Bernoulli **23**(2), 2017, 1202–1232
DOI: 10.3150/15-BEJ775

Marginal likelihood and model selection for Gaussian latent tree and forest models

MATHIAS DRTON[1,*], SHAOWEI LIN[2], LUCA WEIHS[1,**] and PIOTR ZWIERNIK[3]

[1]*Department of Statistics, University of Washington, Seattle, WA, USA.*
E-mail: [*]*md5@uw.edu;* [**]*lucaw@uw.edu*
[2]*Institute for Infocomm Research, Singapore. E-mail: lins@i2r.a-star.edu.sg*
[3]*Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain.*
E-mail: piotr.zwiernik@upf.edu

Gaussian latent tree models, or more generally, Gaussian latent forest models have Fisher-information matrices that become singular along interesting submodels, namely, models that correspond to subforests. For these singularities, we compute the real log-canonical thresholds (also known as stochastic complexities or learning coefficients) that quantify the large-sample behavior of the marginal likelihood in Bayesian inference. This provides the information needed for a recently introduced generalization of the Bayesian information criterion. Our mathematical developments treat the general setting of Laplace integrals whose phase functions are sums of squared differences between monomials and constants. We clarify how in this case real log-canonical thresholds can be computed using polyhedral geometry, and we show how to apply the general theory to the Laplace integrals associated with Gaussian latent tree and forest models. In simulations and a data example, we demonstrate how the mathematical knowledge can be applied in model selection.

Keywords: algebraic statistics; Gaussian graphical model; latent tree models; marginal likelihood; multivariate normal distribution; singular learning theory

1. Introduction

Graphical models based on trees are particularly tractable, which makes them useful tools for exploring and exploiting multivariate stochastic dependencies, as first demonstrated by [3]. More recent work develops statistical methodology for extensions that allow for inclusion of latent variables and in which the graph may be a forest, that is, a union of trees over disjoint vertex sets [2,14,18]. These extensions lead to a new difficulty in that the Fisher-information matrix of a latent tree model is typically singular along submodels given by subforests. As explained in [20], such singularity invalidates the mathematical arguments that lead to the Bayesian information criterion (BIC) of [16], which is widely used to guide model selection algorithms that infer trees or forests [7]. Indeed, the BIC will generally no longer share the asymptotic behavior of Bayesian methods; see also [6], Section 5.1. Similarly, Akaike's information criterion may no longer be an asymptotically unbiased estimator of the expected Kullback–Leibler divergence that it is designed to approximate [20–22].

In this paper, we study the large-sample behavior of the marginal likelihood in Bayesian inference for Gaussian tree/forest models with latent variables, with the goal of obtaining the math-

ematical information needed to evaluate a generalization of BIC proposed in [5]. As we review below, this information comes in the form of so-called real log-canonical thresholds (also known as stochastic complexities or learning coefficients) that appear in the leading term of an asymptotic expansion of the marginal likelihood. We begin by more formally introducing the models that are the object of study.

Let $Z = (Z_u)_{u \in U}$ be a random vector whose components are indexed by the vertices of an undirected tree $T = (U, E)$ with edge set E. Via the paradigm of graphical modeling [11], the tree T induces a *Gaussian tree model* $\mathbf{N}(T)$ for the joint distribution of Z. The model $\mathbf{N}(T)$ is the collection of all multivariate normal distributions on $\mathbb{R}^U$ under which Z_u and Z_v are conditionally independent given $Z_C = (Z_w : w \in C)$ for any choice of two nodes u, v and a set $C \subset U \setminus \{u, v\}$ such that C contains a node on the (unique) path between u and v. For two nodes $u, v \in U$, let $\overline{uv}$ be the set of edges on the path between u and v. It can be shown that a normal distribution with correlation matrix $R = (\rho_{uv})$ belongs to $\mathbf{N}(T)$ if and only if

$$\rho_{uv} = \prod_{e \in \overline{uv}} \rho_e, \tag{1.1}$$

where $\rho_e := \rho_{xy}$ when e is the edge incident to x and y. Indeed, for three nodes $v, w, u \in U$ the conditional independence of Z_v and Z_w given Z_u is equivalent to $\rho_{vw} = \rho_{uv}\rho_{uw}$; compare also [14], page 4359.

In this paper, we are concerned with latent tree models in which only the tree's leaves correspond to observed random variables. So let $V \subset U$ be the set of leaves of tree $T = (U, E)$. Then the *Gaussian latent tree model* $\mathbf{M}(T)$ for the distribution of the subvector $X := (Z_v : v \in V)$ is the set of all V-marginals of the distributions in $\mathbf{N}(T)$. The object of study in our work is the parametrization of the model $\mathbf{M}(T)$. Without loss of generality, we may assume that the latent variables Z_a at the inner nodes $a \in U \setminus V$ have mean zero and variance one. Moreover, we assume that the observed vector X has mean zero. Then, based on (1.1), the distributions in $\mathbf{M}(T)$ can be parametrized by the variances ω_v for each variable X_v, $v \in V$, and the edge correlations ω_e, $e \in E$.

Our interest is in the marginal likelihood of model $\mathbf{M}(T)$ when the variance and correlation parameters are given a prior distribution with smooth and everywhere positive density. Following the theory developed by [20], we will derive large-sample properties of the marginal likelihood by studying the geometry of the fibers (or preimages) of the parametrization map.

Example 1.1. Suppose T is a star tree with one inner node a that is connected to each one of three leaves, labelled 1, 2 and 3. A positive definite correlation matrix $R = (\rho_{vw}) \in \mathbb{R}^{V \times V}$ is the correlation matrix of a distribution in model $\mathbf{M}(T)$ if

$$R = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix} = \begin{pmatrix} 1 & \omega_{a1}\omega_{a2} & \omega_{a1}\omega_{a3} \\ \omega_{a1}\omega_{a2} & 1 & \omega_{a2}\omega_{a3} \\ \omega_{a1}\omega_{a3} & \omega_{a2}\omega_{a3} & 1 \end{pmatrix} \tag{1.2}$$

for a choice of the three correlation parameters $\omega_{a1}, \omega_{a2}, \omega_{a3} \in [-1, 1]$ that are associated with the three edges of the tree.

Now suppose that $R = (\rho_{vw})$ is indeed the correlation matrix of a distribution in $\mathbf{M}(T)$ and that $\rho_{vw} \neq 0$ for all $v \neq w$. Then modulo a sign change that corresponds to negating the latent variable at the inner node a, the parameters can be identified uniquely using the identities

$$\omega_{a1}^2 = \frac{\rho_{12}\rho_{13}}{\rho_{23}}, \qquad \omega_{a2} = \frac{\rho_{12}}{\omega_{a1}}, \qquad \omega_{a3} = \frac{\rho_{13}}{\omega_{a1}}.$$

Hence, the fiber of the parametrization is finite, containing two points.

If instead the correlations between the leaves are zero, then this identifiability breaks down. If R is the identity matrix with $\rho_{12} = \rho_{13} = \rho_{23} = 0$, then every vector $(\omega_{a1}, \omega_{a2}, \omega_{a3}) \in [-1, 1]^3$ that lies in the set

$$\{\omega_{a1} = \omega_{a2} = 0\} \cup \{\omega_{a1} = \omega_{a3} = 0\} \cup \{\omega_{a2} = \omega_{a3} = 0\}$$

satisfies (1.2). The fiber of the identity matrix is thus the union of three line segments that form a one-dimensional semi-algebraic set with a singularity at the origin where the lines intersect.

Remark 1.2. Some readers may be more familiar with rooted trees with directed edges and model specifications based on the Markov properties for directed graphs or structural equations. However, these are equivalent to the setup considered here, as can be seen by applying the so-called trek rule [17]. Our later results also apply to Bayesian inference in graphical models associated with directed trees.

Suppose φ is a smooth and positive density that defines a prior distribution on the parameter space $\Omega = (0, \infty)^V \times [-1, 1]^E$ of the Gaussian latent tree model $\mathbf{M}(T)$. Let $\mathbf{X}_n = (X^{(1)}, \ldots, X^{(n)})$ be a sample consisting of n independent and identically distributed random vectors in $\mathbb{R}^V$, and write $L(\mathbf{M}(T)|\mathbf{X}_n)$ for the marginal likelihood of $\mathbf{M}(T)$. If $\mathbf{X}_n$ is generated from a distribution $q \in \mathbf{M}(T)$ and $n \to \infty$, then it holds that

$$\log L\big(\mathbf{M}(T)|\mathbf{X}_n\big) - \sum_{i=1}^{n} \log q\big(X^{(i)}\big) = -\frac{\lambda_q^T}{2} \log n + \big(\mathfrak{m}_q^T - 1\big) \log\log n + O_p(1), \tag{1.3}$$

where $\lambda_q^T \geq 0$ is a rational number smaller than or equal to the dimension of the model $\mathbf{M}(T)$. The number $\mathfrak{m}_q^T$ is an integer greater than or equal to 1. More detail on how (1.3) follows from results in [20] is given in Section 2. In this paper, we derive formulas for the pair $(\lambda_q^T, \mathfrak{m}_q^T)$ from (1.3), which will be seen to depend on the pattern of zeros in the correlation matrix of the distribution q.

Let σ_{vv}^* and ρ_{vw}^* be the variances and the correlations of the data-generating distribution q. The point of departure for our work is Proposition 2.3, which clarifies that the pair $(\lambda_q^T, \mathfrak{m}_q^T)$ is also determined by the behavior of the deterministic Laplace integral

$$\int_\Omega e^{-nH_q(\omega)} \varphi(\omega)\, d\omega, \tag{1.4}$$

where the phase function in the exponent is

$$H_q(\omega) = \sum_{v \in V} \left(\omega_v - \sigma^*_{vv}\right)^2 + \sum_{\substack{v,w \in V \\ v \neq w}} \left(\prod_{e \in \overline{vw}} \omega_e - \rho^*_{vw} \right)^2.$$

In the formulation of our results, we adopt the notation

$$\mathrm{RLCT}_\Omega(H_q) := \left(\lambda_q^T, \mathfrak{m}_q^T\right),$$

as λ_q^T is sometimes referred to as real log-canonical threshold and $\mathfrak{m}_q^T$ is the threshold's multiplicity. Our formulas for $\mathrm{RLCT}_\Omega(H_q)$ are stated in Theorem 4.3. The proof of the theorem relies on facts presented in Section 3, which concern models with monomial parametrizations in general. As our formulas show, the marginal likelihood admits non-standard large-sample asymptotics, with λ_q^T differing from the model dimension if q exhibits zero correlations (recall Example 1.1). We describe the zero patterns of q in terms of a subforest F^* with edge set E^*.

Our result for trees generalizes directly to models based on forests. If $F = (U, E)$ is a forest with the set $V \subset U$ comprising the leaves of the subtrees, then we may define a Gaussian latent forest model $\mathbf{M}(F)$ in the same way as for trees. Again we assign a variance parameter ω_v to each node $v \in V$ and a correlation parameter ω_e to each edge $e \in E$. Forming products of correlations along paths, exactly as in (1.1), we obtain again a parametrization of the correlation matrix of a multivariate normal distribution on $\mathbb{R}^V$. In contrast to the case of a tree, there may be pairs of nodes with necessarily zero correlation, namely, when two leaves v and w are in distinct connected components of F. Theorem 4.7 extends Theorem 4.3 to the case of forests. The nonstandard cases arise when the data-generating distribution lies in the submodel defined by a proper subforest F^* of the given forest F.

The remainder of the paper begins with a review of the connection between the asymptotics of the marginal likelihood and that of the Laplace integral in (1.4); see Section 2 which introduces the notion of a real log-canonical threshold (RLCT). Gaussian latent tree/forest models have a monomial parametrization and we clarify in Section 3 how the monomial structure allows for calculation of RLCTs via techniques from polyhedral geometry. In Section 4, these techniques are applied to derive the above mentioned Theorems 4.3 and 4.7. In Section 5, we demonstrate how our results can be used in model selection with Bayesian information criteria (BIC). In a simulation study and an example of temperature data, we compare a criterion based on RLCTs to the standard BIC, which is based on model dimension alone.

2. Background

Consider an arbitrary parametric statistical model $\mathbf{M} = \{P_\theta : \theta \in \Theta\}$, with parameter space $\Theta \subseteq \mathbb{R}^d$. Let each distribution P_θ have density $p(x|\theta)$ and, for Bayesian inference, consider a prior distribution with density $\varphi(\theta)$ on Θ. Writing $\mathbf{X}_n = (X^{(1)}, \ldots, X^{(n)})$ for a sample of size n

from P_θ, the log-likelihood function of $\mathbf{M}$ is

$$\ell(\theta|\mathbf{X}_n) = \sum_{i=1}^{n} \log p\big(X^{(i)}|\theta\big).$$

The key quantity for Bayesian model determination is the integrated or *marginal likelihood*

$$L(\mathbf{M}|\mathbf{X}_n) = \int_\Theta e^{\ell(\theta|\mathbf{X}_n)}\varphi(\theta)\,d\theta. \tag{2.1}$$

As in the derivation of the Bayesian information criterion in [16], our interest is in the large-sample behavior of the marginal likelihood.

Let the sample $\mathbf{X}_n$ be drawn from a true distribution with density q that can be realized by the model, that is, $q(x) = p(x|\theta^*)$ for some $\theta^* \in \Theta$. Then, as we will make more precise below, the asymptotic properties of the marginal likelihood $L(\mathbf{M}|\mathbf{X}_n)$ are tied to those of the Laplace integral

$$Z_n(K_q;\varphi) = \int_\Theta e^{-nK_q(\theta)}\varphi(\theta)\,d\theta, \tag{2.2}$$

where

$$K_q(\theta) = \int \log\frac{q(x)}{p(x|\theta)}q(x)\,dx \tag{2.3}$$

is the Kullback–Leibler divergence between the data-generating distribution q and distributions in the model $\mathbf{M}$. Note that $K_q(\theta) \geq 0$ for all θ, and $K_q(\theta) = 0$ precisely when θ satisfies $p(x|\theta) = p(x|\theta^*)$. For large n, the integrand in (2.2) is equal to $\varphi(\theta)$ if $K_q(\theta) = 0$ and is negligibly small otherwise. Therefore, the main contribution to the integral $Z_n(K_q;\varphi)$ comes from a neighborhood of the zero set

$$\mathcal{V}_\Theta(K_q) = \big\{\theta \in \Theta :\, K_q(\theta) = 0\big\},$$

which we also call the *q-fiber*.

Suppose now that $\Theta \subseteq \mathbb{R}^d$ is a semi-analytic set and that $K_q : \Theta \to [0,\infty)$ is an analytic function with compact q-fiber $\mathcal{V}_\Theta(K_q)$. Suppose further that the prior density φ is a smooth and positive function. Then, under additional integrability conditions, the Main Theorem 6.2 in [20] shows that the marginal likelihood has the following asymptotic behavior as the sample size n tends to infinity:

$$\log L(\mathbf{M}|\mathbf{X}_n) \;=\; \ell\big(\theta^*|\mathbf{X}_n\big) - \frac{\lambda}{2}\log n + (\mathfrak{m}-1)\log\log n + O_p(1). \tag{2.4}$$

In (2.4), λ is a rational number in $[0,d]$, and $\mathfrak{m}$ is an integer in $\{1,\dots,d\}$. The number λ is known as *learning coefficient*, *stochastic complexity* or also *real log-canonical threshold*, and $\mathfrak{m}$ is the associated *multiplicity*. As explained in [20], Chapter 4, the pair $(\lambda,\mathfrak{m})$ also satisfies

$$\log Z_n(K_q;\varphi) = -\frac{\lambda}{2}\log n + (\mathfrak{m}-1)\log\log n + O(1). \tag{2.5}$$

Moreover, the pair $(\lambda, \mathfrak{m})$ can equivalently be defined using the concept of a zeta function as illustrated below; compare also [12].

Definition 2.1 (The real log-canonical threshold). *Let $f : \Theta \to [0, \infty)$ be a non-negative analytic function whose zero set $\mathcal{V}_\Theta(f)$ is compact and nonempty. The zeta function*

$$\zeta(z) = \int_\Theta f(\theta)^{-z/2} \varphi(\theta)\, d\theta, \qquad \mathrm{Re}(z) \le 0, \tag{2.6}$$

can be analytically continued to a meromorphic function on the complex plane. The poles of this continuation are real and positive. Let λ be the smallest pole, known as the real log-canonical threshold (r.l.c.t.) of f, and let $\mathfrak{m}$ be its multiplicity. Since we are interested in both the r.l.c.t. and its multiplicity, we use the notation $\mathrm{RLCT}_\Theta(f; \varphi) := (\lambda, \mathfrak{m})$. When $\varphi(\theta) \equiv 1$, we simply write $\mathrm{RLCT}_\Theta(f)$. Finally, if g is another analytic function with $\mathrm{RLCT}_\Theta(g; \varphi) = (\lambda', \mathfrak{m}')$, then we write $\mathrm{RLCT}_\Theta(f; \varphi) > \mathrm{RLCT}_\Theta(g; \varphi)$ if $\lambda > \lambda'$ or if $\lambda = \lambda'$ and $\mathfrak{m} < \mathfrak{m}'$.

Example 2.2. Suppose $K_q(\theta) = \theta_1^2\theta_2^2$ and $\Theta = [0, 1]^2$. Then the q-fiber $\mathcal{V}_\Theta(K_q)$ is the union of two segments of the coordinate axes. Taking $\varphi \equiv 1$, we have

$$Z_n(K_q; \varphi) = \int_0^1 \int_0^1 e^{-n\theta_1^2\theta_2^2}\, d\theta_1\, d\theta_2.$$

This example is simple enough that $\mathrm{RLCT}_\Theta(K_q)$ can be computed by elementary means. Let $\Phi(z)$ be the distribution function of the standard normal distribution. Then

$$Z_n(K_q; \varphi) = \int_0^1 \sqrt{\frac{\pi}{n\theta_2^2}}\big[\Phi(\sqrt{n}\theta_2) - \Phi(0)\big]\, d\theta_2 = \sqrt{\frac{\pi}{n}} \int_0^{\sqrt{n}} \frac{\Phi(v) - 1/2}{v}\, dv.$$

Integration by parts yields

$$\begin{aligned} Z_n(K_q; \varphi) &= \sqrt{\frac{\pi}{n}} \cdot \left[\log(v)\left(\Phi(v) - \frac{1}{2}\right)\right]_0^{\sqrt{n}} - \frac{1}{\sqrt{n}} \int_0^{\sqrt{n}} \log(v) e^{-v^2}\, dv \\ &= \sqrt{\frac{\pi}{n}} \log(\sqrt{n})\left(\Phi(\sqrt{n}) - \frac{1}{2}\right) + O\big(n^{-1/2}\big). \\ &= \frac{\sqrt{\pi}}{4} \cdot \frac{\log(n)}{\sqrt{n}}\big(1 + o(1)\big). \end{aligned}$$

Taking logarithms, we see that (2.5) holds with $\lambda = 1$ and $\mathfrak{m} = 2$. It follows that $\mathrm{RLCT}_\Theta(K_q) = (1, 2)$. Concerning Definition 2.1, we have that

$$\zeta(z) = \int_0^1 \int_0^1 \big(\theta_1^2\theta_2^2\big)^{-z/2}\, d\theta_1\, d\theta_2 = \frac{1}{(1 - z)^2}$$

for all $z \in \mathbb{C}$ with $\mathrm{Re}(z) \leq 0$. In fact, this holds as long as $\mathrm{Re}(z) < 1$. The meromorphic continuation of $\zeta(z)$ given by $1/(1-z)^2$ has one pole at $\lambda = 1$. The pole has multiplicity $\mathfrak{m} = 2$ confirming that $\mathrm{RLCT}_\Theta(K_q) = (1, 2)$.

In this paper, we are concerned with Gaussian models for which we may assume, without loss of generality, that all distributions are centered. So let the data-generating distribution q be the multivariate normal distribution $\mathcal{N}(0, \Sigma^*)$, with positive definite $k \times k$ covariance matrix $\Sigma^* = (\sigma^*_{ij})$. Further, let $p(\cdot|\theta)$ be the density of the distribution $\mathcal{N}(0, \Sigma(\theta))$ with positive definite $k \times k$ covariance matrix $\Sigma(\theta) = (\sigma_{ij}(\theta))$. Then

$$K_q(\theta) = \frac{1}{2}\left(\mathrm{tr}\big(\Sigma(\theta)^{-1}\Sigma^*\big) - k - \log\left(\frac{\det \Sigma^*}{\det \Sigma(\theta)}\right)\right).$$

For fixed positive definite Σ^*, the function

$$\Phi \mapsto \frac{1}{2}\left(\mathrm{tr}\big(\Phi^{-1}\Sigma^*\big) - k - \log\left(\frac{\det \Sigma^*}{\det \Phi}\right)\right)$$

has a full rank Hessian at $\Phi = \Sigma^*$. Hence, in a neighborhood of Σ^*, we can both lower-bound and upper-bound K_q by positive multiples of the function

$$\tilde{K}_q(\theta) = \sum_{i \leq j} \big(\sigma_{ij}(\theta) - \sigma^*_{ij}\big)^2.$$

It follows that $\mathrm{RLCT}_\Theta(K_q; \varphi) = \mathrm{RLCT}_\Theta(\tilde{K}_q; \varphi)$; compare [20], Remark 7.2. For our study of Gaussian latent tree (and forest) models, it is convenient to change coordinates to correlations and consider the function

$$H_q(\theta) = \sum_{i=1}^{k} \big(\sigma_{ii}(\theta) - \sigma^*_{ii}\big)^2 + \sum_{i<j} \big(\rho_{ij}(\theta) - \rho^*_{ij}\big)^2, \tag{2.7}$$

where ρ^*_{ij} and $\rho_{ij}(\theta)$ are the correlations obtained from Σ^* or $\Sigma(\theta)$; so, for example, $\rho^*_{ij} = \sigma^*_{ij}/\sqrt{\sigma^*_{ii}\sigma^*_{jj}}$. Since

$$\mathrm{RLCT}_\Theta\big(K_q(\theta); \varphi\big) = \mathrm{RLCT}_\Theta\big(H_q(\theta); \varphi\big), \tag{2.8}$$

our discussion of latent tree models may thus start from the following fact.

Proposition 2.3. *Let $T = (U, E)$ be a tree with set of leaves $V \subset U$. Let $\Omega = (0, \infty)^V \times [-1, 1]^E$ be the parameter space for the Gaussian latent tree model $\mathbf{M}(T)$, the parameters being the variances ω_v, $v \in V$, and the correlation parameters ω_e, $e \in E$. Suppose the (data-generating) distribution q is in $\mathbf{M}(T)$ and has variances $\sigma^*_{vv} > 0$ and a positive definite correla-*

tion matrix with entries ρ^*_{vw}. *Then* $\mathrm{RLCT}_\Omega(K_q;\varphi) = \mathrm{RLCT}_\Omega(H_q;\varphi)$, *where*

$$H_q(\omega) = \sum_{v\in V}\bigl(\omega_v - \sigma^*_{vv}\bigr)^2 + \sum_{\substack{v,w\in V\\ v\neq w}}\Biggl(\prod_{e\in\overline{vw}}\omega_e - \rho^*_{vw}\Biggr)^2. \tag{2.9}$$

3. Monomial parametrizations

According to Proposition 2.3, the asymptotic behavior of the marginal likelihood of a Gaussian latent tree model is determined by the real log-canonical threshold of the function H_q in (2.9). This function is a sum of squared differences between monomials formed from the parameter vector ω and constants determined by the data-generating distribution q. In this section, we formulate general results on the real log-canonical thresholds for such monomial parametrizations, which also arise in other contexts [15,24].

Specifically, we treat functions of the form

$$H(\omega) = \sum_{i=1}^{k}\bigl(\omega^{u_i} - c^*_i\bigr)^2, \qquad \omega\in\Omega, \tag{3.1}$$

with domain $\Omega\subseteq\mathbb{R}^d$. Here, $c^*_1,\ldots,c^*_k\in\mathbb{R}$ are constants and each monomial $\omega^{u_i} := \omega_1^{u_{i1}}\cdots\omega_d^{u_{id}}$ is given by a vector of nonnegative integers $u_i = (u_{i1},\ldots,u_{id})$. Special cases of this setup are the *regular* case with $H(\omega) = \omega_1^2+\cdots+\omega_d^2$, and the *quasi-regular* case of [23], in which the vectors u_i have pairwise disjoint supports and all $c^*_i = 0$.

Let r be the number of summands on the right-hand side of (3.1) that have $c^*_i\neq 0$. Without loss of generality, assume that $c^*_1,\ldots,c^*_r\neq 0$ and $c^*_{r+1} = \cdots = c^*_k = 0$. Furthermore, suppose that $\omega_1,\ldots,\omega_s$ are the parameters appearing in the monomials $\omega^{u_1},\ldots,\omega^{u_r}$, that is, $\bigcup_{i=1}^r\{j : u_{ij}>0\} = \{1,\ldots,s\}$. If $H(\omega)=0$ then $\omega_i\neq 0$ for all $i=1,\ldots,s$. Moreover, if the zero set $\mathcal{V}_\Omega(H) = \{\omega\in\Omega : H(\omega)=0\}$ is compact, then each one of the parameters $\omega_1,\ldots,\omega_s$ is bounded away from zero on $\mathcal{V}_\Omega(H)$. (Clearly, the zero set of the function H_q from Proposition 2.3 is compact.)

Now define the *nonzero part* H^1 of H as

$$H^1(\omega_1,\ldots,\omega_s) := \sum_{i=1}^{r}\bigl(\omega^{u_i} - c^*_i\bigr)^2 \tag{3.2}$$

and the *zero part* H^0 of H as

$$H^0(\omega_{s+1},\ldots,\omega_d) := \sum_{i=r+1}^{k}\prod_{j=s+1}^{d}\omega_j^{2u_{ij}}. \tag{3.3}$$

Definition 3.1. *The* Newton polytope $\Gamma(H^0)$ *of the zero part* H^0 *is the convex hull of the points* $(u_{ij} : s+1 \le j \le d) \in \mathbb{R}^{d-s}$ *for* $i = r+1, \dots, k$. *The* Newton polyhedron *of* H^0 *is the polyhedron*

$$\Gamma_+\big(H^0\big) := \big\{x + y \in \mathbb{R}^{d-s} : x \in \Gamma\big(H^0\big),\, y \in [0,\infty)^{d-s}\big\}.$$

Let $\mathbf{1} = (1, \dots, 1) \in \mathbb{R}^{d-s}$ *be the vector of all ones. Then the* $\mathbf{1}$-distance *of* $\Gamma_+(H^0)$ *is the smallest* $t \in \mathbb{R}$ *such that* $t\mathbf{1} \in \Gamma_+(H^0)$. *The associated multiplicity is the co-dimension of the (inclusion-minimal) face of* $\Gamma_+(H^0)$ *containing* $t\mathbf{1}$.

We say that $A \subseteq \mathbb{R}^d$ is a *product of intervals* if $A = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d]$ with $a_i < b_i \in \mathbb{R} \cup \{-\infty, \infty\}$. The following is the main result of this section. It is proved in Appendix A.

Theorem 3.2. *Suppose that* Ω *is a product of intervals, and let* Ω_1 *and* Ω_0 *be the projections of* Ω *onto the first* s *and the last* $d-s$ *coordinates, respectively. Let* H *be the sum of squares from* (3.1) *and assume that the zero set* $\{\omega \in \Omega : H(\omega) = 0\}$ *is nonempty and compact. Let* $\varphi : \Omega \to (0,\infty)$ *be a smooth positive function that is bounded above on* Ω. *Then*

$$\mathrm{RLCT}_\Omega(H; \varphi) = (\lambda_0 + \lambda_1, \mathfrak{m}),$$

where λ_1 *is the co-dimension of* $\mathcal{V}_{\Omega_1}(H^1) = \{\omega \in \Omega_1 : H^1(\omega) = 0\}$ *in* $\mathbb{R}^s$, *and* $1/\lambda_0$ *is the* $\mathbf{1}$*-distance of the Newton polyhedron* $\Gamma_+(H^0)$ *with associated multiplicity* $\mathfrak{m}$. *Here,* $\lambda_0 = 0$ *and* $\mathfrak{m} = 1$ *if* H *has no zero part, that is,* $s = d$.

Remark 3.3. In order to compute the co-dimension of $\mathcal{V}_{\Omega_1}(H^1)$, one may consider one orthant at a time and take logarithms (accounting for signs). This turns the equations $H^1(\omega) = 0$ into linear equations in $\log \omega_1, \dots, \log \omega_s$.

Example 3.4. If $H(\omega) = \omega_1^2 + \cdots + \omega_d^2$ and $\Omega = \mathbb{R}^d$, then (2.2) is a Gaussian integral and it is clear (cf. (2.5)) that $\mathrm{RLCT}_\Omega(H) = (d, 1)$. The Newton polytope for $H^0 = H$ is the convex hull of the canonical basis vectors of $\mathbb{R}^d$. The Newton polyhedron of H has $\mathbf{1}$-distance $1/d$ with multiplicity 1. The same is true whenever

$$H(\omega_1, \dots, \omega_d) = \omega_1^2 + \cdots + \omega_d^2 + \text{“higher even order terms”}. \tag{3.4}$$

Example 3.5. Earlier, we have shown that on $\Omega = [0,1]^2$ the function $H(\omega) = \omega_1^2\omega_2^2$ has $\mathrm{RLCT}_\Omega(H) = (1, 2)$; recall Example 2.2. The function has no nonzero part. Its Newton polytope consists of a single point, namely, $(1, 1)$. The Newton polyhedron is $[1,\infty)^2$. Clearly, the $\mathbf{1}$-distance of the Newton polyhedron is 1. Since the ray spanned by $\mathbf{1}$ meets the Newton polyhedron in the vertex $(1, 1)$, the multiplicity is 2, as it had to be according to our earlier calculation.

Example 3.6. Consider the function

$$H(\omega) = (\omega_1\omega_2 - 1)^2 + \omega_1^2\omega_3^2 + \omega_2^2\omega_3^2 + \omega_3^2\omega_4^2$$

on $\Omega = [-2,2]^4$. The nonzero part is $H^1(\omega_1,\omega_2) = (\omega_1\omega_2 - 1)^2$ and the zero part is $H^0(\omega_3,\omega_4) = 2\omega_3^2 + \omega_3^2\omega_4^2$. With $\Omega_1 = [-2,2]^2$, the co-dimension of $\mathcal{V}_{\Omega_1}(H^1)$ is $\lambda_1 = 1$. The Newton polytope of H^0 is the convex hull of $(1,0)$ and $(1,1)$. The Newton polyhedron of H^0 is $[1,\infty)\times[0,\infty)$. Hence, $\lambda_0 = 1$ and $\mathfrak{m} = 1$. Note that while the point $(1,1)$ is a vertex of the Newton polytope, it lies on a one-dimensional face of the Newton polyhedron. In conclusion, $\mathrm{RLCT}_\Omega(H) = (2,1)$.

4. Gaussian latent tree and forest models

Let $T = (U,E)$ be a tree with set of leaves V. By Proposition 2.3, our study of the marginal likelihood of the Gaussian latent tree model $\mathbf{M}(T)$ turns into the study of the function

$$H_q(\omega) = \sum_{v\in V}\big(\omega_v - \sigma^*_{vv}\big)^2 + \sum_{\substack{v,w\in V\\ v\neq w}}\Big(\prod_{e\in\overline{vw}}\omega_e - \rho^*_{vw}\Big)^2. \tag{4.1}$$

Since $\sigma^*_{vv} > 0$ for all $v\in V$, the split of H_q into its zero and nonzero part depends solely on the zero pattern among the correlations ρ^*_{vw} of the data-generating distribution q. Furthermore, from the form of the parametrization in (1.1), it is clear that zero correlations can arise only if one sets $\omega_e = 0$ for one or more edges e in the edge set E. For a fixed set $E_0\subseteq E$, the set of parameter vectors $\omega\in\Omega$ with $\omega_e = 0$ for all $e\in E_0$ parametrizes the forest model $\mathbf{M}(F_0)$, where F_0 is the forest obtained from T by removing the edges in E_0. In this submodel, $\rho_{vw}\equiv 0$ if and only if v and w lie in two different connected components of F_0.

It is possible that two different subforests induce the same pattern of zeros among the correlations of the data-generating distribution q. However, there is always a unique minimal forest $F^*(q) = (U^*,E^*)$ inducing this zero pattern, and we term $F^*(q)$ the *q-forest*. Put differently, the q-forest $F^*(q)$ is obtained from T by first removing *all* edges $e\in\overline{uv}$ for all pairs of nodes $u,v\in U$ that can have zero correlation under q and then removing all inner nodes of T that have become isolated. Isolated leaf nodes are retained so that $V\subseteq U^*$. In the remainder of this section, we take $E_0 = E\setminus E^*$ to be the set of edges whose removal defines $F^*(q)$. We write $v\sim w$ if v and w are two leaves in V that are joined by a path in the q-forest $F^*(q)$.

Example 4.1. Let T be the quartet tree in Figure 1(a). Let q have $\rho^*_{12}\neq 0$ but $\rho^*_{vw} = 0$ for all other $\{v,w\}\subseteq V = \{1,2,3,4\}$. The q-forest $F^*(q)$ is obtained by removing the edges in $E_0 = \{\{a,b\},\{b,3\},\{b,4\}\}$. Inner node b becomes isolated and is removed as well. The forest $F^*(q)$ thus has the five nodes in the set $U^* = \{1,2,3,4,a\}$, and the two edges in the set $E^* = \{\{1,a\},\{2,a\}\}$; see Figure 1(b).

Moving on to the decomposition of the function from (4.1), recall that we divide the parameter vector ω into coordinates $(\omega_1,\dots,\omega_s)$ that never vanish on the q-fiber $\mathcal{V}_\Omega(H_q)$ and the remaining part $(\omega_{s+1},\dots,\omega_d)$. In our case, $(\omega_1,\dots,\omega_s)$ consists of all ω_v for $v\in V$ and ω_e for $e\in E^*$ and

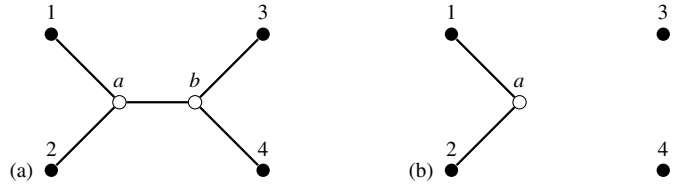


Figure 1. (a) A quartet tree T; (b) the q-forest from Example 4.1.

$(\omega_{s+1}, \ldots, \omega_d)$ consists of ω_e for $e \in E_0 = E \setminus E^*$. Moreover,

$$H_q^1(\omega_1, \ldots, \omega_s) = \sum_{v \in V} \left(\omega_v - \sigma_{vv}^*\right)^2 + \sum_{\substack{v,w \in V \\ v \neq w, v \sim w}} \left(\prod_{e \in \overline{vw}} \omega_e - \rho_{vw}^*\right)^2 \tag{4.2}$$

and

$$H_q^0(\omega_{s+1}, \ldots, \omega_d) = \sum_{v \nsim w} \prod_{e \in \overline{vw} \cap E_0} \omega_e^2. \tag{4.3}$$

The Gaussian latent tree model $\mathbf{M}(T)$ given by a tree T with set of leaves V and edge set E has dimension

$$\dim \mathbf{M}(T) = |V| + |E| - l_2,$$

where l_2 denotes the number of degree two nodes in T. Similarly, the model given by a forest F with set of leaves V and edge set E has dimension

$$\dim \mathbf{M}(F) = \sum_{i=1}^{r} \dim \mathbf{M}(T_i) = |V| + |E| - l_2,$$

where $T_1, \ldots, T_r$ are the trees defined by the connected components of F and l_2 is again the number of degrees two nodes.

Example 4.2. The q-forest F^* from Example 4.1 has $\dim \mathbf{M}(F^*) = 4 + 2 - 1 = 5$. The dimensions for the trees in the forest F^* are $\dim \mathbf{M}(T_1^*) = 3$, $\dim \mathbf{M}(T_2^*) = 1$, and $\dim \mathbf{M}(T_3^*) = 1$; the trees T_2^* and T_3^* each contain only a single node.

The following theorem provides the real log-canonical thresholds of Gaussian latent tree models. The proof of theorem is given in Appendix B.

Theorem 4.3. *Let $T = (U, E)$ be a tree with set of leaves $V \subset U$, and let q be a distribution in the Gaussian latent tree model $\mathbf{M}(T)$. Write $\Omega = (0, \infty)^V \times [-1, 1]^E$ for the parameter space of $\mathbf{M}(T)$, and let $F^*(q) = (U^*, E^*)$ be the q-forest. If $\varphi : \Omega \to (0, \infty)$ is a smooth positive function*

that is bounded above on Ω*, then the function* H_q *from* (4.1) *has*

$$\mathrm{RLCT}_{\Omega}(H_q;\varphi) = \left(\dim \mathbf{M}\big(F^*(q)\big) + \frac{\sum_{e\in E\setminus E^*} w(e)}{2}, 1 + l_2'\right),$$

where $w(e) = |e \cap U^*| \in \{0,1,2\}$ *is the number of nodes that* e *shares with* $F^*(q)$*, and* l_2' *is the number of nodes in* T *that have degree two and are not in* U^**.*

Theorem 4.3 implies in particular that the pair $(\lambda_q^T, \mathfrak{m}_q^T)$ depends on q only through the forest $F^*(q)$ and we write

$$\lambda_{F^*(q),T} := \lambda_q^T, \qquad \mathfrak{m}_{F^*(q),T} := \mathfrak{m}_q^T.$$

Example 4.4. In Example 4.1, $\dim \mathbf{M}(F^*) = 5$ (cf. Example 4.2) and $\sum_{e\in E_0} w(e) = 3$. Hence, the real log-canonical threshold $\lambda_{F^*(q),T}$ is $13/2$, which translates into a coefficient of $13/4$ for the $\log n$ term in the asymptotic expansion of the log-marginal likelihood. Note that the threshold $13/2$ is smaller than $\dim \mathbf{M}(T) = 9$, making the latent tree model behave like a lower-dimensional model.

Example 4.5. Suppose T has two leaves, labelled 1 and 2, and one inner node a, which then necessarily has degree two. If q is a distribution under which the random variables at the two leaves are uncorrelated, then we have

$$H_q(\omega) = \big(\omega_1 - \sigma_{11}^*\big)^2 + \big(\omega_2 - \sigma_{22}^*\big)^2 + (\omega_{1a}\omega_{2a})^2.$$

Using the calculation from Example 2.2 or Example 3.5, we see that $\mathrm{RLCT}_{\Omega}(H_q) = (3,2)$. When applying Theorem 4.3, the q-forest F^* has the leaves 1 and 2 isolated and $\dim \mathbf{M}(F^*) = 2$. Since $l_2' = 1$ and each one of the two removed edges satisfies $w(e) = 1$, the formula from Theorem 4.3 yields $\mathrm{RLCT}_{\Omega}(H_q) = (3,2)$, as it should.

Remark 4.6. Note that if T has an (inner) node of degree two, then we can contract one of the edges the node is adjacent to obtain a tree $\tilde{T}$ with $\mathbf{M}(\tilde{T}) = \mathbf{M}(T)$. Repeating such edge contraction it is always possible to find a tree with all inner nodes of degree at least three that defines the same model as the original tree T. Moreover, in applications such as phylogenetics, the trees of interesting do not have nodes of degree two, in which case the multiplicity in RLCT is always equal to one.

In the model selection problems that motivate this work, we wish to choose between different forests. We thus state an explicit result for forests in the below Theorem 4.7. For a forest F, we define q-forests in analogy to the definition we made for trees. In other words, we apply the previous definitions to each tree appearing in the connected components of F and then form the union of the results. Similarly, the proof of Theorem 4.7 is obtained by simply applying Theorem 4.3 to each connected component of the given forest F.

Theorem 4.7. *Let $F = (U, E)$ be a forest with the set of leaves $V \subset U$, and let q be a distribution in the Gaussian latent forest model $\mathbf{M}(F)$. Write $\Omega = (0, \infty)^V \times [-1, 1]^E$ for the parameter space of $\mathbf{M}(F)$, and let $F^*(q) = (U^*, E^*)$ be the q-forest. If $\varphi : \Omega \to (0, \infty)$ is a smooth positive function that is bounded above on Ω, then the function H_q from (4.1) has*

$$\mathrm{RLCT}_\Omega(H_q; \varphi) := \big(\lambda_q^F, \mathfrak{m}_q^F\big) = \left(\dim \mathbf{M}\big(F^*(q)\big) + \frac{\sum_{e \in E \setminus E^*} w(e)}{2}, 1 + l_2'\right),$$

where $w(e) = |e \cap U^*| \in \{0, 1, 2\}$ is the number of nodes that e shares with $F^*(q)$, and l_2' is the number of nodes in F that have degree two and are not in U^*.

As in Theorem 4.3, the pair $(\lambda_q^F, \mathfrak{m}_q^F)$ depends on q only through the forest $F^*(q)$ and we write

$$\lambda_{F^*(q),F} := \lambda_q^F, \qquad \mathfrak{m}_{F^*(q),F} := \mathfrak{m}_q^F.$$

Remark 4.8. Fix a forest $F = (U, E)$ with leaves $V \subset U$, and let $F^* = (U^*, E^*)$ be any subforest of F with the same leaves (any $F^*(q)$ is of this form). Let d_F and d_{F^*} be such that $d_F(u)$ is the degree of u in F for all $u \in U$ and similarly for d_{F^*}. Note that

$$\sum_{e \in E \setminus E^*} w(e) = \sum_{u \in U^*} \big(d_F(u) - d_{F^*}(u)\big).$$

From this and our prior formula for $\dim \mathbf{M}(F^*)$, we have that

$$\lambda_{F^*,F} = \big|U^*\big| + \big|E^*\big| - l_2 + \frac{1}{2} \sum_{u \in U^*} \big(d_F(u) - d_{F^*}(u)\big),$$

where l_2 is the number of degree 2 nodes in F^*. Computing $\lambda_{F^*,F}$ can now easily be done in linear time in the size of F, that is, in $O(|U| + |E|) = O(|U|)$ time, under the assumption that we have stored F and F^* as adjacency lists and there is a map, with $O(1)$ access time, associating vertices in F^* with those in F. In computational practice, we found that the prior two conditions are trivial to guarantee. In particular, note that if F and F^* are stored as adjacency lists we may simply loop over these lists, taking $O(|U| + |E| + |U^*| + |E^*|) = O(|U|)$ time, and precompute d_F, d_{F^*}, l_2, $|U^*|$, and $|E^*|$. Computing $\lambda_{F^*,F}$ is then simply a matter of summing over $u \in U^*$ and using the precomputed values of $d_F(u)$ and $d_{F^*}(u)$, taking $O(U^*)$ time. Similarly, noting that $l_2' = \sum_{u \in U \setminus U^*} 1_{[d_F(u)=2]}$, we have that $\mathfrak{m}_{F^*,F}$ can also be computed in linear time in the size of F.

5. Singular BIC for latent Gaussian tree models

In this section, we consider the model selection problem of inferring the forest F underlying a Gaussian latent forest model $\mathbf{M}(F)$ based on a sample of independent and identically distributed observations $\mathbf{X}_n = (X^{(1)}, \ldots, X^{(n)})$. To this end, we consider Bayesian information criteria that

are inspired by the developed large-sample theory for the marginal likelihood $L(\mathbf{M}(F)|\mathbf{X}_n)$. Note that for all the following simulations the space of models we consider implicitly include only forests and trees without degenerate degree 2 nodes; as described in Remark 4.6, this results in an RLCT whose multiplicity is always 1.

As stated in (1.3) and (2.4), the RLCTs found in Section 4 give the coefficients for logarithmic terms that capture the main differences between the log-marginal likelihood and the log-likelihood of the true data-generating distribution q. Let $\hat{q}_F$ be the maximum likelihood estimator of q in the Gaussian latent forest model $\mathbf{M}(F)$. By the results of [4], if $q \in \mathbf{M}(F)$ and $n \to \infty$, then

$$\sum_{i=1}^{n} \bigl[\log \hat{q}_F\bigl(X^{(i)}\bigr) - \log q\bigl(X^{(i)}\bigr)\bigr] = O_p(1)$$

and thus, by (2.4), we also have

$$\begin{aligned} &\log L\bigl(\mathbf{M}(F)|\mathbf{X}_n\bigr) \\ &\quad= \sum_{i=1}^{n} \log \hat{q}_F\bigl(X^{(i)}\bigr) - \frac{\lambda_{F^*(q),F}}{2} \log n + (\mathfrak{m}_{F^*(q),F} - 1) \log\log n + O_p(1). \end{aligned} \tag{5.1}$$

The pair $(\lambda_{F^*(q),F}, \mathfrak{m}_{F^*(q),F})$ on the right-hand side still depends on the unknown data-generating distribution q through the forest $F^*(q)$. However, the pair is a discontinuous function of q and plugging in the MLE $\hat{q}_F$ has little appeal. Instead, we will consider a criterion proposed by [5], in which one averages over the possible values of $(\lambda_{F',F}, \mathfrak{m}_{F',F})$ for all subforests F' of F. As in [5], we refer to the resulting model selection score as the 'singular Bayesian information criterion', or sBIC for short.

We briefly describe how sBIC is computed. Let $\mathcal{F}$ be the set of forests in the model selection problem, which we assume to contain the empty forest $F_\varnothing = (V, \varnothing)$. Note that every forest $F \in \mathcal{F}$ has set of leaves V. For forest $F \in \mathcal{F}$ with subforest $F' \in \mathcal{F}$, let $(\lambda_{F',F}, \mathfrak{m}_{F',F})$ be the pair from (5.1) when the distribution q has F' as q-forest, that is, $F^*(q) = F'$. Theorem 4.7 gives the value of this RLCT pair. Define

$$L'_{F'F} = n^{-\lambda_{F'F}/2} (\log n)^{\mathfrak{m}_{F'F}-1} \prod_{i=1}^{n} \hat{q}_F\bigl(X^{(i)}\bigr), \tag{5.2}$$

which is a proxy for the marginal likelihood $L(\mathbf{M}(F)|\mathbf{X}_n)$ obtained by exponentiating the right hand side of (5.1) and omitting the $O_p(1)$ remainder. For each $F \in \mathcal{F}$, the sBIC of model $\mathbf{M}(F)$ is defined as $\log x_F$, where $(x_F : F \in \mathcal{F})$ is the unique positive solution to the equation system

$$\sum_{F' \subseteq F} \bigl(x_F - L'_{F'F}\bigr) x_{F'} = 0, \qquad F \in \mathcal{F}. \tag{5.3}$$

The system (5.3) is triangular and can be solved by recursively solving univariate quadratic equations. The starting point is the case when F is the empty forest $F_\varnothing$, for which $F' = F_\varnothing$ is the only possible q-forest and (5.3) gives $x_{F_\varnothing}(x_{F_\varnothing} - L'_{F_\varnothing F_\varnothing}) = 0$. The sBIC of the model

$\mathbf{M}(F_\varnothing)$ is thus $\log L'_{F_\varnothing F_\varnothing}$, which coincides with the usual BIC as the relevant RLCT is given by $\lambda_{F_\varnothing F_\varnothing} = \dim \mathbf{M}(F_\varnothing) = |V|$ and $\mathfrak{m}_{F_\varnothing F_\varnothing} = 1$. When the forest F is nonempty, the sBIC and the BIC of $\mathbf{M}(F)$ differ.

In [5], sBIC is motivated by considering weighted averages of the approximations $L'_{F'F}$, with the weights being data-dependent. Furthermore, it is shown that the sBIC of $\mathbf{M}(F)$ differs from $\log L(\mathbf{M}(F)|\mathbf{X}_n)$ by an $O_p(1)$ remainder whenever data are generated from a distribution $q \in \mathbf{M}(F)$, even if q lies in a strict submodel $\mathbf{M}(F^*) \subset \mathbf{M}(F)$. The same is true for BIC only if $q \in \mathbf{M}(F)$ does not belong to any strict submodel (i.e., all edge and path correlations are nonzero and F equals the q-forest F^*). In what follows, we explore the differences between the RLCT-based sBIC and the dimension-based BIC in two simulation studies and on a temperature data set.

5.1. Simulation studies

The first task we consider is selection a subforest of a given tree T, where each subforest as well the tree T share a set of leaves V, or in other words, each subforest is a q-forest for some $q \in \mathbf{M}(T)$. When ordering edge sets by inclusion, the set of all subforests of T becomes a poset that we denote by $\mathcal{P}_T$. The poset is a lattice with the empty graph (with $|V|$ isolated nodes) as minimal element and the tree T as maximal element. To select a forest, we optimize BIC and sBIC, respectively, over the set $\mathcal{P}_T$. Maximum likelihood estimates are computed with an EM algorithm, in which we repeatedly maximize the conditional expectation of the complete-data log-likelihood function of forest models $\mathbf{N}(F)$ for a random vector Z comprising both the observed variables at the leaves in V and the latent variables at the inner nodes of F; recall the notation from the Introduction.

As a concrete example, we choose T to be the tree in Figure 2(a). We generate data from a distribution q that lies in $\mathbf{M}(T)$ but under which the third leaf is independent from all other leaves. The corresponding q-forest F^* is depicted on Figure 2(b). We choose q to have covariance matrix

$$\Sigma^* = \begin{bmatrix} 1 & 0.13 & 0 & 0.22 & 0.36 \\ 0.13 & 1 & 0 & 0.22 & 0.13 \\ 0 & 0 & 1 & 0 & 0 \\ 0.22 & 0.22 & 0 & 1 & 0.22 \\ 0.36 & 0.13 & 0 & 0.22 & 1 \end{bmatrix} \tag{5.4}$$

which is obtained by taking all edge correlations equal to 0.6. We then generate a random sample of size n from $N(\mathbf{0}, \Sigma^*)$ and pick the best model with respect to the BIC and the best model with respect to sBIC. For each considered choice of a sample size n, this procedure is repeated 100 times.

The poset $\mathcal{P}_T$ comprises 34 possible forests/models. In Figures 3–5, we display the lattice structure of $\mathcal{P}_T$ overlaid with a heat map of how frequently the models were chosen at the particular sample size. The subforest/submodels are labeled from 1 to 34 with 1 corresponding to the complete independence model and 34 corresponding to $\mathbf{M}(T)$, where T is the tree in Figure 2(a).

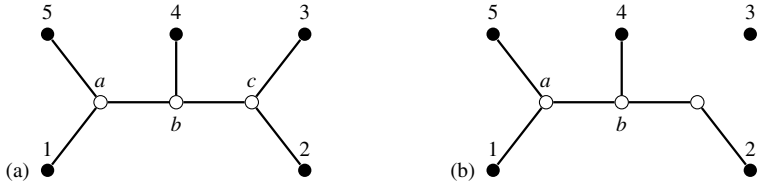


Figure 2. (a) A tree with five leaves; (b) one of its subforests.

If we order the edges as $\{a, 1\}$, $\{a, 5\}$, $\{a, b\}$, $\{b, 4\}$, $\{b, c\}$, $\{c, 2\}$, $\{c, 3\}$ and use $\{0, 1\}$-vectors to indicate the presence of edges then the submodels are:

1:	0000000	**10**:	1101110	**19**:	1011101		
2:	1100000	**11**:	1011110	**20**:	0111101	**28**:	0110111
3:	1011000	**12**:	0111110	**21**:	1111101	**29**:	1110111
4:	0111000	**13**:	1111110	**22**:	0000011	**30**:	0001111
5:	1111000	**14**:	1010101	**23**:	1100011	**31**:	1101111
6:	1010110	**15**:	0110101	**24**:	1011011	**32**:	1011111
7:	0110110	**16**:	1110101	**25**:	0111011	**33**:	0111111
8:	1110110	**17**:	0001101	**26**:	1111011	**34**:	1111111
9:	0001110	**18**:	1101101	**27**:	1010111		

In particular, the smallest true model is model 13.

Figures 3–5 show that the standard dimension-based BIC tends to select too small models that do not contain the data-generating distribution q. In particular, BIC never selects the full

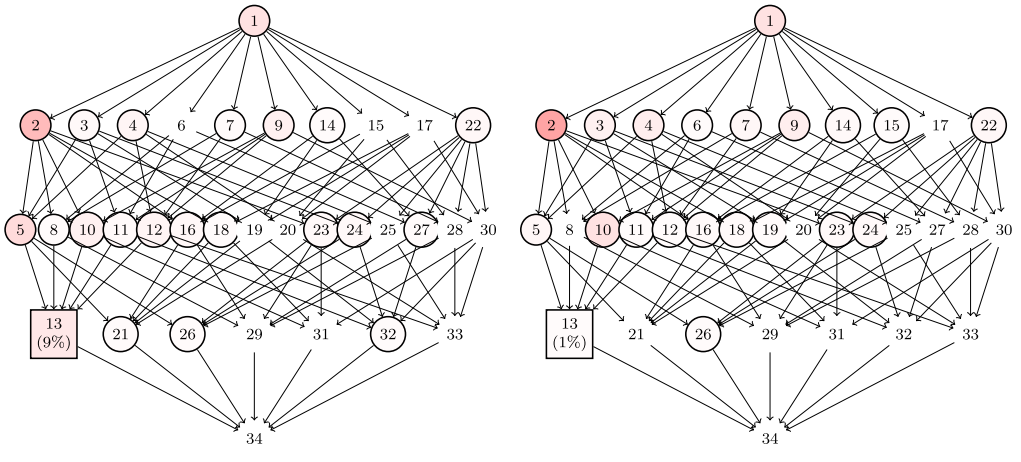


Figure 3. Results from 100 simulations with true covariance matrix given by (5.4) for $n = 25$ (sBIC left, BIC right). Darker color corresponds to higher selection frequency. The square node 13 is the smallest true model and includes the selection frequency. Models never chosen are without border.

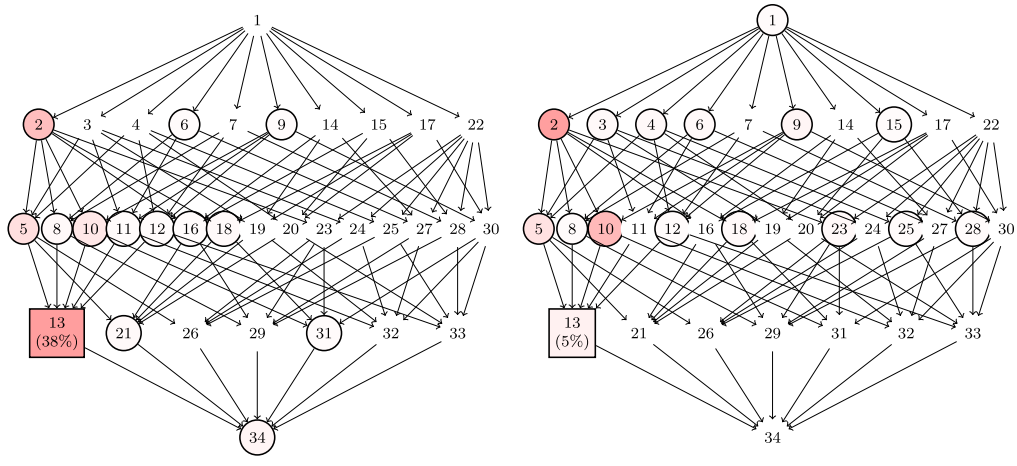


Figure 4. Results from 100 simulations as for Figure 3 but with sample size $n = 75$.

tree model 34. The RLCT-based sBIC, on the other hand, invokes a milder penalty, occasionally selects the tree model 34, and more frequently selects the smallest true model 13. Indeed, already for $n = 75$, sBIC selects the true model more often than any other model. On the other hand, the regular BIC procedure selects too simple a model also when the sample size is increased to $n = 125$.

Next, we consider examples with 10 and 11 leaves, in which case the number of considered models is still tractable. Writing $m := |V|$ for the number of leaves, the lattice $\mathcal{P}_T$ has depth $m-1$ with the complete independence model having depth 0 and the maximal element $\mathbf{M}(T)$ having

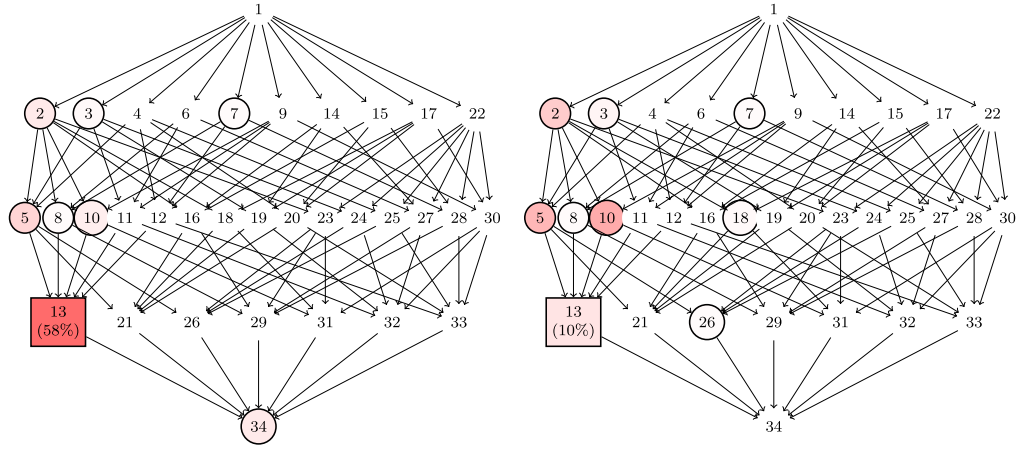


Figure 5. Results from 100 simulations as for Figure 3 but with sample size $n = 125$.

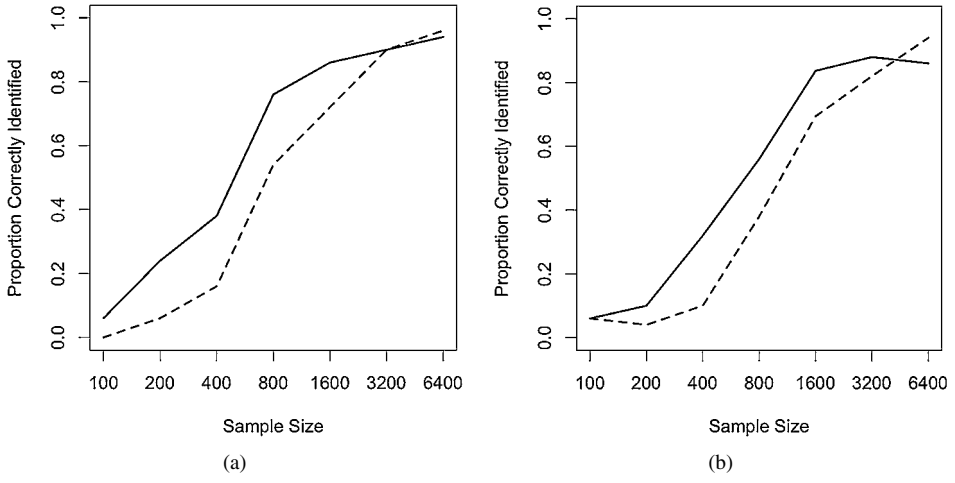


Figure 6. Proportion of times, out of 50 simulations, sBIC (solid line) and BIC (dashed line) select the smallest true model for different samples sizes (displayed on a logarithmic scale). (a) $m = 10$ leaves. (b) $m = 11$ leaves.

depth $m - 1$. Since the penalty in BIC is always at least the penalty in sBIC, it holds trivially that BIC will select the smallest true model more often than sBIC when the smallest true model is at depth 0; the converse is true if the smallest true model is at depth $m - 1$. We thus focus on the middle depth and randomly choose 50 trees $T_1, \ldots, T_{50}$ with corresponding randomly chosen subforests $F_1, \ldots, F_{50}$ each at depth $\lfloor \frac{m-1}{2} \rfloor$. From each subforest which we pick $q_i \in \mathbf{M}(F_i)$ by setting all edge correlations to 0.6 and all leaf variances to 1; note that F_i equals the q-forest $F^*(q_i)$. From each q_i, we generate a dataset of a fixed size n and compare the proportion of times that sBIC and BIC correctly identify the smallest true model $\mathbf{M}(F_i)$ for $1 \leq i \leq 50$. The results of these simulations are summarized in Figure 6. We see that sBIC outperforms BIC for smaller sample sizes with BIC marginally overtaking sBIC in very large samples.

Remark 5.1. In the simulations, we evaluated the quality of the forests found by BIC and sBIC through the proportion of times the chosen forest matched the truth exactly. An exact match is a very strong requirement and one may instead wish to compute the average distance, based on some metric, between selected forest and the truth. Unfortunately, the most natural metrics in our setting are NP-hard to compute and can only be approximated in general [9,10].

5.2. Temperature data

We consider a dataset consisting of average daily temperature values on 310 days from 37 cities across North America, South America, Africa and Europe. The data was sourced from the National Climatic Data Center and compiled in a readily available format by the average daily temperature archive of the University of Dayton [19]. In order to decorrelate and localize the data we first perform a seasonality adjustment where we regress each observed time series of temperature

values on a sinusoid corresponding to the seasons and retain only the residuals. We then consider only the differences of average temperatures on consecutive days reducing the number of data points to $n = 309$.

In the simulations of Section 5.1, we performed an exhaustive search over the lattice of all considered forests, a strategy which quickly becomes infeasible when increasing the number of observed variables beyond the low teens. Thus, in order to do model selection with the 37 observed nodes described above, we need to formulate an approximate sBIC. There are a plurality of possible heuristic strategies for producing this approximation involving combinations of greedy search, truncation of the considered model space, and simulated annealing. An in-depth exploration of these strategies and their relative performance is beyond the scope of this paper, instead we will show the results of using one such method as a proof of concept.

Our selection strategy, which we call a pruned chain search, has the following form:

(1) Generate an approximate maximum likelihood trivalent tree structure T.

(2) Prune the model space of considered forests to only consider a single decreasing path in the poset $\mathcal{P}_T$ starting at T and ending with the empty forest.

(3) Compute the sBIC (or BIC) for models in the pruned space and select the highest scoring model.

Note that after (2) the number of considered models will equal to the number of observed variables making computation tractable for many observed nodes. We accomplish (1) using a version of the structural EM algorithm proposed by [8]. To produce the decreasing path of models in (2), we start with T and iteratively select subforests in a greedy fashion:

(a) Suppose that after the mth iteration we have constructed the decreasing chain $\mathcal{C}_m$ of forests $T = F_0 \supset F_1 \supset F_2 \supset \cdots \supset F_m$.

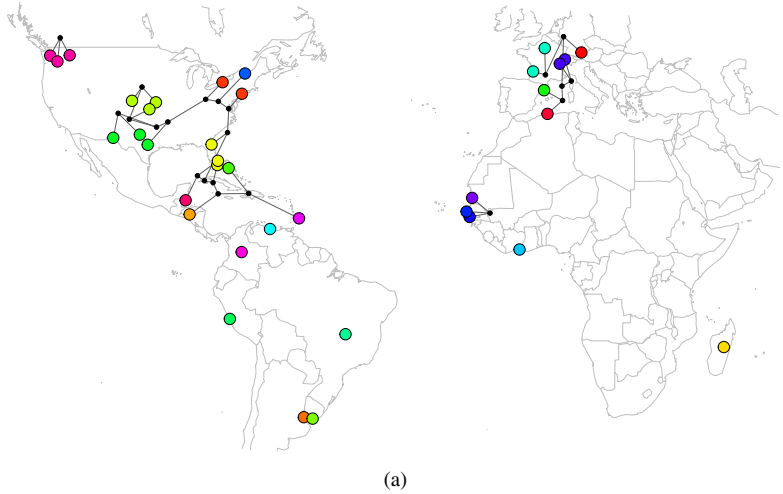
(b) If F_m is the empty forest then we are done.

(c) Otherwise, we extend $\mathcal{C}_m$ to $\mathcal{C}_{m+1}$ by adding to it the forest with largest BIC-penalized log-likelihood (with log-likelihood maximized using the EM algorithm described in Section 5.1) among all maximal subforests of F_m.

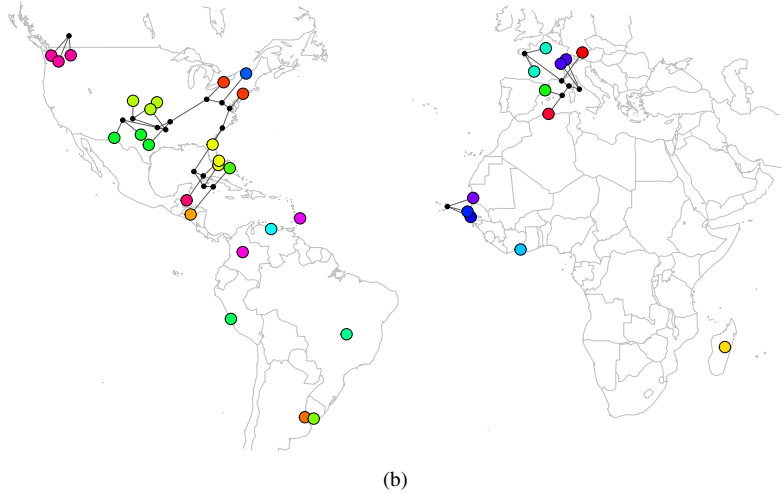
We present the results of applying above selection procedure to the temperature data in Figure 7. Note that the models selected by the sBIC and BIC are quite similar with the majority of the connections following our physical intuition that geographically adjacent cities should have similar temperature fluctuations while further separated cities should be essentially uncorrelated. For instance, all three cities in Washington, USA are connected to each other but to no other cities. The one difference between the model selected by sBIC and that selected by the BIC is the connection of Barbados to the component containing the Bahamas in the sBIC graph. The distance between these nodes is just far enough to place this connection on the border between spurious and reasonable. As in the simulation experiments, we observe sBIC's ability to select larger models.

6. Conclusion

Real log-canonical thresholds and associated multiplicities quantify the large-sample properties of the marginal likelihood in Bayesian approaches to model selection. In this paper, we computed



(a)



(b)

Figure 7. The models selected by sBIC and BIC pruned chain search. Each colored node represents an observed node (nodes with the same color are from the same country or US state) and the black nodes correspond to latent variables. The position of colored nodes corresponds to the city from where the data was collected. (a) Model chosen by sBIC. (b) Model chosen by BIC.

these RLCTs for Gaussian latent tree and forest models; the main results being Theorems 4.3 and 4.7. Our computations relied on the fact that the considered tree and forest models have a monomial parametrization, which allows one to apply methods from polyhedral geometry that we presented in Theorem 3.2.

Knowing RLCTs makes it possible to apply a 'singular Bayesian information criterion' (sBIC) that was recently proposed by [5]. RLCTs provide refined information about the marginal likeli-

hood and our simulations show that, at least in smaller problems, the sBIC outperforms the usual BIC of [16] that is defined using model dimension alone. As an exhaustive search over all models becomes quickly infeasible as the number of observed variables increases, we demonstrated, by example of a temperature dataset, how the sBIC might be approximated and applied to larger problems. In particular, we combined the structural EM of [8] with a greedy search methodology to reduce the number of considered models to a small collection for which the sBIC can be readily computed.

Appendix A: Proof of Theorem 3.2

Let H be the function from (3.1). By assumption, the 'prior' $\varphi : \Omega \to (0, \infty)$ is bounded above and $\mathcal{V}_\Omega(H) = \{\omega \in \Omega : H(\omega) = 0\}$ is compact. Since φ is smooth and positive, φ is bounded away from zero on $\mathcal{V}_\Omega(H)$ and any compact neighborhood of this zero set. The poles of the zeta function in (2.6) can be shown to be the same for all such choices of φ, and we have $\mathrm{RLCT}_\Omega(H; \varphi) = \mathrm{RLCT}_\Omega(H)$.

Our proof of Theorem 3.2 now proceeds in three steps:

Step 1. Show that $\mathrm{RLCT}_\Omega(H) = \mathrm{RLCT}_\Omega(H^0 + H^1)$, where H^0, H^1 are the zero and nonzero parts of H that are defined in (3.2) and (3.3).

Step 2. Show that $\mathrm{RLCT}_{\Omega_1}(H^1) = (\lambda_1, 1)$, where $\lambda_1 = \operatorname{codim} \mathcal{V}_{\Omega_1}(H^1)$.

Step 3. Show that $\mathrm{RLCT}_{\Omega_0}(H^0) = (\lambda_0, \mathfrak{m})$, where λ_0 is the $\mathbf{1}$-distance of the Newton polyhedron $\Gamma_+(H^0)$ and $\mathfrak{m}$ is the multiplicity (recall Definition 3.1).

Since H^0 and H^1 are functions of disjoint sets of coordinates and $\Omega = \Omega_0 \times \Omega_1$ is a Cartesian product, it follows from Remark 7.2(3) in [20] and the above Steps 1–3 that

$$\mathrm{RLCT}_\Omega\big(H^0 + H^1\big) = \big(\lambda_0 + \lambda_1, (\mathfrak{m} + 1) - 1\big) = (\lambda_0 + \lambda_1, \mathfrak{m}),$$

which is the claim of Theorem 3.2.

Before moving on to Step 1, we make a definition. Let $f, g : \Omega \to [0, \infty)$ be two nonnegative functions with common zero set $\mathcal{V}_\Omega(g) = \mathcal{V}_\Omega(f)$. Then f and g are *asymptotically equivalent*, we write $f \sim g$, if there exist two constants $c, C > 0$ and a neighborhood W of $\mathcal{V}_\Omega(g) = \mathcal{V}_\Omega(f)$ such that

$$cf(\omega) \le g(\omega) \le Cf(\omega) \tag{A.1}$$

for all $\omega \in W \cap \Omega$. Note that $\sim$ is indeed an equivalence relation. According to Remark 7.2(1) in [20], $f \sim g$ implies $\mathrm{RLCT}_\Omega(f) = \mathrm{RLCT}_\Omega(g)$.

A.1. Step 1

First, note that $\mathrm{RLCT}_\Omega(H) = \mathrm{RLCT}_{W \cap \Omega}(H)$ for any neighborhood W of the compact zero set $\mathcal{V}_\Omega(H)$. Choose W sufficiently small such that $\omega_1, \ldots, \omega_s$ are bounded away from zero on

$W \cap \Omega$. Next, by definition of the index r in Section 3, we have that $H = H^1 + H^{01}$, where

$$H^{01} = \sum_{i=r+1}^{k} \omega^{2u_i}.$$

When viewed as functions restricted to $W \cap \Omega$, we have $H^0 \sim H^{01}$ because

$$\left(\min_i \prod_{j=1}^{s} \omega_j^{2u_{ij}}\right) H^0 \le H^{01} \le \left(\max_i \prod_{j=1}^{s} \omega_j^{2u_{ij}}\right) H^0$$

and $\omega_1, \ldots, \omega_s$ are bounded above and bounded away from zero on the compactum $W \cap \Omega$. It follows that $\mathrm{RLCT}_\Omega(H) = \mathrm{RLCT}_\Omega(H^0 + H^1)$ because $H^{01} \sim H^0$ implies that $H^1 + H^0 \sim H^1 + H^{01} = H$.

A.2. Step 2

To complete Step 2, we will prove the following result.

Proposition A.1. *Suppose that H satisfies (3.1) with all $c_i^* \neq 0$, that is, H is equal to its nonzero part. Let $\mathcal{V}_\Omega(H)$ be the zero set of H on Ω. Then*

$$\mathrm{RLCT}_\Omega(H) = \big(\mathrm{codim}\, \mathcal{V}_\Omega(H), 1\big).$$

Before turning to the proof, we exemplify the application of Proposition A.1.

Example A.2. Let $\Omega = [0,1] \times [0,1]$ be the unit square in $\mathbb{R}^2$, and consider two functions $g_1(\omega) = (\omega_1 - \omega_2)^2$ and $g_2(\omega) = (\omega_1 + \omega_2)^2$. The zero set of either function is a line in $\mathbb{R}^2$. When restricting to Ω, the zero set $\mathcal{V}_\Omega(g_1)$ is a line segment and of co-dimension one. The zero set $\mathcal{V}_\Omega(g_2)$, on the other hand, consists only of the origin and is of co-dimension two. We have $\mathrm{RLCT}_\Omega(g_1) = (1,1)$ but $\mathrm{RLCT}_\Omega(g_2) = (2,1)$.

To prove Proposition A.1, note first that when H is equal to its nonzero part and $\mathcal{V}_\Omega(H)$ is compact, $\mathrm{RLCT}_\Omega(H)$ is equal to the RLCT of H over a compact set on which all coordinates of the argument ω are bounded away from zero. Partition this compactum into the intersections with each one of the 2^d orthants in $\mathbb{R}^d$. Then $\mathrm{RLCT}_\Omega(H)$ is the minimum RLCT in any orthant. Similarly, the co-dimension of $\mathcal{V}_\Omega(H)$ is the minimum of any co-dimension obtained from intersection with an orthant. We may thus consider one orthant at a time. Changing signs as needed to make all coordinates positive, the following lemma becomes applicable.

Lemma A.3. *Let $W = [a_1, b_1] \times \cdots \times [a_s, b_s]$ with $0 < a_i < b_i < \infty$. Let $\log W = [\log a_1, \log b_1] \times \cdots \times [\log a_s, \log b_s]$. If H satisfies (3.1) with all $c_i^* > 0$ and $\mathcal{V}_W(H)$ is nonempty, then*

$$\mathrm{RLCT}_W(H) = \mathrm{RLCT}_{\log W}\left(\sum_{i=1}^{r} \big(u_i^T \omega - \log c_i^*\big)^2\right).$$

The result follows from a change of coordinates and an argument about asymptotic equivalence that has been used in other contexts. We include the proof of the lemma for sake of completeness.

Proof of Lemma A.3. Change coordinates via the substitution $\tilde{\omega} = \log(\omega)$, where the logarithm is applied entry-wise. Since the Jacobian of this transformation is bounded above and bounded away from zero on W, it may be ignored in the computation of the RLCT, and thus

$$\mathrm{RLCT}_W(H) = \mathrm{RLCT}_{\log W}\left(\sum_{i=1}^{r}\left(e^{u_i^T\omega} - e^{\log c_i^*}\right)^2\right).$$

Since W, and thus also $\log W$, is compact, each of the r linear combinations $u_i^T\omega$ takes its values in a compact set. Restricted to this compact set, the function

$$h_1(x) = \sum_{i=1}^{r}\left(e^{x_i} - e^{\log c_i^*}\right)^2$$

is asymptotically equivalent to the sum of squares

$$h_2(x) = \sum_{i=1}^{r}\left(x_i - \log c_i^*\right)^2,$$

as can been seen by a quadratic Taylor approximation to h_1 around the point $(\log c_1^*, \ldots, \log c_r^*)$. Since asymptotically equivalent functions have the same RLCT, the claim is proven. □

By an application of Lemma A.3, the proof of Proposition A.1 reduces to an analysis of sums of squares of linear forms, that is, functions of the form

$$g(\omega) = \sum_{i=1}^{r}\left(u_i^T\omega - C_i^*\right)^2 \tag{A.2}$$

with $C_i^* \in \mathbb{R}$ and $u_i \in \mathbb{R}^d$. Proposition A.1 thus follows from Proposition A.4 below. Note that $\mathcal{V}_\Omega(g)$ is a polyhedron, which we assume to be nonempty.

Proposition A.4. *If $g : \Omega \to [0, \infty)$ is a sum of squares of linear forms as in* (A.2) *and Ω is a product of intervals, then* $\mathrm{RLCT}_\Omega(g) = (\mathrm{codim}\,\mathcal{V}_\Omega(g), 1)$.

Proof. By [12], Propositions 2.5 and 3.2, or also [20], Remark 2.14, $\mathrm{RLCT}_\Omega(g)$ is the minimum of local thresholds $\mathrm{RLCT}_{\Omega(x)}(g)$ over $x \in \mathcal{V}_\Omega(g)$. Here, each set $\Omega(x) = W(x) \cap \Omega$, where $W(x)$ is a sufficiently small neighborhood of x. We will show that $\mathrm{RLCT}_{\Omega(x)}(g) = (\mathrm{codim}\,\mathcal{V}_\Omega(g), 1)$ for $x \in \mathcal{V}_\Omega(g)$, which implies our claim.

Consider any point $x \in \mathcal{V}_\Omega(g)$. By translation, we may assume without loss of generality that $x = 0$ and $g(\omega) = \sum_i (u_i^T\omega)^2$. We may then take the neighborhood $\Omega(0)$ to be equal to $\{\omega \in \Omega : \max_i |\omega_i| \le \varepsilon\}$ for sufficiently small $\varepsilon > 0$.

When partitioning $\Omega(0)$ into orthants, the co-dimension of $\mathcal{V}_\Omega(g)$ is the minimum of the codimensions of the intersection between $\mathcal{V}_\Omega(g)$ and each one of the orthants. Furthermore, $\mathrm{RLCT}_{\Omega(0)}(g)$ is equal to the smallest RLCT of g over any of these orthants. Therefore, changing the signs of the coordinates ω_i as needed, we are left with checking that $\mathrm{RLCT}_{\Omega_+}(g)$ is given by the codimension of $\mathcal{V}_{\Omega_+}(g)$ for $\Omega_+ = \{\omega \in \Omega : 0 \le \omega_i \le \varepsilon \text{ for all } i = 1, \ldots, d\}$ and $g(\omega) = \sum_i (u_i^T \omega)^2$.

Case 1. If $\mathcal{V}_{\Omega_+}(g)$ intersects the interior of Ω_+, then we may pick any point x_+ in this intersection and consider Ω_+ as a neighborhood of x_+. After a change of coordinates, we have $g(\omega) = \omega_1^2 + \cdots + \omega_s^2$, where s is the co-dimension of $\mathcal{V}_{\Omega_+}(g)$. By Example 3.4, $\mathrm{RLCT}_{\Omega_+}(g) = (s, 1)$, which was to be shown.

Case 2. Suppose now that $\mathcal{V}_{\Omega_+}(g)$ is contained in the boundary of Ω_+. Since the zero set of g on all of $\mathbb{R}^d$ is a linear space, $\mathcal{V}_{\Omega_+}(g)$ is in fact a face of Ω_+, and each $u_i^T \omega$ is a supporting hyperplane of Ω_+. In particular, after appropriate sign changes, we may assume that $u_i^T \omega \ge 0$ on Ω_+. The co-dimension of $\mathcal{V}_{\Omega_+}(g)$ is equal to the number, say s, of facets of Ω_+ containing it. Without loss of generality, we may assume that these facets are given by $\omega_1 = 0$, $\omega_2 = 0, \ldots,$ $\omega_s = 0$. This implies that all u_i have nonzero entries only in the first s coordinates. We now show that when restricted to Ω_+, the functions $g(\omega)$ and $f(\omega) = \omega_1^2 + \cdots + \omega_s^2$ are asymptotically equivalent; recall (A.1).

To show that on Ω_+, the function g can be bounded from below by a positive multiple of f, note that the fact that $u_i^T \omega \ge 0$ on Ω_+ implies that all u_i have nonnegative entries. Hence,

$$\sum_{i=1}^{r} \left(u_i^T \omega\right)^2 = \sum_{i=1}^{r} \left(\sum_{j=1}^{s} u_{ij} \omega_j \right)^2 \ge \sum_{j=1}^{s} \left(\sum_{i=1}^{r} u_{ij}^2 \right) \omega_j^2,$$

where the inequality is obtained by expanding squares and dropping the mixed terms, which are nonnegative. If $\sum_{i=1}^{r} u_{ij}^2 = 0$ for some index j then $u_{ij} = 0$ for all i, which contradicts the fact that $\omega_j = 0$ for all $\omega \in \mathcal{V}_{\Omega_+}(g)$. Thus,

$$c = \min \left\{ \sum_{i=1}^{r} u_{ij}^2 : 1 \le j \le s \right\} > 0,$$

and $g(\omega) \ge c f(\omega)$ for all $\omega \in \Omega_+$.

To prove that g can be bounded above by a multiple of f, note that all $u_i^T \omega$ are nonnegative on Ω_+, and thus

$$\sum_{i=1}^{r} \left(u_i^T \omega\right)^2 \le \left(\sum_{i=1}^{r} u_i^T \omega \right)^2 = \left(\sum_{i=1}^{r} \sum_{j=1}^{s} u_{ij} \omega_j \right)^2.$$

Let $u_{+j} = \sum_i u_{ij}$ and $u_{++} = \sum_j u_{+j}$. Then, since all u_i have nonnegative entries, Jensen's inequality implies that

$$\left(\sum_{i=1}^{r} \sum_{j=1}^{s} u_{ij} \omega_j \right)^2 = u_{++}^2 \left(\sum_{j=1}^{s} \frac{u_{+j}}{u_{++}} \omega_j \right)^2 \le u_{++} \max\{u_{+j} : 1 \le j \le s\} \sum_{i=1}^{s} \omega_i^2.$$

Since g is asymptotically equivalent to $f(\omega) = \omega_1^2 + \cdots + \omega_s^2$, we have $\mathrm{RLCT}_{\Omega_+}(g) = \mathrm{RLCT}_{\Omega_+}(f)$. Let $\Omega'_+ = [-\varepsilon, \varepsilon]^s \times [0, \varepsilon]^{d-s}$. Then

$$\int_{\Omega_+} \left(\omega_1^2 + \cdots + \omega_s^2\right)^{-z/2} d\omega = 2^{-s} \int_{\Omega'_+} \left(\omega_1^2 + \cdots + \omega_s^2\right)^{-z/2} d\omega.$$

Hence, $\mathrm{RLCT}_{\Omega_+}(f) = \mathrm{RLCT}_{\Omega'_+}(f)$. From Case 1, we know that $\mathrm{RLCT}_{\Omega'_+}(f) = (s, 1)$. Putting it all together, we have shown that $\mathrm{RLCT}_{\Omega_+}(g) = (s, 1)$.

□

A.3. Step 3

The remaining step amounts to proving the following result, which concerns the case where the considered function H is equal to its zero part.

Proposition A.5. *Let Ω be a compact product of intervals containing the origin, and let $\Gamma_+(H)$ be the Newton polyhedron of the function $H(\omega) = \sum_i \omega^{2u_i}$. Then*

$$\mathrm{RLCT}_{\Omega}(H) = (\lambda, \mathfrak{m}),$$

where $1/\lambda$ is the **1***-distance of $\Gamma_+(H)$ and $\mathfrak{m}$ is its multiplicity.*

Proof. Note that H is invariant under sign changes. Hence, $\mathrm{RLCT}_{\Omega}(H) = \mathrm{RLCT}_{\Omega'}(H) = \mathrm{RLCT}_{\Omega\cup\Omega'}(H)$ when Ω' is obtained from Ω by changing the signs of any subset of the coordinates $\omega_1, \ldots, \omega_d$. Forming the unions of Ω and its reflected versions shows that in order to prove Proposition A.5, we may assume that the origin is an interior point of Ω. The claim now follows from Theorem 8.6 in [1], see also [12], Section 4, and by Remark A.6 below. □

Remark A.6. When the origin is in the interior of Ω, the function $H(\omega) = \sum_i \omega^{2u_i}$ has $\mathrm{RLCT}_{\Omega}(H) = \mathrm{RLCT}_{\Omega(0)}(H)$ for any small neighborhood $\Omega(0)$ of the origin. Indeed, as mentioned in the proof of Proposition A.4, $\mathrm{RLCT}_{\Omega}(H)$ is the minimum of local RLCTs of H in small neighborhoods $\Omega(x)$ of points $x \in \Omega$. If $x \neq 0$, then some of the variables, say $\omega_1, \ldots, \omega_s$, are bounded away from zero on a sufficiently small neighborhood $\Omega(x)$. Substituting these variables by $\omega_1 - x_1, \ldots, \omega_s - x_s$, respectively, in H, we get a new function H_x for which $\mathrm{RLCT}_{\Omega(0)}(H) = \mathrm{RLCT}_{\Omega(x)}(H_x)$. Now, $0 \leq H_x \leq H$ near x. Consequently, $\mathrm{RLCT}_{\Omega(x)}(H_x) \leq \mathrm{RLCT}_{\Omega(x)}(H)$. We conclude that $\mathrm{RLCT}_{\Omega(0)}(H) \leq \mathrm{RLCT}_{\Omega(x)}(H)$.

Appendix B: Proof of Theorem 4.3

Let $T = (U, E)$ be a tree with set of leaves V, and let q be a distribution in the latent tree model $\mathbf{M}(T)$, which has parameter space $\Omega = (0, \infty)^V \times [-1, 1]^E$. We are to compute $\mathrm{RLCT}_{\Omega}(H_q)$ for the function H_q from (4.1), where ω_v^* and ρ_{vw}^* are the variances and correlations of the distribution q. The basic idea of this proof follows [24].

First, observe that Theorem 3.2 is applicable to this problem. Indeed, H_q has the form from (3.1) and the q-fiber $\mathcal{V}_\Omega(H_q)$ is compact. Compactness holds because $H_q(\omega) = 0$ implies that $\omega_v = \omega_v^*$ for all $v \in V$, and all edge correlations ω_e, $e \in E$, are in the compact interval $[-1, 1]$.

Now, let $F^* := F^*(q) = (U^*, E^*)$ be the q-forest, and let $H_q^1(\omega_1, \ldots, \omega_s)$ be the nonzero part of H_q given in (4.2). The set $\mathcal{V}_{\Omega_1}(H_q^1)$ is equal to the q-fiber under the model $\mathbf{M}(F^*)$; recall that Ω_1 is the projection of Ω onto the first s coordinates. We deduce that $\operatorname{codim} \mathcal{V}_{\Omega_1}(H_q^1) = \dim \mathbf{M}(F^*)$, which gives the value of λ_1 in Theorem 3.2. It remains to show that the zero part $H_q^0(\omega_{s+1}, \ldots, \omega_d)$ defined in (4.3) satisfies

$$(\lambda_0, \mathfrak{m}) = \operatorname{RLCT}_{\Omega_0}\big(H_q^0\big) = \left(\frac{1}{2}\sum_{e \in E_0} w(e), 1 + l_2'\right), \tag{B.1}$$

where Ω_0 is the projection of Ω onto the last $d - s$ coordinates, $E_0 = E \setminus E^*$ is the set of edges that appear in T but not in F^*, and l_2' is the number of degree two nodes of T that are not in U^*.

The zero part of H_q is the sum of squares of the monomials

$$\prod_{e \in \overline{vw} \cap E_0} \omega_e, \qquad v, w \in V, v \nsim w; \tag{B.2}$$

recall that $v \nsim w$ if there is no path between v and w in the q-forest $F^* = (U^*, E^*)$. The edge set E_0 can be partitioned into sets $E_{01}, \ldots, E_{0t}$ such that each E_{0i} defines a tree $S_i = (U_i, E_{0i})$ that has the set of nodes $L_i := U_i \cap U^*$ as leaves. In other words, the set of leaves L_i of tree S_i comprises precisely those nodes that belong to both S_i and the q-forest F^*. For example, in Figure 1, we have $t = 1$ and S_1 is the tree with one inner node b and three leaves $a, 3, 4$. As a further example, consider the tree and q-forest in Figure 8(a) and (b), for which we form two subtrees S_1 and S_2 with edge sets $E_{01} = \{\{a, 3\}\}$ and $E_{02} = \{\{a, 4\}\}$, as shown in Figure 8(c) and (d). In this second example, the sets of leaves are $L_1 = \{a, 3\}$ and $L_2 = \{a, 4\}$, illustrating that the sets $L_1, \ldots, L_t$ need not be disjoint.

Consider now the function $\check{H}_q^0$ given by the sum of squares of the monomials

$$\prod_{e \in \overline{uu'}} \omega_e, \qquad i \in [t], u, u' \in L_i, u \neq u', \tag{B.3}$$

where $[t] = \{1, \ldots, t\}$ and $\overline{uu'}$ refers to the unique path between u and u' in tree S_i. Each monomial listed in (B.3) is also listed in (B.2). To see this, observe that two distinct nodes $u, u' \in L_i$

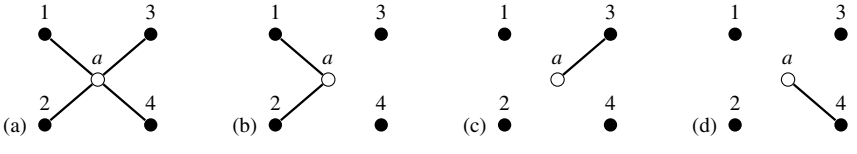


Figure 8. (a) Star tree; (b) q-forest when ρ_{12}^* is the only nonzero correlation; (c), (d) subtrees formed from the removed edges.

belong to distinct connected components in F^*. If we take $v \in V$ from one of the two connected components and $w \in V$ from the other, then the monomial they define in (B.2) is equal to the monomial that u and u' define in (B.3). Moreover, by the definition of the trees S_i, every monomial listed in (B.2) is the product of monomials from (B.3). It follows that the Newton polyhedra $\Gamma_+(H_q^0)$ and $\Gamma_+(\tilde{H}_q^0)$ are equal and hence $\mathrm{RLCT}_{\Omega_0}(H_q^0) = \mathrm{RLCT}_{\Omega_0}(\tilde{H}_q^0)$ (cf. Proposition A.5).

Let f_i be the sum of squares of the monomials in (B.3) that are associated with pairs of distinct nodes u and u' in the set of leaves L_i of the tree S_i. No two trees S_i and S_j for $i \neq j$ share an edge. Hence, the two sums of squares f_i and f_j depend on different subvectors of ω. Since $\tilde{H}_q^0 = f_1 + \cdots + f_t$, it follows from (2.5) that

$$\mathrm{RLCT}_{\Omega_0}\big(H_q^0\big) = \sum_{i=1}^{t} \mathrm{RLCT}_{\Omega_0}(f_i) - (0, t-1); \tag{B.4}$$

see also Remark 7.2(3) in [20]. If T has no nodes of degree two, that is, $l_2 = l_2' = 0$, then the same is true for the each tree S_i. Lemma B.1 below then implies that

$$\mathrm{RLCT}_{\Omega_0}(f_i) = \left(\frac{|L_i|}{2}, 1\right). \tag{B.5}$$

Since the nodes in L_i lie in F^*, we have

$$\sum_{i=1}^{t} |L_i| = \sum_{e \in E_0} w(e), \tag{B.6}$$

where $w(e) \in \{0, 1, 2\}$ is the number of nodes of e that lie in the q-forest F^*. Combining (B.4)–(B.6), we obtain (B.1) and have thus proven Theorem 4.3 in the case of $l_2 = 0$ nodes of degree two. The case with nodes of degree two follows the same way applying Lemma B.2 instead of Lemma B.1.

Lemma B.1. *Let $S = (V, E)$ be a tree with set of leaves L and all inner nodes of degree at least three. Let f be the sum of squares of the monomials*

$$\prod_{e \in \overline{vw}} \omega_e, \qquad v, w \in L, v \neq w. \tag{B.7}$$

If Ω is a neighborhood of the origin, then

$$\mathrm{RLCT}_{\Omega}(f) = \left(\frac{|L|}{2}, 1\right).$$

Proof. If $|L| = 2$, then S has a single edge and no inner nodes. In this case, f is the square of a single variable and it is clear $\mathrm{RLCT}_{\Omega}(f) = (1, 1) = (|L|/2, 1)$. In the remainder of this proof, we assume that $|L| \geq 3$.

By Proposition A.5, it suffices to compute the $\mathbf{1}$-distance and its multiplicity for the Newton polyhedron $\Gamma_+(f) \subset \mathbb{R}^E$. By Definition 3.1, the polyhedron $\Gamma_+(f)$ is determined by the exponent vectors of the monomials in (B.7). Each exponent vector is the incidence vector for a path between a pair of leaves. In other words, each pair of two distinct leaves v and w defines a vector $u \in \mathbb{R}^E$ with $u_e = 1$ if $e \in \overline{vw}$ and $u_e = 0$ otherwise. Write $\mathcal{U}$ for the set of all these $\binom{|L|}{2}$ vectors.

Let E_L be the set of terminal edges of S, that is, the $|L|$ edges that are incident to a leaf. We claim that every point x in the Newton polyhedron $\Gamma_+(f)$ satisfies

$$\sum_{e \in E_L} x_e \geq 2 \tag{B.8}$$

and that the inequality defines a facet of $\Gamma_+(f)$. Indeed, if $x \in \mathcal{U}$ then $\sum_{e \in E_L} x_e = 2$ because every path between two leaves in L includes precisely two edges in E_L. It is then clear that (B.8) holds for all points $x \in \Gamma_+(f)$. Moreover, by [13], Lemma 1, the span of $\mathcal{U}$ is all of $\mathbb{R}^E$. Hence, the affine hull of $\mathcal{U}$ is the hyperplane given by $\sum_{e \in E_L} x_e = 2$, and we conclude that (B.8) defines a facet of $\Gamma_+(f)$.

Since $|E_L| = |L|$, inequality (B.8) implies that the $\mathbf{1}$-distance of $\Gamma_+(f)$ is at least $2/|L|$. We claim that it is equal to $2/|L|$. In fact, we will show that the vector $\frac{2}{|L|}\mathbf{1}$ not only lies in the Newton polyhedron but also in the Newton polytope $\Gamma(f)$, that is, the vector is a convex combination of the incidence vectors in $\mathcal{U}$. To prove this, we construct a set of paths $\mathcal{P}$ in the tree S such that (i) each element of $\mathcal{P}$ is a path between leaves of S, (ii) $\mathcal{P}$ contains precisely $|L|$ paths and (iii) every edge of S is covered by exactly two paths of $\mathcal{P}$. The construction implies our claim because the average of the incidence vectors of the paths in $\mathcal{P}$ is equal to $\frac{2}{|L|}\mathbf{1}$.

Let S^* be any trivalent tree that has the same set of leaves L as S and that can be obtained from S^* by edge contraction. Here, a tree is trivalent if each inner node has degree three. We will use induction on the number of leaves to show that a set of paths $\mathcal{P}$ with the desired properties (i)–(iii) exists. Figure 9 shows an example.

If S^* has three $|L| = 3$ leaves, then there is a single inner node and each path between two leaves has two edges. We may simply take $\mathcal{P}$ to be the set of all the three paths that exist between pairs of leaves. This provides the induction base.

In the induction step, pick two leaves v and w of the tree S^* that are joined by a path with two edges $\{v, a\}$ and $\{a, w\}$. The node a is an inner node of S^*. Remove the two edges and the two leaves to form a subtree S^{**}, in which a becomes a leaf. Then S^{**} has $|L| - 1$ leaves and, by

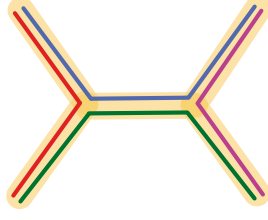


Figure 9. An example of a system of paths such that each edge of a trivalent tree is covered by exactly two paths.

the induction hypothesis, there is a set of paths $\mathcal{P}^{**}$ that satisfies properties (i)–(iii) with respect to S^{**}. In particular, $|\mathcal{P}^{**}| = |L| - 1$. Now, precisely two paths in $\mathcal{P}^{**}$ have the node a as an endpoint. Extend one of them by adding the edge $\{a, v\}$ and extend the other by adding $\{a, w\}$. This gives two paths between leaves of S^*. All other paths in $\mathcal{P}^{**}$ are already paths between leaves of S^*. Add one further path, namely, (v, a, w), and denote the resulting collection of $|L|$ paths by $\mathcal{P}^*$. Clearly, the set $\mathcal{P}^*$ satisfies properties (i)–(iii) with respect to S^*. Contracting each path in $\mathcal{P}^*$ by applying the edge contractions that transform S^* into S, we obtain a system of paths $\mathcal{P}$ that satisfies properties (i)–(iii) with respect to S.

Finally, note that in the construction we just gave we can ensure that $\mathcal{P}$ includes a given path between two leaves in L. Hence, the vector $\frac{2}{|L|}\mathbf{1}$ can be written as a convex combination of vertices of $\Gamma(f)$ such that a given vertex x get positive weight. It follows that $\frac{2}{|L|}\mathbf{1}$ lies in the interior of the Newton polytope and thus the multiplicity $\mathfrak{m}$ is 1. □

The next result generalizes the previous lemma to the case of trees with nodes of degree 2. We remark Example 2.2 is a special case of this generalization. It matches the case where the tree S has two leaves and one inner node, which is then necessarily of degree two.

Lemma B.2. *Let $S = (V, E)$ be a tree with set of leaves L, and let f be the sum of squares of the monomials*

$$\prod_{e \in \overline{vw}} \omega_e, \qquad v, w \in L, v \neq w. \tag{B.9}$$

If Ω is a neighborhood of the origin, then

$$\mathrm{RLCT}_{\Omega}(f) = \left(\frac{|L|}{2}, 1 + l_2\right),$$

where l_2 is the number of (inner) nodes of S that have degree two.

Proof. Suppose a is an inner node of degree two, and that a is incident to the two edges $e = \{a, b\}$ and $f = \{a, c\}$. Then any path connecting to leaves in L either uses both e and f or neither e nor f. Hence, if x is the incidence vector of a path between two leaves in L, then $x_e = x_f$. It follows that the affine hull of Newton polytope generated by the path incidence vectors is no longer a hyperplane but an affine space of dimension $|E| - 1 - l_2$.

Proceeding exactly as in the proof of Lemma B.1, we see that it still holds that the $\mathbf{1}$-distance of the Newton polyhedron $\Gamma_+(f)$ is $2/|L|$. Similarly, the ray spanned by $\mathbf{1}$ still meets $\Gamma_+(f)$ in the relative interior of the Newton polytope $\Gamma(f)$. However, since the co-dimension of the Newton polytope is now $1 + l_2$, we have $\mathrm{RLCT}_{\Omega}(f) = (|L|/2, 1 + l_2)$. □

Acknowledgements

This work was partially supported by the European Union 7th Framework Programme (PIOF-GA-2011-300975), the U.S. National Science Foundation (DMS-1305154), the U.S. National

Security Agency (H98230-14-1-0119), and the University of Washington's Royalty Research Fund. The United States Government is authorized to reproduce and distribute reprints. We are thankful to the referee for constructive remarks.

References

[1] Arnol'd, V.I., Guseĭn-Zade, S.M. and Varchenko, A.N. (1988). *Singularities of Differentiable Maps. Vol. II. Monographs in Mathematics* **83**. Boston, MA: Birkhäuser. MR0966191
[2] Choi, M.J., Tan, V.Y.F., Anandkumar, A. and Willsky, A.S. (2011). Learning latent tree graphical models. *J. Mach. Learn. Res.* **12** 1771–1812. MR2813153
[3] Chow, C.K. and Liu, C.N. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inform. Theory* **14** 462–467.
[4] Drton, M. (2009). Likelihood ratio tests and singularities. *Ann. Statist.* **37** 979–1012. MR2502658
[5] Drton, M. and Plummer, M. (2013). A Bayesian information criterion for singular models. Available at arXiv:1309.0911.
[6] Drton, M., Sturmfels, B. and Sullivant, S. (2009). *Lectures on Algebraic Statistics. Oberwolfach Seminars* **39**. Basel: Birkhäuser. MR2723140
[7] Edwards, D., de Abreu, G. and Labouriau, R. (2010). Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. *BMC Bioinformatics* **11** 1–13.
[8] Friedman, N., Ninio, M., Pe'er, I. and Pupko, T. (2002). A structural EM algorithm for phylogenetic inference. *J. Comput. Biol.* **9** 331–353.
[9] Hein, J., Jiang, T., Wang, L. and Zhang, K. (1996). On the complexity of comparing evolutionary trees. *Discrete Appl. Math.* **71** 153–169. MR1420297
[10] Hickey, G., Dehne, F., Rau-Chaplin, A. and Blouin, C. (2008). SPR distance computation for unrooted trees. *Evol. Bioinform.* **4** 17–27.
[11] Lauritzen, S.L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. New York: Oxford Univ. Press. MR1419991
[12] Lin, S. (2011). Asymptotic approximation of marginal likelihood integrals. Available at arXiv:1003.5338.
[13] Mihaescu, R. and Pachter, L. (2008). Combinatorics of least-squares trees. *Proc. Natl. Acad. Sci. USA* **105** 13206–13211. MR2443724
[14] Mossel, E., Roch, S. and Sly, A. (2013). Robust estimation of latent tree graphical models: Inferring hidden states with inexact parameters. *IEEE Trans. Inform. Theory* **59** 4357–4373. MR3071334
[15] Rusakov, D. and Geiger, D. (2005). Asymptotic model selection for naive Bayesian networks. *J. Mach. Learn. Res.* **6** 1–35. MR2249813
[16] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014
[17] Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, Prediction, and Search*, 2nd ed. Cambridge, MA: MIT Press. MR1815675
[18] Tan, V.Y.F., Anandkumar, A. and Willsky, A.S. (2011). Learning high-dimensional Markov forest distributions: Analysis of error rates. *J. Mach. Learn. Res.* **12** 1617–1653. MR2813149
[19] University of Dayton, Environmental protection agency average daily temperature archive. Available at http://academic.Udayton.Edu/kissock/http/Weather/default.Htm. Accessed 2015-09-20.
[20] Watanabe, S. (2009). *Algebraic Geometry and Statistical Learning Theory. Cambridge Monographs on Applied and Computational Mathematics* **25**. Cambridge: Cambridge Univ. Press. MR2554932
[21] Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* **11** 3571–3594. MR2756194
[22] Watanabe, S. (2010). Equations of states in singular statistical estimation. *Neural Netw.* **23** 20–34.

[23] Yamada, K. and Watanabe, S. (2012). Statistical learning theory of quasi-regular cases. *IEICE Transactions on Fundamentals of Electronics*, *Communications and Computer Sciences* **95** 2479–2487.
[24] Zwiernik, P. (2011). An asymptotic behaviour of the marginal likelihood for general Markov models. *J. Mach. Learn. Res.* **12** 3283–3310. MR2877601

Received December 2014 and revised September 2015