 4.1. The notation (3:2,2,4) in this table denotes the naive Bayesian model with 3 hidden states and 3 feature nodes with 2, 2 and 4 states respectively. The search on these 336 different models took about 25 minutes on Matlab 6 R13, Windows XP Pro, Pentium 1.6MHz mobile, 512MB system. The average is 4.5 seconds per model and the actual time depends on the number of network and joint space parameters of the model, see Table 4.1. Another search, on all three-features naive Bayesian models with less than 10 states, took 17 hours and resulted in finding of 49 degenerate models in addition to 17 found by the previous search. Of these models, only one model has non-exact KZ-bound, as shown at the bottom of Table 4.1. Third, we explored naive Bayesian network with four feature nodes and less than 6 hidden states. The degenerate models found are also shown at the bottom of Table 4.1.

The results of this search suggests a number of conjectures on the behavior of effective dimensionality of naive Bayesian models and on the tightness of the KZ-bound. In particular, we observe that the model is always degenerate when it has more parameters than the number of joint space

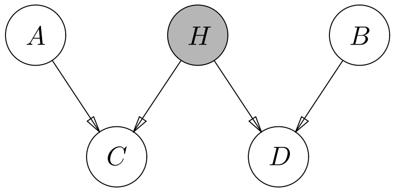


Figure 4.2. W-structure. Node H is hidden.

parameters and one of the feature nodes have more states than the hidden node. This is intuitive because in such cases the hidden node does not have enough states to pass the information from this special feature node to other feature nodes. Furthermore, in these cases, the effective dimensionality is correctly predicted by the KZ-bound. In contrast, in some cases, the model is degenerate even when all feature nodes have less states than the hidden nodes, and these cases are unpredictable by KZ-bound.

We have also explored naive Bayesian models with n feature nodes having the same number of states each. We have searched networks with up to 10 binary features nodes and up to 20 hidden states, and for 5 node networks with up to 5 states in hidden and feature nodes. The results of these searches are negative; Except for the known (3:2,2,2,2) network, we found no degenerate naive Bayesian networks with the same number of states in the feature nodes. We are tempted to speculate that this might be true in general.

4.3.2 Evaluation: Other Models

The applicability of our procedure is not limited to the naive Bayesian models. In this subsection we report effective dimensionality computations for a number of latent models considered by other authors.

First, we evaluate the effective dimensionality of the W-structure on binary nodes (Geiger, Heckerman & Meek, 1996), Figure 4.2. The evaluation of the effective dimensionality took 0.43 seconds and we confirm that the effective dimensionality of this model is 9, while the number of model parameters is 11 and the number of joint space parameters is 15. When the number of hidden states is 3 or more the effective dimensionality of the W-structure is 10 (tested for up to 100 hidden states).

Second, we considered the hierarchical latent class model with 5 binary feature nodes and 3 hidden nodes with 5, 3 and 3 states (Kocka & Zhang, 2002). The evaluation of the effective dimensionality took 40 seconds and we found that the effective dimensionality of this model is 23 while the number of model parameters is 41 and the number of joint space parameters is 31. This evaluation agrees with the theoretical result due to (Kocka & Zhang, 2002).

Finally, in order to demonstrate the efficiency of this approach we have evaluated the effective dimensionality of the ALARM network, when some of the nodes are made hidden. The ALARM network has been build by medical experts for monitoring patients in intensive care (Beinlich, Suermondt, Chavez & Cooper, 1989). This Bayesian network of 37 nodes connected by 46 edges consists of 509 model parameters and its joint parameter space contains $1.7 \cdot 10^{16}$ parameters. Disregarding the medical meaning of the nodes, we set nodes `KINKEDTUBE` and `CATECHOL` to be hidden and evaluated the effective dimensionality of the resulted network. It was evaluated to be 494 instead of 509 as the number of model parameters may suggest, and it took our program 18 seconds to compute this result.

4.4 Automatic Marginal Likelihood Approximation for Singular Statistics

In this section we describe our algorithm for the automatic evaluation of the coefficients λ and m in the asymptotic approximation of the logarithm of marginal likelihood, as described in Section 4.2. The algorithm, presented on Figure 4.3, follows and generalizes the method used in Chapter 3 for naive Bayesian networks and it uses the `desing` library for resolution of singularities developed by (Bodnar & Schicho, 2000).

The validity of the presented algorithm follows from Lemmas 10 and 11 presented in Section 3.4.2. These lemmas allows us to simplify the log-likelihood function (step 6) and approximate it by a simple polynomial p (step 7.1). This makes the subsequent resolution of singularities (step 7.2.1) possible and feasible. The validity of steps 7.2.1-3 is due to the results of (Watanabe, 2001).

This algorithm is implemented in the Maple computer system and available from the web-site `http://www.technion.ac.il/~rusakov/resolution/`. The package includes also auxiliary procedures for determining the maximum likelihood parameters, detection of the singular/regular statistics type, and approximation of the general marginal likelihood integrals given by polynomial log-likelihood

Input:

G_M - Bayesian network structure.

Y - averaged sufficient statistics of the sample.

Algorithm:

1. Find a single ω_{ML} - maximum likelihood model parameters given Y. Let $\theta_{ML} = \theta(w_{ML})$, Eq. 4.4.
2. Solve $\theta(\omega) = \theta_{ML}$ to get the set of all maximum likelihood parameters, $W_{ML} = \{\omega | \theta(\omega) = \theta_{ML}\}$.
3. Find the deepest singularity(ies) of W_{ML}, W_0.
4. If W_{ML} has no singularities, return $\lambda = -\left[ds - dim(W_{ML})\right]/2$ and $m = 1$.
5. If W_0 includes varieties of dimension greater than 0, choose a random representative point for each such variety.
6. Simplify θ as a function of ω by linear combination of existing θ.
7. For each ω_0 in W_0 do
   7.1. Consider polynomial $p = \sum[\theta(\omega) - \theta(\omega_0)]^2$.
   7.2. Evaluate asymptotic approximation coefficients λ and m of integral $I[N] = \int e^{-Np(\omega)} d\omega$:
      7.2.1. Resolve singularities at origin of variety defined by p using `desing` package.
      7.2.2. Use the mappings created by `desing` to find poles of integral $J(\lambda) = \int p^{\lambda} d\omega$.
      7.2.3. Greatest pole λ and its multiplicity m give the coefficients of $\ln N$ and $\ln \ln N$ in approximation to $I[N]$, Eq. 4.6.
8. Return the greatest pair of coefficients (λ, m) evaluated at step 7.

Figure 4.3. Algorithm for asymptotic approximation of the marginal likelihood.

functions (not necessarily arising from Bayesian networks).

4.4.1 Evaluation

Our algorithm was validated on a number of naive Bayesian models and on non Bayesian network models, where the log-likelihood function is a polynomial.

We have run the algorithm and found the asymptotic approximations for marginal likelihood for singular statistics given a naive Bayesian network, as specified in the upper part of Table 4.2. These results agree with the theoretical approximations as specified by Theorem 8, including the type 2 singularity of the two nodes network, which has a particular non-standard form (namely, a $\ln \ln N$ term in the approximation). The test's run time is between a few seconds to a few minutes to compute the approximation parameters for each network, depending on the size of the network. The time in consecutive runs may be significantly smaller because our algorithm implements hashing of the computed resolution results. On the larger networks, the time complexity of the resolution of singularities algorithm may become quite large.

Table 4.2: Experimental results with Algorithm 4.3. Legend: "Marginal likelihood" - the marginal likelihood under evaluation. In the upper part of the table specified by number of nodes in the given naive Bayesian model and type of given statistics. In the lower part specified directly by log-likelihood f function, Eq. 4.5. "λ, m" - asymptotic approximation parameters, Eq. 4.6.

Marginal likelihood	λ, m	Time
NB model, 1 feature node.	$-1/2$, 1	0.6 sec
NB, $n = 2$, Y - sing. type 1.	$-3/2$, 1	0.8 sec
NB, $n = 2$, Y - sing. type 2.	$-3/2$, 3	1.3 sec
NB, $n = 3$, Y - regular.	$-7/2$, 1	3.5 sec
NB, $n = 3$, Y - sing. type 1.	$-5/2$, 1	0.8 sec
NB, $n = 3$, Y - sing. type 2.	-2, 1	6.1 sec
NB, $n = 4$, Y - regular.	$-9/2$, 1	20 sec
NB, $n = 4$, Y - sing. type 1.	$-7/2$, 1	20 sec
NB, $n = 4$, Y - sing. type 2.	$-5/2$, 1	56 sec
NB, $n = 5$, Y - regular.	$-11/2$, 1	4.5 min
NB, $n = 5$, Y - sing. type 1.	$-9/2$, 1	4.6 min
NB, $n = 5$, Y - sing. type 2.	—	> 2 hrs.
$f = \sum_{i \neq j; i,j=1}^{3} w_i^2 w_j^2$	$-3/4$, 1	6 sec
$f = \sum_{i \neq j; i,j=1}^{4} w_i^2 w_j^2$	-1, 1	26 sec
$f = \sum_{i \neq j; i,j=1}^{5} w_i^2 w_j^2$	$-5/4$, 1	108 sec
$f = \sum_{i \neq j; i,j=1}^{6} w_i^2 w_j^2$	$-3/2$, 1	26 min

Steps 7.2.1-3 of algorithm 4.3 are implemented in a separate procedure which can be applied to evaluate the maximum likelihood of models with polynomial log-likelihood function. As an example of such computation we consider the asymptotic evaluation of $I[N] = \int_{W_\epsilon} e^{-N \sum_{i \neq j} w_i^2 w_j^2} dw$ in n variables $w_1, \ldots, w_n$. The evaluation of the asymptotic approximation of the logarithm of this integral (Table 4.2) agrees with the formula $\ln I[N] = -\frac{n}{4} \ln N + O(1)$ developed in Section 2.4.3.

4.5 Discussion

The contribution of this chapter is twofold. First, it generalizes, enhances and implements the effective dimensionality computations algorithm which is suitable for practical use. The usage of the algorithm is demonstrated and a number of previously unknown degenerate latent class models are reported. The current implementation is available on-line and has already been used by several researchers.

Second, we have described and implemented an algorithm that computes analytic asymptotic approximations of marginal likelihood integrals that can not be approximated by the classical formulas.

For this task we have established connections with algorithms from different fields of research. For now this algorithm is slow, due to the high time complexity of resolution of singularities procedure. Still, we were able to validate the algorithm on a number of latent class models. We view our algorithm and its implementation as an ongoing effort and we hope that the development of improved resolution of singularities algorithms will make our procedure more widely applicable. The current algorithm and its implementation can serve as a research tool to develop closed form solutions for non-standard marginal likelihood integrals.

Chapter 5

Parameter Priors for Discrete Bayesian Networks

In this chapter we investigate parameter priors for discrete Bayesian network models without hidden variables. It was shown in previous works that a Dirichlet prior on the parameters of a discrete Bayesian network is inevitable assuming global and local parameter independence for all possible complete network structures. A similar result for Gaussian network models hinted that the assumption of local independence may be redundant.

Herein, we prove that the local independence assumption is necessary in order to dictate a Dirichlet prior on the parameters of a discrete Bayesian network. We explicate the minimal set of assumptions needed to dictate a Dirichlet prior, and we derive the functional form of prior distributions that arise under the global independence assumption alone. A version of this chapter has been published as (Rusakov & Geiger, 2001).

5.1 Introduction

A directed graphical model is a representation of a family of joint probability distributions for a collection of random variables via a Directed Acyclic Graph (DAG). In particular, each node in the DAG corresponds to a random variable, and the lack of an edge between two nodes represents a conditional independence assumption. A specific joint probability distribution is represented by a given directed graphical model together with specific values for the set of associated parameters. The DAG along with such a distribution is called a Bayesian network.

Similar to previous chapters, we focus on learning - the process of updating both the parameters and the structure of a Bayesian network based on data (Buntine, 1994; Heckerman, Geiger & Chickering, 1995). An alternative to using a large sample approximations to marginal likelihood, that is required for Bayesian model selection, is a *closed form* evaluation of marginal likelihood, which is possible under the assumption of specific *conjugate* prior distributions on network parameters. In particular, for learning of Bayesian networks for a set of observed discrete random variables, the assumptions of global and local parameter independence for all network structures, Dirichlet distribution on network parameters, and some other technical assumptions were made (Spiegelhalter & Lauritzen, 1990; Cooper & Herskovits, 1992; Dawid & Lauritzen, 1993). It was later shown that the assumption of global and local parameter independence for all nodes in every complete network structure dictates that the only possible prior parameter distribution for discrete DAG models is a Dirichlet prior (Heckerman, Geiger & Chickering, 1995; Geiger & Heckerman, 1997).

In contrast, for Gaussian DAG models, which consist of a recursive set of linear regression models, it was shown that global independence alone dictates that the only feasible parameter prior is the Normal-Wishart distribution, assuming models with at least three nodes (Geiger & Heckerman, 2002). It was thus natural to hypothesize that the proofs for discrete and continuous case can be unified and, as a result, the assumption of local independence will turn out to be redundant also in the characterization of the Dirichlet distribution.

This chapter shows that, while global independence dictates a Normal-Wishart prior for Gaussian DAG models with more than 3 nodes, global independence alone does not dictate a Dirichlet prior for discrete DAG models with more than 3 nodes. We provide a minimal set of assumptions needed to dictate a Dirichlet prior and, in addition, we specify the class of discrete probability distributions, which is larger than the Dirichlet family, that arise under global independence assumption alone via a solution of a new set of functional equations.

The rest of this chapter is organized in the following way. Section 5.2 defines again the concept of Bayesian networks and specifies the assumptions made for learning particular network without hidden variables from data. In Section 5.3, the family of admissible parameter priors arising from our assumptions is analyzed for the simple two nodes network. Section 5.4 extends the analysis of admissible parameter priors for multiple nodes networks. In this section it is explicitly shown that the class of admissible prior parameter distribution for discrete Bayesian networks without hidden variables, under the global independence assumption, is larger than Dirichlet. The minimal set of assumptions required to guarantee a Dirichlet prior is explicated in Section 5.5, which is followed by a discussion of the results in Section 5.8. Most of the technical details and formal proofs are presented in Section 5.6.

5.2 Assessment of Parameter Priors for DAG Models

Let $m(s, \mathcal{F}_s)$ denote a DAG model for a set of discrete variables $X = \{X_1, \ldots, X_n\}$ each associated with a set of possible values D_i. In the discrete DAG model, the local distributions $p(x_i|\mathbf{pa}_i^s)$ are specified by multinomial parameters $\theta_m^i = \{\theta_{x_i|\mathbf{pa}_i^s}|x_i, \mathbf{pa}_i^s\}$. We assume that the θ_m^i's do not overlap, and that $\theta_m = \langle\theta_m^1, \theta_m^2, \ldots, \theta_m^n\rangle$. Recall that, the complete DAG models can describe any joint distribution on X and, therefore, such models can also be parameterized by the set of joint space parameters $\theta_X = \{\theta_x|x \in \mathbf{D}\}$, where $\theta_x = p(X = x)$.

Moreover, we can switch from one complete discrete DAG model parameterization to another and to the parameterization by the set of joint space parameters. Namely, for any two complete models m_1 and m_2 and for any probability density function (p.d.f.) $\mathfrak{p}_1$ on Θ_{m_1} the corresponding p.d.f. on Θ_{m_2} is

$$\mathfrak{p}_2(\theta_{m_2}) = J_{m_1\to m_2}(\theta_{m_2})\mathfrak{p}_1(\theta_{m_1}) = J^{-1}_{m_2\to m_1}(\theta_{m_1})\mathfrak{p}_1(\theta_{m_1}) = J_2(\theta_{m_2})\mathfrak{p}(\theta_X(\theta_{m_2})) \tag{5.1}$$

where $\theta_{m_1} = f_{12}^{-1}(\theta_{m_2})$, $J_{m_1\to m_2}$ and $J_{m_2\to m_1}$ are the Jacobian determinants of transformations from parameters of model m_1 to parameters of model m_2 (transformation f_{12}^{-1}) and vice versa (transformation f_{12}), and $J_2(\theta_{m_2})$ is the Jacobian determinant of $\theta_X(\theta_{m_2})$, which is the transformation from the joint space parameters to the parameters of m_2. These Jacobian determinants are related by the equality $J_{m_1\to m_2}(\theta_{m_2}) = J^{-1}_{m_2\to m_1}(\theta_{m_1}(\theta_{m_2})) = J_2(\theta_{m_2})J_1^{-1}(\theta_{m_1}(\theta_{m_2}))$, where $J_1(\theta_{m_1})$ is the Jacobian determinant of the transformation $\theta_X(\theta_{m_1})$ from the joint space parameters to the parameters of m_1. [1]

[1] We use the term "transformation" from x to y to denote the process of change of variables from x to y. Therefore the actual function that defines this transformation is the function from $Range(y)$ to $Range(x)$.

Given a complete network m with topological order of nodes $X_{i_1}, \ldots, X_{i_n}$, the Jacobian determinant of transformation from joint space parameters θ_X to network parameters θ_m is (Heckerman, Geiger & Chickering, 1995)

$$J_m = \prod_{j=1}^{n-1} \prod_{x_{i_1},\ldots,x_{i_j}} \left[\theta_{x_{i_j}|x_{i_1},\ldots,x_{i_{j-1}}}\right]^{[\prod_{l=j+1}^{n} r_{i_l}]-1}, \tag{5.2}$$

where r_i is the number of states of variable X_i. Note that the Jacobian determinant J_m is non-zero for all parameters inside the domain of θ_m.

In the Bayesian approach for learning of Bayesian networks from data, the parameters θ_m and the model hypothesis m_h are uncertain but the parametric families are known. To compute the posterior probability of a model hypothesis m_h in closed form one must access an a priori probability distribution for the parametric family Θ_m. The specific form of prior distribution can be either assumed directly, e.g., Dirichlet, or indirectly, by requiring some properties from prior distribution. Two common properties that are requested from prior distributions are *global* and *local* parameter independence (Geiger & Heckerman, 2002).

Note that a prior distribution on joint space parameters, $p(\theta_X)$ induces the distributions of local parameters for each complete model $p(\theta_m|m)$ via the change of parameters formula (Eq. 5.1).

Assumption 1 (Global parameter independence) *The prior parameter distribution $p(\theta_X)$ induce such parameter priors on every complete DAG model $m(s, \mathcal{F}_s)$ for X that parameters θ_m are* globally independent, *which means that parameters θ_m^i of every node X_i are mutually independent, namely, $p(\theta_m|m) = \prod_{i=1}^{n} p(\theta_m^i|m)$ where $\theta_m^i = \{\theta_{i|\mathbf{pa}_i^s}|\mathbf{pa}_i^s\}$.*

Assumption 2 (Local parameter independence) *The prior parameter distribution $p(\theta_X)$ induce such parameter priors on every complete DAG model $m(s, \mathcal{F}_s)$ for X that parameters θ_m^i of every node X_i are* locally independent, *which means that subsets $\{\theta_{x_i|\mathbf{pa}_i^s}|x_i\}$ of θ_m^i are mutually independent for different values of $\mathbf{pa}_i^s$, namely, $p(\theta_m^i|m) = \prod_{\mathbf{pa}_i^s \in \mathbf{D}_{\downarrow \mathbf{Pa}_i^s}} p(\{\theta_{x_i|\mathbf{pa}_i^s}|x_i\}|m)$.*

We refer to the joint space parameters θ_X as being globally (or locally) independent, if their prior distribution satisfies global (local) independence assumption for all complete networks on X. We explicitly assume the regularity of parameter distributions.

Assumption 3 (Parameter Priors Regularity) *The probability distribution functions $p(\theta_X)$ on the joint space parameters and the corresponding p.d.f.s on model parameters are positive and twice differentiable.*

Heckerman, Geiger and Chickering (1995) prove the following theorem that links between parameter independence and Dirichlet distribution:

Theorem 14 (Heckerman, Geiger & Chickering, 1995) *Let m_1 and m_2 be two complete discrete DAG models for X with variable orderings $(X_1, \ldots, X_n)$ and $(X_n, X_1, \ldots, X_{n-1})$, respectively. If parameters of m_1 and m_2 satisfy global independence (A1) and local independence (A2) then $p(\theta_X)$ is Dirichlet, where θ_X are joint space multinomial parameters for X.*

It was suggested in (Geiger & Heckerman, 1997) and (Geiger & Heckerman, 2002) that a Dirichlet prior distribution may arise even under a weaker set of conditions, say under the assumption of global independence for all complete networks, similarly to the result for Gaussian DAG models (Geiger & Heckerman, 2002).

In the following sections we investigate prior distributions for discrete DAG models that satisfy global independence requirement, and show that this class is strictly larger than class of Dirichlet distributions for all discreet models. In addition, we specify the minimal set of global and local independence assumptions that is needed to ensure the Dirichlet prior.

5.3 Globally Independent Priors for Two Node Networks

We commence by analyzing the globally independent prior distributions for parameters of complete two node network. The results and techniques developed in this section will be the basis for advanced derivations in the case of multiple node networks.

Consider the following complete two node network, as shown on Figure 5.1, with variables X,Y having n and k states accordingly. This network is complete and it is capable of describing any multinomial distribution of two random variables. Any multinomial distribution, which is described by a set of parameters $\{\theta_{ij}\}$ for $1 \leq i \leq n$, $1 \leq j \leq k$, can be described by this network by specifying $\theta_{i\cdot} = \sum_{j=1}^{k} \theta_{ij}$ and $\theta_{j|i} = \theta_{ij}/\theta_{i\cdot}$ for $1 \leq i \leq n$ and $1 \leq j \leq k$.

We are interested in finding a functional form of prior distributions $p(\{\theta_{ij}\})$ that satisfy a global independence assumption for all complete network for $\{X, Y\}$, namely, for the networks $X \rightarrow Y$

Figure 5.1. Two node complete DAG model for random variables X and Y.

(shown above) and $X \leftarrow Y$. Such distributions should satisfy the following two functional equations, that describe global independence assumption:

$$
\begin{aligned}
p(\{\theta_{ij}\}) &= J_1^{-1}(\{\theta_{i\cdot}, \theta_{j|i}\}) \cdot f_1(\{\theta_{i\cdot}\}) \cdot g_1(\{\theta_{j|i}\}) \\
p(\{\theta_{ij}\}) &= J_2^{-1}(\{\theta_{\cdot j}, \theta_{i|j}\}) \cdot f_2(\{\theta_{\cdot j}\}) \cdot g_2(\{\theta_{i|j}\})
\end{aligned}
\tag{5.3}
$$

where J_1, J_2 are the Jacobian determinants from transformations of joint space parameters to the parameters of first and second networks (Eq. 5.2), and where the parameters $\theta_{\cdot j}$, $\theta_{i|j}$ of the second network ($X \leftarrow Y$) are defined similarly to the parameters $\theta_{i\cdot}$ and $\theta_{j|i}$ of the first network ($X \rightarrow Y$).

We obtain the following result.

Theorem 15 *Any distribution on joint space parameters $\{\theta_{ij}\}$ that satisfies the regularity assumption (A3) and the global independence assumption (A1) for all complete DAG models for $\{X, Y\}$, i.e., satisfies Eq. 5.3, is of the form*

$$
p(\{\theta_{ij}\}) = C \left[\prod_{i=1}^{n} \prod_{j=1}^{k} \theta_{ij}^{\alpha_{ij}} \right] H\left(\left\{ \frac{\theta_{ij}\theta_{i+1,j+1}}{\theta_{i+1,j}\theta_{i,j+1}} | 1 \leq i \leq n-1, 1 \leq j \leq k-1 \right\} \right) \tag{5.4}
$$

where α_{ij} are arbitrary constants, $H()$ is an arbitrary positive, twice-differentiable function of $(n-1) \cdot (k-1)$ variables, and C is a normalization constant.

Theorem 15 implies that for two node discrete DAG models global independence assumption alone does not guarantee Dirichlet prior distribution of network parameters. Section 5.4 will present the general result for a multiple node discrete DAG model.

5.3.1 The Approach to the Proof of Theorem 15

The proof of Theorem 15 is based on the direct solution of Eq. 5.3 and is explicated in Section 5.6.2. In this section we transform Eq. 5.3 to a simpler form.

Let the variables $\{y_i | 1 \leq i \leq n-1\}$ and $\{z_{ji} | 1 \leq i \leq n, 1 \leq j \leq k-1\}$ correspond to the sets

$\{\theta_{i\cdot}\}$ and $\{\theta_{j|i}\}$ of network parameters of Eq. 5.3. We define,

$$\begin{aligned} &y_n = 1 - \textstyle\sum_{i=1}^{n-1} y_i; && z_{ki} = 1 - \textstyle\sum_{j=1}^{k-1} z_{ji}, \quad 1 \le i \le n \\ &x_j = \textstyle\sum_{i=1}^{n} z_{ji} y_i, \quad 1 \le j \le k; && w_{ji} = \tfrac{z_{ji} y_i}{x_j}, \quad 1 \le j \le k,\ 1 \le i \le n. \end{aligned} \tag{5.5}$$

Note that $x_k = 1 - \sum_{j=1}^{k-1} x_j$ and $w_{jn} = 1 - \sum_{i=1}^{n-1} w_{ji}$ (for $1 \le j \le k$). Here, $\{x_j\}$ corresponds to $\{\theta_{\cdot j}\}$ and $\{w_{ji}\}$ corresponds to $\{\theta_{i|j}\}$. We also use the following notation

$$\begin{aligned} &Y = (y_1, \ldots, y_{n-1}), \quad && Z_i = (z_{1,i}, \ldots, z_{k-1,i}), \quad && Z = (Z_1, \ldots, Z_n), \\ &X = (x_1, \ldots, x_{k-1}), && W_j = (w_{j,1}, \ldots, w_{j,n-1}), && W = (W_1, \ldots, W_k). \end{aligned} \tag{5.6}$$

Thus, for example,

$$g(Z) \equiv g(z_{1,1}, \ldots, z_{k-1,1}, \ldots, z_{1,i}, \ldots, z_{k-1,i}, \ldots, z_{1,n}, \ldots, z_{k-1,n})$$

and

$$f(W) \equiv f(\frac{z_{11}y_1}{x_1}, \ldots, \frac{z_{1,n-1}y_{n-1}}{x_1}, \ldots, \frac{z_{j1}y_1}{x_j}, \ldots, \frac{z_{j,n-1}y_{n-1}}{x_j}, \ldots, \frac{z_{k1}y_1}{x_k}, \ldots, \frac{z_{k,n-1}y_{n-1}}{x_k}).$$

Eq. 5.3 can now be expressed as

$$J_1^{-1}(Y, Z) \cdot F(Y) \cdot g(Z) = J_2^{-1}(X, W) \cdot G(X) \cdot f(W). \tag{5.7}$$

We use the fact that $J_1(Y, Z) = \prod_{i=1}^{n} y_i^{k-1}$ and $J_2(X, W) = \prod_{j=1}^{k} x_j^{n-1}$ (Eq. 5.2), and, therefore, J_1 and J_2 depend only on Y and X, respectively. Let $\hat{F} = \ln(J_1^{-1} \cdot F)$, $\hat{g} = \ln g$, $\hat{G} = \ln(J_2^{-1} \cdot G)$ and $\hat{f} = \ln f$. Taking the logarithm of Eq. 5.8, which is possible under Assumption 3, we get

$$\hat{F}(Y) + \hat{g}(Z) = \hat{G}(X) + \hat{f}(W). \tag{5.8}$$

Note that the only free variables in Eq. 5.8 are $y_1, \ldots, y_{n-1}$ and z_{ji}, $1 \le j \le k-1$, $1 \le i \le n$. All other variables appearing in Eq. 5.8 are defined by Eqs. 5.5 and 5.6. Alternatively, we can also consider any y_{j_1} to be a dependent variable instead of y_n as long as $\sum_{j=1}^{n} y_j = 1$, in which case we remain with the same functional equation, and, similarly, we can consider $z_{j_1 i}$ to be dependent variables instead of z_{ki}, as long as $\sum_{j=1}^{k} z_{ji} = 1$. These observations are particularly apparent when recalling the probabilistic origin of this equation. Furthermore, we may consider $x_1 \ldots, x_{k-1}$ and $w_{ji} = z_{ji} y_i / x_j$, $1 \le i \le n-1$, $1 \le j \le k$ to be free variables and Y and Z dependent. This fact allows us to use any form of solution for $\hat{F}$ or $\hat{g}$ as a solution for $\hat{G}$ or $\hat{f}$, and vice versa.

The solution of Eq. 5.8, which is presented in Section 5.6.2, uses Assumption 3 and is based on the technique of reduction of functional equations to partial differential equations (Aczel, 1966; page 324).

5.4 Globally Independent Priors for Multiple Node Networks

Consider a complete DAG model m on n discrete nodes $X = \{X_1, \ldots, X_n\}$, each having $r_1, \ldots, r_n$ values respectively. Such models describe in various ways, according to the factorization given by the network structure, a set of multinomial joint space parameters $\theta_X = \{\theta_{x_1,\ldots,x_n} | \langle x_1, \ldots, x_n\rangle\}$ via a set of local multinomial parameters of model $\theta_m = \{\theta_{i|\mathbf{pa}_i^s} | i, \mathbf{pa}_i^s\}$. In this section we generalize the results of Section 5.3 and investigate the prior distributions on the joint space parameters θ_X (and therefore on θ_m) that satisfy the global independence assumption (A1).

Assumption 1 means that $p(\theta_X)$ must satisfy the following $n!$ functional equations:

$$p(\theta_X) = J_I^{-1} \cdot \prod_{j=1}^{n} f_{I,j}(\{\theta_{x_{i_j}|x_{i_1},\ldots,x_{i_{j-1}}}\}), \quad \text{for all } I = \langle i_1, \ldots, i_n\rangle \text{ permutations on } \langle 1, \ldots, n\rangle, \tag{5.9}$$

where $f_{I,j}()$ are some functions on $\{\theta_{x_{i_j}|x_{i_1},\ldots,x_{i_{j-1}}} | \langle x_{i_1}, \ldots, x_{i_j}\rangle\}$ expressed in terms of θ_X, and J_I denotes the Jacobian determinant of the transformation from joint space parameters to the parameters of the complete Bayesian network with topological order of nodes specified by I. These Jacobian determinants can be incorporated inside $f_{I,j}$, since they are a function of $\{\theta_{x_{i_j}|x_{i_1},\ldots,x_{i_{j-1}}}\}$, Eq. 5.2.

In order to solve Eq. 5.9 we use the solution for two nodes networks as given by Theorem 15. Consider two discrete random variables $\mathbf{Y}_i = \{Y_i, Y\}$, where $Y_i = X_i$ and $Y = X_1 \times \ldots \times X_{i-1} \times X_{i+1} \times \ldots \times X_n$. Let us denote the joint space parameters of distribution on X by θ_X and the joint space parameters of $\mathbf{Y}_i$ by $\theta_{\mathbf{Y}_i}$. We claim the following

Lemma 16 *Given that $p(\theta_X)$ satisfies the regularity assumption (A3), then θ_X are globally independent (A1) if and only if $\theta_{\mathbf{Y}_i}$ are globally independent (A1) for all $i = 1, \ldots, n$.*

The proof of this lemma is given in Section 5.6.3.

Figure 5.2: An example of the relation between a three nodes complete network and the corresponding two nodes network. The parameters of model M_1 are globally independent if and only if the parameters of model M_2 are globally independent.

Applying Lemma 16 and Theorem 15, we conclude that any $p(\theta_X)$ that satisfies global independence assumption (A1) should satisfy the following n equations (for $i = 1, \ldots, n$):

$$p(\theta_X) = C_i \cdot \left[\prod_{x\in\mathbf{D}} \theta_x^{\alpha_{x,i}}\right] \cdot H_i\left(\left\{\frac{\theta_r\theta_{r'''}}{\theta_{r'}\theta_{r''}} | r, r', r'', r''' \in \mathbf{D}, \text{ s.t. } \begin{array}{c} r'_i = r'''_i = r_i + 1 = r''_i + 1 \\ r''_{\bar{i}} = r'''_{\bar{i}} = r_{\bar{i}} + 1 = r'_{\bar{i}} + 1 \end{array}\right\}\right) \tag{5.10}$$

where r_i and $r_{\bar{i}}$ denote the values of vector r on X_i and $X \setminus X_i$, respectively and the $''+1''$ operation denotes the operation of taking the next value of the corresponding vector in any fixed order. (We assume here alphabetical order for convenience, but the sets of parameters of H are equivalent under all orders.)

For example, consider globally independent parameter priors for Bayesian networks on three binary nodes $\{X_1, X_2, X_3\}$. According to Lemma 16 such parameter priors should result in the globally independent parameter priors for complete networks on $\{X_1, X_2 \times X_3\}$, $\{X_2, X_1 \times X_3\}$ and $\{X_3, X_1 \times X_2\}$. Applying Theorem 15 for these three networks, we get that any globally independent prior for $\{X_1, X_2, X_3\}$ satisfy

$$\begin{aligned}
p(\theta_X) &= C_1 \cdot \left[\prod_{x\in\{0,1\}^3} \theta_x^{\alpha_{x,1}}\right] \cdot H_1\left(\tfrac{\theta_{000}\theta_{101}}{\theta_{100}\theta_{001}}, \tfrac{\theta_{001}\theta_{110}}{\theta_{101}\theta_{010}}, \tfrac{\theta_{010}\theta_{111}}{\theta_{110}\theta_{011}}\right) && \text{from gl. ind. for } \{X_1, X_2 \times X_3\}, \\
p(\theta_X) &= C_2 \cdot \left[\prod_{x\in\{0,1\}^3} \theta_x^{\alpha_{x,2}}\right] \cdot H_2\left(\tfrac{\theta_{000}\theta_{011}}{\theta_{010}\theta_{001}}, \tfrac{\theta_{001}\theta_{110}}{\theta_{011}\theta_{100}}, \tfrac{\theta_{100}\theta_{111}}{\theta_{110}\theta_{101}}\right) && \text{from gl. ind. for } \{X_2, X_1 \times X_3\}, \\
p(\theta_X) &= C_3 \cdot \left[\prod_{x\in\{0,1\}^3} \theta_x^{\alpha_{x,3}}\right] \cdot H_3\left(\tfrac{\theta_{000}\theta_{011}}{\theta_{001}\theta_{010}}, \tfrac{\theta_{010}\theta_{101}}{\theta_{011}\theta_{100}}, \tfrac{\theta_{100}\theta_{111}}{\theta_{101}\theta_{110}}\right) && \text{from gl. ind. for } \{X_3, X_1 \times X_2\}.
\end{aligned} \tag{5.11}$$

The correspondence between three nodes networks and the corresponding two nodes networks is illustrated in Figure 5.2.

Lemma 16 states that all globally independent priors on θ_X are solutions of Eq. 5.10. This equation is solved by the method presented in Section 5.3.1, namely, by transforming all the multiplications in the equation into sums by taking the logarithm of the equation and its variables (possible due to the prior regularity assumption A3), and the application of the following lemma:

Lemma 17 *Consider the system of m functional equations:*

$$\begin{array}{l} f(x_1,\ldots,x_n) = \sum_{i=1}^n \alpha_{1i}x_i + h_1(\vec{b}_{11}\vec{x},\vec{b}_{12}\vec{x},\ldots,\vec{b}_{1k_1}\vec{x}) \\ f(x_1,\ldots,x_n) = \sum_{i=1}^n \alpha_{2i}x_i + h_2(\vec{b}_{21}\vec{x},\vec{b}_{22}\vec{x},\ldots,\vec{b}_{2k_2}\vec{x}) \\ \vdots \\ f(x_1,\ldots,x_n) = \sum_{i=1}^n \alpha_{mi}x_i + h_m(\vec{b}_{m1}\vec{x},\vec{b}_{m2}\vec{x},\ldots,\vec{b}_{mk_m}\vec{x}) \end{array} \tag{5.12}$$

where $f, h_1, \ldots, h_m$ are unknown functions, α_{ji} unknown constants and $\vec{b}_{ji}$ arbitrary (given) n-dimensional vectors. Eq. 5.12 expresses the fact that function f can be represented as a linear sum of its variables plus arbitrary function of any given form: $h_1, \ldots, h_m$. For applications in this paper, $\vec{b}_{ji} \in \{-1, 0, 1\}^n$ and $k_1 = k_2 = \ldots = k_m$.

The general solution for f in Eq. 5.12 is

$$f(x_1,\ldots,x_n) = \sum_{i=1}^n \alpha_i x_i + h(\vec{b}_1\vec{x},\ldots,\vec{b}_l\vec{x}), \tag{5.13}$$

where h is an arbitrary function, $\{\alpha_i\}$ are arbitrary constants and $\vec{b}_1, \ldots, \vec{b}_l$ is the basis of the linear space $\bigcap_{i=1}^m B_i$, where B_i is a linear space spanned by $\vec{b}_{i1}, \ldots, \vec{b}_{ik_i}$.

The proof of this lemma is given in Section 5.6.4.

In the following subsections we give an exact functional form of globally independent prior distributions for binary valued network and show the class of globally independent distributions for general networks. Our results demonstrate that the global independence assumption alone is not sufficient to ensure a Dirichlet prior.

5.4.1 Binary-Valued Networks

Consider n binary-valued random variables: $X = \{X_1, \ldots, X_n\}$, that can be described by 2^n variables, $\{\theta_{\vec{x}}\}_{\vec{x}\in\{0,1\}^n}$ (of them $2^n - 1$ are independent). The following theorem specifies the functional form of any distribution on θ_X that satisfies the global independence assumption.

Theorem 18 *Any prior distribution of θ_X, where $X = \{X_1, \ldots, X_n\}$ are binary random variables, that satisfies the regularity and the global independence assumptions (A1,A3) for X, i.e., satisfies Eq. 5.9, is of the form*

$$p(\theta_X) = C \cdot \left[\prod_{x \in \{0,1\}^n} \theta_x^{\alpha_x} \right] \cdot H \left(\frac{\prod_{x \in \{0,1\}^n,\ x \text{ contains even number of ones}} \theta_x}{\prod_{x \in \{0,1\}^n,\ x \text{ contains odd number of ones}} \theta_x} \right) \tag{5.14}$$

where H is an arbitrary positive, twice-differentiable function, α_x are arbitrary constants and C is a normalization constant.

The proof is given in Section 5.6.5. For example, the prior for three nodes binary networks (solution of Eq. 5.11) is

$$p(\{\theta_x\}_{x \in \{0,1\}^3}) = C \cdot \left[\prod_{x \in \{0,1\}^3} \theta_x^{\alpha_x} \right] \cdot H \left(\frac{\theta_{000}\theta_{011}\theta_{101}\theta_{110}}{\theta_{001}\theta_{010}\theta_{100}\theta_{111}} \right)$$

and the prior for four node binary networks is

$$p(\{\theta_x\}_{x \in \{0,1\}^4}) = C \cdot \left[\prod_{x \in \{0,1\}^4} \theta_x^{\alpha_x} \right] \cdot H \left(\frac{\theta_{0000}\theta_{0011}\theta_{0101}\theta_{0110}\theta_{1001}\theta_{1010}\theta_{1100}\theta_{1111}}{\theta_{0001}\theta_{0010}\theta_{0100}\theta_{1000}\theta_{0111}\theta_{1011}\theta_{1101}\theta_{1110}} \right).$$

The latest is a concrete example showing that for discrete DAG models on more than three nodes the global independence assumption alone does not dictate the Dirichlet prior. This stands in contrast to a similar result for Gaussian DAG networks, where global independence alone dictates a conjugate (Normal-Wishart) prior (Geiger & Heckerman, 2002).

5.4.2 The General Network

As the number of nodes and their values grows, the functional equations that describe admissible distributions for arbitrary n-node network becomes more complex. For any specific network, however, the globally independent prior distribution can be computed using Eq. 5.10 and Lemma 17.

The following theorem generalizes the solution for binary valued networks. We believe that it states the unique solution of Eq. 5.9, but we are able to prove only that it is a valid solution.

Theorem 19 *Consider n discrete random variables $X = \{X_1, \ldots, X_n\}$ receiving values from $\mathbf{D} = D_1 \times \ldots \times D_n$, and described by joint space parameters $\theta_X = \{\theta_x | x \in \mathbf{D}\}$.*

Any $f(\theta_X)$ of the form given below gives a globally independent prior on θ_X:

$$f(\theta_X) = C \cdot \left[\prod_{x \in \mathbf{D}} \theta_x^{\alpha_x} \right] \cdot H \left(\{a_{x_1,x_2}(\theta_x) | x_1, x_2 \in \mathbf{D},\ \forall i, 1 \leq i \leq n, x_1[i] \neq x_2[i]\} \right) \tag{5.15}$$

where C is a normalization constant and H is an arbitrary function with arguments

$$a_{x_1,x_2}(\theta_x) = \frac{\prod_{x \in A_{0,x_1,x_2}} \theta_x}{\prod_{x \in A_{1,x_1,x_2}} \theta_x}, \tag{5.16}$$

where $A_{0,x_1,x_2} = \{x \in \mathbf{D}\ |x \text{ contains even number of entries from } x_1, \text{ the rest from } x_2\}$ *and where* $A_{1,x_1,x_2} = \{x \in \mathbf{D}\ |x \text{ contains odd number of entries from } x_1, \text{ the rest from } x_2\}$. *Namely, the arguments for H are of the same form as in Eq. 5.14, with vectors x_1 and x_2 playing roles of zeros and ones. (Note that not all of $a_{x_1,x_2}(\theta_x)$ are independent.)*

Proof: The theorem follows directly from the fact that Eq. 5.15 satisfies Eq. 5.9. ■

Eq. 5.15 provides a form (with inter-dependencies among parameters of H) of the admissible distributions for a general network under global independence assumption. For any specific set of X that can have values from $\mathbf{D}$ the symbolic form of admissible distributions can be specified by solving the appropriate form of Eq. 5.10 using Lemma 17. For example, for three ternary variables X_1, X_2, X_3 global independence on all complete network structures dictates a parameter prior of the form:

$$f(\{\theta_x\}) = C \cdot \prod_{x \in \{1,2,3\}^3} \theta_x^{\alpha_x} \cdot H \left(\begin{array}{c} \frac{\theta_{111}\theta_{122}\theta_{212}\theta_{221}}{\theta_{112}\theta_{121}\theta_{211}\theta_{222}}, \frac{\theta_{121}\theta_{132}\theta_{222}\theta_{231}}{\theta_{122}\theta_{131}\theta_{221}\theta_{232}}, \frac{\theta_{111}\theta_{122}\theta_{312}\theta_{321}}{\theta_{112}\theta_{121}\theta_{311}\theta_{322}}, \frac{\theta_{121}\theta_{132}\theta_{322}\theta_{331}}{\theta_{122}\theta_{131}\theta_{321}\theta_{332}}, \\ \\ \frac{\theta_{112}\theta_{123}\theta_{213}\theta_{222}}{\theta_{113}\theta_{122}\theta_{212}\theta_{223}}, \frac{\theta_{122}\theta_{133}\theta_{223}\theta_{232}}{\theta_{123}\theta_{132}\theta_{222}\theta_{233}}, \frac{\theta_{112}\theta_{123}\theta_{313}\theta_{322}}{\theta_{113}\theta_{122}\theta_{312}\theta_{323}}, \frac{\theta_{122}\theta_{133}\theta_{323}\theta_{332}}{\theta_{123}\theta_{132}\theta_{322}\theta_{333}} \end{array} \right)$$

Note that in terms of Eq. 5.15, the arguments of h are $a_{111,222}$, $a_{121,232}$, $a_{111,322}$, $a_{121,332}$, $a_{112,223}$, $a_{122,233}$, $a_{112,323}$ and $a_{122,333}$. The matlab code for computing the symbolic form of such distributions for an arbitrary X is provided in Section 5.7.

5.5 Dirichlet Priors: The Minimal Set of Assumptions

We have shown in previous sections that global independence alone is not sufficient to ensure a Dirichlet prior on the network parameters. The natural question is: "What is the minimal set of independence requirements that dictate a Dirichlet prior?". This section gives the answer to this question. We start by providing some additional results that link between global independence in various networks.

Definition [global independence for node parameters] *Parameters $\theta_m^i = \{\theta_{x_i|\mathbf{pa}_i^s}\}$ of node X_i in a DAG model $m(s, \mathcal{F}_s)$ are said to be* globally independent *if $p(\theta_m|m) = p(\theta_m^i|m)p(\theta_m \setminus \theta_m^i|m)$.*

Lemma 20 *Let m_1 be a complete n-node network with topological order of nodes $X_{i_1}, \ldots, X_{i_n}$, $\{i_1, \ldots, i_n\} = \{1, \ldots, n\}$ and let m_2 be another complete network, with order $X_{j_1}, \ldots, X_{j_n}$ where $\{j_1, \ldots, j_n\} = \{1, \ldots, n\}$. Given $i_k = j_k$ and $\{i_1, \ldots, i_{k-1}\} = \{j_1, \ldots, j_{k-1}\}$, it holds that $\theta_{m_1}^{i_k}$ are globally independent if and only if $\theta_{m_2}^{j_k}$ are globally independent.*

Proof: Since m_1, m_2 are arbitrary, only one direction of the "iff" need to be proved. Suppose $\theta_{m_1}^{i_k}$ are globally independent, then

$$p(\theta_{m_2}|m_2) = J^{-1}_{\theta_{m_1} \to \theta_{m_2}} p(\theta_{m_1}|m_1) = J^{-1}_{\theta_{m_1} \to \theta_{m_2}} p(\theta_{m_1}^{i_k}|m_1) p(\theta_{m_1} \setminus \theta_{m_1}^{i_k}|m_1), \tag{5.17}$$

where $\theta_{m_1} = f(\theta_{m_2})$ is an appropriate transformation, and $J_{\theta_{m_1} \to \theta_{m_2}}$ is the Jacobian determinant of this transformation. Since X_{j_k} have the same parents in both networks, then $f_{\downarrow \theta_{m_1}^{i_k}}(\theta_{m_2}) = \theta_{m_2}^{j_k}$, and $f_{\downarrow \theta_{m_1} \setminus \theta_{m_1}^{i_k}}(\theta_{m_2})$ depends only on the parameters $\theta_{m_2} \setminus \theta_{m_2}^{j_k}$. Therefore the Jacobian determinant $J_{\theta_{m_1} \to \theta_{m_2}}$ depends only on $\theta_{m_2} \setminus \theta_{m_2}^{j_k}$. We have:

$$p(\theta_{m_2}|m_2) = f_1(\theta_{m_2}^{j_k}) f_2(\theta_{m_2} \setminus \theta_{m_2}^{j_k}) \tag{5.18}$$

where $f_1(\theta_{m_2}^{j_k}) = p(\theta_{m_2}^{j_k}|m_1)$ and $f_2(\theta_{m_2} \setminus \theta_{m_2}^{j_k}) = J_{\theta_{m_1} \to \theta_{m_2}} p(f_{\downarrow \theta_{m_1} \setminus \theta_{m_1}^{i_k}}(\theta_{m_2} \setminus \theta_{m_2}^{j_k})|m_1)$. Thus θ_{m_2} are globally independent. ■

We now present the main result of this section:

Theorem 21 *Let $X = \{X_1, \ldots, X_n\}$ be a set of random variables receiving values from $\mathbf{D} = D_1 \times \ldots \times D_n$. Let $m_1(s_1, \mathcal{F}_{s_1})$ be an arbitrary, complete DAG model for X with topological order of nodes $X_{i_1}, \ldots, X_{i_n}$, and let $m_2(s_2, \mathcal{F}_{s_2})$ be another complete DAG model for X, with order $X_{j_1}, \ldots, X_{j_n}$, such that $j_n = i_1$.*

If $p(\theta_x)$ satisfies the regularity assumptions (A3) and

- **[MA1]** *Parameters of X_{i_1} in m_1 are globally independent, i.e.,*

$$p(\theta_{m_1}|m_1) = p(\theta_{m_1}^{i_1}|m_1) p(\theta_{m_1} \setminus \theta_{m_1}^{i_1}|m_1) \tag{5.19}$$

- **[MA2]** *Parameters of X_{j_n} in m_2 are globally and locally independent, i.e.,*

$$p(\theta_{m_2}|m_2) = p(\theta_{m_2} \setminus \theta_{m_2}^{j_n}|m_2) \prod_{\mathbf{pa}_{j_n}^{s_2} \in \mathbf{D}_{\downarrow \mathbf{Pa}_{j_n}^{s_2}}} p(\theta_{X_{j_n}|\mathbf{pa}_{j_n}^{s_2}}|m_2), \tag{5.20}$$

where $\theta_{X_i|\mathbf{pa}_i^s} = \{\theta_{x_i|\mathbf{pa}_i^s}|x_i\}$, *then $p(\theta_X)$ is Dirichlet and the conditions [MA1], [MA2] are minimal in sense that the elimination of any one of these two conditions extends the class of admissible priors beyond Dirichlet.*

The theorem states that among all the global and local independence assumptions used by previous authors, one actually need only two assumptions: global independence for the network parameters for the first node in some network, and global and local independence for the same node in other network where this node is the last node.

Proof: The minimality of these two assumptions is straightforward, since eliminating any of them will allow any distribution of the form 5.19 or 5.20. Using Lemma 20 we assume without the loss of generality that two DAG models under consideration are models with node orders $X_n, X_1, \ldots, X_{n-1}$ and $X_1, \ldots, X_n$ respectively. By treating nodes $X_1, \ldots, X_{n-1}$ as a single super node for the random variable $Y = X_1 \times X_2 \times \ldots \times X_{n-1}$, the problem reduces to determining prior distributions for two nodes networks with global independence for all configurations and local independence for the last node in one network, that is a Dirichlet distribution (Section 5.6.6). ■

5.6 Proofs and Derivations

5.6.1 Lemmas

In this section we prove several lemmas that are needed for solving equations presented in this chapter.

Lemma 22 *The general solution of the following partial differential equation for $f(x_1, \ldots, x_n)$,*

$$f_{x_i} - f_{x_j} = \frac{\alpha}{x_i} + \frac{\beta}{x_j} \tag{5.21}$$

is given by

$$f(x_1, \ldots, x_n) = \alpha \ln x_i - \beta \ln x_j + h(x_i + x_j, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n) \tag{5.22}$$

where h is an arbitrary differentiable function having $n-1$ arguments.

Proof: Let $s = x_i + x_j$ and $t = x_i - x_j$. Thus, $f_{x_i} = f_s + f_t$, $f_{x_j} = f_s - f_t$. Hence, after a change of variables, the differential equation becomes

$$f_t = \frac{\alpha}{s+t} + \frac{\beta}{s-t}$$

Integrating with respect to (wrt) t and changing back to the original variables yields the desired solution. ■

Lemma 23 *Let $f(x_1,\ldots,x_n)$ be a twice-differentiable function from $(0,1)^n$ to $(0,1)$. If for all $1 \leq i < j \leq n$, f can be represented as*

$$f(x_1,\ldots,x_n) = a_i \ln x_i + a_j \ln x_j + h_{ij}(x_i + x_j, x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_{j-1},x_{j+1},\ldots,x_n)$$

where h_{ij} are arbitrary twice differentiable functions from $(0,1)^{n-1}$ to $(0,1)$, then

$$f(x_1,\ldots,x_n) = g(\sum_{i=1}^{n} x_i) + \sum_{i=1}^{n} a_i \ln x_i \tag{5.23}$$

where g is an arbitrary twice-differentiable function from $(0,1)$ to $(0,1)$.

Proof: We prove this lemma by induction. We prove that for every $l = 2,\ldots,n$, and for every permutation of the indices of $x_1,\ldots,x_n$, it holds that

$$f(x_1,\ldots,x_n) = h(\sum_{i=1}^{l} x_i, x_{l+1},\ldots,x_n) + \sum_{i=1}^{l} a_i \ln x_i \tag{5.24}$$

where h is an arbitrary twice differentiable function from $(0,1)^{n-l+1}$ to $(0,1)$. The function h_l depends on the permutation, although this fact is not reflected in our notation. The base case $l = 2$ is assumed by the lemma and the case $l = n$ needs to be proved.

We assume that the induction hypothesis holds for the standard order of variables $x_1,\ldots,x_n$ (Eq. 5.24) and for the order $x_1,\ldots,x_{l-1},x_{l+1},x_l,x_{l+2},\ldots,x_n$,

$$f(x_1,\ldots,x_n) = g(\sum_{i=1}^{l-1} x_i + x_{l+1}, x_l, x_{l+2},\ldots,x_n) + \sum_{i=1}^{l-1} b_i \ln x_i + b_{l+1} \ln x_{l+1} \tag{5.25}$$

From Eqs. 5.24 and 5.25 we get

$$h(\sum_{i=1}^{l} x_i, x_{l+1},\ldots,x_n) = g(\sum_{i=1}^{l-1} x_i + x_{l+1}, x_l, x_{l+2},\ldots,x_n) + \sum_{i=1}^{l-1} (b_i - a_i) \ln x_i - a_l \ln x_l + b_{l+1} \ln x_{l+1}. \tag{5.26}$$

Differentiating Eq. 5.26 with respect to x_j for $j = 1,\ldots,l-1$ we get

$$h_1(\sum_{i=1}^{l} x_i, x_{l+1},\ldots,x_n) = g_1(\sum_{i=1}^{l-1} x_i + x_{l+1}, x_l, x_{l+2},\ldots,x_n) + \frac{b_j - a_j}{x_j}, \tag{5.27}$$

and differentiation of Eq. 5.26 with respect to x_l gives

$$h_1(\sum_{i=1}^{l} x_i, x_{l+1},\ldots,x_n) = g_2(\sum_{i=1}^{l-1} x_i + x_{l+1}, x_l, x_{l+2},\ldots,x_n) - \frac{a_l}{x_l}. \tag{5.28}$$

Combining Eqs. 5.27 and 5.28 we get (for all $j = 1, \ldots, l-1$)

$$g_2(\sum_{i=1}^{l-1} x_i + x_{l+1}, x_l, x_{l+2}, \ldots, x_n) - g_1(\sum_{i=1}^{l-1} x_i + x_{l+1}, x_l, x_{l+2}, \ldots, x_n) = \frac{b_j - a_j}{x_j} + \frac{a_l}{x_l}. \quad (5.29)$$

Varying x_j and x_{l+1} together, such that $\sum_{i=1}^{l-1} x_i + x_{l+1}$ remains constant, changes the right side of the equation while keeping the left side constant. Therefore $b_j - a_j = 0$ for all $j = 1, \ldots, l-1$. Let us now introduce two new variables $t = \sum_{i=1}^{l-1} x_i + x_{l+1}$ and $s = \sum_{i=1}^{l+1} x_i$, and denote the variables $x_{l+2}, \ldots, x_n$ by $\bar{x}$. Eq. 5.29 becomes

$$g_2(t, s-t, \bar{x}) - g_1(t, s-t, \bar{x}) = \frac{a_l}{s-t}$$

and by integrating both parts by t we get

$$g(t, s-t, \bar{x}) = a_l \ln(s-t) + B(s, \bar{x}). \quad (5.30)$$

Consequently, by switching back to the original variables and substituting Eq. 5.30 into Eq. 5.25 we get

$$f(x_1, \ldots, x_n) = B(\sum_{i=1}^{l+1} x_i, x_{l+2}, \ldots, x_n) + \sum_{i=1}^{l+1} b_i \ln x_i,$$

with $b_l = a_l$, which is the required induction step.■

Lemma 24 *Let* $\mathbf{t} = \langle t_1, \ldots, t_n \rangle$ *be* n *dimensional vector of variables (or values) and let* $\tilde{\mathbf{t}}$ *denote the first* $n-1$ *elements of* $\mathbf{t}$*, i.e.* $\tilde{\mathbf{t}} = \langle t_1, \ldots, t_{n-1} \rangle$*. In addition, let* $\Delta_m = \{\langle \alpha_1, \ldots, \alpha_m \rangle | \alpha_i > 0, \sum_{i=1}^{m} \alpha_i < 1\}$ *denote the* m *dimensional unit simplex.*

Consider the following general type of functional equation:

$$f(t_1, \ldots, t_n) = h_i\left[\mathbf{t}, f_i(g_i(\mathbf{t}))\right], \quad i = 1, \ldots, k \quad (5.31)$$

where $f, f_1, \ldots, f_k$ *are unknown functions and* $\mathbf{t} = \langle t_1, \ldots, t_n \rangle \in D \subseteq \mathbb{R}^n$ *are independent variables. Suppose* D *is such that* $\{\mathbf{t} | \mathbf{t} > 0, \sum_{j=1}^{n} t_j = 1\} \subset D$*.*

Consider the same equation, with $t_1, \ldots, t_n$ *being dependent, with* $t_n = 1 - \sum_{j=1}^{n-1} t_j$*:*

$$\tilde{f}(\tilde{\mathbf{t}}) = h_i\left[q(\tilde{\mathbf{t}}), f_i(g_i(q(\tilde{\mathbf{t}})))\right], \quad i = 1, \ldots, k \quad (5.32)$$

where $q(\tilde{\mathbf{t}}) = \langle t_1, \ldots, t_{n-1}, 1 - \sum_{j=1}^{n-1} t_j \rangle$ *and* $\tilde{\mathbf{t}} \in \Delta_{n-1}$*.*

The relationship between the solutions of Eqs. 5.31 and 5.32 is:

1. *For any solution* $f(\mathbf{t})$ *of Eq. 5.31,* $\tilde{f}(\tilde{\mathbf{t}}) = f(q(\tilde{\mathbf{t}}))$ *is a solution for Eq. 5.32.*

2. *If* h_i *and* g_i *are scale-independent on* D*, i.e.,* $h_i(a\mathbf{t}, b) = h_i(\mathbf{t}, b)$ *and* $g_i(a\mathbf{t}) = g_i(\mathbf{t})$ *for all* $i = 1, \ldots, k$*,* $a, b \in \mathbb{R}$ *and* $\mathbf{t}, a\mathbf{t} \in D$*, then any solution* $\tilde{f}$ *of Eq. 5.32 is of the form* $\tilde{f}(\tilde{\mathbf{t}}) = f(q(\tilde{\mathbf{t}}))$*, where* f *is a solution for Eq. 5.31.*

Proof, Part I: Since f is a solution for Eq. 5.31, there exist functions $f_1, \ldots, f_k$ such that Eq. 5.31 holds for all $\mathbf{t} \in D$. We have, for $i = 1, \ldots, k$:

$$\tilde{f}(\tilde{\mathbf{t}}) = f(q(\tilde{\mathbf{t}})) = h_i(q(\tilde{\mathbf{t}}), f_i(g_i(q(\tilde{\mathbf{t}})))) \tag{5.33}$$

for all $\tilde{\mathbf{t}} \in \Delta_{n-1}$. Thus $\tilde{f}$ is a solution of Eq. 5.32.

Part II: We first show that for any solution $\tilde{f}$ for Eq. 5.32, $f(\mathbf{t}) = \tilde{f}(\tilde{\mathbf{t}}/\sigma)$ is a solution for Eq. 5.31, where $\sigma = \sum_{j=1}^{n} t_j$. Indeed, since $\tilde{f}$ is a solution for Eq. 5.32 there exist functions $f_1, \ldots, f_k$ such that Eq. 5.32 holds for all $\tilde{\mathbf{t}} \in \Delta_{n-1}$ (and for all $\mathbf{t} \in D$, $\tilde{\mathbf{t}}/\sigma \in \Delta_{n-1}$). We have, for $1 \le i \le k$:

$$f(\mathbf{t}) = \tilde{f}(\tilde{\mathbf{t}}/\sigma) = h_i(q(\tilde{\mathbf{t}}/\sigma), f_i(g_i(q(\tilde{\mathbf{t}}/\sigma)))) = h_i(\mathbf{t}/\sigma, f_i(g_i(\mathbf{t}/\sigma))) = h_i(\mathbf{t}, f_i(g_i(\mathbf{t}))) \tag{5.34}$$

for all $\mathbf{t} \in D$.

It is easy to check that $\tilde{f}(\tilde{\mathbf{t}}) = f(q(\tilde{\mathbf{t}}))$. We have, for $1 \le i \le k$:

$$f(q(\tilde{\mathbf{t}})) = \tilde{f}(\tilde{\mathbf{t}}/1) = \tilde{f}(\tilde{\mathbf{t}}). \tag{5.35}$$

Thus any solution for Eq. 5.32 can be expressed via a solution of Eq. 5.31. ■

5.6.2 The Proof of Theorem 15

As explicated in Section 5.3.1, the proof of Theorem 15 is reduced to showing that Eq. 5.4 is the solution of Eq. 5.8, which is equivalent to Eq. 5.3 under the prior regularity assumption (A3).

We now solve Eq. 5.8. Let

$$\begin{array}{ll}
\hat{F}_i(Y) = \frac{\partial \hat{F}(Y)}{\partial y_i} & 1 \le i \le n-1 \\
\hat{F}_{i_1 i_2}(Y) = \frac{\partial^2 \hat{F}(Y)}{\partial y_{i_1} \partial y_{i_2}} & 1 \le i_1, i_2 \le n-1 \\
\hat{g}_{ji}(Z) = \frac{\partial \hat{g}(Z)}{\partial z_{ji}} & 1 \le i \le n,\ 1 \le j \le k-1 \\
\hat{g}_{(j_1 i_1)(j_2 i_2)}(Z) = \frac{\partial^2 \hat{g}(Z)}{\partial z_{j_1 i_1} \partial z_{j_2 i_2}} & 1 \le i_1, i_2 \le n,\ 1 \le j_1, j_2 \le k-1
\end{array} \tag{5.36}$$

and similarly for $\hat{G}$ and $\hat{f}$. Taking partial derivatives of X and W gives

$$
\begin{array}{lll}
\frac{\partial x_j}{\partial y_i} = z_{ji} - z_{jn} & 1 \le j \le k, & 1 \le i \le n-1 \\
\frac{\partial x_m}{\partial z_{ji}} = \begin{cases} y_i & j = m \\ 0 & j \ne m \end{cases} \quad , \quad \frac{\partial x_k}{\partial z_{ji}} = -y_i & 1 \le m, j \le k-1, & 1 \le i \le n \\
\frac{\partial w_{ji}}{\partial y_l} = \frac{\partial \frac{z_{ji} y_i}{x_j}}{\partial y_l} = \begin{cases} -\frac{z_{ji} y_i (z_{jl} - z_{jn})}{x_j^2} & i \ne l \\ -\frac{z_{ji} y_i (z_{jl} - z_{jn})}{x_j^2} + \frac{z_{ji}}{x_j} & i = l \end{cases} & 1 \le j \le k, & 1 \le i, l \le n-1 \\
\frac{\partial w_{ji}}{\partial z_{ml}} = \frac{\partial \frac{z_{ji} y_i}{x_j}}{\partial z_{ml}} = \begin{cases} -\frac{z_{ji} y_i y_l}{x_j^2} & i \ne l, m = j \\ -\frac{z_{ji} y_i y_l}{x_j^2} + \frac{y_i}{x_j} & i = l, m = j \\ 0 & m \ne j \end{cases} & 1 \le j, m \le k-1, & 1 \le i, l \le n, \ i \ne n \\
\frac{\partial w_{ki}}{\partial z_{ml}} = \frac{\partial \frac{z_{ki} y_i}{x_k}}{\partial z_{ml}} = \begin{cases} \frac{z_{ki} y_i y_l}{x_k^2} & i \ne l \\ \frac{z_{ki} y_i y_l}{x_k^2} - \frac{y_i}{x_k} & i = l \end{cases} & 1 \le m \le k-1, & 1 \le i, l \le n, \ i \ne n
\end{array}
\tag{5.37}
$$

(Note, that there is no error in line 3: $\frac{\partial \frac{z_{ki} y_i}{x_k}}{\partial y_l}$ indeed can be written in the given form). Additional calculations give:

$$
\begin{array}{ll}
\frac{\partial \hat{G}}{\partial y_i} = \sum_{j=1}^{k-1} \left[(z_{ji} - z_{jn}) \hat{G}_j \right] & 1 \le i \le n-1 \\
\frac{\partial \hat{G}}{\partial z_{ji}} = y_i \hat{G}_j & 1 \le i \le n, \ 1 \le j \le k-1 \\
\frac{\partial \hat{f}}{\partial y_i} = \sum_{j=1}^{k} \left(\sum_{l=1}^{n-1} \left[-\frac{z_{jl} y_l (z_{ji} - z_{jn})}{x_j^2} \hat{f}_{jl} \right] + \frac{z_{ji}}{x_j} \hat{f}_{ji} \right) & 1 \le i \le n-1 \\
\frac{\partial \hat{f}}{\partial z_{ji}} = \sum_{l=1}^{n-1} \left[-\frac{z_{jl} y_l y_i}{x_j^2} \hat{f}_{jl} + \frac{z_{kl} y_l y_i}{x_k^2} \hat{f}_{kl} \right] + \frac{y_i}{x_j} \hat{f}_{ji} - \frac{y_i}{x_k} \hat{f}_{ki} & 1 \le i \le n-1, \ 1 \le j \le k-1 \\
\frac{\partial \hat{f}}{\partial z_{jn}} = \sum_{l=1}^{n-1} \left[-\frac{z_{jl} y_l y_n}{x_j^2} \hat{f}_{jl} + \frac{z_{kl} y_l y_n}{x_k^2} \hat{f}_{kl} \right] & 1 \le j \le k-1
\end{array}
\tag{5.38}
$$

Let $\hat{C}_j = \sum_{l=1}^{n-1} \left[-\frac{z_{jl} y_l}{x_j^2} \hat{f}_{jl} \right]$. By taking the logarithm and then a derivative wrt y_i $(1 \le i \le n-1)$ of Eq. 5.8, we get,

$$
\hat{F}_i(Y) = \sum_{j=1}^{k-1} \left[(z_{ji} - z_{jn}) \hat{G}_j(X) \right] + \sum_{j=1}^{k} \left[(z_{ji} - z_{jn}) \hat{C}_j(x_j, W) + \frac{z_{ji}}{x_j} \hat{f}_{ji}(W) \right] \tag{5.39}
$$

By taking the logarithm and then a derivative wrt z_{ji} $(1 \le i \le n-1, 1 \le j \le k-1)$ of Eq. 5.8, we get,

$$
\hat{g}_{ji}(Z) = y_i \hat{G}_j(X) + y_i \hat{C}_j(x_j, W) - y_i \hat{C}_k(x_k, W) + \frac{y_i}{x_j} \hat{f}_{ji}(W) - \frac{y_i}{x_k} \hat{f}_{ki}(W) \tag{5.40}
$$

and by the same operation wrt z_{jn} $(1 \le j \le k-1)$:

$$
\hat{g}_{jn}(Z) = y_n \hat{G}_j(X) + y_n \hat{C}_j(x_j, W) - y_n \hat{C}_k(x_k, W) \tag{5.41}
$$

From Eqs. 5.40 and 5.41 we get: $(1 \le i \le n-1, 1 \le j \le k-1)$

$$\frac{1}{y_i}\hat{g}_{ji}(Z) - \frac{1}{y_n}\hat{g}_{jn}(Z) = \frac{1}{x_j}\hat{f}_{ji}(W) - \frac{1}{x_k}\hat{f}_{ki}(W) \tag{5.42}$$

Solving Eq. 5.42 for $\frac{1}{x_j}\hat{f}_{ji}(W)$ and substitution into Eq. 5.41 gives:

$$\sum_{l=1}^{n}\frac{z_{jl}}{x_j}\hat{g}_{jl}(Z) = \hat{G}_j(X) - \sum_{l=1}^{n-1}(w_{jl}-w_{kl})\frac{1}{x_k}\hat{f}_{kl}(W) \tag{5.43}$$

Simplifying Eq. 5.39 by using Eqs. 5.41 and 5.42, and recalling $z_{ki} = 1 - \sum_{j=1}^{k-1} z_{ji}$ we get (for $1 \le i \le n-1$):

$$\hat{F}_i(Y) = \sum_{j=1}^{k-1}\left(\frac{z_{ji}}{y_i}\hat{g}_{ji}(Z) - \frac{z_{jn}}{y_n}\hat{g}_{jn}(Z)\right) + \frac{1}{x_k}\hat{f}_{ki}(W) \tag{5.44}$$

Multiplying Eq. 5.44 by $(w_{ji} - w_{ki})$, taking the sum of the resulting equations and substitution $\sum_{l=1}^{n-1}(w_{jl}-w_{kl})\frac{1}{x_k}\hat{f}_{kl}(W)$ from Eq. 5.43 gives (for $1 \le j \le k-1$):

$$\sum_{l=1}^{n-1}(w_{jl}-w_{kl})\hat{F}_l(Y) = \sum_{l=1}^{n-1}(w_{jl}-w_{kl})\left[\sum_{m=1}^{k-1}\left(\frac{z_{ml}}{y_l}\hat{g}_{ml}(Z) - \frac{z_{mn}}{y_n}\hat{g}_{mn}(Z)\right)\right] + \hat{G}_j(X) - \sum_{l=1}^{n}\frac{z_{jl}}{x_j}\hat{g}_{jl}(Z) \tag{5.45}$$

After some simplifications (recall that $w_{ji} = \frac{z_{ji}y_i}{x_j}$) we get:

$$\sum_{l=1}^{n-1}(w_{jl}-w_{kl})\hat{F}_l(Y) = \sum_{l=1}^{n}\left[\left(\frac{z_{jl}}{x_j} - \frac{z_{kl}}{x_k}\right)\sum_{m=1}^{k-1} z_{ml}\hat{g}_{ml}(Z)\right] + \hat{G}_j(X) - \sum_{l=1}^{n}\frac{z_{jl}}{x_j}\hat{g}_{jl}(Z) \tag{5.46}$$

Taking a derivative by z_{ji} we get:

$$\begin{aligned}
&\sum_{l=1}^{n-1}\left(-\frac{z_{jl}y_l y_i}{x_j^2} - \frac{z_{kl}y_l y_i}{x_k^2}\right)\hat{F}_l(Y) + \frac{y_i}{x_j}\hat{F}_i(Y) + \frac{y_i}{x_k}\hat{F}_i(Y) = \\
&\quad \sum_{l=1}^{n}\left[\left(\frac{z_{jl}}{x_j} - \frac{z_{kl}}{x_k}\right)\sum_{m=1}^{k-1} z_{ml}\hat{g}_{(ml)(ji)}(Z)\right] + \left(\frac{z_{ji}}{x_j} - \frac{z_{ki}}{x_k}\right)\hat{g}_{ji}(Z) \\
&\quad + \sum_{l=1}^{n}\left[\left(-\frac{z_{jl}y_i}{x_j^2} - \frac{z_{kl}y_i}{x_k^2}\right)\sum_{m=1}^{k-1} z_{ml}\hat{g}_{ml}(Z)\right] + \left(\frac{1}{x_j} + \frac{1}{x_k}\right)\sum_{m=1}^{k-1} z_{mi}\hat{g}_{mi}(Z) \\
&\quad + y_i\hat{G}_{jj}(X) - \sum_{l=1}^{n}\frac{z_{jl}}{x_j}\hat{g}_{(jl)(ji)}(Z) + \sum_{l=1}^{n}\frac{z_{jl}y_i}{x_j^2}\hat{g}_{jl}(Z) - \frac{1}{x_j}\hat{g}_{ji}(Z)
\end{aligned} \tag{5.47}$$

Substituting $z_{ji} = \frac{1}{k}$ (and thus $x_j = \frac{1}{k}$, $w_{ji} = y_i$) for $1 \le i \le n$, $1 \le j \le k$ we get $(1 \le i \le n-1)$:

$$-\sum_{l=1}^{n-1} y_l\hat{F}_l(Y) + \hat{F}_i(Y) = \frac{1}{y_i}C_i + A \tag{5.48}$$

where

$$C_i = \frac{1}{k}\sum_{m=1}^{k-1}\hat{g}_{mi}(\frac{1}{k}) - \frac{1}{2k}\sum_{l=1}^{n}\hat{g}_{(jl)(ji)}(\frac{1}{k}) - \frac{1}{2}\hat{g}_{ji}(\frac{1}{k}) \tag{5.49}$$
$$A = -\sum_{l=1}^{n}\sum_{m=1}^{k-1}\hat{g}_{ml}(\frac{1}{k}) + \frac{1}{2k}\hat{G}_{jj}(\frac{1}{k}) + \frac{1}{2}\sum_{l=1}^{n}\hat{g}_{jl}(\frac{1}{k})$$

where C_i's and A are computed for some value of j.

In case $n = 2$, ($Y = \{y\}$, $\hat{F}_1(Y) = \hat{F}'(y)$) we have only one equation

$$-y\hat{F}'(y) + \hat{F}'(y) = \frac{1}{y}C_1 + A \tag{5.50}$$

Thus $F(y) = Cy^{C_1}(1-y)^{-(A+C_1)}$ (i.e. F is of Dirichlet form).

In case $n \geq 3$, by subtracting two Eqs. 5.48 for $1 \leq i_1, i_2 \leq n-1$, $i_1 \neq i_2$, we get:

$$\hat{F}_{i_1}(Y) - \hat{F}_{i_2}(Y) = \frac{C_{i_1}}{y_{i_1}} - \frac{C_{i_2}}{y_{i_2}} \tag{5.51}$$

The general solution for Eqs. 5.51 is (by Lemmas 22 and 23):

$$\hat{F}(Y) = h(\sum_{i=1}^{n-1} y_i) + \sum_{i=1}^{n-1} C_i \ln y_i \tag{5.52}$$

Substituting the solution into Eq. 5.48, and solving for h, we get:

$$F(Y) = C\prod_{i=1}^{n} y_i^{C_i} \tag{5.53}$$

where C is an arbitrary constant, $y_n = 1 - \sum_{i=1}^{n-1} y_i$ and $C_n = -A - \sum_{i=1}^{n-1} C_i$.

Similarly, $G(X) = B\prod_{j=1}^{k} x_j^{B_j}$, and by substitutions $y_i = \frac{1}{n}$, $x_j = \frac{1}{n}z_{j\cdot}$, and $w_{ji} = \frac{z_{ji}}{z_{j\cdot}}$, where $z_{j\cdot} = \sum_{i=1}^{n} z_{ji}$, Eq. 5.46 transforms to:

$$\sum_{l=1}^{n-1}\left(\frac{z_{jl}}{z_{j\cdot}} - \frac{z_{kl}}{z_{k\cdot}}\right)(C_l - C_n) - \left(\frac{B_j}{z_{j\cdot}} - \frac{B_k}{z_{k\cdot}}\right) = \sum_{l=1}^{n}\left[\left(\frac{z_{jl}}{z_{j\cdot}} - \frac{z_{kl}}{z_{k\cdot}}\right)\sum_{m=1}^{k-1} z_{ml}\hat{g}_{ml}(Z)\right] - \sum_{l=1}^{n}\frac{z_{jl}}{z_{j\cdot}}\hat{g}_{jl}(Z) \tag{5.54}$$

By substitution we see that $\hat{g}_p(Z) = \sum_{i=1}^{n}\sum_{j=1}^{k}\alpha_{ji}\ln z_{ji}$ is a particular solution of Eq. 5.54. From Eqs. 5.49 and 5.54, the coefficients $\{\alpha_{ji}\}$, $\{C_i\}$ and $\{B_j\}$ satisfy: $C_i = \sum_{m=1}^{k}\alpha_{mi}$, $B_j = \sum_{l=1}^{n}\alpha_{jl}$, $B_k = \frac{1}{k}\sum_{m=1}^{k} B_m$ and, by symmetry, $C_n = \frac{1}{n}\sum_{l=1}^{n} C_l$ (see Section 5.6.2 for derivations)

We must now solve the homogeneous first-order partial differential equation:

$$\sum_{l=1}^{n}\left[\left(\frac{z_{jl}}{z_{j\cdot}} - \frac{z_{kl}}{z_{k\cdot}}\right)\sum_{m=1}^{k-1} z_{ml}\hat{g}_{ml}(Z)\right] - \sum_{l=1}^{n}\frac{z_{jl}}{z_{j\cdot}}\hat{g}_{jl}(Z) = 0 \tag{5.55}$$

Multiplying each equation by $z_{j\cdot}$, taking the sum by j, $1 \leq j \leq k-1$ and simplifying we get:

$$\sum_{l=1}^{n} z_{kl} \sum_{m=1}^{k-1} z_{ml}\hat{g}_{ml}(Z) = 0 \tag{5.56}$$

so Eq. 5.55 transforms to (for $1 \leq j \leq k-1$):

$$\sum_{l=1}^{n} z_{jl} \sum_{m=1}^{k-1} z_{ml}\hat{g}_{ml}(Z) = \sum_{l=1}^{n} z_{jl}\hat{g}_{jl}(Z) \tag{5.57}$$

Substitution $\hat{g}(Z) = \bar{g}(\bar{Z})$, where $\bar{z}_{ji} = \frac{z_{ji}}{z_{j+1,i}}$ for $1 \leq j \leq k-1$, $1 \leq i \leq n$ gives:

$$z_{ji}\hat{g}_{ji}(Z) = \begin{cases} \bar{z}_{ji}\bar{g}_{ji}(\bar{Z}) + \frac{z_{ji}}{z_{ki}}\bar{z}_{k-1,i}\bar{g}_{k-1,i}(\bar{Z}) & j = 1 \\ -\bar{z}_{j-1,i}\bar{g}_{j-1,i}(\bar{Z}) + \bar{z}_{ji}\bar{g}_{ji}(\bar{Z}) + \frac{z_{ji}}{z_{ki}}\bar{z}_{k-1,i}\bar{g}_{k-1,i}(\bar{Z}) & 2 \leq j \leq k-1 \end{cases} \tag{5.58}$$

and

$$\sum_{m=1}^{k-1} z_{ml}\hat{g}_{ml}(Z) = \frac{1}{z_{kl}}\bar{z}_{k-1,l}\bar{g}_{k-1,l}(\bar{Z}) \tag{5.59}$$

Plugging Eqs. 5.58 and 5.59 into Eq. 5.57 gives

$$\begin{array}{ll} \sum_{l=1}^{n} \bar{z}_{1l}\bar{g}_{1l}(\bar{Z}) = 0 & j = 1 \\ \sum_{l=1}^{n} \bar{z}_{jl}\bar{g}_{jl}(\bar{Z}) - \sum_{l=1}^{n} \bar{z}_{j-1,l}\bar{g}_{j-1,l}(\bar{Z}) = 0 & 2 \leq j \leq k-1 \end{array} \tag{5.60}$$

Thus for all j, $(1 \leq j \leq k-1)$:

$$\sum_{l=1}^{n} \bar{z}_{jl}\bar{g}_{jl}(\bar{Z}) = 0 \tag{5.61}$$

Additional substitution $\bar{g}(\bar{Z}) = \tilde{g}(\tilde{Z})$, where

$$\begin{array}{ll} \tilde{z}_{ji} = \frac{\bar{z}_{ji}}{\bar{z}_{j,i+1}} & 1 \leq i \leq n-1,\ 1 \leq j \leq k-1 \\ \tilde{z}_{jn} = \bar{z}_{jn} & 1 \leq j \leq k-1 \end{array} \tag{5.62}$$

gives (for $1 \leq j \leq k-1$):

$$\tilde{z}_{jn}\tilde{g}_{jn}(\tilde{Z}) = 0 \tag{5.63}$$

The general solution to Eq. 5.63 is

$$\tilde{g}(\tilde{Z}) = h(\{\tilde{z}_{ji} | 1 \leq i \leq n-1,\ 1 \leq j \leq k-1\}) + C_h \tag{5.64}$$

where h is an arbitrary differentiable function of $(n-1) \times (k-1)$ variables and C_h is an arbitrary constant.

Thus the general solution for g of Eq. 5.8 is:

$$g(Z) = C\left[\prod_{i=1}^{n}\prod_{j=1}^{k} z_{ji}^{\alpha_{ji}}\right] H\left(\left\{\frac{z_{ji}z_{j+1,i+1}}{z_{j+1,i}z_{j,i+1}} | 1 \leq i \leq n-1, 1 \leq j \leq k-1\right\}\right) \tag{5.65}$$

where H is an arbitrary differentiable function of $(n-1)\times(k-1)$ variables and C is an arbitrary constant.

Using $z_{ji} = \theta_{j|i} = \frac{\theta_{ij}}{\theta_{i\cdot}}$, $y_i = \theta_{i\cdot}$ (for $1 \leq i \leq n$, $1 \leq j \leq k$), as well as properties of the particular solution of Eq. 5.54 ($C_i = \sum_{m=1}^{k}\alpha_{mi}$) we get from Eqs. 5.53 and 5.65:

$$p(\{\theta_{ij}\}) = C\left[\prod_{i=1}^{n}\prod_{j=1}^{k} \theta_{ij}^{\alpha_{ij}}\right] H\left(\left\{\frac{\theta_{ij}\theta_{i+1,j+1}}{\theta_{i+1,i}\theta_{i,j+1}} | 1 \leq i \leq n-1, 1 \leq j \leq k-1\right\}\right) \tag{5.66}$$

This solves Eq. 5.3 and completes the proof of Theorem 15. ■

Some technical details

Here we show that $\hat{g}_p(Z) = \sum_{i=1}^{n}\sum_{j=1}^{k}\alpha_{ji}\ln z_{ji}$ is a particular solution of Eq 5.54, where the coefficients $\{\alpha_{ji}\}$ should satisfy some constrains imposed by Eq 5.54 and Eq 5.49. Note that $\frac{\partial \hat{g}_p}{\partial z_{ji}}(Z) = \frac{\alpha_{ji}}{z_{ji}} - \frac{\alpha_{ki}}{z_{ki}}$.

Substituting $\hat{g}_p(Z)$ into right hand side of Eq 5.54, we get:

$$\begin{aligned}
&\sum_{l=1}^{n}\left[\left(\frac{z_{jl}}{z_{j\cdot}} - \frac{z_{kl}}{z_{k\cdot}}\right)\sum_{m=1}^{k-1} z_{ml}\left(\frac{\alpha_{ml}}{z_{ml}} - \frac{\alpha_{kl}}{z_{kl}}\right)\right] - \sum_{l=1}^{n}\frac{z_{jl}}{z_{j\cdot}}\left(\frac{\alpha_{jl}}{z_{jl}} - \frac{\alpha_{kl}}{z_{kl}}\right)\\
&= \sum_{l=1}^{n}\left[\left(\frac{z_{jl}}{z_{j\cdot}} - \frac{z_{kl}}{z_{k\cdot}}\right)\left(\sum_{m=1}^{k-1}\alpha_{ml} - (1-z_{kl})\frac{\alpha_{kl}}{z_{kl}}\right)\right] - \sum_{l=1}^{n}\left(\frac{\alpha_{jl}}{z_{j\cdot}} - \frac{z_{jl}\alpha_{kl}}{z_{j\cdot}z_{kl}}\right)\\
&= \sum_{l=1}^{n}\left[\left(\frac{z_{jl}}{z_{j\cdot}} - \frac{z_{kl}}{z_{k\cdot}}\right)\sum_{m=1}^{k}\alpha_{ml}\right] - \sum_{l=1}^{n}\frac{z_{jl}\alpha_{kl}}{z_{j\cdot}z_{kl}} + \sum_{l=1}^{n}\frac{\alpha_{kl}}{z_{k\cdot}} - \sum_{l=1}^{n}\frac{\alpha_{jl}}{z_{j\cdot}} + \sum_{l=1}^{n}\frac{z_{jl}\alpha_{kl}}{z_{j\cdot}z_{kl}}\\
&= \sum_{l=1}^{n-1}\left[\left(\frac{z_{jl}}{z_{j\cdot}} - \frac{z_{kl}}{z_{k\cdot}}\right)\sum_{m=1}^{k}\alpha_{ml}\right] - \left((1-\frac{z_{jn}}{z_{j\cdot}}) - (1-\frac{z_{kn}}{z_{k\cdot}})\right)\sum_{m=1}^{k}\alpha_{mn}\\
&\quad - \left(\frac{\sum_{l=1}^{n}\alpha_{jl}}{z_{j\cdot}} - \frac{\sum_{l=1}^{n}\alpha_{kl}}{z_{k\cdot}}\right)\\
&= \sum_{l=1}^{n-1}\left[\left(\frac{z_{jl}}{z_{j\cdot}} - \frac{z_{kl}}{z_{k\cdot}}\right)\sum_{m=1}^{k}\alpha_{ml}\right] - \left(\frac{\sum_{l=1}^{n-1}z_{jl}}{z_{j\cdot}} - \frac{\sum_{l=1}^{n-1}z_{kl}}{z_{k\cdot}}\right)\sum_{m=1}^{k}\alpha_{mn}\\
&\quad - \left(\frac{\sum_{l=1}^{n}\alpha_{jl}}{z_{j\cdot}} - \frac{\sum_{l=1}^{n}\alpha_{kl}}{z_{k\cdot}}\right)\\
&= \sum_{l=1}^{n-1}\left[\left(\frac{z_{jl}}{z_{j\cdot}} - \frac{z_{kl}}{z_{k\cdot}}\right)(\sum_{m=1}^{k}\alpha_{ml} - \sum_{m=1}^{k}\alpha_{mn})\right] - \left(\frac{\sum_{l=1}^{n}\alpha_{jl}}{z_{j\cdot}} - \frac{\sum_{l=1}^{n}\alpha_{kl}}{z_{k\cdot}}\right)
\end{aligned} \tag{5.67}$$

Thus, from Eq 5.67 and Eq 5.54, we get:

$$\begin{aligned}
&C_i - C_n = \sum_{m=1}^{k}\alpha_{mi} - \sum_{m=1}^{k}\alpha_{mn} && 1 \leq i \leq n-1\\
&B_j = \sum_{l=1}^{n}\alpha_{jl} && 1 \leq j \leq k
\end{aligned} \tag{5.68}$$

If $\hat{g}_p(Z)$ is indeed a particular solution to Eq 5.54 then $\{\alpha_{ji}\}$ and $\{C_i\}$ should satisfy Eq 5.49. Substituting $\hat{g}_p(Z)$ into Eq 5.49 and using Eq 5.68, we get $(1 \leq i \leq n-1)$:

$$\begin{aligned} C_i &= \textstyle\sum_{m=1}^{k-1}(\alpha_{mi} - \alpha_{ki}) + \frac{k}{2}(\alpha_{ji} + \alpha_{ki}) - \frac{k}{2}(\alpha_{ji} - \alpha_{ki}) = \sum_{m=1}^{k} \alpha_{mi} \\ A &= -k \textstyle\sum_{l=1}^{n} \sum_{m=1}^{k-1}(\alpha_{ml} - \alpha_{kl}) - \frac{k}{2}(B_j + B_k) + \frac{k}{2}\sum_{l=1}^{n}(\alpha_{jl} - \alpha_{kl}) \\ &= -k \textstyle\sum_{m=1}^{k}(B_m - B_k) - kB_k \end{aligned} \tag{5.69}$$

Thus, $C_n = \sum_{m=1}^{k} \alpha_{mn}$ (from Eqs 5.68,5.69) and using the fact that $A = -\sum_{l=1}^{n} C_l$ (from the definition of C_n, after Eq 5.53), we get:

$$k \textstyle\sum_{m=1}^{k}(B_m - B_k) + kB_k = \sum_{l=1}^{n} C_l \tag{5.70}$$

Thus $B_k = \frac{1}{k}\sum_{m=1}^{k} B_m$ and, by symmetry, $C_n = \frac{1}{n}\sum_{l=1}^{n} C_l$.

5.6.3 The Proof of Lemma 16

The 'only if' part of the proof follows directly from the correspondence between parameters θ_X and $\theta_{\mathbf{Y}_i}$ and the definition of global independence (A1) and the fact that global independence holds for the networks with node ordering: $X_i, X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ and $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n, X_i$. In fact, the assumption that θ_X is globally independent for all network structures is redundant, it is enough to assume global independence for only two networks: with X_i being first and last node respectively.

For the 'if' part of the lemma, we apply the 'only if' part of the lemma and Theorem 15 and conclude that any $p(\theta_X)$ that satisfies the global independence assumption (A1) satisfies the following n equations (for $i = 1, \ldots, n$):

$$p(\theta_X) = C_i \cdot \left[\prod_{x \in \mathbf{D}} \theta_x^{\alpha_{x,i}}\right] \cdot H_i\left(\left\{\frac{\theta_r \theta_{r'''}}{\theta_{r'}\theta_{r''}} | r, r', r'', r''' \in \mathbf{D}, \text{ s.t. } \begin{array}{l} r'_i = r'''_i = r_i + 1 = r''_i + 1 \\ r''_{\bar{i}} = r'''_{\bar{i}} = r_{\bar{i}} + 1 = r'_{\bar{i}} + 1 \end{array}\right\}\right) \tag{5.71}$$

where r_i and $r_{\bar{i}}$ denote the values of vector r on X_i and $X \setminus X_i$, respectively and the $''+1''$ operation denotes the operation of taking the next value of the corresponding vector in any fixed order. This is exactly the same equation as Eq. 5.10.

Consider a solution $f(\theta_X)$ of Eq. 5.71. We show that $f(\theta_X)$ satisfies Eq. 5.9 for an arbitrary $I = \langle i_1, \ldots, i_n \rangle$. Indeed, for any order $I = \langle i_1, \ldots, i_n \rangle$ of X, $f(\theta_X)$ satisfies Eq. 5.71 for $i = i_n$, namely,

$$f(\theta_X) = C_{i_n} \cdot \left[\prod_{d \in \mathbf{D}} \theta_x^{\alpha_{x,i_n}} \right] \cdot H_{i_n} \left(\left\{ \frac{\theta_r \theta_{r'''}}{\theta_{r'} \theta_{r''}} | r, r', r'', r''' \in \mathbf{D},\ s.t. \begin{array}{l} r'_{i_n} = r'''_{i_n} = r_{i_n} + 1 = r''_{i_n} + 1 \\ r''_{\bar{i}_n} = r'''_{\bar{i}_n} = r_{\bar{i}_n} + 1 = r'_{\bar{i}_n} + 1 \end{array} \right\} \right),$$

where, similar to Eq. 5.71, r_i and $r_{\bar{i}}$ denote the values of vector r on X_i and $X \setminus X_i$, respectively and the $'' + 1''$ operation denotes the operation of taking the next value of the corresponding vector in any fixed order. Changing the variables to θ_I (variables of complete network with topological order of nodes specified by I) yields

$$f(\theta_X) = C_{i_n} \cdot \prod_{j=1}^{n} \left[\prod_{\tilde{r} \in D_{i_j}} \prod_{r \in D_{i_1} \times \ldots \times D_{i_{j-1}}} \theta_{\tilde{r}|r}^{\beta_{\tilde{r},r,j}} \right] \cdot H_{i_n} \left(\left\{ \frac{\theta_{\tilde{r}|r} \theta_{\tilde{r}^+|r^+}}{\theta_{\tilde{r}^+|r} \theta_{\tilde{r}|r^+}} \middle| \begin{array}{l} \tilde{r}, \tilde{r}^+ \in D_{i_n} \\ r, r^+ \in D_{i_1} \times \ldots \times D_{i_{n-1}} \\ \tilde{r}^+ = \tilde{r} + 1,\ r^+ = r + 1 \end{array} \right\} \right),$$

where $\beta_{\tilde{r},r,j} = \sum_{\hat{r} \in \mathbf{D},\ \text{s.t. } \hat{r}[i_j] = \tilde{r}, \langle \hat{r}[i_1], \ldots, \hat{r}[i_{j-1}] \rangle = r} \alpha_{\hat{r},i_n}$. Note that $H()$ depends only on the variables of the last node X_{i_n}. This concludes the proof, since we have shown that $f(\theta_X)$ factorizes according to arbitrary complete network order I, yielding that θ_X are globally independent. ■

5.6.4 The Proof of Lemma 17

The proof is given for $m = 2$ and the general case follows by induction. Suppose that we have a following system of functional equations:

$$\begin{array}{l} f(x_1, \ldots, x_n) = \sum_{i=1}^n \alpha_i x_i + h(\vec{a}_1^T \vec{x}, \ldots, \vec{a}_k^T \vec{x}) \\ f(x_1, \ldots, x_n) = \sum_{i=1}^n \beta_i x_i + g(\vec{b}_1^T \vec{x}, \ldots, \vec{b}_{k'}^T \vec{x}) \end{array} \tag{5.72}$$

By comparing the two equations we have:

$$h(\vec{a}_1^T \vec{x}, \ldots, \vec{a}_k^T \vec{x}) = \sum_{i=1}^{n} (\beta_i - \alpha_i) x_i + g(\vec{b}_1^T \vec{x}, \ldots, \vec{b}_{k'}^T \vec{x}) \tag{5.73}$$

It is easy to see that $V = \{\vec{a}^T \vec{x} | \vec{a} \in \mathbb{R}^n, \vec{x} = (x_1, \ldots, x_n)^T\}$ is a vector space and that $V_h = span(\vec{a}_1^T \vec{x}, \ldots, \vec{a}_k^T \vec{x})$ and $V_g = span(\vec{b}_1^T \vec{x}, \ldots, \vec{b}_{k'}^T \vec{x})$ are subspaces of V. We can now change the variables $x_1, \ldots, x_n$ to n-independent variables $y_1, \ldots, y_n$ where $y_i = \vec{\gamma}_i^T \vec{x}$ such that:

$$\begin{array}{l} V_h \bigcap V_g = span(y_1, \ldots, y_{k^*}) \\ V_h = span(y_1, \ldots, y_{k^*}, y_{k^*+1}, \ldots, y_k) \\ V_g = span(y_1, \ldots, y_{k^*}, y_{k+1}, \ldots, y_{k+k'-k^*}) \\ V = span(y_1, \ldots, y_n) \end{array} \tag{5.74}$$

where $k* = dim(V_h \bigcap V_g)$. Let Γ_h be the $n \times k$ matrix with columns $\vec{\gamma}_1, \ldots, \vec{\gamma}_k$, let Γ_g be the $n \times k'$ matrix with columns $\vec{\gamma}_1, \ldots, \vec{\gamma}_{k^*}, \vec{\gamma}_{k+1}, \ldots, \vec{\gamma}_{k+k'-k^*}$ and let Γ denote the $n \times n$ matrix with columns $\vec{\gamma}_1, \ldots, \vec{\gamma}_n$. Eq. 5.73 now transforms to:

$$\hat{h}(y_1, \ldots, y_{k^*}, y_{k^*+1}, \ldots, y_k) = \sum_{i=1}^{n} \delta_i y_i + \hat{g}(y_1, \ldots, y_{k^*}, y_{k+1}, \ldots, y_{k+k'-k^*}) \tag{5.75}$$

where

$$\begin{aligned} &\hat{h}(t_1, \ldots, t_k) = h((\Gamma_h^+ \vec{a}_1)^T \vec{t}, \ldots, (\Gamma_h^+ \vec{a}_k)^T \vec{t}) \\ &\hat{g}(t_1, \ldots, t_{k'}) = g((\Gamma_g^+ \vec{b}_1)^T \vec{t}, \ldots, (\Gamma_g^+ \vec{b}_{k'})^T \vec{t}) \\ &\vec{\delta} = \Gamma^{-1}(\vec{\beta} - \vec{\alpha}) \end{aligned} \tag{5.76}$$

and $\Gamma_h^+ \vec{a}_i$ (or $\Gamma_g^+ \vec{b}_i$) are coordinates of $\vec{a}_i^T \vec{x}$ (or $\vec{b}_i^T \vec{x}$) in the corresponding basis of y's. For any matrix B, B^+ denotes pseudo-inverse, $B^+ = (B^T B)^{-1} B^T$, which is well defined since Γ_h, Γ_g are of full rank and $n > k, k'$ (since $rank(B^T B) = rank(B)$, see Ben-Israel and Greville, 1974; page 20).

Equation 5.75 can be further transformed to:

$$\tilde{h}(y_1, \ldots, y_{k^*}, y_{k^*+1}, \ldots, y_k) = \sum_{i=k+k'-k^*+1}^{n} \delta_i y_i + \tilde{g}(y_1, \ldots, y_{k^*}, y_{k+1}, \ldots, y_{k+k'-k^*}) \tag{5.77}$$

where

$$\begin{aligned} &\tilde{h}(t_1, \ldots, t_k) = \hat{h}(t_1, \ldots, t_k) - \textstyle\sum_{i=1}^{k} \delta_i t_i \\ &\tilde{g}(t_1, \ldots, t_{k'}) = \hat{g}(t_1, \ldots, t_{k'}) + \textstyle\sum_{i=1}^{k'-k^*} \delta_{k+i} t_{k^*+i} \end{aligned} \tag{5.78}$$

Now it is clear from Eq. 5.77 that $\delta_i = 0$ for $i = k + k' - k^* + 1, \ldots, n$ and functions $\tilde{h}$ and $\tilde{g}$ depend only on variables $y_1, \ldots, y_{k^*}$. Returning to the original variables $x_1, \ldots, x_n$, we obtain the solution to the original system 5.72, which is:

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n} \lambda_i x_1 + \bar{h}(\vec{\gamma}_1^T \vec{x}, \ldots, \vec{\gamma}_{k^*}^T \vec{x}) \tag{5.79}$$

where $\{\lambda_i\}_{i=1}^{n}$ are arbitrary constants, $\bar{h}$ is an arbitrary function of k^* variables, and $\vec{\gamma}_1^T \vec{x}, \ldots, \vec{\gamma}_{k^*}^T \vec{x}$ is the basis for $V_h \bigcap V_g$. ■

5.6.5 The Proof of Theorem 18

Lemma 16 state that in order to find all distributions that are dictated by the global independence assumption it is sufficient to find a general solution for Eq. 5.10. This equation satisfies the scale-independence conditions of Lemma 24, and therefore its general solution can be found by treating $\{\theta_{\vec{x}}\}$ as independent variables.

Eq. 5.10 for the binary X can be written as (the normalization constant is inside H_i):

$$p(\{\theta_{\vec{x}}\}_{\vec{x}\in\{0,1\}^n}) = \prod_{\vec{x}\in\{0,1\}^n} \theta_{\vec{x}}^{\alpha_{\vec{x},i}} \cdot H_i\left(\left\{\frac{\theta_{[\vec{x}\setminus x_i]=\vec{y},x_i=0}\theta_{[\vec{x}\setminus x_i]=\vec{y}+1,x_i=1}}{\theta_{[\vec{x}\setminus x_i]=\vec{y},x_i=1}\theta_{[\vec{x}\setminus x_i]=\vec{y}+1,x_i=0}} | \vec{y}\in\{0,1\}^{n-1}, |\vec{y}|\neq 1\right\}\right) \tag{5.80}$$

Taking the logarithm of both sides of Eq. 5.80, and of arguments of H_i, and changing the variables to $\{\ln\theta_{\vec{x}}\}_{\vec{x}\in\{0,1\}^n}$, we get Eq. 5.12 with:

$$\vec{b}_{ij}[m] = \begin{cases} 1 & [\vec{x}\setminus x_i] = \vec{y}, x_i = 0 \; or \; [\vec{x}\setminus x_i] = \vec{y}+1, x_i = 1 \\ -1 & [\vec{x}\setminus x_i] = \vec{y}, x_i = 1 \; or \; [\vec{x}\setminus x_i] = \vec{y}+1, x_i = 0 \\ 0 & otherwise \end{cases} \tag{5.81}$$

where $1 \leq i \leq n$, $0 \leq j \leq 2^{n-1}-2$, $0 \leq m \leq 2^n - 1$ and $\vec{y}, \vec{x}$ denote the binary representations of j and m respectively. For example $\vec{b}_{10} = [1\ -1\ -1\ 1\ 0\ 0\ 0\ 0\ \ldots\ 0]$ and $\vec{b}_{21} = [0\ 1\ 0\ -1\ -1\ 0\ 1\ 0\ \ldots\ 0]$, where the vectors $\vec{b}$ are indexed from left to right. The application of Lemma 17 and the lemma that now follows (for $k = n$) concludes the proof. ■

Lemma 25 *Given $\vec{b}_{ij}$ as specified by Eq. 5.81,*

$$\bigcap_{i=1}^{k} B_i = A_k, \quad for\ k = 2, \ldots, n \tag{5.82}$$

where $B_i = span(\vec{b}_{i,0}, \ldots, \vec{b}_{i,2^{n-1}-2})$ and $A_k = span(\vec{a}_{k,0}, \ldots, \vec{a}_{k,2^{n-i}-1})$, s.t.

$$\vec{a}_{k,\vec{y}}[\vec{x}] = \begin{cases} 1 & \vec{x}_{\perp[k+1,\ldots,n]} = \vec{y}, \vec{x}^2_{\perp[1,\ldots,k]}\%2 = 0 \\ -1 & \vec{x}_{\perp[k+1,\ldots,n]} = \vec{y}, \vec{x}^2_{\perp[1,\ldots,k]}\%2 = 1 \\ 0 & otherwise. \end{cases} \tag{5.83}$$

We use $\vec{y}$ and $\vec{x}$ to denote binary representation of indexes: $0 \leq \vec{y} \leq 2^{n-k}-1$ and $0 \leq \vec{x} \leq 2^n - 1$.

Proof: By induction on k.

Induction Basis: Let $\mathcal{M}_i$ denote the matrix with rows equal to basis vectors of B_i, i.e., with rows $\{\vec{b}_{i,j}\}_{j=0}^{2^{n-1}-2}$. In particular, for the induction basis, consider $\mathcal{M}_1$, $\mathcal{M}_2$ matrices with rows $\{\vec{b}_{1,j}\}_{j=0}^{2^{n-1}-2}$ and $\{\vec{b}_{2,j}\}_{j=0}^{2^{n-1}-2}$ respectively:

$$\mathcal{M}_1 = \begin{bmatrix} \vec{b}_{10} \\ \vec{b}_{11} \\ \vec{b}_{12} \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & \ldots & 0 \\ \vdots & & & & & & & & & & \end{bmatrix} \tag{5.84}$$

and

$$\mathcal{M}_2 = \begin{bmatrix} \vec{b}_{20} \\ \vec{b}_{21} \\ \vec{b}_{22} \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & \dots & 0 \\ \vdots \end{bmatrix}. \tag{5.85}$$

Note that the same pattern as in the first two rows, repeats itself with 4-column shift for all $2^{n-1}-1$ rows of $\mathcal{M}_1$ and $\mathcal{M}_2$ except the very last row. Considering the matrix $\mathcal{M} = [\mathcal{M}_1^T | \mathcal{M}_2^T]^T$, with row space equal to $B_1 + B_2$, we can see that it is exactly of rank $3 \cdot 2^{n-2} - 2$, since rows $\{\vec{b}_{1,j}\}_{j=0}^{2^{n-1}-2}$ and $\{\vec{b}_{2,j}\}_{j=1,3,\dots,2^{n-1}-3}$ are linearly independent and $\vec{b}_{1,j} = \vec{b}_{2,j}$ for $j = 0, 2, \dots, 2^{n-1}-2$. Thus, $dim(B_1 \cap B_2) = dim(B_1) + dim(B_2) - dim(B_1 + B_2) = (2^{n-1}-1) + (2^{n-1}-1) - (3 \cdot 2^{n-2} - 2) = 2^{n-2}$ and the basis for $B_1 \cap B_2$ is specified by $\vec{a}_{2,i} = \vec{b}_{1,2i}$, for $i = 0, \dots, 2^{n-2}-1$.

Induction Step: We demonstrate the induction step for $k = 3$. The proof for arbitrary k is given at the end. Let $\mathcal{M}_{A_k}$ denote the matrix with rows equal to basis vectors of $A_k = \bigcap_{i=1}^{k} B_i$. Consider matrices $\mathcal{M}_{A_2}$, $\mathcal{M}_3$ with rows $\{\vec{a}_{2,j}\}_{j=0}^{2^{n-2}-1}$ and $\{\vec{b}_{3,j}\}_{j=0}^{2^{n-1}-2}$ respectively. From the induction basis,

$$\mathcal{M}_A = \begin{bmatrix} \vec{a}_{20} \\ \vec{a}_{21} \\ \vec{a}_{22} \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & \dots & 0 \\ \vdots \end{bmatrix} \tag{5.86}$$

and

$$\mathcal{M}_3 = \begin{bmatrix} \vec{b}_{30} \\ \vec{b}_{31} \\ \vec{b}_{32} \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & -1 & 1 & 0 & \dots & 0 \\ \vdots \end{bmatrix}. \tag{5.87}$$

The first 4-row pattern of $\mathcal{M}_3$ repeats itself with 8 column shift for all $2^{n-1}-1$ rows of $\mathcal{M}_3$ (the last pattern is truncated to 3 rows). For a general $\mathcal{M}_k$ matrix - the pattern of the first 2^{k-1} rows repeats with a shift of 2^k columns, which is twice the length of a non-zero segment in $\{\vec{a}_{k-1,i}\}$.

It follows that for the matrix $\mathcal{M} = [\mathcal{M}_A^T|\mathcal{M}_3^T]^T$, the rows $\{\vec{b}_{3,j}\}_{j=0}^{2^{n-1}-2}$ and $\{\vec{a}_{2,j}\}_{j=1,3,5,\ldots}^{2^{n-2}-1}$ are linearly independent and $\vec{a}_{2,j} = \vec{b}_{3,4j} - \vec{b}_{3,4j+2} + \vec{a}_{2,j+1}$, for $j = 0, 2, \ldots, 2^{n-2}-2$. Thus, $dim(A_2 \cap B_3)$ is equal to number of dependent rows in $\mathcal{M}$, i.e., 2^{n-3}, and the basis for A_3 is $\vec{a}_{3,j} = \vec{a}_{2,2j} - \vec{a}_{2,2j+1} = \vec{b}_{3,2j} - \vec{b}_{3,2j+2}$, for $j = 0, \ldots, 2^{n-3}-1$.

For the general $k = 3, \ldots, n$, the rows $\{\vec{b}_{k,j}\}_{j=0}^{2^{n-1}-2}$ and $\{\vec{a}_{k,j}\}_{j=1,3,\ldots,2^{n-(k-1)}-1}$ are linearly independent and $\vec{a}_{k,j} = \vec{a}_{k-1,2j} - \vec{a}_{k-1,2j+1} = \sum_{l=0}^{2^{k-2}-1}(-1)^l \vec{b}_{k,2j+2l}$ for $j = 0, \ldots, 2^{n-k}-1$. ■

5.6.6 The Proof of Theorem 21 for Two Node Networks

In this section we complete the proof of Theorem 21 by solving the following system of functional equations that correspond to the assumptions of Theorem 21 for arbitrary two node network (similar to Eq. 5.3):

$$\begin{aligned} p(\{\theta_{ij}\}) &= f_1(\{\theta_{i\cdot}\}_{i=1,\ldots,n-1}) g_1(\{\theta_{j|i}\}_{i=1,\ldots,n}^{j=1,\ldots,k-1}) \\ p(\{\theta_{ij}\}) &= f_2(\{\theta_{\cdot j}\}_{j=1,\ldots,k-1}) \prod_{j=1}^k h_j(\{\theta_{i|j}\}_{i=1,\ldots,n-1}) \end{aligned} \tag{5.88}$$

(The Jacobians are incorporated inside $f_1, f_2, g_1, h_1, \ldots, h_k$).

Any solution p that satisfies Eq. 5.88 satisfies also Eq. 5.3 and thus can be written in the form given by Theorem 15 (Eq. 5.4). We have, from Eqs. 5.88 and 5.4,

$$C\left[\prod_{i=1}^n \prod_{j=1}^k \theta_{ij}^{\alpha_{ij}}\right] H\left(\left\{\frac{\theta_{ij}\theta_{i+1,j+1}}{\theta_{i+1,i}\theta_{i,j+1}}|_{1\le i\le n-1}^{1\le j\le k-1}\right\}\right) = f_2(\{\theta_{\cdot j}\}_{j=1,\ldots,k-1}) \prod_{j=1}^k h_j(\{\theta_{i|j}\}_{i=1,\ldots,n-1}). \tag{5.89}$$

Changing the variables on the left hand side to $\{\theta_{\cdot j}\}$ and $\{\theta_{i|j}\}$, we have:

$$C\prod_{j=1}^k \left[x_j^{\sum_{i=1}^n \alpha_{ij}} \prod_{i=1}^n w_{ji}^{\alpha_{ij}}\right] H\left(\left\{\frac{w_{ji}w_{j+1,i+1}}{w_{j+1,i}w_{j,i+1}}|_{1\le i\le n-1}^{1\le j\le k-1}\right\}\right) = f_2(\{x_j\}_{j=1}^{k-1}) \prod_{j=1}^k h_j(\{w_{ji}\}_{i=1}^{n-1}) \tag{5.90}$$

where $x_j = \theta_{\cdot j}$ and $w_{ji} = \theta_{i|j}$ (similar to Section 5.3.1). Clearly $f_2(\{x_j\}_{j=1,\ldots,k-1}) = \prod_{j=1}^k x_j^{\sum_{i=1}^n \alpha_{ij}}$. Eliminating $\{x_i\}$, incorporating C inside $H()$, $\{w_{ji}\}$ inside h_j, and taking the logarithm, we get

$$\tilde{H}\left(\left\{\frac{w_{ji}w_{j+1,i+1}}{w_{j+1,i}w_{j,i+1}}|_{1\le i\le n-1}^{1\le j\le k-1}\right\}\right) = \sum_{j=1}^k \tilde{h}_j(\{w_{ji}\}_{i=1,\ldots,n-1}) \tag{5.91}$$

where

$$\begin{aligned} \tilde{H}(t_{1,1}, \ldots, t_{(k-1),(n-1)}) &= \ln\left[C \cdot H(t_{1,1}, \ldots, t_{(n-1),(k-1)})\right] \\ \tilde{h}_j(t_{j,1}, \ldots, t_{j,n-1}) &= \ln\left[\prod_{i=1}^n t_{ji}^{-\alpha_{ij}} h_j(t_{j,1}, \ldots, t_{j,n-1})\right] \qquad 1 \le j \le k. \end{aligned} \tag{5.92}$$

Changing the variables to $\hat{w}_{ji} = \ln(w_{ji}/w_{j,i+1})$ (this mapping is one-to-one, since $w_{jn} = 1 - \sum_{i=1}^{n-1} w_{ji}$), we obtain

$$\hat{H}\left(\left\{\hat{w}_{ji} - \hat{w}_{j+1,i}\Big|_{1\le i\le n-1}^{1\le j\le k-1}\right\}\right) = \sum_{j=1}^{k} \hat{h}_j(\{\hat{w}_{ji}\}_{i=1,\ldots,n-1}) \tag{5.93}$$

where $\hat{H}$ and $\hat{h}$ are such that

$$\begin{aligned}\tilde{H}(t_{1,1},\ldots,t_{(k-1),(n-1)}) &= \hat{H}(\ln t_{1,1},\ldots,\ln t_{(k-1),(n-1)})\\ \tilde{h}_j(t_{j,1},\ldots,t_{j,n-1}) &= \hat{h}_j(\ln \tfrac{t_{j,1}}{t_{j,2}},\ldots,\ln\tfrac{t_{j,n-2}}{t_{j,n-1}},\ln\tfrac{t_{j,n-1}}{1-\sum_{i=1}^{n-1}t_{ji}}).\end{aligned} \tag{5.94}$$

Taking derivatives of $\hat{H}()$ and $\hat{h}()$ wrt $\{\hat{w}_{ji}\}$ we get (for $1 \le i \le n$)

$$\begin{array}{ll}\frac{\partial\hat{H}}{\partial t_{ji}} = \frac{\partial\hat{h}_j}{\partial\hat{w}_{ji}} & j=1\\ \frac{\partial\hat{H}}{\partial t_{ji}} - \frac{\partial\hat{H}}{\partial t_{j-1,i}} = \frac{\partial\hat{h}_j}{\partial\hat{w}_{ji}} & 2\le j\le k-1\\ -\frac{\partial\hat{H}}{\partial t_{j-1,i}} = \frac{\partial\hat{h}_j}{\partial\hat{w}_{ji}} & j=k,\end{array} \tag{5.95}$$

where $t_{ji} = \frac{\hat{w}_{ji}}{\hat{w}_{j+1,i}}$ (for $1 \le i \le n-1$, $1 \le j \le k-1$) denote the arguments of $\hat{H}$. Summing Eqs.5.95 we get (for all $1 \le i \le n$):

$$\sum_{j=1}^{k} \frac{\partial\hat{h}_j}{\partial\hat{w}_{ji}}(\hat{w}_{j1},\ldots,\hat{w}_{j,n-1}) = 0 \tag{5.96}$$

Since the arguments of all $\hat{h}_j$'s are distinct, we get

$$\frac{\partial\hat{h}_j}{\partial\hat{w}_{ji}}(\hat{w}_{j1},\ldots,\hat{w}_{j,n-1}) = c_{ji}, \quad 1\le j\le k,\ 1\le i\le n-1 \tag{5.97}$$

where c_{ji} are arbitrary constants such that $\sum_{j=1}^{k} c_{ji} = 0$. From Eq. 5.97 we obtain

$$\hat{h}_j(\hat{w}_{j1},\ldots,\hat{w}_{j,n-1}) = c_j + \sum_{i=1}^{n-1} c_{ji}\hat{w}_{ji}, \quad 1\le j\le k. \tag{5.98}$$

and changing back to the original variables yields

$$h_j(w_{j1},\ldots,w_{j,n-1}) = \beta_j \prod_{i=1}^{n} w_{ji}^{\beta_{ji}}, \quad 1\le j\le k, \tag{5.99}$$

where $b_j, b_{j1},\ldots,b_{j,n}$ are appropriate constants. Combining the results of Eq. 5.99, Eq. 5.90 and Eq. 5.88, we obtain

$$p(\{\theta_{ij}\}) = C'\left[\prod_{i=1}^{n}\prod_{j=1}^{k}\theta_{ij}^{\gamma_{ij}}\right] \tag{5.100}$$

i.e., $\{\theta_{ij}\}$ are distributed Dirichlet. ■

5.7 Matlab Code

This section presents a matlab code for determining the symbolic form of an arbitrary globally independent distribution. This code is also available from `http://www.technion.ac.il/~rusakov`.

```
function [Mjoint] = joint_pairwise_space(nodes);
% Input: 'nodes' - list of number of states of each variable.
% Output: 'Mjoint' - each column corresponds to one argument of
%                    the arbitrary function.
%         Mjoint[i,j] describes the role of \theta_i in argument #j.
%                     1  - \theta_i is in the nominator of argument #j
%                     -1 - \theta_i is in the denominator of argument #j
%                     0  - \theta_i doesn't participate in argument #j
% Description agreement:
% States are counted from 0.
% State A_{n1,n2,...,n_n} will be described in place
% \sum_{i=1}^{i=n} n_i*tailprod(i)
% where tailprod(i) = \prod_{j=i+1}^{j=n} nodes[j]
n = size(nodes,2);
tailprod = fliplr(cumprod(fliplr(nodes))./fliplr(nodes))';
Mjoint = [];
for p = 1:n
  % construct descriptions of pairwise solutions:
  % node #p vs. all other nodes.
  M = [];
  scount = prod(nodes)/nodes(p);
  for i = 1:nodes(p)-1;
    state = zeros(size(nodes));
    state(p) = i-1;
    for j = 1:scount-1;
      % building states
```

```
      nstate = next_state(nodes,state,[p]);
      state1 = state;
      state1(p) = i;
      nstate1 = nstate;
      nstate1(p) = i;
      % building description.
      desc = zeros(1,prod(nodes));
      desc(state*tailprod + 1) = 1;
      desc(nstate1*tailprod + 1) = 1;
      desc(state1*tailprod + 1) = -1;
      desc(nstate*tailprod + 1) = -1;
      M = [M, desc'];
      % moving to next state.
      state = nstate;
    end
  end
  if isempty(Mjoint)
    Mjoint = M;
  else
    m = null( [Mjoint, M], 'r' );
    if isempty(m)
      disp('no free function');
      return
    end
    m = m(1:size(Mjoint,2),:);
    Mjoint = Mjoint*m;
  end
end
return
function [out] = next_state(nodes,state,exclude_list);
```

```
out = state;
for i = size(nodes,2):-1:1;
  if isempty(exclude_list) | sum(exclude_list == i) == 0
    out(i) = out(i) + 1;
    if out(i) >= nodes(i);
      out(i) = 0;
    else
      break;
    end
  end
end
return
```

5.8 Discussion

This chapter investigates the prior parameter distributions for discrete DAG models that arise under different independence conditions. It is shown that global independence alone can not ensure the Dirichlet prior. We have also explicitly defined the class of prior distributions that satisfy global independence assumption for binary valued networks and provided an algorithmic way to compute the functional form of admissible distributions for any discrete network.

In addition, the minimal set of global and concurrent (on the same network) local independence assumptions that result in Dirichlet prior was determined and a number of functional equations were solved. Among the open topics, the question of explicit functional form of general solution to Eq. 5.9 remains unanswered. This, however, may be of little practical importance, since the application of Eq. 5.10 and Lemma 17 can provide a solution for any specific set of random variables X.

All the results presented in this chapter were obtained under the assumption of local parameter distributions being twice differentiable and everywhere positive (A3). It is plausible that this assumption, of twice differentiability and being everywhere positive, for the probability density functions can be removed using the techniques presented in (Járai, 1998).

Chapter 6

A Bayesian LOD Score for Linkage Analysis of Complex Diseases

This chapter presents an application of Bayesian model selection to the problem of genetic linkage analysis. The objective of genetic linkage analysis is to locate genes responsible for a genetic diseases. The emphasis of genetic linkage analysis is rapidly shifting from the task of locating genes responsible for Mendelian diseases, for which one major gene causes the disease, to identifying a collection of genes and their interactions which govern what is termed complex diseases. We propose a Bayesian averaging approach based on a known method called *MMLS-C*. The difference is in the way the likelihood of data, given a specific position of a disease gene, is computed. Given a certain inheritance model, either recessive or dominant, we compute the marginal likelihood of data by averaging over all penetrance values for the given penetrance model. We have implemented this approach, called Maximizing Bayesian *LOD* score (*MBLOD*), within SUPERLINK (Fishelson & Geiger, 2002), and demonstrated via simulations its advantages. A version of this chapter has been included in (Fishelson, Rusakov & Geiger, 2004).

6.1 Introduction

The emphasis of genetic linkage analysis is rapidly shifting from the task of locating single genes causing Mendelian diseases to locating a collection of disease genes and their interactions which govern what is termed complex diseases. One method that has been pursued in the genetics literature for analyzing complex diseases is to compute the likelihood of odds (*LOD* score) assuming two inheritance models of the disease: recessive or dominant, each with 50% penetrance, commit to the model that produces the higher maximum score, and correct for multiple tests by subtracting 0.3 from the maximum score (Hodge, Abreu & Greenberg, 1997; Greenberg, Abreu & Hodge, 1998). This method, called *MMLS-C*, has been shown to work well in simulations (Greenberg, Abreu & Hodge, 1998; Abreu, Greenberg & Hodge, 1999; Abreu, Greenberg & Hodge, 1999; Greenberg & Abreu, 2001) and is used in real genetic linkage studies (Barrett & et. al., 2001; Silverman & et. al., 2002).

In this chapter, we propose a Bayesian averaging approach based on *MMLS-C*. The difference is in the way the likelihood of data, given a specific position of a disease gene, is computed. According to the Bayesian averaging approach proposed herein, the likelihood of data given an inheritance model, either recessive or dominant, is computed by averaging the likelihood of data given this model over different penetrance values, using a flat prior. As in *MMLS-C*, we commit to the model that produces the higher maximum score, and correct for multiple tests by subtracting 0.3 from the maximum score. We have implemented this approach, which we call Maximizing Bayesian *LOD* score (*MBLOD*), within SUPERLINK (Fishelson & Geiger, 2002), a genetic linkage analysis program based on Bayesian networks, and demonstrated via simulations the advantages of Bayesian averaging in the context of genetic linkage analysis.

Our evaluation methodology follows that of (Greenberg, Abreu & Hodge, 1998; Durner, Vieland & Greenberg, 1999). We generated pedigree data assuming various complex models of inheritance in which the disease is caused by two, three, or four genes. The location of one of the genes is recovered using our method and previous methods. We have shown that *MMLS-C* performs well on inheritance models that have not been tested before, and on larger families than used in previous simulation studies. For the complex inheritance models we tested, *MBLOD* has higher predictive power than *MMLS-C* and *NPL*, on the average.

The implementation of *MMLS-C* and *MBLOD* into SUPERLINK makes this program readily available to analyze a variety of complex genetic diseases. Its ability to perform computations on larger pedigrees than other programs (Fishelson & Geiger, 2002) raises the hope that the additional predictive power gained due to larger pedigrees and due to the *MBLOD* approach proposed herein will help decipher additional inheritance mechanisms of complex diseases.

The rest of this chapter is organized as follows. Sections 6.2 and 6.3 introduce selected background on genetic linkage analysis. Section 6.4 describes the Bayesian averaging method we propose. In Section 6.5 we describe complex inheritance models used for simulation studies. Experimental results are reported in Section 6.6. Finally, Section 6.7 summarizes and outlines future work.

6.2 Genetic Linkage Analysis

A *locus* is the location on the chromosome of a *gene* or a *marker*. Common markers are short tandem repeats of two, three, or four nucleotides. Each gene, or marker, can appear in several different forms, or *alleles*. The different alleles of markers are the different number of repeats. In diploid organisms, such as humans, each individual carries two copies of each autosomal locus, one copy inherited from each parent. The sequence of alleles at different loci inherited by an individual from one parent is called a haplotype, and the two haplotypes of an individual constitute this individual's genotype. When genotypes are measured, the result is a list of unordered pairs of alleles, one pair for each locus. The standard measurement procedures do not distinguish between the paternal and maternal alleles at each locus.

A *recombination* is said to have occurred between two loci, if an haplotype of an individual contains alleles that resided in different haplotypes in the individual's parent. The probability that a recombination occurs between two loci is called the *recombination fraction* and is denoted by θ. The recombination fraction is a measure of the distance between two loci. For a comprehensive background on human genetic linkage analysis consult (Terwilliger & Ott, 1994; Lange, 1997; Ott, 1999).

The input for a genetic linkage problem consists of family trees, called pedigrees, phenotype description of people in the pedigree at the trait studied (affected, unaffected, or unknown), and partial unordered genotype information for people at marker loci. The output of genetic linkage analysis is the most probable location of a disease gene under study. This is traditionally done by placing a hypothetical disease locus somewhere within a map of markers and computing the probability of data given that specific location which is represented as the recombination fractions between the disease locus and its closest flanking markers. The position which produces the highest probability (and its vicinity) is the location declared to contain a disease gene, provided the result is statistically significant.

Pedigree data is an assignment e of values to a subset E of marker and disease phenotype variables. The probability of data $Pr(E = e \mid \theta)$ is computed by assigning a specific value to each measured variable, and then summing the product of all the probability tables just described over all possible values for the rest of the variables. The likelihood function $Pr(E = e \mid \theta)$ is maximized w.r.t. the recombination fraction θ which determines the possible positions of a disease gene. The variables and probability tables described above define a Bayesian network (Friedman, Geiger & Lotner, 2000; Fishelson & Geiger, 2002).

The significance of the best recombination fraction $\theta = \theta_1$ is normally determined using the LOD score defined for an inheritance model M (say recessive M_r or dominant M_d) along with a given penetrance probability α:

$$LOD(\theta_1) = \log_{10} \mathcal{L} = \log_{10} \frac{P(Data|\theta_1, \alpha, M)}{P(Data|\theta = 0.5, \alpha, M)} \tag{6.1}$$

where $\mathcal{L}$ is the likelihood ratio. The classical practice is to declare that linkage has been found between a disease gene and a location on a chromosome when the LOD score is above 3. The threshold of 3 approximately corresponds to a p-value of 0.05 for detecting true linkage (Ott, 1999). Genetics texts and papers warn from performing LOD score calculations with several different inheritance models and penetrances, and choosing the highest outcome, because in such practice a LOD score of 3 no longer corresponds to a p-value of 0.05, often yielding unacceptably many erroneous proposals for disease genes' locations. Such errors are said to be of type I.

Almost all analyses of linkage are done by comparing the probability of pedigree data assuming a specific disease gene location, $\theta = \theta_1$, versus the probability of the null hypothesis, $\theta = 0.5$, that the disease gene is on a different chromosome, or far on the same chromosome. There is some effort however to introduce a Bayesian approach which requires the specification of a prior on the true location of the disease genes (Vieland, 1998). The interpretation of the results of a linkage study in the Bayesian framework is simply the probability of linkage as a function of the location on the genetic map. Nevertheless, this method is not widely used in linkage studies.

6.3 Linkage Analysis for Complex Diseases

Complex diseases are diseases controlled by several genes, whose influence may or may not interact with external factors such as, say, obesity. For complex diseases the problem of type I errors in standard linkage analysis is magnified considerably. One normally does not know which inheritance model and penetrance values to use, and computing *LOD* scores for many different models will often yield eventually a high but meaningless *LOD* score.

There are several methods to detect linkage that try to avoid the model-based approach. Among these methods, which are called *non-parametric*, are Affected Sib-Pairs (Suarez & Van Eerdewegh, 1984), Affected Sibships (Ethier & Hodge, 1985), Affected Pedigree Members (APM) (Weeks & Lange, 1988), and Non-Parametric Linkage (NPL) (Kruglyak, Daly, Reeve-Daly & Lander, 1996). These methods are based on the fact that if a disease gene is located close on the chromosome to some marker, than, for all models of inheritance, it segregates with that marker to offsprings more often then random segregation. For example, two siblings share on the average one allele at each position on the chromosome. However, if one observes that *affected* siblings share on the average more than one allele at a specific location, one may conclude that a disease gene resides near that marker. Non-parametric methods are appealing because one does not have to specify a model of inheritance or penetrances, and yet, suspected areas for the location of disease genes are determined. Indeed these methods, and especially *NPL* using the GENEHUNTER software (Kruglyak, Daly, Reeve-Daly & Lander, 1996), are currently the leading methods in the study of complex diseases.

The simplicity of these non-parametric methods comes at a statistical cost since they do not use the entire data. They only use affected individuals, neglecting information due to healthy members of the pedigrees.

AGH (Abreu, Greenberg & Hodge, 1999) show that *NPL* scores follow a normal distribution, which allows one to transform *NPL* to *LOD* scores via $NPL^2/(2\ln 10) \approx LOD$ for sufficiently large samples. We refer to the transformed *NPL* score as *adjusted score* for *NPL*.

GAH (Greenberg, Abreu & Hodge, 1998) proposed the following method. Find the maximum *LOD* score for a given map of genetic markers assuming a dominant inheritance model with 50% penetrance. Repeat the same computation assuming a recessive inheritance model with 50% penetrance. Commit to the model that scores higher and use the *LOD* score of the selected model minus 0.3. This method is termed *MMLS-C* for Maximizing Maximum Lod Score using a Correction. The computed value is referred to as the *adjusted score* for *MMLS-C*, and is denoted by Z. The correction of 0.3, which equals $\log_{10} 2$, compensates for testing two models rather than one (Hodge, Abreu & Greenberg, 1997). The correction of $\log_{10} g$ for g tests is a standard correction (Ott, 1999). In other words, a *LOD* score of 3.3 obtained using two tests gives the same or better level of type I errors than a *LOD* score of 3 obtained using one test.

Extensive simulations have been performed demonstrating convincingly that the simple model-based method *MMLS-C* has stronger power to detect disease genes even for complex diseases, than the described non-parametric methods (Greenberg, Abreu & Hodge, 1998; Abreu, Greenberg & Hodge, 1999; Durner, Vieland & Greenberg, 1999; Greenberg & Abreu, 2001). It has been shown for several complex inheritance models that when testing for linkage, the inheritance model at the linked locus being examined is important in the detection of linkage, and not the overall inheritance model of the disease. The effect of the other loci can be subsumed in the reduced penetrance with relatively little loss of power to detect linkage (Abreu, Greenberg & Hodge, 1999). This is the apparent reason *MMLS-C* is so successful, albeit the 50% penetrance value selected by *MMLS-C* is a bit arbitrary.

6.4 Averaging of Penetrances

We propose to take uncertainty regarding both inheritance model and penetrance value into account in the prediction of disease loci position, as follows.

We consider a class of inheritance models $\mathcal{M}$, currently consisting of recessive M_R and dominant M_D models. For each inheritance model $M \in \mathcal{M}$, we specify a prior probability $\mu(\alpha|M)$ of penetrance values given the model M. Herein we select a flat prior for the penetrances given an inheritance model, namely, $\mu(\alpha|M_d) = \mu(\alpha|M_r) = 1$ for $0 \leq \alpha \leq 1$. The likelihood of the data given an inheritance model and a specific location of a disease gene is computed as follows:

$$P(Data|\theta, M) = \int P(Data|\theta, \alpha, M)\mu(\alpha|M)d\alpha \tag{6.2}$$

Eq. 6.2 is based on the assumption that the prior $\mu(\alpha|\theta, M) = \mu(\alpha|M)$ does not depend on the true location of a disease gene θ. We now define the following *BLOD* score for evaluating linkage for any particular disease locus for a specific model of inheritance:

$$BLOD(\theta, M) = \log_{10} \frac{P(Data|\theta, M)}{P(Data|\theta = 0.5, M)}, \tag{6.3}$$

where $P(Data|\theta, M)$ is computed via Eq. 6.1. Note that we fall short of specifying a prior for M and for θ, hence this is not a fully Bayesian approach.

Our approach, called *MBLOD*, computes the Maximum *BLOD* score assuming two inheritance models, recessive or dominant, commits to the inheritance model that produced the higher maximum *BLOD*, and corrects for multiple testing by subtracting 0.3 from the maximum *BLOD*. The value computed is referred to as the *adjusted score* for *MBLOD*.

Note that *BLOD* score generalizes the standard *LOD* score computations. Had we placed all the mass on a specific penetrance value α, then the *BLOD* score would have reduced to the LOD score for the specified inheritance model with the specified penetrance.

From a user's perspective our method is non-parametric in the sense that a user need not specify a model of inheritance or penetrances, however, the computations done are clearly model-based, taking inheritance models and penetrances into account.

A *LOD* score of 3 is the standard threshold for declaring linkage. It is based on the well known fact that $P(\mathcal{L} > 10^3) < 10^{-3}$ where $\mathcal{L}$ is the likelihood ratio. The same reasoning yields a threshold of 3 for *BLOD*, provided that the probability of the null hypothesis can be written as

$$P(Data|\theta = 0.5) = \int P(Data|\theta = 0.5, \alpha)\mu(\alpha)d\alpha.$$

6.5 Inheritance Models

We generated data under different complex disease inheritance models. The first nine are two-locus models (Figure 6.1). The next four are three-locus and four-locus disease models, which generalize the respective two-locus models. For all models presented below we describe the condition necessary for an individual to be affected with probability α, which is the penetrance value.

Additive2 Model (Add2). A total of at least two mutant alleles at the two disease loci are required for a person to be affected.

Additive3 Model (Add3). A total of at least three mutant alleles at the two disease loci are required for a person to be affected.

Additive4 Model (Add4, also termed Epistatic Model RR). A total of four mutant alleles at the two disease loci are required for a person to be affected. This model corresponds to the known phenomenon of *gene duplication*.

Epistatic Model DD. At least one mutant allele at each of the two disease loci is required for a person to be affected.

Epistatic Model DR. At least one mutant allele at the linked disease locus and two mutant alleles at the unlinked disease locus are required for a person to be affected.

Epistatic Model RD. Two mutant alleles at the linked disease locus and at least one mutant allele at the unlinked disease locus are required for a person to be affected.

Complementary Action Model R+R. At least two mutant alleles at one of the two disease loci are required for a person to be affected. This model corresponds to the known phenomenon of *complementary gene action*.

Model D+D. At least one mutant allele at one of the two disease loci is required for a person to be affected. This model also corresponds to the phenomenon of *complementary gene action*.

Suppression Model R⊕R. Two mutant alleles at one of the two disease loci and at most one mutant allele at the other disease locus are required for a person to be affected. This model corresponds to the phenomenon of *suppression*.

The following four models are three-locus and four-locus disease models.

Gene Duplication Model ($\mathbf{R^3}$, $\mathbf{R^4}$). Two mutant alleles at each of the three, or four, disease loci are required for a person to be affected.

Complementary Gene Action Model ($\sum_3 \mathbf{R}$, $\sum_4 \mathbf{R}$). At least two mutant alleles at one of the three, or four, disease loci are required for a person to be affected.

6.6 Experimental Results

The experiments compare the predictive power of *MBLOD* versus *MMLS-C* and *NPL* on a variety of realistic simulated data sets, and show its superiority for most generating models examined. Run

Add2

0	0	α
0	α	α
α	α	α

Add3

0	0	0
0	0	α
0	α	α

Add4 (RR)

0	0	0
0	0	0
0	0	α

DD

0	0	0
0	α	α
0	α	α

DR

0	0	0
0	0	α
0	0	α

RD

0	0	0
0	0	0
0	α	α

R+R

0	0	α
0	0	α
α	α	α

D+D

0	α	α
α	α	α
α	α	α

R$\oplus$R

0	0	α
0	0	α
α	α	0

Figure 6.1: Penetrance tables for the two-locus inheritance models. The rows correspond to the linked disease locus, and the columns correspond to the unlinked disease locus. Both relate to the following phenotypes in the respective order: *hh, hd, dd* (*h* represents the normal allele and *d* the mutant allele).

times of *MBLOD* are 2-5 fold slower compared to *MMLS-C*, which may limit its applicability for large data sets. All experiments use SUPERLINK (Fishelson & Geiger, 2002) for *MBLOD* and *MMLS-C* scores and GENEHUNTER (Kruglyak, Daly, Reeve-Daly & Lander, 1996) for *NPL* scores via the option NPL_{all}.

Data was generated using a simulation program, retaining only families which included at least two affected siblings. In all generating models, one of the disease loci is linked to the marker map at $\theta = 0.01$, and the other disease loci are unlinked. The frequencies of the of mutant alleles at the disease loci were set so that the population prevalence is approximately 1%. For all models, the penetrance $\alpha = 1.0$ was used. For each inheritance model in each experiment, 100 sets of 20 random families were simulated. The inheritance models along with the parameters used for generating the data are termed generating models (GMs).

Several sets of experiments were performed. In the first experiment (Tables 6.1, 6.2), we used all 13 models listed in Section 6.5 to analyze pedigree data consisting of nuclear families with 2 to 5 children. The analysis used one marker with 2 alleles of equal frequencies. In the next two experiments, 6 inheritance models, $\mathbf{RR}, \mathbf{RD}, \mathbf{DR}, \mathbf{DD}, \mathbf{R+R}, \mathbf{R \oplus R}$, were used. In the second experiment (Table 6.3, Figure 6.3), pedigree data consisted of three-generation families of 12 individuals. The third experiment (Tables 6.3, 6.4) was based on multipoint analysis with 5 markers, each with 2 alleles of equal frequencies, with recombination of 0.1 between the markers. Pedigree

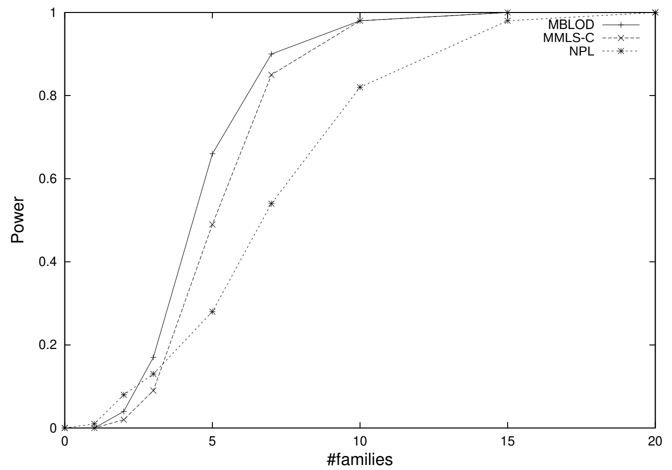


Figure 6.2: Power of each method for the **R+R** model as a function of the number of nuclear families segregating the disease. Each point is based on 100 data sets of x nuclear families.

data for this experiment consisted of nuclear families. For each experiment we present the average adjusted LOD-score computed by each analysis method, and the power to detect linkage achieved by each one. Power is defined as the percentage of data sets in which the given statistic reached a given significance level. For each analysis, the average adjusted score, denoted by Z, is computed as an average over 100 data sets.

Tables 6.1-6.4 and Figure 6.3 show that the power and average adjusted score (Z) are usually highest using *MBLOD* and lowest using *NPL*.

In the fourth experiment we used the **R+R** Model to test the power increase of *MBLOD*, *MMLS-C*, and *NPL*, as the number of nuclear families segregating the disease increased. Nuclear families with progeny generated according to the negative-binomial distribution with mean 2.8 and variance 5.3 (Cavalli-Sforza & Bodmer, 1971) were analyzed using one marker with 5 alleles of equal frequencies. Figure 6.2 summarizes the results showing that starting from 3 families, *MBLOD* retained the highest power for all levels of evidential support.

We repeated the fourth experiment for a semi-multiplicative model with three disease loci. If an individual is homozygous for the mutant allele in all three disease loci, then the penetrance is α^3. If an individual is homozygous for the mutant allele in only two of the disease loci, then the penetrance is α^2. In all other cases the penetrance is zero. A model in this spirit may be used for explaining the inheritance of the short-segment Hirschsprung disease (Gabriel, Salomon,

Table 6.1: Average Z scores for different Generating Models (GMs), for one-marker analysis on nuclear family data, using three types of analysis methods. Standard deviations are given in parenthesis.

GM	*MBLOD*	*MMLS-C*	*NPL*
Additive2:	**4.01**(1.86)	3.65(1.64)	2.35(1.13)
Additive3	**1.74**(1.27)	1.25(1.06)	1.54(0.91)
Epistatic (RR)	**4.35**(1.69)	3.97(1.54)	2.46(1.18)
Epistatic (DD)	**2.47**(1.27)	2.09(1.10)	2.13(1.00)
Epistatic (DR)	**3.80**(1.75)	3.51(1.50)	2.21(0.93)
Epistatic (RD)	**3.16**(1.51)	2.97(1.36)	2.34(0.98)
Complementation (R+R)	**3.91**(1.81)	3.66(1.64)	2.31(1.10)
Complementation (D+D)	**2.21**(1.52)	1.50(1.15)	1.90(1.11)
Suppression $\mathbf{(R \oplus R)}$	**4.00**(1.77)	3.80(1.62)	2.39(1.08)
$\mathbf{R^3}$	**5.28**(1.63)	4.81(1.41)	2.35(0.96)
$\mathbf{R^4}$	**5.40**(1.71)	4.92(1.49)	2.47(1.10)
$\sum_{\mathbf{3}} \mathbf{R}$	**5.21**(1.82)	4.77(1.54)	2.50(0.97)
$\sum_{\mathbf{4}} \mathbf{R}$	**5.38**(1.59)	5.03(1.39)	2.53(1.02)

Pelet, Angrist, Amiel, Fornage, AttiiBitach, Olson, Hofstra, Buys, Steffann, Munnich, Lyonnet & Chakravarti, 2002). The power results for *NPL*, *MMLS-C*, and *MBLOD*, assuming $\alpha = 0.8$, were 0.32, 0.60, and 0.58, respectively, for 20 families, and for 40 families, all methods reached practically the same power of 0.94, 0.96, and 0.96.

A final experiment tested whether the rate of type I errors introduced by the *MMLS-C* and *MBLOD* methods is indeed under the required level of 0.001, as guaranteed by theory. We generated 10,000 data sets each consisting of 20 nuclear families with 2-11 children, assuming dominant inheritance model with 50% penetrance and assuming the disease locus is *unlinked*. There were only two data sets for which the adjusted score for *MBLOD* exceeded the value 3, leading to a measured type I error rate of 0.0002 in this sample, far below the requirement. *MMLS-C* and *NPL* erred only once.

6.7 Discussion

In this chapter we defined the *MBLOD* method for genetic linkage analysis and tested it on simulated data sets. The results confirm that *MMLS-C* should be preferred over *NPL* in analyzing complex diseases. Furthermore, *MBLOD* provided higher power than *MMLS-C* for two-locus disease models (Tables 6.2 and 6.4, Figure 6.3), at the cost of larger run time. For three-locus and four-locus models

Table 6.2. Power to achieve a given Z value, for one-marker analysis on nuclear family data.

	Z = 2.0			Z = 3.0			Z = 4.0		
GM	*MBLOD*	*MMLS-C*	*NPL*	*MBLOD*	*MMLS-C*	*NPL*	*MBLOD*	*MMLS-C*	*NPL*
Add2	**0.87**	0.83	0.56	**0.65**	0.60	0.24	**0.47**	0.42	0.06
Add3	**0.37**	0.24	0.31	**0.20**	0.07	0.06	**0.04**	0.02	0
RR	**0.93**	0.9	0.62	**0.77**	0.70	0.27	**0.54**	0.48	0.08
DD	**0.6**	0.51	0.48	**0.35**	0.19	0.20	**0.14**	0.05	0.07
DR	**0.84**	0.82	0.56	**0.66**	0.63	0.18	**0.47**	0.38	0.05
RD	0.74	**0.75**	0.63	**0.54**	0.47	0.24	**0.32**	0.26	0.08
R+R	**0.87**	0.85	0.53	**0.66**	0.62	0.18	**0.43**	0.35	0.07
D+D	**0.49**	0.29	0.41	**0.30**	0.10	0.17	**0.12**	0.06	0.04
R⊕R	**0.86**	0.85	0.62	**0.70**	0.68	0.22	**0.45**	0.44	0.08
$\mathbf{R^3}$	**0.99**	0.98	0.56	**0.91**	0.90	0.24	**0.77**	0.69	0.06
$\mathbf{R^4}$	**0.97**	**0.97**	0.61	**0.92**	0.91	0.26	**0.8**	0.71	0.07
$\sum_3 \mathbf{R}$	0.95	**0.96**	0.67	**0.88**	0.86	0.25	**0.76**	0.7	0.06
$\sum_4 \mathbf{R}$	**0.99**	**0.99**	0.64	0.92	**0.93**	0.29	**0.79**	0.78	0.11

Table 6.3: Average Z scores for different Generating Models, using three types of analysis methods. Standard deviations are given in parenthesis. The first 3 columns refer to one-marker analysis on 3-generation family data. The next 3 columns refer to five-marker analysis on nuclear family data.

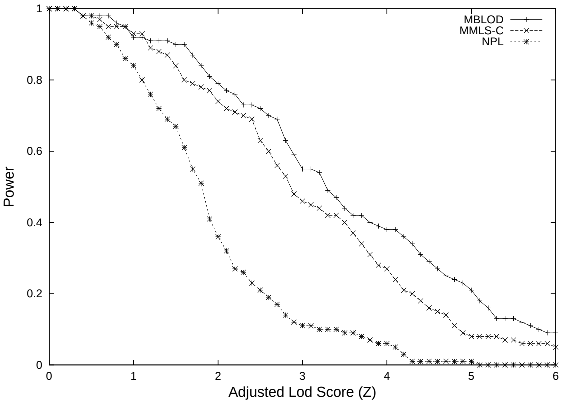
	Experiment 2			Experiment 3		
GM	*MBLOD*	*MMLS-C*	*NPL*	*MBLOD*	*MMLS-C*	*NPL*
Epistatic (RR)	**3.53**(1.72)	3.08(1.50)	1.90(0.96)	**8.72**(2.27)	7.80(1.83)	4.66(1.22)
Epistatic (DD)	1.45(1.17)	1.32(1.04)	**1.57**(0.84)	**6.21**(1.74)	4.30(1.65)	4.53(1.23)
Epistatic (DR)	**3.59**(1.65)	3.19(1.42)	1.84(0.78)	**5.52**(2.15)	5.21(1.94)	4.62(1.35)
Epistatic (RD)	**2.46**(1.41)	2.19(1.24)	1.75(0.95)	**4.18**(1.82)	3.81(1.70)	3.20(0.88)
Compl. (R+R)	**1.99**(1.32)	1.69(1.30)	1.36(1.07)	**8.83**(2.11)	7.92(1.82)	4.79(1.23)
Suppr. (R⊕R)	**2.26**(1.41)	1.93(1.29)	1.35(0.88)	**8.43**(1.90)	7.57(1.74)	**8.43**(1.90)

Table 6.4. Power to achieve a given Z value, for five-marker analysis on nuclear families data.

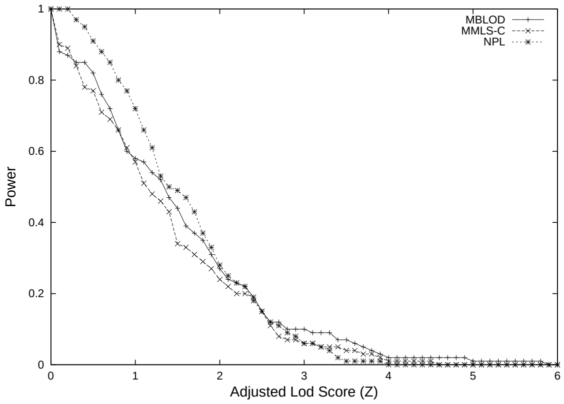
	Z = 2.0			Z = 3.0			Z = 4.0		
GM	*MBLOD*	*MMLS-C*	*NPL*	*MBLOD*	*MMLS-C*	*NPL*	*MBLOD*	*MMLS-C*	*NPL*
RR	0.99	**1.0**	**1.0**	0.98	**0.99**	0.93	**0.97**	**0.97**	0.67
DD	**0.99**	0.9	**0.99**	**0.97**	0.81	0.91	**0.9**	0.59	0.62
DR	0.94	0.94	**0.99**	0.9	0.88	**0.91**	**0.85**	0.78	0.69
RD	0.87	0.86	**0.91**	**0.7**	0.68	0.61	**0.56**	0.41	0.2
R+R	0.99	**1.0**	**1.0**	**0.99**	**0.99**	0.96	**0.99**	0.98	0.69
R⊕R	**1.0**	**1.0**	0.98	**1.0**	**1.0**	0.91	**0.98**	0.97	0.71

tested, *MMLS-C* and *MBLOD* had similar (high) power.

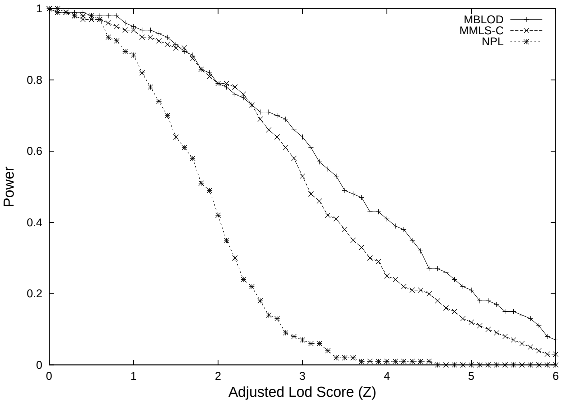
The flat prior chosen for the penetrance α in the *MBLOD* method can be replaced by other prior distributions. A particularly appealing choice is a Beta prior $\mu(\alpha) \propto \alpha^{a-1}(1-\alpha)^{b-1}$, where $n_0 = a + b$ is the imaginary sample size and $a/(a+b)$ is the maximum of the Beta distribution. The Beta distribution is a special case of the Dirichlet distribution which serves as a conjugate prior for multinomial sampling. Choosing large values for n_0 sharpens the peak around the maximum and decreases the variance, reflecting higher confidence of the user regarding the prior penetrance α. Choosing large values for the hyper-parameter a and making b equal to a is asymptotically equivalent to asserting that $\alpha = 0.5$ with certainty. If the number of loci known to cause a complex disease is above a certain number c, one may want to place the maximum of $\mu(\alpha)$, say at c^{-1}, or another function of c, with a value for n_0 that depends on the evidence supporting the constant c. Such priors have not been incorporated nor tested so far. Furthermore, one may use more than one penetrance value for each inheritance model, in which case a more sophisticated choice of a Dirichlet prior can be introduced.



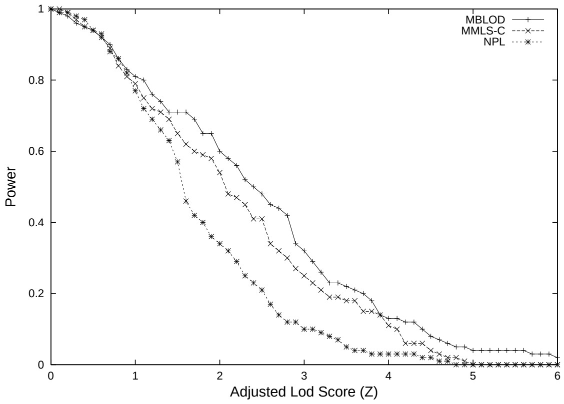
(a) RR Model



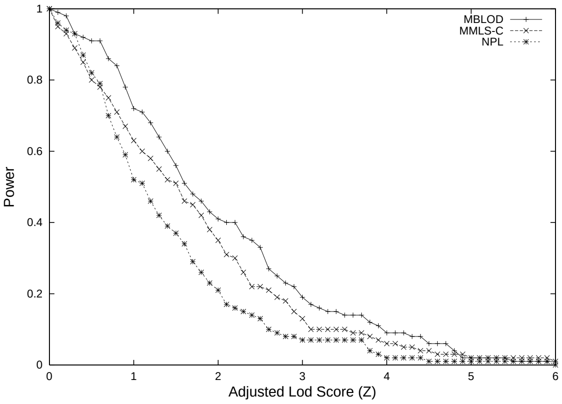
(b) DD Model



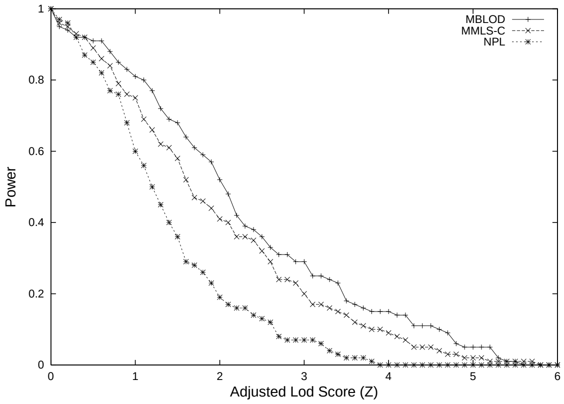
(c) RR Model



(d) DD Model



(e) R+R Model



(f) R⊕R Model

Figure 6.3: Power curves to achieve a given Z value for the Epistatic RR, Epistatic DD, Epistatic DR, Epistatic RD, Complementation R+R, and Suppression R⊕R Models, for one-marker analysis on 3-generation family data. In all models, *MBLOD* shows the highest power for $Z \geq 3$.

References

Abhyankar, S. S. (1990). *Algebraic Geometry for Scientists and Engineers.* Number 35 in Mathematical Surveys and Monographs. American Mathematical Society.

Abreu, P., Greenberg, D. A., & Hodge, S. E. (1999). Direct power comparisons between simple lod scores and npl scores for linkage analysis in complex diseases. *The American Journal of Human Genetics*, *65*(3), 847–857.

Aczél, J. (1966). *Lectures on Functional Equations and Their Applications.* Number 19 in Mathematics in Science and Engineering. Academic Press.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Atiyah, M. (1970). Resolution of singularities and division of distributions. *Communications on Pure and Applied Mathematics*, *13*, 145–150.

Barndorff-Nielsen, O. E. (1978). *Information and Exponential Families In Statistical Theory.* Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons.

Barndorff-Nielsen, O. E. & Cox, D. R. (1994). *Inference and Asymptotics.* Monographs on Statistics and Applied Probability. Chapman & Hall.

Barrett, S. & et. al. (2001). An autosomal genomic screen for autism. *The American Journal of Human Genetics*, *105*(8), 609–615.

Beinlich, I., Suermondt, G., Chavez, R., & Cooper, G. (1989). The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings of the Second European Conference on Artificial Intelligence in Medicine.*

Ben-Israel, A. & Greville, T. N. E. (1974). *Generalized Inverses: Theory and Applications.* John Wiley & Sons.

Bleistein, N. & Hendelsman, R. A. (1975). *Asymptotic Expansions of Integrals.* Holt, Rinehart and Winston.

Bodnar, G. & Schicho, J. (2000). Automated resolution of singularities for hypersurfaces. *Journal of Symbolic Computation*, (30), 401–428.

Bravo, A., Encinas, S., & Villamayor, O. (2002). A simplified proof of desingularization and applications. http://arxiv.org/abs/math.AG/0206244.

Buntine, W. L. (1994). Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, *2*, 159–225.

Cavalli-Sforza, L. L. & Bodmer, W. F. (1971). *The genetics of human populations.* San Francisco: Freeman.

Cheeseman, P. & Stutz, J. (1995). Bayesian classification (AutoClass): Theory and results. In U. Fayyad, G. Piatesky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 153–180). AAAI Press.

Chickering, D. M. & Heckerman, D. (1997). Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables. *Machine Learning*, *29*(2-3), 181–212.

Chor, B., Hendy, M. D., Holland, B. R., & Penny, D. (2000). Multiple maxima of likelihood in phylogenetic trees: An analytic approach. *Molecular Biology and Evolution*, *17*(10), 1529–1541.

Cooper, G. F. & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, *9*(4), 309–347.

Dawid, A. P. & Lauritzen, S. L. (1993). Hyper markov laws in the statistical analysis of decomposable graphical models. *Annals of Statistics*, *21*(3), 1272–1317.

DeGroot, M. H. (1970). *Optimal Statistical Decisions.* McGraw-Hill Book Company.

Durner, M., Vieland, V., & Greenberg, D. A. (1999). Further evidence for the increased power of lod scores compared with nonparametric methods. *The American Journal of Human Genetics*, *64*(1), 281–289.

Ethier, S. N. & Hodge, S. E. (1985). Identity-by-descent analysis of sibship configurations. *The American Journal of Human Genetics*, *22*, 263–272.

Fishelson, M. & Geiger, D. (2002). Exact genetic linkage computations for general pedigrees. *Bioinformatics*, *18*(Suppl. 1), S189–S198.

Fishelson, M., Rusakov, D., & Geiger, D. (2004). A Bayesian LOD score for linkage analysis of complex diseases. http://www.cs.technion.ac.il/~rusakov/archive/texts/Fishelson-Rusakov-Geiger.ps.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, *29*(2-3), 131–163.

Friedman, N., Geiger, D., & Lotner, N. (2000). Likelihood computation with value abstraction. In *Proc. 16th Conf. on Uncertainty in Artificial Intelligence (UAI).*

Gabriel, B. S., Salomon, R., Pelet, A., Angrist, M., Amiel, J., Fornage, M., AttiiBitach, T., Olson, J. M., Hofstra, R., Buys, C., Steffann, J., Munnich, A., Lyonnet, S., & Chakravarti, A. (2002). Segregation at three loci explains familial and population risk in hirschsprung disease. *Nat. Genet*, *31*(1), 89–93.

Geiger, D. & Heckerman, D. (1997). A characterization of the dirichlet distribution through global and local parameter independence. *Annals of Statistics*, *25*(3), 1344–1369.

Geiger, D. & Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Annals of Statistics*, *30*(5), 1412–1440.

Geiger, D., Heckerman, D., King, H., & Meek, C. (2001). Stratified exponential families: Graphical models and model selection. *Annals of Statistics*, *29*(2), 505–529.

Geiger, D., Heckerman, D., & Meek, C. (1996). Asymptotic model selection for directed networks with hidden variables. In Horvitz, E. & Jensen, F. (Eds.), *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, (pp. 283–290). Morgan Kaufmann Publishers, Inc.

Greenberg, D. A. & Abreu, P. (2001). Determining trait locus position from multipoint analysis: accuracy and power of three different statistics. *Genet Epidemiol*, *21*(4), 299–314.

Greenberg, D. A., Abreu, P., & Hodge, S. E. (1998). The power to detect linkage in complex disease by means of simple lod-score analyses. *The American Journal of Human Genetics*, *63*(3), 870–879.

Haughton, D. (1988). On the choice of a model to fit data from an exponential family. *Annals of Statistics*, *16*(1), 342–355.

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, *20*(3), 197–243.

Hironaka, H. (1964). Resolution of singularities of an algebraic variety over a field of characteristic zero. *Annals of Mathematics*, *7*(1,2), 109–326.

Hodge, S. E., Abreu, P., & Greenberg, D. A. (1997). Magnitude of type i error when single-locus linkage analysis is maximized over models: a simulation study. *The American Journal of Human Genetics*, *60*(4), 217–227.

Járai, A. (1998). Regularity property of the functional equation of the dirichlet distribution. *Aequationes Mathematicae*, *56*, 37–46.

Kass, R. E. & Vos, P. W. (1997). *Geometrical Foundations of Asymptotic Inference*. Wiley Series in Probability and Statistics. John Wiley & Sons.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya, Series A*, *62*(1).

Kocka, T. & Zhang, N. L. (2002). Dimension correction for hierarchical latent class models. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-02).*

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., & Lander, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *The American Journal of Human Genetics*, *58*(6), 1347–1363.

Lang, S. (1993). *Complex Analysis* (3rd ed.). Springer-Verlag.

Lange, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis.* New York: Springer-Verlag.

Lauritzen, S. L. (1996). *Graphical Models.* Number 17 in Oxford Statistical Science Series. Clarendon Press.

Malgrange, B. (1966). *Ideals of Differentiable Functions.* Oxford University Press.

Murray, J. (1984). *Asymptotic Analysis.* Number 48 in Applied Mathematical Sciences. Springer-Verlag.

Murray, M. K. & Rice, J. W. (1993). *Differential Geometry and Statistics.* Number 48 in Monographs and Statistics and Applied Probability. Chapman and Hall.

Olver, F. W. J. (1974). *Asymptotics and Special Functions.* Computer Science and Applied Mathematics: A Series of Monographs and Textbooks. Academic Press.

Ott, J. (1999). *Analysis of Human Genetic Linkage.* Baltimore, Maryland: Johns Hopkins University Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann.

Rusakov, D. & Geiger, D. (2001). On parameter priors for discrete DAG models. In Jaakkola, T. & Richardson, T. (Eds.), *Artificial Intelligence and Statistics 2001: Proceedings of the Eighth International Workshop*, (pp. 280–286). Morgan Kaufmann Publishers.

Rusakov, D. & Geiger, D. (2002). Asymptotic model selection for naive Bayesian networks. In Darwiche, A. & Friedman, N. (Eds.), *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-02).*

Rusakov, D. & Geiger, D. (2003). Automated analytic asymptotic evaluation of the marginal likelihood for latent models. In Kjaerulff, U. & Meek, C. (Eds.), *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence (UAI-03).*

Rusakov, D. & Geiger, D. (2004). Asymptotic model selection for naive Bayesian networks. Accepted to Journal of Machine Learning Research. source at http://www.technion.ac.il/˜rusakov/.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Settimi, R. & Smith, J. Q. (1998). On the geometry of Bayesian graphical models with hidden variables. In Cooper, G. F. & Moral, S. (Eds.), *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, (pp. 472–479). Morgan Kaufmann Publishers, Inc.

Settimi, R. & Smith, J. Q. (1999). On the geometry and model selection of Bayesian directed graphs with isolated hidden nodes. http://galton.uchicago.edu/˜settimi/research/ resreport.html.

Settimi, R. & Smith, J. Q. (2000). Geometry, moments and conditional independence trees with hidden variables. *Annals of Statistics*, *28*, 1179–1205.

Silverman, E. K. & et. al. (2002). Genome-wide linkage analysis of severe, early-onset chronic obstructive pulmonary disease: airflow obstruction and chronic bronchitis phenotypes. *Human Molecular Genetics*, *11*(6), 623–632.

Spiegelhalter, D. J. & Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, *20*, 579–605.

Spirtes, P., Richardson, T., & Meek, C. (1997). The dimensionality of mixed ancestral graphs. Technical Report CMU-PHIL-83, Philosophy Department, Carnegie Mellon University.

Suarez, B. K. & Van Eerdewegh, P. (1984). A comparison of three affected-sib-pair scoring methods to detect hla-linked disease susceptibility genes. *The American Journal of Human Genetics*, *18*, 135–146.

Terwilliger, J. D. & Ott, J. (1994). *Handbook of Human Genetic Linkage.* Baltimore, Maryland: Johns Hopkins University Press.

Vieland, J. V. (1998). Bayesian linkage analysis, or: How i learned to stop worrying and love the posterior probability of linkage. *The American Journal of Human Genetics*, *63*(4), 947–954.

Watanabe, S. (2001). Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, *13*(4), 899–933.

Weeks, D. E. & Lange, K. (1988). The affected-pedigree member method of linkage analysis. *The American Journal of Human Genetics*, *42*(2), 315–326.

Wong, R. (1989). *Asymptotic Approximations of Integrals.* Computer Science and Scientific Computing. Academic Press.

המחקר נעשה בהנחיית פרופ'/ח' דן גייגר
בפקולטה למדעי המחשב

הכרת תודה

אני מודה לטכניון על התמיכה הכספית הנדיבה בהשתלמותי

רשתות בייסיאניות: בחירת מודלים ושימושים

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת תואר
דקטור לפילוסופיה

דמיטרי רוסקוב

הוגש לסנט הטכניון — מכון טכנולוגי לישראל

אדר תשס"ד חיפה מרץ 2004

רשתות בייסיאניות: בחירת מודלים ושימושים

דמיטרי רוסקוב