Algebraic tests of general Gaussian latent tree models

Dennis Leung & Mathias Drton

University of Southern California and University of Washington

Neural Information Processing Systems Foundation

Goals

- Derive polynomial characterizations for the covariance matrices of identifiable Gaussian latent tree models
- Leverage these characterizations to hypothesis-test the validity of a possibly large Gaussian latent tree model.

1 Gaussian latent-tree models

- Given an undirected tree $T = (V, E)$ with a node set V and an edge set E, a subset of nodes $\mathbf{X} = \{X_1, \dots, X_m\} \subset V$ corresponds to m observed variables and its complement $V \backslash \mathbf{X}$ corresponds to latent (unobserved) variables.
- $\mathcal{M}_{\mathbf{X}}(T)$ (T-Gaussian latent tree model on $\mathbf{X}$): All marginal distributions for $\mathbf{X}$ induced by all $|V|$-variate Gaussian distributions respecting the pairwise Markov property of T.
- Each latent node in $V \backslash \mathbf{X}$ must have a minimal degree of 3 for model identifiability (Choi et al., 2011):

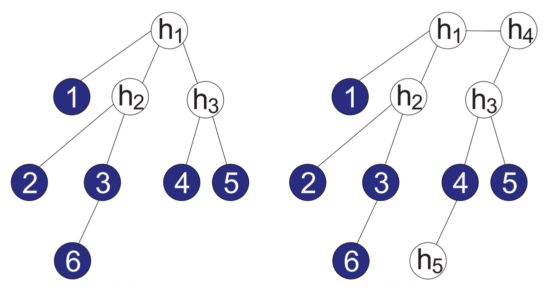


Figure 1: (Choi et al., 2011) Shaded nodes are observed and unshaded nodes are hidden. (a) An identifiable tree. (b) A non-identifiable tree because h_4 and h_5 have degrees less than 3.

2 T-induced pseudo-metric on $\mathbf{X}$

Suppose $w : E \longrightarrow \mathbb{R}_{\geq 0}$ is any function that assigns non-negative weights to the edges in E, and let $ph_T(p, q)$ be the set of edges on the unique path that connects X_p and X_q in T. One can define a pseudo-metric $\delta_w : \mathbf{X} \times \mathbf{X} \longrightarrow \mathbb{R}_{\geq 0}$ by

$$\delta_w(X_p, X_q) = \begin{cases} \sum_{e \in ph_T(p,q)} w(e) & : p \neq q, \\ 0 & : p = q. \end{cases}$$

This is known as a T-induced pseudo-metric on $\mathbf{X}$. We have the following main result:

Theorem 1 (Extension of Corollary 1 in Shiers et al. (2016)). *Suppose $\delta : \mathbf{X} \times \mathbf{X} \longrightarrow \mathbb{R}_{\geq 0}$ is a pseudo-metric defined on $\mathbf{X}$. Let $\delta_{pq} = \delta(X_p, X_q)$ for any $p, q \in \{1, \dots, m\}$ for simplicity. Then δ is a T-induced pseudo-metric if and only if for any four distinct $1 \leq p, q, r, s \leq m$ such that $ph_T(p,q) \cap ph_T(r,s) = \emptyset$,*

$$\delta_{pq} + \delta_{rs} \leq \delta_{pr} + \delta_{qs} = \delta_{ps} + \delta_{qr}, \tag{1}$$

and for any three distinct $1 \leq p, q, r \leq m$,

$$\delta_{pq} + \delta_{qr} = \delta_{pr} \tag{2}$$

if $ph_T(p,r) = ph_T(p,q) \cup ph_T(q,r)$.

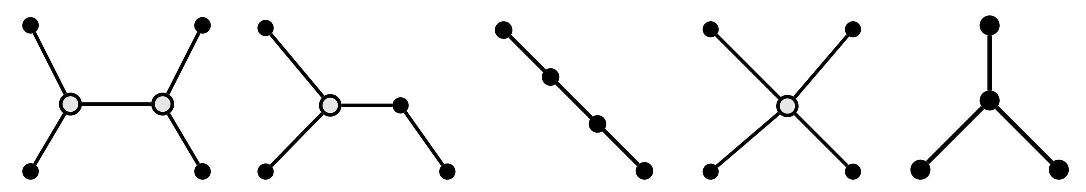
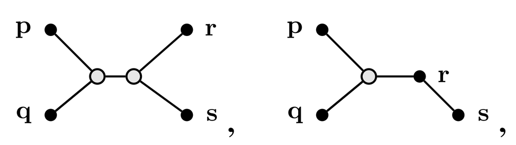


Figure 2: (Leung and Drton, 2018) The solid circles are the observed nodes and grey open circles are the latent nodes.

To illustrate the idea of the theorem, Figure 2 lists out all possible configurations for the minimal subtree induced by any given set of four observed nodes on a tree T (where any node of degree ≤ 2 and not among the considered observed nodes is suppressed). For example, if the four observed nodes have the leftmost two configurations in Figure 2 and are distributed as in



then condition (1) is apparent. Condition (2) can be understood in a similar manner by considering all possible configurations for a minimal subtree induced by any given set of *three* observed nodes on the tree T.

3 Polynomial characterization

Let ρ_{pq} be the Pearson correlation of the pair (X_p, X_q) for any $1 \leq p \neq q \leq m$. The pairwise Markov property implies that

$$\rho_{pq} = \prod_{(u,v) \in ph_T(X_p, X_q)} \rho'_{uv},$$

where ρ'_{uv} is the Pearson correlation between a pair of nodes u and v in V. Theorem 1 implies a characterization for the covariance matrix of a random vector $\mathbf{X}$ in the model $\mathcal{M}_{\mathbf{X}}(T)$; refer to Leung and Drton (2018, Corollary 2.2). To illustrate, let $\mathcal{Q}$ be the set of all unordered quadruples of points $\{p, q, r, s\}$ from $\{1, \dots, m\}$ such that exactly one of the three path pairs in

$$(ph_T(p,q), ph_T(r,s)), \quad (ph_T(p,r), ph_T(q,s)), \quad (ph_T(p,s), ph_T(q,r))$$

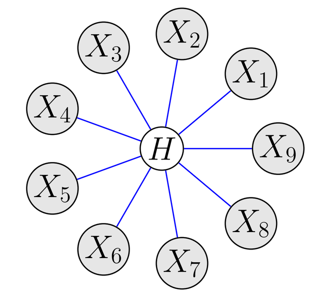
gives an empty set when the union of its two components is taken. In other words, $\mathcal{Q}$ contains all quadruples of points whose induced subtree has the leftmost three configurations in Figure 2. Given $\{p, q, r, s\} \in \mathcal{Q}$, we write $\{p, q\}|\{r, s\} \in \mathcal{Q}$ to indicate that $\{p, q, r, s\}$ belongs to $\mathcal{Q}$ in such a way that the two paths of edges $ph_T(p,q)$ and $ph_T(r,s)$ have an empty intersection. In particular, condition (1) in Theorem 1 implies the following (necessary) conditions for the covariance matrix of $\mathbf{X}$, $\Sigma = (\sigma_{pq})_{1 \leq p,q \leq m}$, to belong to the model $\mathcal{M}_{\mathbf{X}}(T)$:

1. For any $\{p, q\}|\{r, s\} \in \mathcal{Q}$, $\sigma^2_{pr}\sigma^2_{qs} - \sigma^2_{pq}\sigma^2_{rs} \leq 0$ and $\sigma_{pr}\sigma_{qs} - \sigma_{ps}\sigma_{qr} = 0$.
2. For any $\{p, q, r, s\} \notin \mathcal{Q}$, $\sigma_{ps}\sigma_{qr} - \sigma_{pr}\sigma_{qs} = \sigma_{pq}\sigma_{rs} - \sigma_{pr}\sigma_{qs} = 0$.

4 Testing a star tree model

4.1 The star tree model

A star tree model, i.e. a Gaussian latent tree model of the tree



(eight observed nodes and one latent node in this picture), is equivalent to a single factor model with m observed variables $\mathbf{X} = \{X_1 \dots, X_m\}$ described by the linear system of equations

$$X_p = \mu_p + \beta_p H + \epsilon_p, \qquad 1 \leq p \leq m, \tag{3}$$

where μ_p is the mean of X_p, $H \sim N(0,1)$ is a latent variable, β_p is the loading coefficient for variable X_p, and $\epsilon_p \sim N(0, \sigma^2_{p,\epsilon})$ is the idiosyncratic error for variable X_p. In terms of the tree structure no quadruples belong to $\mathcal{Q}$. Hence to test if a dataset comes from a Gaussian star tree model one can test whether for any four points $\{p, q, r, s\} \in \{1, \dots m\}$,

$$\sigma_{ps}\sigma_{qr} - \sigma_{pr}\sigma_{qs} = \sigma_{pq}\sigma_{rs} - \sigma_{pr}\sigma_{qs} = 0.$$

Note that the two polynomials are respectively $\det(\Sigma_{pq,sr})$ and $\det(\Sigma_{ps,qr})$.

4.2 New testing methodology

Assume $\mathbf{X}$ has mean 0 and let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be i.i.d. draws from the distribution of $\mathbf{X}$. Due to the independence of samples, the polynomial $\det(\Sigma_{pq,sr})$ can be estimated unbiasedly with the differences

$$Y_{i,(pq)(sr)} := X_{p,i}X_{s,i}X_{q,i+1}X_{r,i+1} - X_{p,i}X_{r,i}X_{q,i+1}X_{s,i+1}, \qquad i = 1, \dots, n-1,$$

where the subscripts in $Y_{i,(pq)(sr)}$ is indicative of the row and column indices for the submatrix $\Sigma_{pq,sr}$. If we arrange all the polynomials in $\{\det(\Sigma_{pq,sr}), \det(\Sigma_{ps,qr})\}_{\{p,q,r,s\} \in \{1,\dots,m\}}$ into a $2\binom{m}{4}$-vector Θ, and correspondingly arrange the estimates $\{Y_{i,(pq)(sr)}, Y_{i,(ps)(qr)}\}_{\{p,q,r,s\} \in \{1,\dots,m\}}$ into a $2\binom{m}{4}$-vector $\mathbf{Y}_i$ for each i, then the central limit theorem for *1-dependent* sums ensures that for a sufficiently large sample size n we have the distributional approximation

$$\sqrt{n-1}(\bar{\mathbf{Y}} - \Theta) \approx_d N(0, \Upsilon),$$

where $\bar{\mathbf{Y}} = (n-1)^{-1}\sum_{i=1}^{n-1}\mathbf{Y}_i$ and $\Upsilon = \text{Cov}[\mathbf{Y}_1, \mathbf{Y}_1] + 2\text{Cov}[\mathbf{Y}_1, \mathbf{Y}_2]$. The latter limiting covariance matrix will not degenerate to a singular matrix even if the underlying covariance matrix for $\mathbf{X}$ has a lot of zeros, unlike a previous testing approach taken by Shiers et al. (2016) which is susceptible to such singularity issues.

Since $\Theta = 0$ when the star tree model is the true generating mechanism, we propose to use a scaled version of the computable sup-norm quantity

$$\sqrt{n-1}\|\bar{\mathbf{Y}}\|_\infty$$

as a test statistic for the model validity. A recent advance in high-dimensional Gaussian approximation theory (Chernozhukov et al., 2013) suggests that the asymptotic distribution of this test statistic can be well-approximated with a multiplier bootstrapping technique even when the dimension m is large compared to the sample size n; refer to (Leung and Drton, 2018) for the discussion therein.

4.3 Simulation Results

We experimented with our new testing methodology via simulations, with data generated from the one-factor model in (3) for both $(m, n) = (20, 250)$ and $(m, n) = (20, 500)$. The parameter values are as follows: Both loadings β_1 and β_2 are taken to be 10, while the other loadings are independently generated based on a normal distribution with mean 0 and variance 0.2. The error variances $\sigma^2_{p,\epsilon}$ all equal $1/3$. Our testing methodology is compared with the classical likelihood ratio (LR) test. This is a "near-singular" model since many entries in the covariance matrix are close to zero. Figure 3 shows the empirical test sizes.

The resulting plots highlight the advantages of our proposed testing method based on the sup-norm test statistic. As n increases, the empirical test size of our test leans closer to the 45° line. This is in contrast to the performance of the LR test which rejects the true model (3) all too often, even as n increases. Our approach based on the unbiased polynomial estimates is not subject to non-standard limiting behaviors that plague the LR test when the parameter values lean close to singularities of the parameter space (Drton, 2009).

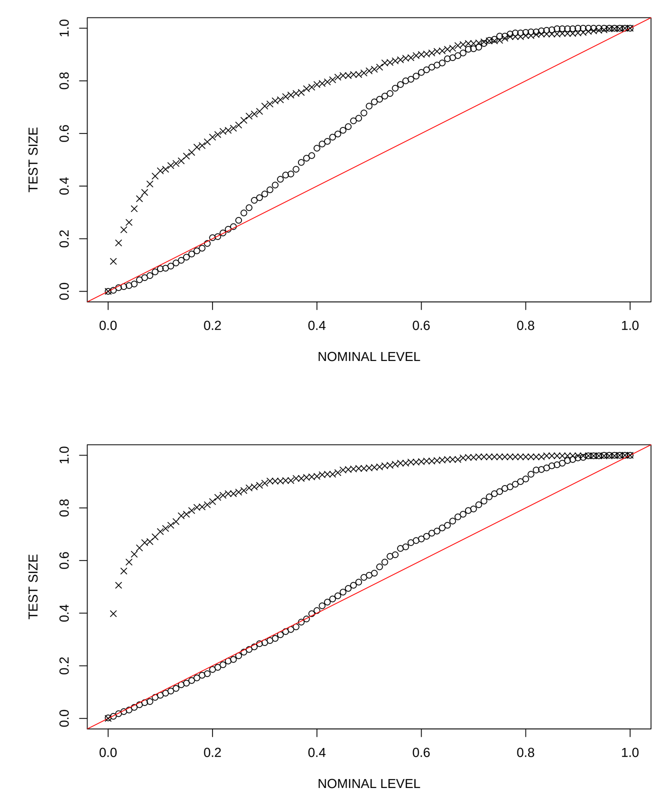


Figure 3: Empirical test sizes vs nominal test levels based on 500 experiments. Upper panels: $(m, n) = (20, 250)$. Lower panels: $(m, n) = (20, 500)$. Open circles: Test based on our sup-norm test statistic. Crosses: Likelihood ratio test.

References

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41(6):2786–2819, 2013.

Myung Jin Choi, Vincent Y. F. Tan, Animashree Anandkumar, and Alan S. Willsky. Learning latent tree graphical models. *J. Mach. Learn. Res.*, 12:1771–1812, 2011.

Mathias Drton. Likelihood ratio tests and singularities. *Ann. Statist.*, 37(2):979–1012, 2009.

Dennis Leung and Mathias Drton. Algebraic tests of general gaussian latent tree models. *NIPS*, 2018.

N. Shiers, P. Zwiernik, J. A. D. Aston, and J. Q. Smith. The correlation space of Gaussian latent tree models and model selection without fitting. *Biometrika*, 103(3):531–545, 2016.