Journal of Machine Learning Research 6 (2005) 1-35 Submitted 7/03; Revised 2/04; Published 1/05

Asymptotic Model Selection for Naive Bayesian Networks

Dmitry Rusakov RUSAKOV@CS.TECHNION.AC.IL
Dan Geiger DANG@CS.TECHNION.AC.IL
Computer Science Department
Technion - Israel Institute of Technology
Haifa, 32000, Israel

Editor: David Madigan

Abstract

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. Our work 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.

Keywords: Bayesian networks, asymptotic model selection, Bayesian information criterion (BIC)

1. Introduction

Statisticians are often faced with the problem of choosing the appropriate model that best fits a given set of observations. One example of such problem is the choice of structure in learning of Bayesian networks (Heckerman et al., 1995; Cooper and Herskovits, 1992). In such cases 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 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 $P(D|M) = \int_{\Omega} P(D|M,\omega)P(\omega|M)d\omega$. 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:

$$\mathbb{I}[N, Y_D, M] = \int_{\Omega} e^{\mathcal{L}(Y_D, N|\omega, M)}\mu(\omega|M)d\omega, \quad (1)$$

where $\mu(\omega|M)$ is the prior parameter density for model M, and $\mathcal{L}$ is the log-likelihood function of 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 Eq. 1, 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

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

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 $R = R(N,Y_D,M)$ 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 suppressed from our notation in the rest of this paper.

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 (Geiger et al., 2001; Spirtes et al., 1997). 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 $\mathbb{I}[N,Y_D]$ 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 et al., 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 et al., 1996; Settimi and 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 paper 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, is useful in classification of binary vectors into two classes (Friedman et al., 1997). Our results are derived under similar assumptions

to the ones made by Schwarz (1978) and Haughton (1988). In this sense, our paper generalizes the mentioned works, providing valid asymptotic formulas for a new type of marginal likelihood integrals. The resulting asymptotic approximations, presented in Theorem 4, 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. 2) is no longer $-\frac{d}{2}$ but rather a function of the sufficient statistics. An additional result presented in Theorem 5 describes the asymptotic marginal likelihood given a degenerate (missing links) naive Bayesian model; it complements the main result presented by Theorem 4.

The rest of this paper is organized as follows. Section 2 introduces the concept of asymptotic expansions and presents methods of asymptotic approximation of integrals. Section 3 reviews naive Bayesian models and explicates the relevant marginal likelihood integrals for these models. Section 4 states and explains our main results and Section 5 gives a proof outline of Theorem 4 that demonstrates the mathematical techniques used herein. The full proof of our theorems is deferred to Appendices A and B. Section 6 discusses our contributions and outlines future research directions.

2. Asymptotic Approximation of 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 integral $\mathbb{I}[N, Y_D]$ 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.

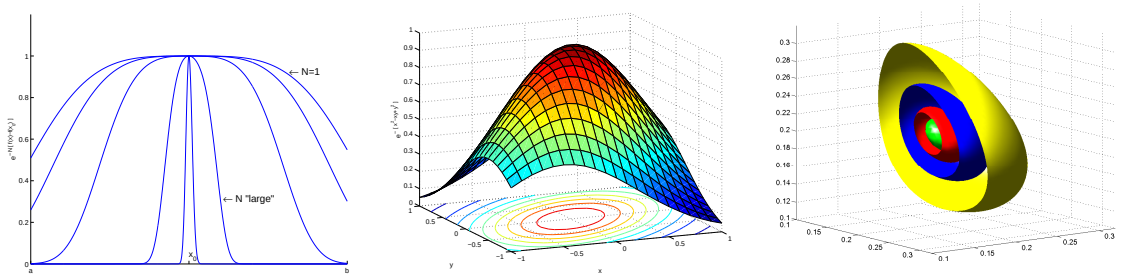
We often approximate $f(z)$ by several terms via an iterative approximation of the error terms. An asymptotic approximation by m terms has the form $f(z) = \sum_{n=1}^{m} a_n g_n(z) + o(g_m(z))$, as $z \to \infty$, where $\{g_n\}$ is an *asymptotic sequence* which means that $g_{n+1}(z) = o(g_n(z))$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

The objective of this paper is deriving asymptotic approximation of marginal likelihood integrals as represented by Eq. 1, which for exponential families have the form

$$\mathbb{I}[N, Y_D] = \int_{\Omega} e^{-N f(\omega, Y_D)} \mu(\omega) d\omega \tag{3}$$



(a) (b) (c)

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

where $f(\omega,Y_D)=-\mathcal{L}(Y_D|\omega)$ is the minus log-likelihood function. 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 statistics. Note that the specific models discussed in this paper are indeed exponential.

Consider Eq. 3 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 1(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 1b,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 1 (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

$$I(N)\sim Ce^{-Nf(u_0)}N^{-d/2}, \tag{4}$$

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. 4 yields the form of BIC score as presented by Eq. 2.

However, in many cases, and, in particular, in the case of naive Bayesian networks to be defined in the next section, the minimum of f is achieved not at a single point in Ω but rather 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.

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 $\mathbb{I}[N, Y_D]$ as $N \to \infty$ computed in a neighborhood of a maximum likelihood point.[1]

Theorem 2 (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

$$J(\lambda) = \int_{f(w)<\varepsilon} f(w)^\lambda \mu(w) dw \quad (Re(\lambda) > 0) \tag{5}$$

where $\varepsilon > 0$ is a sufficiently small constant.[2]

The above 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. 3) can be determined by the poles of

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

evaluated in the neighborhoods W_ε of points w_0 on which f attains its minimum. This claim, which is further developed in Section 5, 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.

Often, however, it is not easy to find the largest pole and multiplicity of $J(\lambda)$ defined by Eq. 5. 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 3 (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 $U_0 = g^{-1}(W_0)$.*

1. Throughout this paper we use styled '$\mathbb{I}$' symbol to denote our particular marginal likelihood integrals rather than standard 'I' symbol that denote general integrals appearing in theorems, examples and auxiliary derivations.
2. 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. (See, e.g., Lang (1993), Section 5.3.)

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 2 and 3 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 3. Select a finite cover of U_α by $U_{\alpha\beta}$, this is possible since closure of each U_α is also compact.

3. Compute the integral $J(\lambda)$ (Eq. 5) 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

$$\begin{aligned} J_{\alpha\beta}(\lambda) &= \textstyle\int_{W_{\alpha\beta}} f(w)^\lambda \mu(w) dw \\ &= \textstyle\int_{U_{\alpha\beta}} f(g_\alpha(w))^\lambda \mu(g_\alpha(u)) |g'_\alpha(u)| du \\ &= \textstyle\int_{U_{\alpha\beta}} a(u)^\lambda u_1^{\lambda k_1} u_2^{\lambda k_2} \ldots u_d^{\lambda k_d} \mu(g_\alpha(u)) \, |g'_\alpha(u)| \, du. \end{aligned} \tag{6}$$

   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 the above method, we conclude this section with an example, approximating the integral

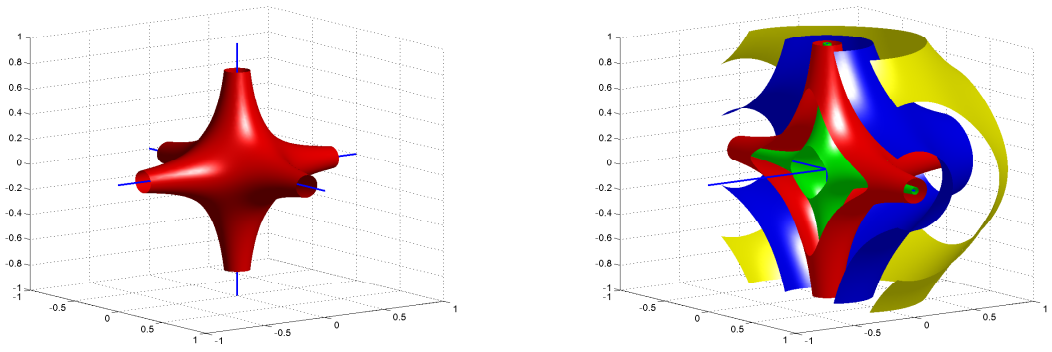
$$I[N] = \int_{-\varepsilon}^{+\varepsilon} \int_{-\varepsilon}^{+\varepsilon} \int_{-\varepsilon}^{+\varepsilon} e^{-N(u_1^2 u_2^2 + u_1^2 u_3^2 + u_2^2 u_3^2)} \, du_1 du_2 du_3 \tag{7}$$

as N tends to infinity. This approximation of $I[N]$ is an important component in establishing our main results. The key properties of the integrand function in Eq. 7 are illustrated in Figure 2.

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

$$J(\lambda) = \int_{-\varepsilon}^{+\varepsilon} \int_{-\varepsilon}^{+\varepsilon} \int_{-\varepsilon}^{+\varepsilon} (u_1^2 u_2^2 + u_1^2 u_3^2 + u_2^2 u_3^2)^\lambda \, du_1 du_2 du_3. \tag{8}$$

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



(a) (b)

Figure 2: 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 1c).

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

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

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

$$8\varepsilon^{4\lambda+3}\int_{(0,1)^2} u_1^{4\lambda+2}u_2^{2\lambda+1}du_1du_2 \;\le\; J(\lambda) \;\le\; 24\varepsilon^{4\lambda+3}\int_{(0,1)^2} u_1^{4\lambda+2}u_2^{2\lambda+1}du_1du_2, \tag{9}$$

yielding

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

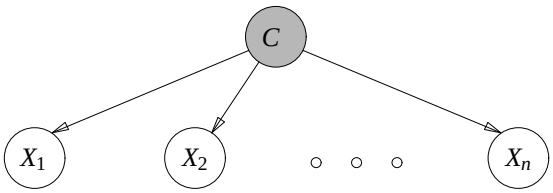


Figure 3: A naive Bayesian model. Class variable C is latent.

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 2, that $I[N]$ defined by Eq. 7 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. 6). In particular, we have established that $k_1 = 4$, $k_2 = 2$, $k_3 = 0$, $a(u) = 1 + u_3^2 + u_2^2 u_3^2$ and $|g'(u)| = u_1^2 u_2$ for the appropriate range under study. The mapping g (of Theorem 3) is the composition of the two transformations we used and is defined via $u_1 = u_1$, $u_2 = u_1 u_2$ and $u_3 = u_1 u_2 u_3$. 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 our theorems we perform a similar process of resolution of singularities producing implicitly the mapping g which is guaranteed to exist according to Theorem 3, and which determines the values of k_i and $|g'(u)|$ needed for evaluation of poles of function $J(\lambda)$ as required by Theorem 2.

3. 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, 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 Bayesian networks (Pearl, 1988) and they are widely used in clustering (Cheeseman and 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, named T, relates these two sets of parameters:

$$\theta_x = t \prod_{i=1}^{n} a_i^{x_i} (1 - a_i)^{1 - x_i} + (1 - t) \prod_{i=1}^{n} b_i^{x_i} (1 - b_i)^{1 - x_i}, \tag{10}$$

and the marginal likelihood integral (Eq. 1) for these models becomes

$$\mathbb{I}[N, Y_D] = \int_{(0,1)^{2n+1}} e^{N \sum_x Y_x \ln \theta_x(\omega)} \mu(\omega) d\omega \tag{11}$$

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.

4. Main Results

This section presents an asymptotic approximation of the integral $\mathbb{I}[N, Y_D]$ (Eq. 11) 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 et al., 2001). Second, Watanabe's approach as explained in Section 2, 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

$$y_{(\delta_1, \ldots, \delta_n)} = t \prod_{i=1}^n a_i^{\delta_i} (1 - a_i)^{1 - \delta_i} + (1 - t) \prod_{i=1}^n b_i^{\delta_i} (1 - b_i)^{1 - \delta_i} \tag{12}$$

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.

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. 12 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) 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 4 (Asymptotic Marginal Likelihood Formula) *Let $\mathbb{I}[N, Y_D]$ (Eqs. 10 and 11) 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,*

$$\begin{aligned} &\mathbb{I}[N, Y_D] = \textstyle\int_{(0,1)^{2n+1}} e^{N \sum_x Y_x \ln \theta_x(\omega)} \mu(\omega) d\omega, \\ &\theta_{(x_1, \ldots, x_n)} = t \textstyle\prod_{i=1}^n a_i^{x_i} (1 - a_i)^{1 - x_i} + (1 - t) \prod_{i=1}^n b_i^{x_i} (1 - b_i)^{1 - x_i}, \end{aligned} \tag{13}$$

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 $N \geq N_0$.*

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

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

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{2n+1}{2} \ln N + O(1), \tag{14}$$

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

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{2n-1}{2} \ln N + O(1), \tag{15}$$

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

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{n+1}{2} \ln N + O(1), \tag{16}$$

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, $Y \in S \setminus S'$),

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{3}{2} \ln N + O(1), \tag{17}$$

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

$$\ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{3}{2} \ln N + 2 \ln \ln N + O(1), \tag{18}$$

and for $n = 1$,

$$(f) \qquad \ln \mathbb{I}[N, Y_D] = N \ln P(Y|\omega_{ML}) - \frac{1}{2} \ln N + O(1), \tag{19}$$

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 paper.

Note that Eq. 15 corresponds to selecting $\lambda_1 = -\frac{2n-1}{2}$ and $m_1 = 1$ in Watanabe's method and Eq. 16 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. 14, 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 4 and its proof is explicated in Appendix B.

Theorem 5 *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,

$$\begin{aligned} &\mathbb{I}[N, Y_D] = \textstyle\int_{(0,1)^{2n+1}} e^{N \sum_x Y_x \ln \theta_x(\omega)} \mu(\omega) d\omega, \\ &\theta_x = \left(t \textstyle\prod_{i=1}^{n-m} a_i^{x_i} (1-a_i)^{1-x_i} + (1-t) \prod_{i=1}^{n-m} b_i^{x_i} (1-b_i)^{1-x_i}\right) \prod_{i=n-m+1}^{n} c_i^{x_i} (1-c_i)^{1-x_i}, \end{aligned} \tag{20}$$

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. 12) 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 $m = n$,

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

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 4. 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.

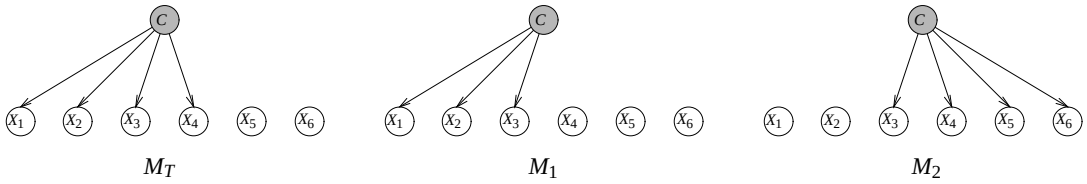


Figure 4: An example of incorrect Bayesian model selection by the standard BIC score. M_T represents the generating model, and M_1, M_2 represent models being compared. If the maximum likelihoods of data given M_1 and M_2 happen to be equal, e.g., for true model parameters $a_1 = 0.75, a_2 = 0.2, a_3 = 0.12, a_4 = 0.17, b_1 = 0.33, b_2 = 0.12, b_3 = 0.07, b_4 = 0.77, a_5 = b_5 = 0.2, a_6 = b_6 = 0.6, t = 0.42$, 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 5), on the other hand, gives an advantage to M_2, reflecting its higher marginal likelihood.

Now, if the maximum likelihoods of the data given model M_1 and given model M_2 happen to be equal, which is possible when X_4 depends strongly on C in M_T (Figure 4), then the standard choice of the model is dictated by the penalty term of the BIC score (Eq. 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 5). 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.

Note that when comparing a naive Bayesian model versus a sub-model, where the data comes from the smaller model, then the standard BIC score may underevaluate the larger model, but this would not lead to an incorrect model selection.

5. Proof Outline of Theorem 4

The proof of Theorem 4 consists of two logical parts. The first part is the proof of claim (a) of Theorem 4 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 1). The proof of Theorem 4a, which reflects standard practice, is provided in Appendix A.2. The second logical part consists of the proofs of claims (b) and (c) of Theorem 4 and requires the advanced techniques of Watanabe (Section 2). 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 focus on one thread of our proof, which demonstrates this method, deferring the full proof to Appendix A.

5.1 Useful Transformations

Decomposing the transformation T from the model parameters (a,b,t) to the joint space parameters θ_x, as defined by Eq. 13, 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 et al., 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:

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

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

$$s = 2t-1, \quad u_i = \frac{a_i - b_i}{2}, \quad x_i = t a_i + (1-t) b_i, \quad i = 1,\ldots,n. \tag{21}$$

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

$$t = (s+1)/2, \quad a_i = x_i + (1-s)u_i, \quad b_i = x_i - (1+s)u_i, \quad i = 1,\ldots,n. \tag{22}$$

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

$$0 \leq x_i \leq 1, \quad -1 \leq s \leq 1, \quad -x_i \leq (1-s)u_i \leq 1 - x_i, \quad x_i - 1 \leq (1+s)u_i \leq x_i. \tag{23}$$

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. We now express $\nu_{ij\ldots k}$ using the model parameters (a,b,t) via

$$\nu_{ij\ldots k} = t a_i a_j \ldots a_k + (1-t) b_i b_j \ldots b_k. \tag{24}$$

Using Eq. 22, we rewrite Eq. 24 obtaining

$$\begin{aligned} &\nu_i = x_i, \quad \nu_{ij} = x_i x_j + (1-s^2) u_i u_j, \\ &\nu_{ijk} = x_i x_j x_k + (1-s^2)(x_i u_j u_k + u_i x_j u_k + u_i u_j x_k) - 2s(1-s^2) u_i u_j u_k \\ &\nu_{12\ldots r} = x_1 x_2 \cdots x_r + \textstyle\sum_{i=2}^{r} p_i(s) \, (\sum \text{“products of } i \; u'\text{s and } r-i \; x'\text{s''}) \end{aligned} \tag{25}$$

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

$$z_i = \nu_i, \quad z_{ij} = \nu_{ij} - \nu_i\nu_j, \quad z_{ijk} = \nu_{ijk} - \nu_i\nu_j\nu_k - z_{ij}\nu_k - z_{ik}\nu_j - z_{jk}\nu_i, \quad \text{etc.} \tag{26}$$

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

$$z_i = x_i, \quad z_{ij} = p_2(s)u_iu_j, \quad \ldots, \quad z_{12\ldots r} = p_r(s)u_1u_2\ldots u_r \tag{27}$$

obtained by combining Eqs. 25 and 26. 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 and 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 $z_{12\ldots r} = \mathbb{E}[\prod_{i=1}^r (X_i - \mathbb{E}[X_i])]$.

5.2 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 6 *Let $\mathbb{I}[N,Y_D]$ be defined by Eq. 13, 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

$$\tilde{\mathbb{I}}[N,Y] = \int_U e^{-Nf(\theta[x,u,s])}dx\,du\,ds \tag{28}$$

where

$$\begin{aligned} &f(x,u,s) = f_Y - \textstyle\sum_{i=1}^{2^n} Y_i \ln\theta_i[x,u,s], \\ &\theta[x,u,s] = (T_3 \circ T_2)[x,u,s], \quad \theta_{2^n}[x,u,s] = 1 - \textstyle\sum_{i=1}^{2^n-1}\theta_i[x,u,s], \end{aligned} \tag{29}$$

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 A*3*, namely,* $Y_D = Y$ *for* $N \geq N_0$.

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

$$\ln \mathbb{I}[N,Y_D] = N f_Y + \ln \tilde{\mathbb{I}}[N,Y] + O(1) \tag{30}$$

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 $e^{N f_Y}$ with the constants independent on N and Y_D. Eq. 30 expresses this fact in a logarithmic scale. ■

Lemma 7 *Consider* $\tilde{\mathbb{I}}[N,Y]$ *and* $f(x,u,s)$ *as defined in Lemma 6 (Eqs. 28 and 29). 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*

$$\mathbb{J}[N,Y] = \max_{p_0 \in U_0} \mathbb{J}_{p_0}[N,Y] \quad \text{and} \quad \mathbb{J}_{p_0}[N] = \int_{U_\varepsilon \cap U} e^{-N \sum_I (z_I(x,u,s) - z'_I)^2} dx\,du\,ds, \tag{31}$$

where $z_I(x,u,s)$ *is the I-th coordinate of* $z(x,u,s) = T_2[x,u,s]$, 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 (A*2*) and* $Y \in \Upsilon_0$*, then*

$$\ln \tilde{\mathbb{I}}[N,Y] = \ln \mathbb{J}[N,Y] + O(1) \quad \text{for all } N > 1. \tag{32}$$

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 Appendix A.1.

Lemmas 6 and 7 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 *A*3).

5.3 Analysis of Type 2 Singularity

We now focus on the proof of Theorem 4c that deals with the singular points in S'. Let $Y \in S'$. Our starting point in proving Theorem 4c is integral $\mathbb{J}[N,Y]$ (Eq. 31), which by Lemmas 6 and 7 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

$$\gamma_j = \sum_{\delta \in \{0,1\}^n, \text{s.t. } \delta_j = 1} Y_{(\delta_1,\ldots,\delta_n)}, \qquad j = 1,\ldots,n. \tag{33}$$

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. 33.

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

$$U_0 = \bar{U}_{0-} \cup \bar{U}_{0+} \cup \bigcup_{j=1}^{n} \bar{U}_{0j}, \tag{34}$$

where

$$\begin{aligned}
U_{0-} &= \left\{(x=\gamma,u,s=-1) \mid u_i \in \left(\tfrac{-\gamma_i}{2}, \tfrac{1-\gamma_i}{2}\right),\ i=1,\ldots,n\right\}, \quad W_{0-} = \{(a,b=\gamma,t=0) \mid a_i \in (0,1)\},\\
U_{0+} &= \left\{(x=\gamma,u,s=1) \mid u_i \in \left(\tfrac{\gamma_i-1}{2}, \tfrac{\gamma_i}{2}\right),\ i=1,\ldots,n\right\}, \qquad W_{0+} = \{(a=\gamma,b,t=1) \mid b_i \in (0,1)\},\\
U_{0j} &= \left\{(x=\gamma,u,s) \mid \begin{array}{l} u_i = 0, \forall i \neq j; \\ u_j \in \left(-\tfrac{1}{2},\tfrac{1}{2}\right); s \in (-1,1); \\ -\gamma_j < (1-s)u_j < 1-\gamma_j, \\ \gamma_j - 1 < (1+s)u_j < \gamma_j \end{array}\right\}, \qquad W_{0j} = \left\{(a,b,t) \mid \begin{array}{l} a_i = b_i = \gamma_i, \forall i \neq j; \\ ta_j + (1-t)b_j = \gamma_j \end{array}\right\},
\end{aligned} \tag{35}$$

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 5 along with a representative point from $C1$ through $C5$.

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 the fact that the zero planes intersect (see Figure 5), 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. 31) 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 7. We now treat case $C2$ which demonstrates the main ideas, deferring the other cases to Appendix A.

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 $\phi(x,u,s) = \sum_I [z_I(x'+x,u'+u,s'+s) - z'_I]^2$. Note that $\phi(x,u,s)$ is term

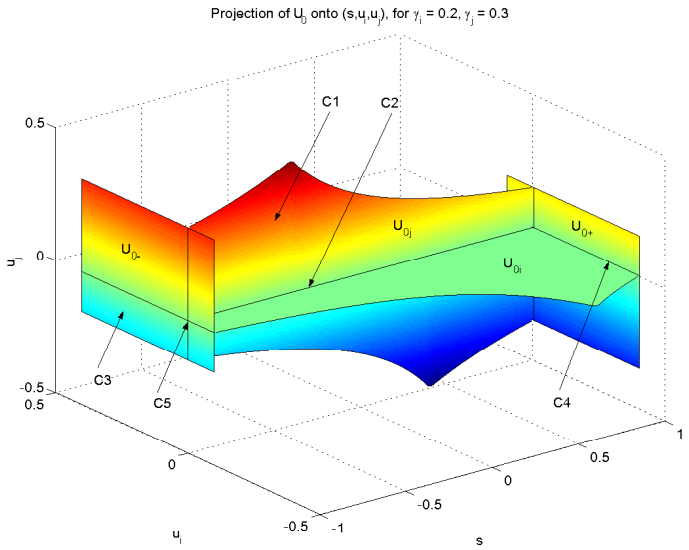


Figure 5: The set U_0 projected on (s,u_i,u_j), for $x_i=\gamma_i=0.2$, $x_j=\gamma_j=0.3$. Examples of points of types $C1$-$C5$ are marked.

in the exponent of the integrand of $\mathbb{J}[N,Y]$ centered around the minimum point (x',u',s'). Using transformation T_2 (Eq. 27), we obtain

$$\begin{aligned}\phi(x,u,s) = &\quad \textstyle\sum_I [z_I(x'+x,u'+u,s'+s)-z'_I]^2 \\ = &\quad \textstyle\sum_i [z_i-z'_i]^2+\sum_{ij,i\neq j}[z_{ij}-z'_{ij}]^2+\sum_{ijk,i\neq j\neq k}[z_{ijk}-z'_{ijk}]^2+\dots \\ = &\quad \textstyle\sum_i [(x'_i+x_i)-x'_i]^2+\sum_{ij,i\neq j}\left[(1-(s'+s)^2)u_iu_j-0\right]^2+\text{“higher order terms”} \\ = &\quad \textstyle\sum_i x_i^2+\sum_{ij,i\neq j}\left[(1-s'^2)u_iu_j-(s+2s')su_iu_j\right]^2+\text{“higher order terms”}.\end{aligned} \tag{36}$$

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. 36. For example, third terms are of form $(z_{ijk}-z'_{ijk})^2=4(s'+s)^2(1-(s'+s)^2)^2u_i^2u_j^2u_k^2\leq 5\varepsilon^2u_i^2u_j^2$ for all $s,u_i,u_j,u_k<\varepsilon$ for ε small enough. Similar bounds can be obtained for all high order terms in Eq. 36. Thus, the principal part of ϕ, that bounds ϕ within the multiplicative constant near zero, is given by

$$\tilde{\phi}(x,u,s)=\sum_i x_i^2+\sum_{ij,i\neq j}u_i^2u_j^2. \tag{37}$$

and $\tilde{\phi}(x,u,s)\leq\phi(x,u,s)\leq 2\tilde{\phi}(x,u,s)$ for all $s,u_i,u_j<\varepsilon$ 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 $\int e^{-N(\sum_i x_i^2+\sum_{ij,i\neq j}u_i^2u_j^2)}dxduds$ in order to get the asymptotic form of integral $\mathbb{J}[N,Y\in S']$ (Eq. 31) 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

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

For $n = 3$, this is precisely the integral evaluated as example in Section 2 which was found to be asymptotically equal to $cN^{-\frac{3}{4}}$. Generalizing the approach demonstrated in the example in Section 2 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 $cN^{-\frac{3n}{4}}$.

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 Appendix A). 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 6 and 7, $\ln \mathbb{I}[N, Y_D] = N \cdot P(Y|\omega_{ML}) - \frac{n+1}{2} \ln N + O(1)$, as claimed by Theorem 4c. ■

6. Discussion

This paper presents an asymptotic approximation of the marginal likelihood of data given a naive Bayesian model with binary variables (Theorem 4). 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 et al., 2001).

The major limitation of Theorem 4 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 4 shows that when comparing the classical BIC score with our adjusted BIC score (Eq. 2 versus Eqs. 15, 16), 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 4.

We conclude with two remarks. First, we note that the adjusted penalty term (Eqs. 15, 16) 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 et al., 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 paper. 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.

Acknowledgments

The second author thanks David Heckerman and Chris Meek for years of collaboration on this subject. An early version of this paper, without proofs and without Theorem 5, has been presented at the 18th UAI Conference (Rusakov and Geiger, 2002).

Appendix A. Proof of Theorem 4 (The Main Theorem)

We start with the proof of Lemma 7, which requires two additional lemmas. Then we proceed with a case by case proof of Theorem 4.

A.1 Proof of Lemma 7

The proof of Lemma 7 uses Lemmas 8 and 9. In particular, Lemma 8 states that a local version of the claim made by Lemma 7 (Eq. 32) holds in the neighborhood of extremum points p_0 under two

additional assumptions denoted by $B1$ and $B2$. Lemma 9 shows that $B1$ and $B2$ hold. Finally, the proof of Lemma 7 elevates the local version to the global claim.

Lemma 8 *Let*

$$\begin{aligned} f(x,u,s) &= f_Y - \textstyle\sum_{i=1}^{2^n} Y_i \ln \theta_i[x,u,s], \\ \theta[x,u,s] &= (T_3 \circ T_2)[x,u,s], \quad \theta_{2^n}[x,u,s] = 1 - \textstyle\sum_{i=1}^{2^n-1} \theta_i[x,u,s], \end{aligned} \tag{38}$$

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*

$$\tilde{\mathbb{I}}_{p_0}[N,Y] = \int_{U_\varepsilon} e^{-N f(x,u,s)} dx\,du\,ds, \tag{39}$$

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

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

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,

$$\ln \tilde{\mathbb{I}}_{p_0}[N,Y] = \ln \mathbb{J}_{p_0}[N] + O(1) \quad \text{for all } N > 1. \tag{40}$$

(The right hand side of Eq. 40 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

$$\eta_1 \sum_I (z_I - z'_I)^2 \;<\; f(z) \;<\; \eta_2 \sum_I (z_I - z'_I)^2, \quad \text{for } z \in \Lambda_\varepsilon, \tag{41}$$

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 $T_2(U_\varepsilon) \subseteq \Lambda_\varepsilon$ and Inequality 41 holds for $z(x,u,s) = T_2(x,u,s)$ for all points (x,u,s) in U_ε. Using Inequality 41 for evaluating $\tilde{\mathbb{I}}_{p_0}[N,Y]$ (Eq. 39) yields

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

Due to Theorem 2, the bounding integrals are asymptotically equivalent up to a multiplicative constant, because the poles and multiplicities of the corresponding $J(\lambda)$ functions (Eq. 5) that determine their asymptotic behavior are the same for any constant multiplies of $\sum (z_I(x,u,s) - z'_I)^2$, and in particular, for the multipliers η_1, η_2 and 1. ■

Lemma 9 *Let $f(u,x,s)$ be as defined by Eq. 38, namely,*

$$f(x,u,s) = f_Y - \sum_{i=1}^{2^n} Y_i \ln \theta_i[x,u,s],$$

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

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

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

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

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

Claim *B*3 follows from the proof of Theorem 12 of (Geiger et al., 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 *B*2 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 *B*3 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 *B*4 follows from claim *B*1 that $\theta' = (Y_1, \ldots, Y_{2^n-1})$ and from Theorem 13 in (Geiger et al., 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 *B*4, that satisfy $\theta' = T_2[x,u,s]$. ■

Proof of Lemma 7: Lemma 8 combined with Lemma 9 establish the asymptotic behavior of $\tilde{\mathbb{I}}[N,Y]$ in the ε_{p_0} neighborhood of a single minimum p_0 (Eq. 40). Now, since U is closed and bounded (Eq. 23), 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 $c_1 > c_2$. 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 7 (Eq. 31). We obtain

$$\mathbb{J}[N,Y] \;\le\; \tilde{\mathbb{I}}[N,Y] \;\le\; \sum_{p_0 \in U_0'} \mathbb{J}_{p_0}[N] \le k \cdot \mathbb{J}[N,Y], \tag{42}$$

where k is the number of points in U_0'. Taking the logarithm of Eq. 42 yields Eq. 32 which establishes Lemma 7. ■

Claims *B*3 and *B*4 of Lemma 9 have not been used in the proof of Lemma 7. These claims are needed in the next section.

A.2 Proof of Theorem 4a (Regular Statistics Case)

Theorem 4a rephrases standard facts regarding asymptotic expansion of integrals around a single extremum point. Recall that Theorem 4a states that if $Y_D = Y$ for $N \ge N_0$, $Y_i > 0$ for $i = 1, \ldots, 2^n$ and $Y \in \Upsilon_0 \setminus S$, then asymptotic approximation of $\ln \mathbb{I}[N, Y_D]$ (Eq. 13) equals $N \ln P(Y|w_{ML}) - \frac{2n+1}{2} \ln N + O(1)$ (Eq. 14). To prove this claim we use Lemma 6 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 1 (Laplace approximation).

We start by noticing that $\mathbb{I}[N,Y_D]$ absolutely converges for any $N \ge 1$ and $Y_D \ge 0$. That is because the integrand function $e^{N\sum_x Y_x \ln \theta_x(w)} = \prod_x \theta_x(w)^{NY_x}$ satisfies $0 \le \theta_x(w)^{NY_x} \le 1$ 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 $N \ge 1$ and any $Y \ge 0$, as required in order to use Lemma 1.

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 7). Since $\tilde{\mathbb{I}}[N,Y]$ converges and Claims *B*1, *B*3 and *B*4 of Lemma 9 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 1 (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 1 to the small neighborhoods of the two minimum points of f and by combining Eq. 30 with the logarithm of sum of two approximations described by Eq. 4 we obtain the Theorem 4a. ■

Theorem 4a 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 1. 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 $t' = 1-t''$.

A.3 Proof of Theorem 4b (Type 1 Singularity)

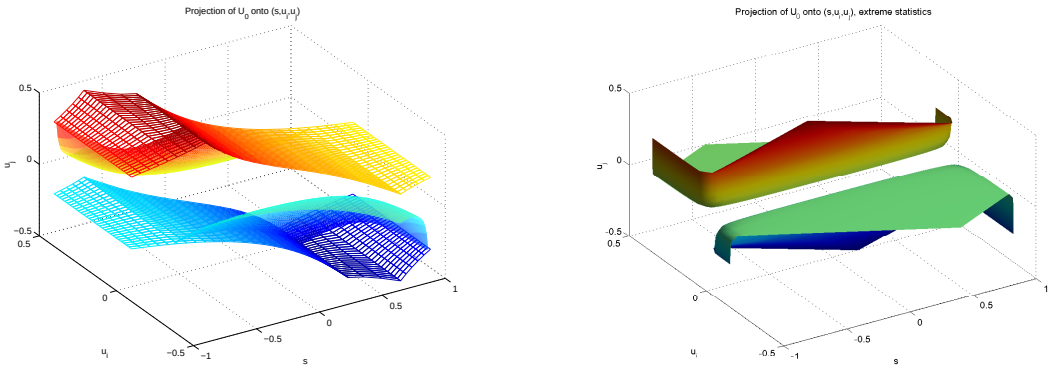
Theorem 4b states that if $Y_D = Y$ for $N \geq N_0$, $Y_i > 0$ for $i = 1,\ldots,2^n$ and $Y \in S \setminus S'$, then $\ln \mathbb{I}[N,Y_D]$ (Eq. 13) is asymptotically equal to $NP(Y|w_{ML}) - \frac{2n-1}{2}\ln N + O(1)$ (Eq. 15). To prove this claim we first employ Lemma 6, which relates $\mathbb{I}[N,Y_D]$ with $\tilde{\mathbb{I}}[N,Y]$ (Eqs. 28 and 30) and Lemma 7, which relates $\tilde{\mathbb{I}}[N,Y]$ with $\mathbb{J}[N,Y]$ (Eqs. 31 and 32). 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 1 (Laplace Approximation) no longer 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. 12) 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

$$U_0 = \left\{ (x,u,s) \in U \;\middle|\; \begin{array}{l} x_i = a_i,\ \forall i = 1,\ldots,n,\ i \neq l,k, \\ x_l = ta_l + (1-t)b_l, \\ x_k = ta_k + (1-t)b_k; \\ u_i = 0,\ \forall i = 1,\ldots,n,\ i \neq l,k, \\ u_l, u_k, s, \text{ such that } (1-s^2)u_l u_k = z'_{lk} \end{array} \right\}. \tag{43}$$

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 6.

We now apply the method of Watanabe, as described in Section 2, 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



(a) (b)

Figure 6: 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 5(b).

origin, yields

$$
\begin{aligned}
\phi(x,u,s) = \;& \textstyle\sum_I \left[z_I(x'+x,u'+u,s'+s) - z'_I\right]^2 \\
= \;& \textstyle\sum_i [z_i(x'+x,u'+u,s'+s) - z'_i]^2 \\
& + \left[z_{lk}(x'+x,u'+u,s'+s) - z'_{lk}\right]^2 \\
& + \textstyle\sum_{i\neq l,k} \left[z_{il}(x'+x,u'+u,s'+s) - z'_{il}\right]^2 + \left[z_{ik}(x'+x,u'+u,s'+s) - z'_{ik}\right]^2 \\
& + \textstyle\sum_{i,j\neq l,k} \left[z_{ij}(x'+x,u'+u,s'+s) - z'_{ij}\right]^2 + \ldots \\
= \;& \textstyle\sum_{i=1}^n [(x'_i + x_i) - x'_i]^2 \\
& + \left[(1-(s'+s)^2)(u'_l+u_l)(u'_k+u_k) - (1-s'^2)u'_l u'_k\right]^2 \\
& + \textstyle\sum_{i\neq l,k} \left[(1-(s'+s)^2)(u'_l+u_l)u_i - 0\right]^2 + \left[(1-(s'+s)^2)(u'_k+u_k)u_i - 0\right]^2 \\
& + \textstyle\sum_{i,j\neq l,k} \left[(1-(s'+s)^2)u_i u_j - 0\right]^2 + \ldots \\
= \;& \textstyle\sum_{i=1}^n x_i^2 \\
& + \left[-2s'u'_l u'_k s + (1-s'^2)u'_k u_l + (1-s'^2)u'_l u_k + \text{“smaller terms”}\right]^2 \\
& + \textstyle\sum_{i\neq l,k} \left[(1-s'^2)u'_l u_i + \text{“smaller terms”}\right]^2 + \left[(1-s'^2)u'_k u_i + \ldots\right]^2 \\
& + \textstyle\sum_{i,j\neq l,k} \left[(1-s'^2)u_i u_j + \text{“smaller terms”}\right]^2 + \ldots \,.
\end{aligned}
\tag{44}
$$

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

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

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

Consequently, in order to approximate the integral $\mathbb{J}_{p_0}[N]$ (Eq. 31) for $p_0=(x',u',s')$ with $s'\neq 0$, it remains to approximate the integral

$$\begin{aligned} &\tilde{\mathbb{J}}_1[N]=\int e^{-N\tilde{\phi}_1(x,u,s)}dxduds, \\ &\text{where}\quad \tilde{\phi}_1(x,u,s)=\textstyle\sum_i x_i^2+\left[\tilde{C}_1s+\tilde{C}_2u_l+\tilde{C}_3u_k\right]^2+\sum_{i\neq l,k}\tilde{c}_iu_i^2 \end{aligned} \tag{45}$$

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. 44) 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

$$\begin{aligned} &\tilde{\mathbb{J}}_2[N]=\int e^{-N\tilde{\phi}_2(x,u,s)}dxduds, \\ &\text{where}\quad \tilde{\phi}_2(x,u,s)=\textstyle\sum_i x_i^2+\left[\hat{C}_1s^2+\hat{C}_2u_l+\hat{C}_3u_k\right]^2+\sum_{i\neq l,k}\hat{c}_iu_i^2, \end{aligned} \tag{46}$$

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 $u'_lu'_k$, u'_k, u'_l and $u'^2_l+u'^2_k$.

From Eq. 45, by changing the coordinates to $v=\tilde{C}_1s+\tilde{C}_2u_l+\tilde{C}_3u_k$, 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 $\int e^{-N[\hat{C}_1s^2+\hat{C}_2u_i+\hat{C}_3u_j]^2}dsdu_idu_j$. The changes of variables $t=(\tilde{C}_2u_i+\tilde{C}_3u_j)/\tilde{C}_1$ transforms the remaining part of $\tilde{\mathbb{J}}_2[N]$ to

$$\tilde{\mathbb{J}}_3[N]=\int_{-\varepsilon_1}^{+\varepsilon_1}\int_{-\varepsilon_2}^{+\varepsilon_2}e^{-N(s^2+t)^2}dsdt.$$

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. 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. 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 $0<s<t<1$ gives 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

$$\frac{1}{2}\int_{0}^{1}\int_{-v}^{v}(v-u)u^{2\lambda}v^{2\lambda}dudv<\int_{0}^{1}\int_{0}^{1}2t(s^2-t^2)^{2\lambda}dsdt<\frac{1}{2}\int_{0}^{2}\int_{-v}^{v}(v-u)u^{2\lambda}v^{2\lambda}dudv. \qquad (47)$$

Computing the lower bound in Eq. 47, we obtain

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

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 6b and Figure 5, to see that as a point $Y\in S\setminus S$ approaches $Y'\in S'$, the zero set for Y depicted by Figure 6 approaches the zero set for Y' depicted in Figure 5.

A.4 Proof of Theorem 4c (Type 2 Singularity)

The outline of the proof of Theorem 4c is presented in Section 5.3 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. 31). The maximal contribution determine, according to Lemmas 6 and 7, the asymptotic behavior of the integral $\mathbb{I}[N,Y_D]$ (Eq. 11) 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') we obtain

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

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 5.3.

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

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

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

$$\begin{aligned}
\phi(x,u,s) = &\quad \textstyle\sum_I [z_I(x'+x,u'+u,s'+s) - z'_I]^2 \\
= &\quad \textstyle\sum_i [z_i(x'+x,u'+u,s'+s) - z'_i]^2 + \sum_{i\neq j}[z_{ij}(x'+x,u'+u,s'+s) - z'_{ij}]^2 \\
&\quad + \textstyle\sum_{i,k\neq j}[z_{ik}(x'+x,u'+u,s'+s) - z'_{ik}]^2 + \text{“higher order terms”} \\
= &\quad \textstyle\sum_i [(x'_i + x_i) - x'_i]^2 + \sum_{i\neq j}\left[(1-(s'+s)^2)(u'_j+u_j)u_i - 0\right]^2 \\
&\quad + \textstyle\sum_{i,k\neq j}\left[(1-(s'+s)^2)u_iu_k - 0\right]^2 + \text{“higher order terms”} \\
= &\quad \textstyle\sum_i x_i^2 + \sum_{i\neq j}\left[\mp 2su'_ju_i \mp 2su_ju_i - s^2u'_ju_i - s^2u_ju_i\right]^2 \\
&\quad + \textstyle\sum_{i,k\neq j}\left[\mp 2su_iu_k - s^2u_iu_k\right]^2 + \text{“higher order terms”} \\
\approx &\quad \textstyle\sum_i x_i^2 + s^2\sum_{i\neq j} u_i^2.
\end{aligned} \tag{48}$$

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

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

The standard change of variables to $u_i = u_1u_i$ 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

$$\begin{aligned}
\phi(x,u,s) = &\quad \textstyle\sum_I [z_I(x'+x,u'+u,s'+s) - z'_I]^2 \\
= &\quad \textstyle\sum_i [z_i(x'+x,u'+u,s'+s) - z'_i]^2 + \sum_{i,j}[z_{ij}(x'+x,u'+u,s'+s) - z'_{ij}]^2 \\
&\quad + \textstyle\sum_{i,j,k}[z_{ijk}(x'+x,u'+u,s'+s) - z'_{ijk}]^2 + \ldots \\
= &\quad \textstyle\sum_i [(x'_i + x_i) - x'_i]^2 + \sum_{i,j}\left[(1-(s'+s)^2)u_iu_j - 0\right]^2 \\
&\quad + \textstyle\sum_{i,j,k}\left[-2(s'+s)(1-(s'+s)^2)u_iu_ju_k - 0\right]^2 + \text{“higher order terms”} \\
= &\quad \textstyle\sum_i x_i^2 + \sum_{i,j}\left[\mp 2su_iu_j - s^2u_iu_j\right]^2 \\
&\quad + \textstyle\sum_{i,j,k}[4su_iu_ju_k + \text{“smaller terms”}]^2 + \text{“higher order terms”} \\
\approx &\quad \textstyle\sum_i x_i^2 + s^2\sum_{i,j} u_i^2u_j^2.
\end{aligned} \tag{49}$$

The higher order terms are bounded by some $s^2u_i^2u_j^2$ term, because of the special form of $p_i(s)$ term in $z_{12\ldots i}$ (Eq. 27). 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 $p_i(s'+s)=s\cdot 1/2(2s'+s)[(1-(s'+s))^{i-1}-(-1)^{i-1}(1+(s'+s))^{i-1}]$. Thus, any high-order term $z^2_{ij\ldots r}(x'+x,u,\pm 1+s)$ is of form

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

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 $\{ij\ldots k\}$. Consequently, this term is bounded by $s^2u_i^2u_j^2$ 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

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

The analysis is similar to the one presented in Section 2, but with additional variable s. Thus the largest pole of $J(\lambda)$ this time 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 $\sum_{l,k}u_l^2u_k^2$ contains only one term and this results in increasing the multiplicity of the pole $\lambda=-1/2$.

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 4 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 6 and 7, $\ln\mathbb{I}[N,Y_D]=NP(Y|w_{ML})-\frac{n+1}{2}\ln N+O(1)$ as claimed. ■

A.5 Proof of Claims (d,e) of Theorem 4 (Case $n=2$)

Claims (d,e) of Theorem 4 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. 13) is asymptotically equal to $NP(Y|w_{ML})-\frac{3}{2}\ln N+O(1)$ (Eq. 17) for $Y\notin S'$ and asymptotically equal to $NP(Y|w_{ML})-\frac{3}{2}\ln N+2\ln\ln N+O(1)$ (Eq. 18) for $Y\in S'$. Similar to the proofs of Claims (b,c), we first employ Lemma 6, which relates $\mathbb{I}[N,Y_D]$ with $\tilde{\mathbb{I}}[N,Y]$ (Eqs. 28 and 30) and Lemma 7, which relates $\tilde{\mathbb{I}}[N,Y]$ with $\mathbb{J}[N,Y]$ (Eqs. 31 and 32). 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 4) 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. 43 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. 45 and 46 without the $\sum_{l\neq i,j}c_lu_l^2$ 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 6 and 7 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. 34 and 35. 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 situation is in $C2$, $C4$ and $C5$ cases, 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, 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. 37. 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. 48). Similarly to the 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. 49). 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, $\ln \mathbb{I}[N,Y] = N f_Y - \frac{3}{2}\ln N + 2\ln\ln N + O(1)$. ■

A.6 Proof of Theorem 4f (Case $n = 1$)

Theorem 4f 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. 13) is asymptotically equal to $NP(Y|w_{ML}) - \frac{1}{2}\ln N + O(1)$ (Eq. 19). Once again, we first employ Lemma 6, which relates $\mathbb{I}[N,Y_D]$ with $\tilde{\mathbb{I}}[N,Y]$ (Eqs. 28 and 30) and Lemma 7, which relates $\tilde{\mathbb{I}}[N,Y]$ with $\mathbb{J}[N,Y]$ (Eqs. 31 and 32). 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 2 the asymptotic form of the integral $J_{p_0}[N] = \int_{U_\varepsilon} e^{-N(z_1(x,u,s)-z_1')^2} dx du ds$ is determined by the poles of $J(\lambda) = \int_{U_\varepsilon} (z_1(x,u,s) - z_1')^{2\lambda} dx du ds$, 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 $N f_Y - \frac{1}{2}\ln N + O(1)$. ■

We can also compute the integral $\mathbb{I}[N,Y_D]$ (Eq. 11) 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 $0 \le a < b \le 1$ (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 1) 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 $x_0 \in [a+\varepsilon, b-\varepsilon]$ and $N \ge 1$. Note that the above approximation of f is only valid for $Y_0 \ne 0,1$ (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 the 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 $\mathbb{I}_1[N,Y] \le (b-a) m^N < c_3 e^{Nf_Y} N^{-1/2}$, 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 $(a,b) \in \{(a,b) | a < Y_0, b > Y_0, b - a > 2\varepsilon > 0\}$. Since the later region has a non-zero Lebesgue measure, it follows that $\mathbb{I}[N,Y] \sim c e^{Nf_Y} N^{-1/2}$ and $\ln \mathbb{I}[N,Y] = Nf_Y - \frac{1}{2} \ln N + O(1)$.

Appendix B. Proof of Theorem 5

Theorem 5 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 degenerate part and

a naive Bayesian part. 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 gives Theorem 5.

Let ψ be the log-likelihood function of the marginal likelihood integral (Eq. 20) for the degenerate binary naive Bayesian network described in Theorem 5. We have

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

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. 20) can be rewritten as

$$\mathbb{I}[N,Y_D] \sim \hat{\mathbb{I}}[N,Y] = \left[\prod_{i=n-m+1}^{n}\int_0^1 c_i^{NY_i}(1-c_i)^{N(1-Y_i)}dc_i\right]\int_{(0,1)^{2n-2m+1}} e^{N\sum_{\tilde{x}} Y_{\tilde{x}}\ln\theta_{\tilde{x}}(\omega)}d\omega. \qquad (50)$$

where $\tilde{x} = (x_1,\ldots,x_{n-m})$. The first m integrals are integrals 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 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

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

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. 50 is exactly of the type analyzed in Theorem 4, and the theorem follows by summing up the contributions of these two parts. ■

References

Shreeram S. Abhyankar. *Algebraic Geometry for Scientists and Engineers*. Number 35 in Mathematical Surveys and Monographs. American Mathematical Society, 1990.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974.

M.F. Atiyah. Resolution of singularities and division of distributions. *Communications on Pure and Applied Mathematics*, 13:145–150, 1970.

Peter Cheeseman and John Stutz. Bayesian classification (AutoClass): Theory and results. In U. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 153–180. AAAI Press, 1995.

Gregory F. Cooper and Edward Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, October 1992.

Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill Book Company, 1970.

Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.

Dan Geiger, David Heckerman, Henry King, and Christopher Meek. Stratified exponential families: Graphical models and model selection. *Annals of Statistics*, 29(2):505–529, 2001.

Dan Geiger, David Heckerman, and Christopher Meek. Asymptotic model selection for directed networks with hidden variables. In Eric Horvitz and Finn Jensen, editors, *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, pages 283–290. Morgan Kaufmann Publishers, Inc., 1996.

Dominique Haughton. On the choice of a model to fit data from an exponential family. *Annals of Statistics*, 16(1):342–355, 1988.

David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

Heisuke Hironaka. Resolution of singularities of an algebraic variety over a field of characteristic zero. *Annals of Mathematics*, 7(1,2):109–326, 1964.

Christine Keribin. Consistent estimation of the order of mixture models. *Sankhya, Series A*, 62(1), February 2000.

Serge Lang. *Complex Analysis*. Springer-Verlag, 3rd edition, 1993.

Steffen L. Lauritzen. *Graphical Models*. Number 17 in Oxford Statistical Science Series. Clarendon Press, 1996.

James D. Murray. *Asymptotic Analysis*. Number 48 in Applied Mathematical Sciences. Springer-Verlag, 1984.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

Dmitry Rusakov and Dan Geiger. Asymptotic model selection for naive Bayesian networks. In Adnan Darwiche and Nir Friedman, editors, *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-02)*, 2002.

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

Raffaella Settimi and Jim Q. Smith. On the geometry of Bayesian graphical models with hidden variables. In Gregory F. Cooper and Serafin Moral, editors, *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 472–479. Morgan Kaufmann Publishers, Inc., 1998.

Raffaella Settimi and Jim Q. Smith. Geometry, moments and conditional independence trees with hidden variables. *Annals of Statistics*, 28:1179–1205, 2000.

Peter Spirtes, T Richardson, and Christopher Meek. The dimensionality of mixed ancestral graphs. Technical Report CMU-PHIL-83, Philosophy Department, Carnegie Mellon University, 1997.

Sumio Watanabe. Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13(4):899–933, 2001.

Roderick Wong. *Asymptotic Approximations of Integrals*. Computer Science and Scientific Computing. Academic Press, 1989.