2. T-induced pseudo-metric on X

Suppose \( W \rightarrow E \rightarrow \mathbb{R}_2 \) is any function that assigns non-negative weights to the edges in \( E \), and let \( p_{\mathbb{R}_2}(\cdot) \) be the set of edges on the unique path that connects \( X_p \) and \( X_q \) in \( T \). One can define a pseudo-metric \( d_T \) on \( X \) by

\[
\delta(x, y) = \sum_{i=1}^{m} e^{-w_i/2}, \quad \text{for} \quad p < q.
\]

This is known as a \( T \)-induced pseudo-metric on \( X \). We have the following main result:

**Theorem 1** (Extension of Corollary 1) in Shers et al. (2016), Suppose \( W : X \rightarrow \mathbb{R}_2 \) is a pseudo-metric defined on \( X \). Let \( d_T(X_p, X_q) = \delta(\mathbb{R}_2, X_p, X_q) \) for any \( p, q \in \{1, \ldots, m\} \). For any four distinct indices \( 1 \leq p, q, r, s \leq m \) such that \( p \neq q \neq r \neq s \), we have

\[
\delta(x, y) = \sum_{i=1}^{m} e^{-w_i/2}, \quad \text{for} \quad p < q.
\]

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

\[
\delta(x, y) = \delta(x, z) + \delta(z, y).
\]

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

\[
X_p = \begin{pmatrix} X_p \\ \vdots \end{pmatrix}, \quad \text{and} \quad \Sigma_p = \begin{pmatrix} \Sigma_p & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & \cdots & \Sigma_p \end{pmatrix}
\]

is a polynomial correlation matrix defined by

\[
\rho_{pq} = \frac{\langle \phi(p, q), \phi(p, r) \rangle}{\sqrt{\langle \phi(p, q), \phi(p, q) \rangle \langle \phi(p, r), \phi(p, r) \rangle}},
\]

where \( \rho_{pq} \) 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 \( X \) in the model \( M \) of \( T \) (Leung and Drton, 2018). To illustrate, let \( Q \) be the set of all unsorted quadruples of points \( \{p, q, r, s\} \) from \( \{1, \ldots, m\} \) such that exactly one of the three path pairs in

\[
(\phi(p, q), \phi(p, r)), \quad (\phi(p, q), \phi(q, r)), \quad (\phi(p, r), \phi(q, r))
\]

gives an empty set when the union of its two components is taken. In other words, \( Q \) contains all quadruples of points whose induced subtree has the most three configurations in Figure 2. Given \( \{p, q, r, s\} \subseteq \mathbb{R}_2 \), we write \( Q_1, Q_2, Q_3 \) for the three path pairs in

\[
(\phi(p, q), \phi(p, r)), \quad (\phi(p, q), \phi(q, r)), \quad (\phi(p, r), \phi(q, r))
\]

for \( 1 \leq p < q < r < s \leq m \).

4.2 New testing methodology

Assume \( X \) has mean 0 and let \( X_1, \ldots, X_n \) be i.i.d. draws from the distribution of \( X \). Due to the independence of samples, the polynomial \( d_T(X_p, X_q) \) can be estimated unbiasedly with the difference

\[
Y_{pq}(X_p, X_q) = X_p \cdot X_q - X_p \cdot X_q - \sum_{i=1}^{m} X_i \cdot X_i - X_{p, q},
\]

where the subscripts in \( Y_{pq}(X_p, X_q) \) indicate the row and column indices for the submatrix \( \Sigma_{pq} \). If we arrange all the polynomials in \( d_T(X_p, X_q) \) into a \( (m^2) \times (m^2) \) matrix and correspondingly arrange the estimates \( (Y_{pq}(X_p, X_q)) \) into a \( (m^2) \times (m^2) \) matrix \( Y \), then the central limit theorem for \( T \)-independent sums ensures that for a sufficiently large sample size \( n \) we have the distributional approximation

\[
\sqrt{n} \left( n^{-1/2} \sum_{i=1}^{m} Y_i - n^{-1/2} \sum_{i=1}^{m} T_i \right) \sim N(0, \Sigma_T),
\]

where \( \Sigma_T \) is the covariance matrix of the \( T \)-independent sum. Up to the leading order term \( o(1) \), the covariance matrix will not degenerate to a singular matrix even if the underlying covariance matrix for \( X \) has a lot of zeroes, unlike a previous testing approach taken by Shers et al. (2016) which is susceptible to such singularity issues.

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

\[
\sqrt{n} \left( n^{-1/2} \sum_{i=1}^{m} Y_i - n^{-1/2} \sum_{i=1}^{m} T_i \right) \sim N(0, \Sigma_T),
\]

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 \( (m, n) \in \{(20, 500), (20, 1000), (50, 500), (50, 1000)\} \) and \( (m, n) = \{(20, 200), (20, 1000)\} \). The parameter values are as follows: Both loadings \( \beta_1 \) and \( \beta_2 \) 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 \( \sigma_i^2 \), equal 1.\( 2 \). One 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 5% 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).