

3 **Estimation of graphical models for skew**
4 **continuous data**

5 **Linh H. Nghiem^{1,3} | Francis K.C.Hui¹ | Samuel**
Müller^{2,3} | A.H. Welsh¹

¹Research School of Finance, Actuarial
Studies and Statistics, Australian National
University

²Department of Mathematics and Statistics,
Macquarie University

³School of Mathematics and Statistics,
University of Sydney

Correspondence

Linh H. Nghiem
Research School of Finance, Actuarial
Studies and Statistics, Australian National
University, Acton ACT 2601 Australia
Email: linh.nghiem@anu.edu.au

Funding information

ARC Discovery Grant DP180100836

We consider a new approach for estimating non-Gaussian undirected graphical models. Specifically, we model continuous data from a class of multivariate *skewed* distributions, whose conditional dependence structure depends on both a precision matrix and a shape vector. To estimate the graph, we propose a novel estimation method based on nodewise regression: we first fit a linear model, then fit a one component projection pursuit regression model to the residual obtained from the linear model, and finally threshold appropriate quantities. Theoretically, we establish error bounds for each nodewise regression and prove consistency of the estimated graph when the number of variables diverges with the sample size. Simulation results demonstrate the superior finite sample performance of our new method to existing methods for estimating Gaussian and non-Gaussian graphical models. We finally demonstrate an application of the proposed method on observations of physicochemical properties of wine.

Keywords – skewness, projection pursuit, neighborhood regression, skew normal distribution

6 **1 | INTRODUCTION**

7 In multivariate analysis, a graphical model is a popular and attractive representation of the re-
8 lationship between random variables. More specifically, given a p -dimensional random vec-

9 tor $\mathbf{Y} = (Y_1, \dots, Y_p)^\top$, the underlying Markov random field is specified by an undirected graph
10 \mathcal{G} (Lauritzen, 1996). This graph consists of a set of vertices $V = \{1, \dots, p\}$ and an edge set
11 $E \subset V \times V$ that represents the conditional dependence structure among the random variables:
12 A pair $(j, k) \in E$ if and only if Y_j and Y_k are conditionally dependent given all the remaining
13 variables. Vertex k is said to be in the neighborhood $\mathcal{N}_j = \{k : (j, k) \in E\}$ of vertex j if the
14 pair (j, k) belongs to the edge set. Undirected graphical models have applications in many ar-
15 eas, ranging from establishing gene regulatory networks (Schäfer and Strimmer, 2005; Werhli
16 et al., 2006), social network analysis (Jamali and Ester, 2010), to functional brain connectivity
17 (Belilovsky et al., 2016; Monti et al., 2017) among many others. Knowledge of the conditional
18 dependence structure can also be used to improve other inferential methods. For example, Yu
19 and Liu (2016) incorporates graphical structure of the covariates to improve estimation of the
20 coefficients in the linear model.

21 When all the random variables are continuous, the problem of estimating undirected graph-
22 ical models from data has largely focused on the multivariate Gaussian distribution. In this set-
23 ting, the conditional dependence structure is completely characterised by the precision matrix,
24 i.e the inverse of the covariance matrix. Specifically, a pair (j, k) belongs to the edge set of the
25 graph if and only if the corresponding element in the precision matrix is not equal to zero. As a
26 result, estimating the Gaussian graphical model is equivalent to estimating its precision matrix,
27 for which there are two main estimation approaches: maximum likelihood estimation based
28 on multivariate Gaussian distribution, and nodewise regression, in which each component Y_j is
29 regressed upon $\mathbf{Y}_{-j} = \{Y_k, k \neq j\}$. When the sample size n is greater than the number of vari-
30 ables p , estimating the Gaussian graphical model usually involves testing the hypothesis that
31 the element of the precision matrix is zero; see for example Drton and Perlman (2004). When
32 p is large, regularization is usually imposed; examples include the graphical lasso in Friedman
33 et al. (2008) and the nodewise lasso in Meinshausen et al. (2006).

34 Despite the popularity of Gaussian graphical models, for many practical applications, the
35 assumption of a multivariate Gaussian distribution is restrictive; many physical processes gen-
36 erate data that are continuous but non-Gaussian, such as bacteria growth (Ghosh et al., 2016),
37 cloud cover formation (Sengupta et al., 2016) among many others. Baba et al. (2004) points out
38 that in the class of elliptical distributions, conditional independence is only possible for the mul-
39 tivariate Gaussian distribution, so it is generally thought to be difficult to relax the aforemen-
40 tioned multivariate Gaussian assumption and still allow conditional independence. To over-
41 come this challenge, Liu et al. (2009), Liu et al. (2012), and Xue et al. (2012) study the undi-
42 rected graphical model for the Gaussian copula family (non-paranormal distributions), where
43 a marginal transformation maps the non-Gaussian data to a latent multivariate Gaussian data
44 and preserves the conditional independence structure. While the Gaussian copula family is
45 richer than the multivariate Gaussian, this class of distribution is still restrictive because its

conditional dependence structure is assumed to be that of a latent multivariate Gaussian distribution. Another approach taken by Fellinghauer et al. (2013) and Voorman et al. (2014) is to assume that data are generated from a multivariate distribution where any conditional distribution of one component given all the others $p(Y_j|\mathbf{Y}_{-j})$ depends on \mathbf{Y}_{-j} only through the conditional mean $E(Y_j|\mathbf{Y}_{-j})$. Hence, estimation of the graphical model can be done through variable selection for the conditional mean, using methods such as random forest (Fellinghauer et al., 2013) and additive models (Voorman et al., 2014). This assumption can be violated if the conditional variance depends on \mathbf{Y}_{-j} , and neither Fellinghauer et al. (2013) nor Voorman et al. (2014) points out a joint distribution of data other than the Gaussian and non-paranormal distributions for which the above assumption holds. In another line of research, Lin et al. (2016), Yuan et al. (2016) and Yu et al. (2018) use a score matching method to study undirected graphical models for the multivariate exponential family distribution. Zhuang et al. (2016) generalizes the exponential family to a class of exponential trace models. Morrison et al. (2017) develops an algorithm for estimating graphical models for continuous non-Gaussian data using transport maps; however, the algorithm is greedy and typically requires very large sample size to recover the true graphical model even with a small number of variables.

In this paper, we take a different approach by considering the problem of graphical model estimation for a class of multivariate skewed distributions, where the conditional independence structure depends critically on the skewness of the distribution. In particular, we model data using the generalized multivariate skew normal distributions (GMSN). A random vector $\mathbf{Y} \in \mathbb{R}^p$ is said to follow a p -variate GMSN distribution, denoted as $\mathbf{Y} \sim \text{GMSN}(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, G)$, if its joint density has the form

$$f(\mathbf{Y}) = 2\phi_p(\mathbf{Y}; \boldsymbol{\xi}, \boldsymbol{\Sigma})G\left\{\boldsymbol{\alpha}^\top \mathbf{D}^{-1}(\mathbf{Y} - \boldsymbol{\xi})\right\}, \quad (1)$$

where $\phi_p(\cdot; \boldsymbol{\xi}, \boldsymbol{\Sigma})$ denotes the p -variate Gaussian density with mean vector $\boldsymbol{\xi}$ and covariance matrix $\boldsymbol{\Sigma}$, $\mathbf{D} = \text{diag}(\sigma_{11}^{1/2}, \dots, \sigma_{pp}^{1/2})$ is a diagonal matrix whose elements equal the square root of the diagonal elements of $\boldsymbol{\Sigma}$, and $G(\cdot)$ is the distribution function of a univariate random variable that satisfies $G(x) = 1 - G(-x)$. Starting from a p -variate Gaussian random vector $\mathbf{U} = (U_1, \dots, U_p)^\top \sim N_p(\boldsymbol{\xi}, \boldsymbol{\Sigma})$, and a univariate random variable V whose distribution function is G , we obtain the GMSN random vector with density (1) from the transformation $\mathbf{Y} = \mathbf{U}1(\boldsymbol{\alpha}^\top \mathbf{D}^{-1}(\mathbf{U} - \boldsymbol{\xi}) < V) - \mathbf{U}1(\boldsymbol{\alpha}^\top \mathbf{D}^{-1}(\mathbf{U} - \boldsymbol{\xi}) > V)$, where $1(\cdot)$ denotes an indicator function. Unlike the copula family that is obtained from applying marginal transformations to \mathbf{U} , the GMSN distribution is obtained from a transformation of all the components of \mathbf{U} simultaneously. If G is the distribution function of a standard normal random variable, then (1) is the density of the multivariate skew normal (MSN) distribution (Azzalini and Dalla Valle, 1996; Azzalini and Capitanio, 1999), which have applications in many areas; see Azzalini (2013) and references therein

81 for a survey of applications. When $\alpha = \mathbf{0}$, the GMSN distribution reduces to the multivariate
82 Gaussian distribution with mean vector ξ and covariance matrix Σ . Figure 1 shows data from
83 a bivariate Gaussian distribution, along with data from the generalized bivariate skew normal
84 distribution with the same covariance matrix but different shape vectors and skewing func-
85 tions.

86 Compared to the multivariate Gaussian distribution and the Gaussian copula family, the
87 GMSN distribution's conditional dependence structure depends both on the precision matrix
88 $\Omega = \Sigma^{-1}$ and on the shape vector α . This characteristic is also shared by the MSN distribu-
89 tion and its variants, such as the extended multivariate skew normal distribution (EMSN). The
90 graphical model for EMSN distribution is studied by Capitanio et al. (2003); although they dis-
91 cuss maximum likelihood estimation for graphical model estimation, they do not discuss asymp-
92 totic properties nor provide theoretical results for the estimator. They further note that maxi-
93 mum likelihood estimation for EMSN is computationally challenging, mostly because it involves
94 all the parameters simultaneously. This is also true for the GMSN distribution; however, it
95 turns out that for identifying the graphical model, we can avoid maximum likelihood estimation.
96 Zareifard et al. (2016) develop a Bayesian method to estimate graphical model of the multivari-
97 ate closed skew normal distribution, which is another generalization of MSN. Noting that the
98 dependence of the graphical model on the shape vector makes estimation of graphs challeng-
99 ing, they impose additional assumptions on the graph such that conditional dependence only
100 depends on the precision matrix (the graph is decomposable). In this paper, we aim to inves-
101 tigate estimation of the graphical model for GMSN without any further assumption and also
102 avoids maximum likelihood estimation; our approach is based instead on nodewise regression,
103 in which we estimate the graphical model by estimating appropriate quantities in the condi-
104 tional expectation of GMSN.

105 We will work under a setting where we assume that the graphical model is sparse. Such an
106 assumption is often made when the number of variables p is large, that is, only a few among
107 $p(p - 1)/2$ pairs (Y_i, Y_j) are conditionally dependent. Applying this condition to the GMSN dis-
108 tribution translates to assuming that both the shape vector α and the precision matrix Ω are
109 sparse. Such assumptions will be made precise in Section 4. Furthermore, we will assume
110 throughout the paper that the number of observations n is greater than the number of vari-
111 ables p . Theoretically, we prove important properties of the conditional distribution and condi-
112 tional expectation of the GMSN distribution, which form the basis for our new nodewise ap-
113 proach to estimate the graphical model from the data. We then establish a theoretical bound
114 on the error for each nodewise regression, and provide a simple thresholding algorithm, which
115 we show consistently estimates the true graphical model. Finally, simulation studies and a real
116 data example demonstrate the performance of the new methodology in estimating the GMSN
117 graphical model compared to that of popular existing methods for estimating Gaussian and

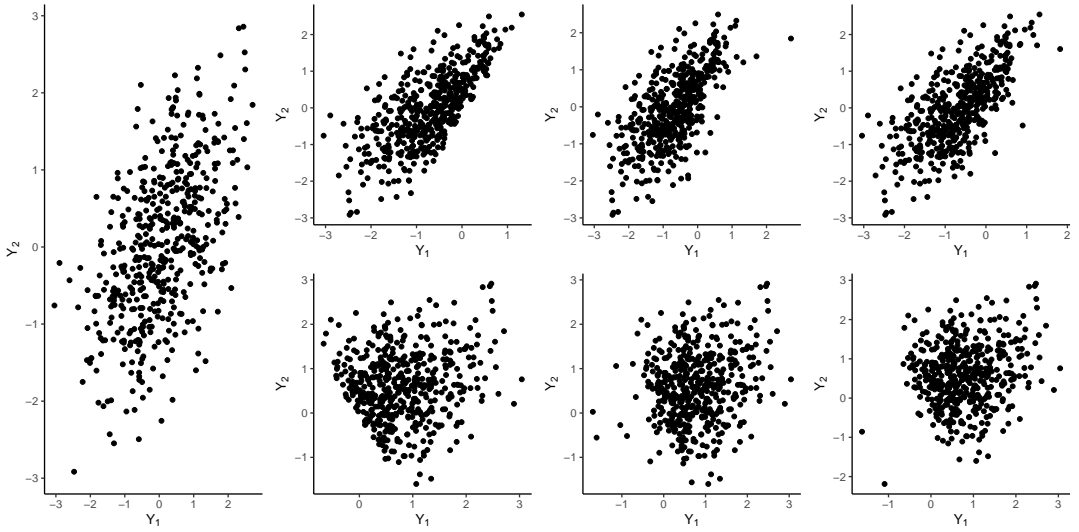


FIGURE 1 Simulated data from the bivariate normal distribution with mean zero vector, variance one and correlation 0.5 (leftmost) and data from several bivariate generalized skew normal distributions. In the first row, all distributions have shape vector $\alpha = (-10, 5)^T$, while in the second row, all distributions have shape vector $\alpha = (10, 5)^T$. From the second to the fourth column, the skewing function is chosen to be the distribution function of the standard normal, standard Cauchy, and equal mixture of two Cauchy distribution having location 2 and -2 and same scale 0.5, respectively.

118 **continuous non-Gaussian graphical models.**

119 The paper is organized as follows. Section 2 establishes properties of the GMSN distribu-
 120 tion, particularly properties related to the conditional expectation. Section 3 proposes a node-
 121 wise method for estimating the graphical model, and Section 4 discusses theoretical results
 122 for the new method **when the true location vector ξ and scale matrix D are assumed to be**
 123 **known**, and Section 5 discusses estimation of ξ and D . Section 6 presents a simulation study
 124 that demonstrates the superior performance of the new nodewise method. Section 7 illus-
 125 trates the application of our new method to a real dataset. Section 8 contains some concluding
 126 remarks. All detailed proofs are deferred to the Appendix.

127 The following notations are used throughout the paper. For any matrix A , we use $a^{(i)}$, a_j ,
 128 and a_{ij} to denote the i th row, the j th column, and the (i, j) -th element of A , respectively. We
 129 let $A^{(-i)}$ and A_{-j} denote the sub-matrix of A with row i th and column j th removed, respectively.
 130 Then the notations $A_{-j}^{(-i)}$ and $a_{-j}^{(i)}$ denote the sub-matrix of A with both row i th and column j th
 131 removed and, the i th row of A with column j th removed, respectively.

2 | THE GENERALIZED MULTIVARIATE SKEW NORMAL DISTRIBUTION

Consider the random vector $\mathbf{Y} = (Y_1, \dots, Y_p)^\top \sim \text{GMSN}(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, G)$ with joint density (1). Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, with elements $\{\omega_{jk}\}, j, k = 1, \dots, p$. The vector $\boldsymbol{\xi} = (\xi_1, \dots, \xi_p)^\top$ and the matrix $\mathbf{D} = \text{diag}(\sigma_{11}^{1/2}, \dots, \sigma_{pp}^{1/2})$ are referred to as the location and scale matrix of \mathbf{Y} respectively, because the random vector $\mathbf{Z} = \mathbf{D}^{-1}(\mathbf{Y} - \boldsymbol{\xi}) \sim \text{GMSN}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\alpha}, G)$, where $\tilde{\boldsymbol{\Sigma}} = \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1}$ is the correlation matrix associated with $\boldsymbol{\Sigma}$. To simplify the notation, we let $\sigma_j = \sigma_{jj}^{1/2}$ and refer to $\{\sigma_j, j = 1, \dots, p\}$ as the scale parameters of \mathbf{Y} . **We also note that when $\boldsymbol{\alpha} \neq \mathbf{0}$, the vector $\boldsymbol{\xi}$ is the mean of the normal part \mathbf{U} , but not the mean vector of \mathbf{Y} ; similarly, $\boldsymbol{\Sigma}$ is the covariance matrix of \mathbf{U} but not the covariance matrix of \mathbf{Y} .**

The following theorem states the condition for any pair (Y_j, Y_k) to be conditionally independent given all the other components of \mathbf{Y} .

Theorem 1 *Let $\mathbf{Y} = (Y_1, \dots, Y_p)^\top \sim \text{GMSN}(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, G)$. Let $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. Then the two components Y_j and Y_k are conditionally independent if and only if $\omega_{jk} = \omega_{kj} = 0$ and $\alpha_j\alpha_k = 0$.*

Theorem 1 implies that conditional dependence in the GMSN distribution depends on both the precision matrix $\boldsymbol{\Omega}$ and the shape vector $\boldsymbol{\alpha}$, but **does not depend on the location $\boldsymbol{\xi}$ and scale matrix \mathbf{D} . In other words, the graphical models for \mathbf{Y} and \mathbf{Z} are identical.** The conditions in Theorem 1 are the same as those obtained by Azzalini and Capitanio (1999, Section 6.3) and Capitanio et al. (2003, Section 5.3) for conditional independence in the MSN and EMSN distribution, respectively. The novelty here is that in the GMSN distribution, these conditions for conditional independence are invariant to the skewing function G .

While we can estimate all the parameters (and hence its graphical model) of the GMSN distribution by maximum likelihood, it is challenging due to a large number of parameters, especially when the skewing function G is unknown and the number of dimensions p is large. Importantly however, it is not necessary to estimate all these parameters for recovering the graphical model of the GMSN distribution; in fact, we only need to determine which elements of $\boldsymbol{\Omega}$ and which elements of $\boldsymbol{\alpha}$ are zero. This leads us to propose a graphical model estimation approach that do not rely on maximum likelihood estimation; instead, we employ a nodewise regression approach and estimate appropriate quantities in the conditional expectations of GMSN to recover the true graphical model.

To form a basis for the proposed estimation method, we establish some properties of the conditional distributions. **Since the graphical models for \mathbf{Y} and \mathbf{Z} are identical, we discuss the properties regarding the conditional distributions for \mathbf{Z} to simplify the notation.**

Theorem 2 *Let $\mathbf{Z} \sim \text{GMSN}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\alpha}, G)$, where $\tilde{\boldsymbol{\Sigma}}$ is a correlation matrix. Then for any skewing function G ,*

166 **(a)** If $\alpha_j = 0$, then the conditional distribution of Z_j given \mathbf{Z}_{-j} is normal with mean $\mathbf{Z}_{-j}^\top \beta_j$ and variance

167
$$\tilde{\sigma}_{jj,-j} = 1 - \tilde{\sigma}_{-j}^{(j)} \left(\tilde{\boldsymbol{\Sigma}}_{-j}^{(-j)} \right)^{-1} \tilde{\sigma}_j^{(-j)}, \text{ where } \beta_j = \left(\tilde{\boldsymbol{\Sigma}}_{-j}^{(-j)} \right)^{-1} \tilde{\sigma}_j^{(-j)}.$$

168 **(b)** In general, the conditional expectation of Z_j given \mathbf{Z}_{-j} is

169
$$E(Z_j | \mathbf{Z}_{-j}) = \mathbf{Z}_{-j}^\top \beta_j + g_j(\mathbf{Z}_{-j}^\top \zeta_j),$$

170 with $\zeta_j = \alpha_{-j} + \alpha_j \beta_j$, and

171
$$g_j(\mathbf{Z}_{-j}^\top \zeta_j) = \tilde{\sigma}_{jj,-j}^{1/2} \frac{E \left\{ \tilde{Z} G \left(\tilde{Z} \alpha_j \tilde{\sigma}_{jj,-j}^{-1/2} + \mathbf{Z}_{-j}^\top \zeta_j \right) \right\}}{E \left\{ G \left(\tilde{Z} \alpha_j \tilde{\sigma}_{jj,-j}^{-1/2} + \mathbf{Z}_{-j}^\top \zeta_j \right) \right\}}, \tilde{Z} \sim N(0, 1).$$

172 **(c)** Consider the random variable $g_j(\mathbf{Z}_{-j}^\top \zeta_j)$. For any odd function $h : \mathbb{R}^{p-1} \rightarrow \mathbb{R}$, i.e $h(\mathbf{x}) = -h(-\mathbf{x})$,
 173 we have $E \left\{ h(\mathbf{Z}_{-j}) g_j \left(\mathbf{Z}_{-j}^\top \zeta_j \right) \right\} = 0$. Specifically, we have

174
$$E \left\{ \mathbf{Z}_{-j} g_j \left(\mathbf{Z}_{-j}^\top \zeta_j \right) \right\} = \mathbf{0}, \text{ and } E \left\{ \left(\mathbf{Z}_{-j} \mathbf{Z}_{-j}^\top \right)^{-1} \mathbf{Z}_{-j} g_j \left(\mathbf{Z}_{-j}^\top \zeta_j \right) \right\} = \mathbf{0}.$$

175 Part (a) of Theorem 2 implies that, if $\alpha_j = 0$, although Z_j is not marginally normal, the conditional
 176 distribution of Z_j given \mathbf{Z}_{-j} is still normal. Also in this case, the function g_j is zero, so the condi-
 177 tional expectation is linear. On the other hand, when $\alpha_j \neq 0$, parts (b) and (c) of Theorem 2 imply
 178 that the conditional expectation has the form of an extended partially linear single index model
 179 (Xia et al., 1999), where the non-linear and linear part are orthogonal. In this case, the function
 180 g_j does not generally have a closed form, unless the skewing function G is the standard normal
 181 distribution function, see Azzalini and Capitanio (1999) for the closed form in that case.

182 Theorem 3 below combines Theorems 1 and 2, relating conditional independence between
 183 any two components of Y to the quantities in the conditional expectations. This result is the
 184 key to the estimation method proposed in the next section.

185 **Theorem 3** Let $\mathbf{Y} = (Y_1, \dots, Y_p)^\top \sim \text{GMSN}(\xi, \boldsymbol{\Sigma}, \alpha, G)$ and define quantities as in Theorem 2. Let
 186 $\beta_j^{(k)}$ and $\zeta_j^{(k)}$ denote the coefficients corresponding to Z_k in β_j and ζ_j , respectively. Consider the
 187 following statements:

188 **(1)** Y_j and Y_k are conditionally independent.

189 **(2)** $\omega_{jk} = 0$ and $\alpha_j \alpha_k = 0$.

190 **(3a)** $\beta_k^{(j)} = \beta_j^{(k)} = 0$.

191 **(3b)** $\zeta_k^{(j)} \zeta_j^{(k)} = 0$.

192 **(3c)** At least one of g_j and g_k is a zero function.

193 Then we have:

194 (i) (1) \iff (2).

195 (ii) (2) \implies (3a), (3b), and (3c).

196 (iii) (3a) and (3b) \implies (2).

197 (iv) (3a) and (3c) \implies (2).

198 Part (i) of Theorem 3 implies that estimating the graphical model for the GMSN distribution
 199 requires us to identify the zero elements in both the precision matrix $\mathbf{\Omega}$ and the shape vector α .
 200 Similar to the Gaussian graphical model, a zero element in $\mathbf{\Omega}$ is equivalent to a zero component
 201 of the coefficient associated with the linear part of the conditional expectation. Next, as sug-
 202 gested by parts (iii) and (iv) of Theorem 3, we have two choices to identify whether the product
 203 of two corresponding elements in the shape vector is zero. The first choice, corresponding to
 204 (iii), is by identifying zero components of the coefficient associated with the non-linear part of
 205 the conditional expectation. The second choice, corresponding to (iv) is by identifying whether
 206 the entire non-linear part is a zero function. Our proposed method in Section 3 is based on
 207 the second choice, because the p -dimensional vector ζ_j is not identifiable from the data if the
 208 true function g_j is zero, but it is comparably more straightforward to assess whether the true
 209 function g_j is zero without relying on ζ_j .

210 3 | NODEWISE REGRESSION FOR GRAPHICAL MODEL ESTIMATION

211 We now turn to the problem of estimating the graphical model from an $n \times p$ data matrix \mathbf{Y} ,
 212 whose rows $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(n)}$ are independent random sample following $\text{GMSN}(\xi, \mathbf{\Sigma}, \alpha, G)$. We are
 213 interested in estimating the neighborhood of each vertex in the graph, and thus the underlying
 214 graphical model. **In the next two sections, we will assume that the location ξ and scale matrix**
 215 **\mathbf{D} are known, and work with the standardized vectors $\mathbf{z}^{(i)} = \mathbf{D}^{-1}(\mathbf{y}^{(i)} - \xi)$, $i = 1, \dots, n$. Let \mathbf{Z} be**
 216 **the $n \times p$ matrix whose j^{th} row is $\mathbf{z}^{(j)}$. Estimation of ξ and \mathbf{D} will be discussed in Section 5. To**
 217 **avoid confusion, we note that for the remaining of the paper (excluding the appendix), we will**
 218 **use capital boldface letter to denote an appropriate matrix.**

219 Based on Theorem 3, we propose the following nodewise algorithm for estimating the graph-
 220 ical model of the GMSN distribution. At the j th iteration, $j = 1, \dots, p$, we consider a regression
 221 with response vector the j th column z_j and $n \times (p - 1)$ matrix \mathbf{Z}_{-j} , which is the matrix \mathbf{Z} with the
 222 j th column removed.

223 *Step 1: Estimating β_j*

224 Since $E\{\mathbf{Z}_{-j}g_j(\mathbf{Z}_{-j}\zeta_j)\} = \mathbf{0}$ by part (c) of Theorem 2, the vector β_j can be estimated by fitting

a linear model without an intercept. In this paper, we use ordinary least squares to estimate β_j . That is

$$\hat{\beta}_j = (\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j})^{-1} \mathbf{Z}_{-j}^\top \mathbf{z}_j. \quad (2)$$

It is important for the linear model to not have an intercept, and the variables should not be centered. If the intercept is included in the model or the variables are centered, the slope estimator estimates $\beta_j + \{\text{Var}(\mathbf{z}_{-j}^{(i)})\}^{-1} \text{Cov}\{\mathbf{z}_{-j}^{(i)}, g_j(\mathbf{z}_{-j}^{(i)}; \zeta_j)\}$, which is not equal to β_j since the covariance term is not zero. On the other hand, if we fit the linear model *without* an intercept, the slope estimator will estimate $\beta_j + \{E(\mathbf{z}_{-j}^{(i)} \mathbf{z}_{-j}^{(i)})\}^{-1} E\{\mathbf{z}_{-j}^{(i)} g_j(\mathbf{Z}_{-j}^\top; \zeta_j)\} = \beta_j$, since $E\{\mathbf{z}_{-j}^{(i)} g_j(\mathbf{z}_{-j}^{(i)}; \zeta_j)\} = 0$ by part (c) of Theorem 2. In Section 4.1, we show that the estimator $\hat{\beta}_j$ is unbiased for β_j and establish a bound on the ℓ_2 error of $\hat{\beta}_j$ when p diverges with n .

Step 2: Estimating the function g_j

By Theorem 3, any pair (Y_j, Y_k) or (Z_j, Z_k) is conditionally independent when at least one of the two functions g_j and g_k is the zero function. Therefore, after estimating the coefficients of the linear term β_j , we proceed to estimate the function g_j and assess whether it is a zero function. Let $\mathbf{r}_j = \mathbf{z}_j - \mathbf{Z}_{-j} \hat{\beta}_j$ denote the residual vector from the linear model fit in Step 1, and $r_j^{(i)}$, $i = 1, \dots, n$ denote the i th element of \mathbf{r}_j . We then fit a one-component projection pursuit regression model with parameters $\mu_{j,0} \in \mathbb{R}$, $\theta_j \in \mathbb{R}^{p-1}$, $\tau_j \in \mathbb{R}$ and (unknown) smooth function v_j as

$$r_j^{(i)} = \mu_{j,0} + \tau_j v_j(\mathbf{z}_{-j}^{(i)} \theta_j) + \varepsilon_j^{(i)}, \quad (3)$$

subject to

$$\|\theta_j\|_2 = 1, \tau_j \geq 0, \frac{1}{n} \sum_{i=1}^n v_j(\mathbf{z}_{-j}^{(i)} \theta_j) = 0, \frac{1}{n} \sum_{i=1}^n v_j^2(\mathbf{z}_{-j}^{(i)} \theta_j) = 1, E(\varepsilon_j^{(i)}) = 0. \quad (4)$$

The constraints in (4) are standard for fitting the projection pursuit regression model (Friedman, 1984), ensuring that the model (3) is estimable from the data. The intercept $\mu_{j,0}$ estimates the overall mean and the scale τ_j estimates the standard deviation of the function g_j , while the function v_j estimates the centered and scaled version of g_j . Since we are interested in whether the function g_j is zero everywhere or not, we focus on the properties of the estimated scalar $\hat{\tau}_j$. In Section 4, we will show that $\hat{\tau}_j$ converges to the true standard deviation of the function g_j , and hence can be used to determine whether g_j is a zero function or not.

While there are multiple approaches to fit the model (3), see for example Friedman and Stuetzle (1981), Friedman et al. (1983) and Friedman (1984), we employ a similar approach to

255 Chen (1991) because it produces an estimator with a convergence rate that does not depend
 256 on the dimension p as $n \rightarrow \infty$. Specifically, we estimate the parameters of the model (3) by
 257 minimizing

$$\sum_{i=1}^n \left\{ r_j^{(i)} - \mu_{j,0} - \tau_j v_j(\mathbf{z}_{-j}^{(i)}; \boldsymbol{\theta}_j) \right\}^2$$

259 subject to constraints (4). The intercept $\mu_{j,0}$ is estimated by $\hat{\mu}_{j,0} = n^{-1} \sum_{i=1}^n r_j^{(i)}$, which is the aver-
 260 age of the residual vector. Next, let $\tilde{\mathbf{r}}_j$ be the vector with elements $\tilde{r}_j^{(i)} = r_j^{(i)} - \hat{\mu}_{j,0}$, $i = 1, \dots, n$.
 261 The optimization procedure then cycles through the following three steps until convergence.
 262 First, for given $\boldsymbol{\theta}_j$ and a real function v_j , we estimate τ_j by

$$\hat{\tau}_j = \frac{1}{n} \left| \sum_{i=1}^n \tilde{r}_j^{(i)} v_j(\mathbf{z}_{-j}^{(i)}; \boldsymbol{\theta}_j) \right| \geq 0.$$

264 Second, for a given τ_j and $\boldsymbol{\theta}_j$, we estimate the function v_j by M_j cubic B-spline functions $\{\psi_k\}$, $k =$
 265 $1, \dots, M_j$ with knots placed at equidistant points in the interval $[0, 1]$. In other words, $v_j(\mathbf{Z}_{-j}; \boldsymbol{\theta}_j) \approx$
 266 $\sum_{k=1}^{M_j} \gamma_k \psi_k(\mathbf{Z}_{-j}; \boldsymbol{\theta}_j)$ where $\gamma_k \in \mathbb{R}$. Let \mathbf{A} be the $n \times M_j$ matrix whose elements (i, k) equal $\psi_k(\mathbf{z}_{-j}^{(i)}; \boldsymbol{\theta}_j)$.
 267 Then the vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_{M_j})^\top$ is estimated by $\hat{\boldsymbol{\gamma}} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \tilde{\mathbf{r}}_j$, and hence the function v_j
 268 is estimated by

$$v_j(\mathbf{z}_{-j}^{(i)}; \boldsymbol{\theta}_j) = \frac{\sum_{k=1}^{M_j} \hat{\gamma}_k \psi_k(\mathbf{z}_{-j}^{(i)}; \boldsymbol{\theta}_j)}{\left[\sum_{i=1}^n \left\{ \sum_{k=1}^{M_j} \hat{\gamma}_k \psi_k(\mathbf{z}_{-j}^{(i)}; \boldsymbol{\theta}_j) \right\}^2 \right]^{1/2}}.$$

270 We choose the number of basis functions M_j using generalized cross-validation. Finally, given
 271 a non-constant function v_j and a scalar τ_j , we estimate $\boldsymbol{\theta}_j$ by the Gauss-Newton method. Let
 272 $\hat{\boldsymbol{\theta}}_{j,0}$ be an initial estimate of $\boldsymbol{\theta}_j$. Then we have

$$v_j(\mathbf{z}_{-j}^{(i)}; \boldsymbol{\theta}_j) \approx v_j(\mathbf{z}_{-j}^{(i)}; \hat{\boldsymbol{\theta}}_{j,0}) + v_j'(\mathbf{z}_{-j}^{(i)}; \hat{\boldsymbol{\theta}}_{j,0}) \mathbf{z}_{-j}^{(i)} (\boldsymbol{\theta}_j - \hat{\boldsymbol{\theta}}_{j,0}).$$

274 Hence, we estimate $\boldsymbol{\theta}_j$ by solving

$$\begin{aligned} \hat{\boldsymbol{\theta}}_j &= \arg \min_{\boldsymbol{\theta}_j} \sum_{i=1}^n \left\{ \tilde{r}_j^{(i)} - v_j(\mathbf{z}_{-j}^{(i)}; \hat{\boldsymbol{\theta}}_{j,0}) + v_j'(\mathbf{z}_{-j}^{(i)}; \hat{\boldsymbol{\theta}}_{j,0}) \mathbf{z}_{-j}^{(i)} \hat{\boldsymbol{\theta}}_{j,0} - v_j'(\mathbf{z}_{-j}^{(i)}; \hat{\boldsymbol{\theta}}_{j,0}) \mathbf{z}_{-j}^{(i)} \boldsymbol{\theta}_j \right\}^2 \\ &= \arg \min_{\boldsymbol{\theta}_j} \left\| \mathbf{Q}(\boldsymbol{\kappa}_j - \mathbf{Z}_{-j}^\top \boldsymbol{\theta}_j) \right\|_2^2, \quad \text{subject to } \|\boldsymbol{\theta}_j\|_2 = 1, \end{aligned}$$

278 where \mathbf{Q} is an $n \times n$ diagonal matrix whose diagonal elements are $\{v_j' (\mathbf{z}_{-j}^{(i)} \hat{\theta}_{j,0})\}^2$, and κ_j is a
 279 $n \times 1$ vector whose i th element equals $\{\hat{r}_j^{(i)} - v_j (\mathbf{z}_{-j}^{(i)} \hat{\theta}_{j,0}) + v_j' (\mathbf{z}_{-j}^{(i)} \hat{\theta}_{j,0}) \mathbf{z}_{-j}^{(i)} \hat{\theta}_{j,0}\} / v_j' (\mathbf{z}_{-j}^{(i)} \hat{\theta}_{j,0})$. The
 280 above minimization problem is a weighted least squares problem, so the solution is the stan-
 281 dardized version of $\hat{\theta}_j = (\mathbf{Z}_{-j}^\top \mathbf{Q} \mathbf{Z}_{-j})^{-1} \mathbf{Z}_{-j}^\top \mathbf{Q} \kappa_j$.

282 **Step 3: Neighborhood selection**

283 **By repeating the two steps above**, we obtain the estimates $\hat{\beta}_j$ and $\hat{\tau}_j$ for all $j = 1, \dots, p$. To
 284 estimate the neighborhood of all the vertices in the graph, we subsequently form two $p \times p$
 285 matrices $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$. For any pair (j, k) with $k \neq j$, the off-diagonal elements of $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$
 286 are defined as $b_{jk}^{(1)} = \max(|\hat{\beta}_j^{(k)}|, |\hat{\beta}_k^{(j)}|)$ and $b_{jk}^{(2)} = \hat{\tau}_j \hat{\tau}_k$. The diagonal elements of both matrices
 287 are set to zero. With this definition, both $\mathbf{B}^{(1)}$ and $\mathbf{B}^{(2)}$ are symmetric. We then estimate Y_j and
 288 Y_k to be conditionally independent if both $b_{jk}^{(1)}$ and $b_{jk}^{(2)}$ are small, i.e when the sum $o_{jk} = (b_{jk}^{(1)})^2 +$
 289 $(b_{jk}^{(2)})^2 \leq T$, with T being a positive threshold **(we note that other norms, such as $|b_{jk}^{(1)}| + |b_{jk}^{(2)}|$,**
 290 **can be used as well)**. The estimated neighborhood for vertex j is then $\hat{N}_j = \{k : o_{jk} > T\}$. In
 291 Section 4.2, we establish the theoretical choice of T for neighborhood consistency.

292 Empirically, we can select the threshold using a procedure similar to that developed by
 293 Mestres et al. (2018). In detail, we select the threshold that maximizes either the path con-
 294 nectivity or the Agglomerative Nesting (AGNES) risk function, which are computed as follows.
 295 Consider a grid of values $T_1 < \dots < T_M$ containing M equidistant points between $\min_{jk} (o_{jk})$
 296 and $\max_{j,k} (o_{jk})$. **Note that when the threshold increases, the estimated graph is more sparse.**
 297 Recall $V = \{1, \dots, p\}$ is the set of vertices of the graph. For a value of threshold T_m in the grid,
 298 let $\hat{G}(T_m)$ denote the corresponding estimated graphical model. For any pair (j, k) with $j \neq k$,
 299 let $\hat{g}_{jk}(T_m)$ be the smallest number of edges connecting V_j to V_k (also known as the geodesic
 300 distance between V_j and V_k) in $\hat{G}(T_m)$. If there is no path linking V_j and V_k , we let $\hat{g}_{jk}(T_m) = \infty$.
 301 Then the mean geodesic distance of $\hat{G}(T_m)$ is computed as

$$302 \quad H(T_m) = \frac{2}{p(p-1)} \sum_{j < k} \hat{g}_{jk}(T_m) I_{\{\hat{g}_{jk}(T_m) < \infty\}}.$$

303 Next, we compute the first order differences $\mathcal{D}(T_m) = H(T_m) - H(T_{m-1})$, and the running aver-
 304 age $\bar{\mathcal{D}}(T_m) = m^{-1} \sum_{i=1}^m H(T_i)$. The path connectivity risk function of the estimated graph $\hat{G}(T_m)$
 305 is defined to be

$$306 \quad \text{PC}(\hat{G}(T_m)) = \left| \frac{\mathcal{D}(T_m)}{\bar{\mathcal{D}}(T_m)} \right|,$$

307 and we choose the threshold T_m that maximizes this risk function. **With this definition, the path**
 308 **connectivity risk is maximized at a strictly positive value, so the complete graph (corresponding**

309 to PC = 0) is never selected.

310 To compute the AGNES risk function, first, we cluster the vertices in the estimated graph-
 311 ical model through the Agglomerative Nesting algorithm. The AGNES risk function of the es-
 312 timated graph, is defined to be the agglomerative coefficient (AC) that measures the average
 313 geodesic distance between a vertex in the graph and its closest cluster of vertices. More de-
 314 tails about how to compute the AGNES risk function, including a subset selection procedure to
 315 reduce the computational cost, can be found in Mestres et al. (2018, Section 3.3).

316 The complete nodewise algorithm for estimating the GMSN graphical model when the true
 317 location and scale are known is summarized in Algorithm 1.

Algorithm 1 Estimating GMSN graphical model with known location ξ and scale \mathbf{D}

Form the standardized data $\mathbf{z}^{(i)} = \mathbf{D}^{-1}(\mathbf{y}^{(i)} - \xi)$, $i = 1, \dots, n$.

for $j = 1, \dots, p$ **do**

Fit the linear model without the intercept of \mathbf{z}_j on \mathbf{Z}_{-j} to obtain $\hat{\beta}_j$ and \mathbf{r}_j .

Fit the projection pursuit regression model (3) subject to constraints (4) to obtain $\hat{\tau}_j$.

end for

Form matrix $\mathbf{B}^{(1)}$ with off-diagonal elements $b_{jk}^{(1)} = \max(|\hat{\beta}_j^{(k)}|, |\hat{\beta}_k^{(j)}|)$.

Form matrix $\mathbf{B}^{(2)}$ with off-diagonal elements $b_{jk}^{(2)} = \hat{\tau}_j \hat{\tau}_k$, $j, k = 1, \dots, p$.

Compute $o_{jk} = (\ell_j^{(k)})^2 + (m_j^{(k)})^2$, $j, k = 1, \dots, p$.

Consider a grid of $T = \{T_1, \dots, T_M\}$ consisting of M equidistant point between $\min_{j,k} (o_{jk})$
 and $\max_{j,k} (o_{jk})$, $j \neq k$.

for each value of T_m in the grid **do**,

Estimate Y_j and Y_k to be conditionally independent when $o_{jk} \leq T_m$ and conditionally
 dependent otherwise.

Form the estimated graph $\hat{\mathcal{G}}(T_m)$.

end for

Finally, choose the optimal threshold \hat{T} and the associated graph that maximizes either the
 path connectivity or AGNES risk function.

318 **4 | THEORETICAL ANALYSES**

319 In this section, we establish theoretical results for the estimation procedure proposed in Sec-
 320 tion 3, assuming the location parameter ξ and scale matrix \mathbf{D} are known. Without loss of gen-
 321 erality, we assume that the data are generated from a GMSN distribution with zero location
 322 and identity scale matrix. Furthermore, we assume $n > p$ and allow the number of variables

323 p to grow with the sample size n with the rate defined later in Section 4.2 We assume that the
 324 underlying graphical model is sparse (Meinshausen et al., 2006), meaning that only a few pairs
 325 (Y_j, Y_k^*) are conditionally dependent. Let $S_\alpha = \{j : \alpha_j \neq 0\}$ and $s_\alpha = |S_\alpha|$. Theorem 3 im-
 326 plies that the sparsity of the graphical model is equivalent to the two following assumptions in
 327 Condition 1.

328 **Condition 1** (a) At least one off-diagonal element of Ω is zero, and (b) $2 \leq s_\alpha < p$.

329 Condition (1a) is analogous to the one imposed for covariance selection in the Gaussian
 330 graphical model (Dempster, 1972). Note that it is a weak condition, since we do not impose any
 331 particular sparsity pattern on the precision matrix, such as restricting the maximum number
 332 of non-zero elements in each row of Ω . For condition (1b), the lower bound $s_\alpha \leq 2$ makes
 333 the corresponding graph different from that of a $N_p(\xi, \Sigma)$ random vector, and the strict upper
 334 bound $s_\alpha < p$ ensures not all the pairs are conditionally dependent. As a result, there are only
 335 s_α functions among g_j in the conditional mean functions, $j = 1, \dots, p$ that are truly non-zero.

336 Furthermore, we impose the following technical conditions:

337 **Condition 2** There exist positive constants C_1 and C_2 such that $0 < C_1 \leq \lambda_{\min}(\Sigma) = \lambda_p(\Sigma) <$
 338 $\lambda_{p-1}(\Sigma) \leq \dots \leq \lambda_{\max}(\Sigma) = \lambda_1(\Sigma) \leq C_2 < \infty$, where $\lambda_d(\Sigma)$ denotes the d^{th} largest eigenvalue of Σ .

339 **Condition 3** For any j , the sequence of conditional expectations $E(z_{ij}|\mathbf{Z}_{-j})$, $i = 1, 2, \dots, n$ is asymp-
 340 totically uniformly integrable.

341 **Condition 4** For $j \in S_\alpha$, there exists a positive constant C_3 such that $\tau_j^2 = \text{Var}\{g_j(\mathbf{z}_{-j}^{(i)}\zeta_j)\} \geq C_3 > 0$
 342 for all $i = 1, \dots, n$.

343 Condition 2 implies that all the eigenvalues of $\Sigma_{-j}^{(-j)}$ are bounded away from zero and bounded
 344 above by a sufficiently large constant. Condition 3 ensures the convergence of marginal mo-
 345 ments of the residuals obtained from the linear model. Finally, Condition 4 implies that the
 346 variance τ_j^2 is zero if and only if the true function g_j is a zero function

347 We first establish the theoretical results for each nodewise regression.

348 4.1 | Error bound for each nodewise regression

349 First, we give the bound on the error of the least squares estimator $\hat{\beta}_j$. Recall that the true
 350 conditional expectation on each node is a linear model when $\alpha_j = 0$, and is a special case of the
 351 extended partially linear single index model with zero product moment between the linear and
 352 non-linear part when $\alpha_j \neq 0$. Hence, when $\alpha_j = 0$, we can expect the same rate of convergence
 353 as in the usual linear regression model. However, when $\alpha_j \neq 0$, the rate of convergence is
 354 expected to be slower due to the presence of the nonlinear part. Theorem 4 establishes the
 355 unbiasedness and the rate of convergence for the least squares estimator in both cases.

356 **Theorem 4** Assume Condition 2 is satisfied. Then

357 1. $\hat{\beta}_j$ is a (marginally) unbiased estimator of β_j .

358 2. $\hat{\beta}_j$ satisfies

$$359 \quad E \left(\|\hat{\beta}_j - \beta_j\|_2^2 \right) = O \left(\frac{n(p-1)^3}{(n-p)^2} + \frac{p-1}{n-p} \right), \quad \text{when } \alpha_j \neq 0,$$

360 and

$$361 \quad E \left(\|\hat{\beta}_j - \beta_j\|_2^2 \right) = O \left(\frac{p-1}{n-p} \right), \quad \text{when } \alpha_j = 0.$$

362 When $\alpha_j = 0$, the true conditional expectation is a linear model, so the least square estimator
 363 $\hat{\beta}_j$ is conditionally unbiased for β_j at any \mathbf{Z}_{-j} , and hence marginally unbiased. However, when
 364 $\alpha_j \neq 0$, since the true conditional expectation contains another non-linear part, the estimator
 365 $\hat{\beta}_j$ is not conditionally unbiased at any \mathbf{Z}_{-j} , but it is marginally unbiased because \mathbf{Z}_{-j} follows a
 366 GMSN distribution. When p is fixed, the rate of convergence is the same as in the linear model
 367 $O(n^{-1})$. However, when p is growing, the non-linear part slows down the rate of convergence.
 368 In fact, estimation consistency is only achieved when $p^3/n \rightarrow 0$, compared to the condition
 369 $p = o(n)$ in the linear model.

370 We next establish the rate of convergence for $\hat{\tau}_j$, the estimated scalar from the one-component
 371 projection pursuit regression model in (3).

372 **Theorem 5** Assume $p^4/n \rightarrow 0$ and Condition 3 is satisfied. Let τ_j denote the true standard deviation
 373 of the function $g_j(\mathbf{z}_{-j}^{(j)}; \zeta_j)$. Then for all $j = 1, \dots, p$, there exists a positive constant C such that

$$374 \quad \lim_{n \rightarrow \infty} P \left(|\hat{\tau}_j - \tau_j| \geq C n^{-2/5} \right) = 0.$$

375 From Theorem 5, we see that the rate of convergence reflects the fact that we estimate the
 376 function g_j nonparametrically without any additional assumption on g_j . Requiring $p^4/n \rightarrow 0$
 377 ensures the conditional expectation of the residual $E(\mathbf{r}_j | \mathbf{Z}_{-j})$ obtained from the linear model
 378 converges to $g_j(\mathbf{Z}_{-j}; \zeta_j)$ in probability, which is needed for the asymptotic results of the projec-
 379 tion pursuit regression to hold (Chen, 1991).

380 4.2 | Graphical model estimation

381 With the estimates for each node, we now establish the following theorem on consistency of
 382 the estimated graphical model.

383 **Theorem 6** Assume $p^5/n \rightarrow 0$ and Conditions 1-4 are satisfied. For any positive sequence δ such
 384 that $\delta \rightarrow 0$, $n^{-1}\delta^{-2}s_\alpha p^3 \rightarrow 0$ and $n^{-1}\delta^2 p^2 \rightarrow 0$, let $\hat{T} = \delta^2 + C_4 p^4 n^{-4/5}$, where C_4 is a positive
 385 constant. Then

$$386 \quad P(\sigma_{jk} > \hat{T}) \rightarrow \begin{cases} 1, & \text{for } (j, k) \in E \\ 0, & \text{for } (j, k) \notin E. \end{cases}$$

387 As a result, letting $\widehat{\mathcal{N}}_j = \{k : \sigma_{jk} > \hat{T}\}$ be the estimated neighborhood for the node j in the graph, we
 388 have $P(\widehat{\mathcal{N}}_j = \mathcal{N}_j) \rightarrow 1$ as $n \rightarrow \infty$.

389 The conditions for Theorem 6 ensure that we have consistency for all p nodewise regres-
 390 sions. Specifically, we require $p^4/n \rightarrow 0$ for the convergence of both $\hat{\beta}_j$ and $\hat{\tau}_j$ to the true pa-
 391 rameter β_j and τ_j respectively, so accounting for all the p regressions requires $p^5/n \rightarrow 0$.

392 5 | ESTIMATION OF LOCATION AND SCALE PARAMETERS

393 In this section, we address the issue of estimating the location ξ and the diagonal scale matrix
 394 $\mathbf{D} = \text{diag}(\sigma_1, \dots, \sigma_p)$. A direct application of Huang et al. (2013, Theorem 2.1) states that the
 395 location and scale parameters of the GMSN distribution are preserved marginally.

396 **Proposition 7 (Huang et al., 2013)** Let $\mathbf{Y} = (Y_1, \dots, Y_p)^\top \sim \text{GMSN}(\xi, \Sigma, \alpha, G)$ with joint density
 397 given in (1). Then each component Y_j marginally follows a generalized univariate skew normal distri-
 398 bution with location ξ_j and scale parameter σ_j for $j = 1, \dots, p$ (potentially with a skewing function
 399 different from G and a shape parameter different from α_j).

400 Proposition 7 implies that we can estimate each pair (ξ_j, σ_j) separately from Algorithm 1
 401 by applying the methods developed for estimating the location and scale parameter of univari-
 402 ate generalized skew normal distributions to each marginal data \mathbf{y}_j . Several such methods are
 403 available in the literature; see the Supporting Information for a brief review. In this article, we
 404 focus on using the following generalized method of moment (GMM) to estimate (ξ_j, σ_j) .

405 Recall that $Z_j = (Y_j - \xi_j)/\sigma_j$, so that Z_j follows a univariate generalized skew normal distri-
 406 bution with location 0 and scale parameter 1. By the invariance property of skew-symmetric
 407 distributions (Wang et al., 2004, Proposition 6), for any even function $T : \mathbb{R} \rightarrow \mathbb{R}$, that is
 408 $T(x) = T(-x)$, we have $T(Z_j) \stackrel{d}{=} T(\tilde{U})$, where \tilde{U} is the standard normal random variable and
 409 $\stackrel{d}{=}$ denotes equality in distribution. Specifically, we have $|Z_j| \sim \chi_1$, where χ_1 denotes the chi-
 410 distribution with one degree of freedom (Azzalini et al., 2010, Section 2.1). Therefore, for $k > 0$,

411 we obtain

$$412 \quad E \{ |Z_j|^k \} = \frac{2^{k/2} \Gamma(k/2 + 1/2)}{\Gamma(1/2)} \triangleq c_k, \quad (5)$$

413 where $\Gamma(\cdot)$ denotes the gamma function. Using (5) with $k = 2$ and replacing the left hand side by
 414 its sample counterpart, we obtain $\hat{\sigma}_j^2 = n^{-1} \sum_{i=1}^n (y_{ij} - \hat{\xi}_j)^2$. Next, let $K \geq 3$ be a positive integer,
 415 and define $v_{jk} = n^{-1} \sum_{i=1}^n |y_{ij} - \xi_j|/\sigma_j - c_k$, so each $v_{jk} = 0$ is an unbiased estimating equation
 416 for (ξ_j, σ_j) . Denoting $\nu_j = (v_{j1}, \dots, v_{jK})^\top$, then the GMM estimator for (ξ_j, σ_j) is defined as

$$417 \quad (\hat{\xi}_j, \hat{\sigma}_j) = \underset{\xi_j, \sigma_j}{\operatorname{argmin}} \nu_j^\top \mathbf{W}^{-1} \nu_j, \text{ subject to } \sigma_j^2 = \frac{1}{n} \sum_{i=1}^n (y_{ij} - \xi_j)^2 \quad (6)$$

418 where \mathbf{W} is a $K \times K$ covariance matrix with elements $w_{km} = \operatorname{Cov}(v_{jk}, v_{jm}) = c_{k+m} - c_k c_m$. Sub-
 419 stituting the constraint to the objective function $\nu_j^\top \mathbf{W}^{-1} \nu_j$ makes it a univariate function of ξ_j .
 420 This GMM estimator is similar to the invariance-based estimating equation (IBEE) method pro-
 421 posed by Azzalini et al. (2010), which only uses the first two absolute moments of Z_j ($k \in \{1, 2\}$),
 422 and the minimum distance characteristic function (MDCF) method of Potgieter and Genton
 423 (2013), which essentially uses all the even moments of Z_j . Finally, we note that these moment-
 424 based estimators are analogous to the regular asymptotically linear estimator and semipara-
 425 metric efficient estimator of Ma et al. (2005).

426 5.1 | Practical issues with current methods

427 The GMM and other methods used to estimate the location and scale parameters possess two
 428 technical challenges. We illustrate the issues in the case \mathbf{Y} follows a multivariate skew normal
 429 distribution with the joint density $f(\mathbf{Y}) = 2\phi_\rho(\mathbf{Y}; \xi, \Sigma) \Phi\{\alpha^\top \mathbf{D}^{-1}(\mathbf{Y} - \xi)\}$, and expect that the
 430 issues will also arise for a more general skewing function G .

431 First, when \mathbf{Y} follows a multivariate skew normal distribution, each component Y_j has a uni-
 432 variate skew normal distribution (Azzalini and Capitanio, 1999) with skewness given by the
 433 j th component of the vector $\delta = (1 + \alpha^\top \tilde{\Sigma} \alpha)^{-1/2} \tilde{\Sigma} \alpha$, where $\tilde{\Sigma}$ is the correlation matrix corre-
 434 sponding to Σ . As demonstrated in Potgieter and Genton (2013), both the IBEE and MDCF
 435 estimators fail for any component that has skewness $\delta_j \rightarrow 0$, since the determinant of the
 436 covariance matrix of the estimators tends to infinity when $n \rightarrow \infty$. In other words, these meth-
 437 ods fail on any marginal component that is close to the normal distribution. Nevertheless, if
 438 a marginal component is not too far from a normal distribution, we expect the sample mean
 439 $\bar{y}_j = n^{-1} \sum_{i=1}^n y_{ij}$ and sample variance $s_j^2 = (n-1)^{-1} \sum_{i=1}^n (y_{ij} - \bar{y}_j)^2$ to be good estimators for ξ_j
 440 and σ_j^2 , respectively. In the context of graphical model estimation, since we impose some spar-

441 city assumptions on Ω and α , it is common for marginal distributions to exhibit a wide range of
 442 skewness, including many with very small skewness. Therefore, when estimating the location
 443 and scale parameter for each marginal component, it is important to compare the estimate ob-
 444 tained from the GMM methods with the sample mean and standard deviation (\bar{y}_j, s_j) .

445 A second challenge arises even when $\delta_j \rightarrow 0$ is that the criterion functions for the esti-
 446 mators usually have multiple local minima (Ma et al., 2005; Azzalini et al., 2010; Potgieter and
 447 Genton, 2013). This also holds true for our GMM approach defined in (6), and the global opti-
 448 mum is not always guaranteed to be the best estimate. Multiple criteria have been proposed
 449 for selecting the best local minimum, see the Supporting Information for details. Combining
 450 two challenges together implies that in practice, we need to select the final estimate for (ξ_j, σ_j)
 451 from possibly more than two candidates, including (\bar{y}_j, s_j) and the local minima from (6), which
 452 is generally a challenging problem.

453 6 | SIMULATION STUDY

454 6.1 | Graph recovery

455 We conduct several simulation studies to explore the performance of the proposed estima-
 456 tion method for estimating the graphical model of the GMSN distribution. Samples of sizes
 457 $n \in \{250, 1000\}$ are generated from a p -variate generalized skew normal distribution, with
 458 $p \in \{40, 60\}$. Two skewing functions are considered, including the standard normal distribu-
 459 tion function, and the Cauchy distribution function. The precision matrix $\Omega = \Sigma^{-1}$ is generated
 460 following either a “hub”, “scale-free” or a “random” structure using the `huge` package (Jiang
 461 et al., 2019), and the number of variables vary over $p \in \{40, 60\}$. Note that to estimate the
 462 graphical model, we have to estimate $p(p-1)/2$ parameters; i.e $p = 40$ corresponds to 780 and
 463 $p = 60$ corresponds to 1770 parameters, respectively. The shape vector α is structured such
 464 that its first $p - s_\alpha$ components are non-zero, and we consider three scenarios for s_α , namely
 465 $s_\alpha \in \{0.1p, 0.3p, 0.5p\}$. Note that the larger the value of s_α , the greater the deviation from
 466 the multivariate Gaussian distribution. The first $s_\alpha/2$ non-zero components of α are randomly
 467 generated from the $U(1, 3)$ distribution, while the last $s_\alpha/2$ non-zero components of α are ran-
 468 domly generated from the $U(10, 20)$ distribution, so the marginal components possesses differ-
 469 ent levels of skewness. The above structure of Ω and α means the true number of conditionally
 470 dependent pairs range from 4% to 33% of the total number of pairs. Finally, each location ξ_j and
 471 scale σ_j are independently generated from the uniform $U(1, 5)$ distribution. In each setting, 500
 472 datasets are generated.

473 We compare the ability of our proposed approach to recover the underlying graphical model
 474 against some common methods for recovering the Gaussian graphical model, including node-
 475 wise linear models, the Graphical Lasso (GLasso) approach of Friedman et al. (2008), and the

476 nodewise lasso approach of (Meinshausen et al., 2006). We also consider two other methods
 477 already developed in the literature for estimating non-Gaussian graphical models, including
 478 the nonparanormal SKEPTIC approach of Liu et al. (2012) and the SPACEJAM method (SJ) of
 479 Voorman et al. (2014). These are computed using the `huge` (Jiang et al., 2019) and `spacejam`
 480 (Voorman, 2013) packages in \mathbb{R} , respectively, both with default settings. Furthermore, we con-
 481 sider two versions of the newly proposed procedure. In the first version, we assume the true
 482 location ξ and scale matrix \mathbf{D} are known, and in the second, we use the estimated location $\hat{\xi}$
 483 and scale $\hat{\mathbf{D}}$. For the j th component, all the candidates for $\hat{\xi}_j$ include the sample mean \bar{y}_j and
 484 the local minima obtained from the GMM approach using $K = 4$; then, we select the candidate
 485 closest to the true ξ_j as the final estimate $\hat{\xi}_j$ and compute $\hat{\sigma}_j = n^{-1} \sum_{i=1}^n (y_{ij} - \hat{\xi}_j)^2$, $j = 1, \dots, p$.

486 Table 1 presents the mean area under the receiver operating characteristic curve (AUC) for
 487 all the settings when Ω has a hub or random structure, and G is the standard normal distribu-
 488 tion function. By using the AUC, we marginalize over all the possible thresholds or regulariza-
 489 tion parameters, so the AUC is able to give us a general picture of the ability of each method
 490 to distinguish conditionally independent from conditionally dependent pairs. A higher AUC
 491 implies a better performance, with $\text{AUC} = 1$ indicating a perfect separation of conditionally in-
 492 dependent from conditionally dependent pairs. The AUC for the other settings shows similar
 493 trends and can be found in the Supporting Information.

494 Table 1 demonstrates that all the estimation methods have worse performance when the
 495 shape vector α has more non-zero components, and that the SPACEJAM method has the worst
 496 performance among all the considered methods. At sample size $n = 250$ and $s_\alpha = 0.1p$, both
 497 versions of the new estimation method have similar performance to the nodewise linear and
 498 nodewise lasso method; this may be attributed to the slow rate of convergence of the projec-
 499 tion pursuit regression. Nevertheless, when the sample size increases to $n = 1000$, the new
 500 estimation method generally outperforms all the other methods. The gain is more remark-
 501 able when the proportion of non-zero components in the shape vector increases, that is for
 502 $s_\alpha \in \{0.3p, 0.5p\}$. When the precision matrix Ω has a hub structure, there is little difference
 503 between the AUC of the new estimation method using true location and scale parameter and
 504 that using the estimates of these parameters. However, when Ω has a random and scale-free
 505 structure (the latter is shown in the Supporting Information), using the true location and scale
 506 gives a substantially higher AUC than using the estimated values.

507 6.2 | Choice of threshold

508 We next evaluate the performance of the path connectivity and the AGNES risk function for
 509 selecting the threshold in the new estimation procedure, with estimated location $\hat{\xi}$ and scale
 510 matrix $\hat{\mathbf{D}}$. The performance metrics include the median true positive rate (TPR) and median

TABLE 1 Mean area under the curve (AUC) of different methods, including nodewise linear (NLinear), graphical Lasso (GLasso), nodewise lasso (NLasso), SKEPTIC, spacejam (SJ), the new method with true location and scale (ξ, \mathbf{D}), and the new method with estimated location and scale ($\hat{\xi}, \hat{\mathbf{D}}$) for recovering the graphical model from the multivariate skew normal distribution where the true precision matrix has a hub and random structure. The best AUC in each setting is highlighted.

n	Ω	p	s_α/p	NLinear	GLasso	NLasso	SKEPTIC	SJ	New method ξ, \mathbf{D}	$\hat{\xi}, \hat{\mathbf{D}}$
250	hub	40	0.1	0.91	0.82	0.90	0.82	0.32	0.94	0.94
			0.3	0.67	0.62	0.71	0.63	0.21	0.70	0.70
			0.5	0.59	0.58	0.66	0.60	0.16	0.61	0.62
		60	0.1	0.87	0.81	0.88	0.82	0.27	0.90	0.90
			0.3	0.63	0.64	0.70	0.65	0.19	0.64	0.65
			0.5	0.56	0.58	0.57	0.58	0.10	0.57	0.60
	random	40	0.1	0.96	0.83	0.92	0.84	0.39	0.96	0.95
			0.3	0.77	0.67	0.73	0.67	0.25	0.80	0.78
			0.5	0.65	0.57	0.61	0.57	0.17	0.72	0.67
		60	0.1	0.94	0.85	0.91	0.85	0.47	0.93	0.93
			0.3	0.71	0.64	0.68	0.64	0.28	0.75	0.72
			0.5	0.61	0.55	0.57	0.55	0.21	0.67	0.62
1000	hub	40	0.1	0.94	0.77	0.86	0.77	0.37	0.98	0.98
			0.3	0.69	0.57	0.66	0.59	0.23	0.80	0.81
			0.5	0.59	0.52	0.60	0.55	0.16	0.70	0.72
		60	0.1	0.90	0.76	0.83	0.77	0.32	0.95	0.95
			0.3	0.65	0.60	0.65	0.63	0.21	0.74	0.75
			0.5	0.56	0.54	0.52	0.55	0.12	0.65	0.69
	random	40	0.1	0.98	0.80	0.91	0.81	0.38	0.99	0.98
			0.3	0.80	0.65	0.73	0.65	0.25	0.91	0.84
			0.5	0.69	0.55	0.61	0.56	0.17	0.82	0.73
		60	0.1	0.95	0.82	0.90	0.83	0.50	0.98	0.96
			0.3	0.75	0.63	0.68	0.63	0.32	0.86	0.78
			0.5	0.65	0.53	0.57	0.54	0.23	0.79	0.69

511 false positive rate (FPR) across the 500 samples. Table 2 below demonstrates the performance
 512 when G is the standard normal distribution function and Ω has a scale-free structure. The re-
 513 sults for other settings are similar and reported in the Supporting Information.

TABLE 2 Performance of the path connectivity and AGNES risk function in selecting the threshold for estimating the graphical model of the multivariate skew normal distribution based on the median true positive rate (TPR) and false positive rate (FPR) when Ω has a random structure. Interquartile ranges are included in parentheses.

p	s_α/p	Metric	$n = 250$		$n = 1000$	
			PC	AGNES	PC	AGNES
40	0.1	TPR	0.90	0.36	0.96	0.41
		FPR	0.22	0.07	0.16	0.00
	0.3	TPR	0.64	0.24	0.71	0.23
		FPR	0.23	0.08	0.18	0.00
	0.5	TPR	0.46	0.17	0.54	0.12
		FPR	0.22	0.08	0.20	0.00
60	0.1	TPR	0.82	0.41	0.93	0.35
		FPR	0.20	0.11	0.19	0.01
	0.3	TPR	0.50	0.26	0.63	0.18
		FPR	0.20	0.11	0.20	0.01
	0.5	TPR	0.36	0.19	0.48	0.10
		FPR	0.22	0.12	0.22	0.01

514 Table 2 demonstrates that the performance of both risk functions improves when the sam-
 515 ple size increases from $n = 250$ to $n = 1000$ in all the considered settings. At $n = 250$, the path
 516 connectivity risk function leads to overfitting with both high true positive rate and false posi-
 517 tive rates. In contrast, the AGNES risk function leads to underfitting with a true positive rate
 518 smaller than one. At $n = 1000$, except in the case of $n = 250$ and $s_\alpha/p = 0.1$, the AGNES risk
 519 function performs better than the path connectivity risk function; while both criteria lead to
 520 true positive rate one (i.e identify all the true conditionally dependent pairs), the AGNES risk
 521 function has lower false positive rates, particularly when $s_\alpha = 0.3p$ and $s_\alpha = 0.5p$.

522 7 | REAL DATA ANALYSIS

523 As an illustration of the new approach for estimating graphical models, we apply the new method-
 524 ology to a dataset consisting of $p = 11$ physicochemical properties of 1599 specimens of the

525 red variant of the Portuguese “Vinho Verde” wine. The data are presented and examined by
526 Cortez et al. (2009) and are publicly available in the UCI Machine Learning Repository. We
527 take the log of the highly skewed variables in the original data, including sulphates amount, to-
528 tal sulfur dioxide, free sulfur dioxide, residual sugar, and chlorides. After transformation, we
529 compute the robust Mahalanobis distance of each observation using the `robustbase` pack-
530 age (Todorov and Filzmoser, 2009), and remove observations with Mahalanobis distances in
531 the top 10%. After pre-processing, we have $n = 1359$ observations. Both the Mardia multivari-
532 ate Gaussian goodness-of-fit tests targeting skewness and kurtosis reject multivariate normal-
533 ity (p -value ≈ 0). Furthermore, all the marginal components have absolute sample skewness
534 ranging from 0.05 (for density) to 0.92 (for fixed acidity), and absolute sample kurtosis ranging
535 from 0.06 (for volatile acidity) to 1.05 (for chlorides).

536 As a baseline, we first treat the data as coming from the multivariate Gaussian distribution
537 and apply the nodewise linear approach to estimate its graphical model. To compare against
538 our more general framework, we then model these $p = 11$ chemical components using the
539 GMSN distribution and estimate the associated graphical model. Denoting the data matrix as
540 \mathbf{Y} , each row $\mathbf{y}^{(i)} \sim \text{GMSN}(\boldsymbol{\xi}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, G)$, $i = 1, \dots, n$. Unlike the simulation study, we do not know
541 the true location in the real data, so we use the following approach to estimate the location
542 and scale parameters. We first conduct the Shapiro-Wilk normality test at significance level
543 $0.05/11$ (to account for multiple testing) for all the 11 marginal components; it turns out that
544 these tests rejects normality for all except for density and pH . Next, we compute the GMM
545 estimator from Section 5 with $K = 4$ for 9 non-normal marginal components as indicated by
546 the Shapiro-Wilk tests. For any component when multiple local minima of GMM appear, we
547 use the model complexity criterion of Azzalini et al. (2010) to select the final estimate. For the
548 two remaining components, i.e density and pH , we use the sample mean and standard devia-
549 tion as the estimates for the location and scale parameters, respectively. Next, we form the
550 standardized data $\hat{\mathbf{Z}}$ with row $\hat{\mathbf{z}}^{(i)} = \hat{\mathbf{D}}^{-1} (\mathbf{y}^{(i)} - \hat{\boldsymbol{\xi}})$, $i = 1, \dots, n$ and applied the newly proposed
551 method as outlined in Algorithm 1 to estimate the graphical model on the $\hat{\mathbf{Z}}$ data. For both the
552 nodewise linear and new estimation methods, we choose the thresholds based on the AGNES
553 risk function because the simulation study demonstrates that it has a lower false positive rate
554 when n is large.

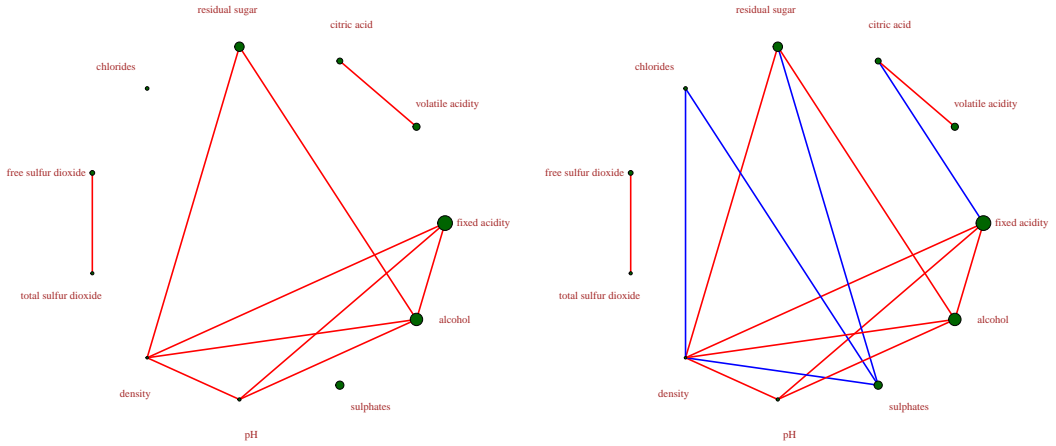


FIGURE 2 Estimated graphical model for the wine dataset based on the linear nodewise (left) and the newly proposed methods (right). Red edges are common edges in the two estimated graphs. Blue edges are the edges that only found in the estimated graph from the newly proposed methods. The size of vertices in the graph is proportional to the absolute value of the sample marginal skewness.

555 Figure 2 shows the estimated graphical models from the nodewise linear and new methods.
 556 Out of 55 possible edges, the linear nodewise method found 10 edges, while the new method
 557 found 15 edges. Noticeably, all the edges found by the linear nodewise method are also found
 558 in the graphical model by the new method, but the new method found additional conditional
 559 dependence pairs on the nodes that have large absolute skewness, such as (residual sugar, sul-
 560 phates). Therefore, for this data application, the new method may be useful in identifying pos-
 561 sible conditional dependence structures beyond those of the Gaussian graphical model.

562 8 | CONCLUDING REMARKS

563 This paper presents steps towards developing methods for estimating graphical models for
 564 continuous, non-normally distributed data that cannot be transformed back to normality. Al-
 565 though it is desirable to study the graphical model for a more general class of multivariate skew
 566 symmetric distribution, it is important to highlight that conditional independence is not always
 567 possible. For example, Baba et al. (2004) showed that conditional independence cannot occur
 568 in the class of elliptical distribution outside of the Gaussian distribution. If we consider the gen-
 569 eralized multivariate skew-elliptical distribution of Genton and Loperfido (2005), whose den-
 570 sity has the form $f(\mathbf{y}^*) = 2f_0(\mathbf{y}^*; \xi, \Sigma)G\{\alpha^\top \mathbf{D}^{-1}(\mathbf{y}^* - \xi)\}$, where f_0 is the density of a p -variate

571 elliptical distribution, conditional independence can only occur in this family when the base
572 distribution f_0 is the density of multivariate Gaussian distribution (i.e, $f_0 = \phi_p$). In this case, we
573 come back to the density of the generalized multivariate skew normal distribution (1) studied
574 in the paper. Future research needs to **address the challenge of estimating location and scale**
575 **parameters, especially for the components with medium skewness**, and make the proposed
576 nodewise method scalable in high dimensional settings. While it is intuitive to replace ordi-
577 nary least squares by a penalized regression estimator, choosing appropriate tuning parame-
578 ter is challenging because the true conditional expectation contains an additional non-linear
579 part. Moreover, fitting a projection pursuit regression in high dimension is computationally ex-
580 pensive, which would require new development in terms of both methodology and software
581 implementation.

582 ACKNOWLEDGEMENT

583 We thank the Associate Editor and two anonymous Reviewers for providing invaluable com-
584 ments that helped us substantially improve the content and presentation of the paper. We
585 thank Cornelis Potgieter for useful discussions as well as for providing us with the relevant
586 code. We thank Peter Radchenko for providing us the code that implements a method to es-
587 timate the high-dimensional single-index model. All authors are supported by the Australian
588 Research Council Discovery Grant DRC180100836.

589 REFERENCES

- 590 Adelchi Azzalini. *The skew-normal and related families*, volume 3. Cambridge University Press, 2013.
- 591 Adelchi Azzalini and Antonella Capitanio. Statistical applications of the multivariate skew normal distribution. *Journal of the*
592 *Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):579–602, 1999.
- 593 Adelchi Azzalini and A Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.
- 594 Adelchi Azzalini, Marc G Genton, and Bruno Scarpa. Invariance-based estimating equations for skew-symmetric distributions.
595 *Metron*, 68(3):275–298, 2010.
- 596 Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional
597 independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.
- 598 Eugene Belilovsky, Gaël Varoquaux, and Matthew B Blaschko. Testing for differences in gaussian graphical models: applica-
599 tions to brain connectivity. In *Advances in Neural Information Processing Systems*, pages 595–603, 2016.
- 600 A Capitanio, A Azzalini, and Elena Stanghellini. Graphical models for skew-normal variates. *Scandinavian Journal of Statistics*,
601 30(1):129–144, 2003.
- 602 Hung Chen. Estimation of a projection-pursuit type regression model. *The Annals of Statistics*, 19(1):142–157, 1991.

- 603 Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining
604 from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.
- 605 Arthur P Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.
- 606 Mathias Drton and Michael D Perlman. Model selection for gaussian concentration graphs. *Biometrika*, 91(3):591–602, 2004.
- 607 Bernd Fellinghauer, Peter Bühlmann, Martin Ryffel, Michael Von Rhein, and Jan D Reinhardt. Stable graphical model esti-
608 mation with random forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis*, 64:
609 132–152, 2013.
- 610 Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Bio-*
611 *statistics*, 9(3):432–441, 2008.
- 612 Jerome H Friedman. Smart user’s guide. Technical report, Stanford University, Lab for Computational Statistics, California,
613 USA, 1984.
- 614 Jerome H Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):
615 817–823, 1981.
- 616 Jerome H Friedman, Eric Grosse, and Werner Stuetzle. Multidimensional additive spline approximation. *SIAM Journal on*
617 *Scientific and Statistical Computing*, 4(2):291–301, 1983.
- 618 Marc G Genton and Nicola MR Loperfido. Generalized skew-elliptical distributions and their quadratic forms. *Annals of the*
619 *Institute of Statistical Mathematics*, 57(2):389–401, 2005.
- 620 Surya K Ghosh, Andrey G Cherstvy, Denis S Grebenkov, and Ralf Metzler. Anomalous, non-gaussian tracer diffusion in crowded
621 two-dimensional environments. *New Journal of Physics*, 18(1):013027, 2016.
- 622 Wen-Jang Huang, Nan-Cheng Su, and Arjun K Gupta. A study of generalized skew-normal distribution. *Statistics*, 47(5):942–
623 953, 2013.
- 624 Mohsen Jamali and Martin Ester. A matrix factorization technique with trust propagation for recommendation in social net-
625 works. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 135–142, 2010.
- 626 Haoming Jiang, Xinyu Fei, Han Liu, Kathryn Roeder, John Lafferty, Larry Wasserman, Xingguo Li, and Tuo Zhao. *huge: High-*
627 *Dimensional Undirected Graph Estimation*, 2019. URL <https://CRAN.R-project.org/package=huge>. R package
628 version 1.3.4.
- 629 Steffen L Lauritzen. *Graphical Models*, volume 17. Clarendon Press, 1996.
- 630 Lina Lin, Mathias Drton, and Ali Shojaie. Estimation of high-dimensional graphical models using regularized score matching.
631 *Electronic journal of statistics*, 10(1):806, 2016.
- 632 Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected
633 graphs. *Journal of Machine Learning Research*, 10(Oct):2295–2328, 2009.
- 634 Han Liu, Fang Han, Ming Yuan, John Lafferty, Larry Wasserman, et al. High-dimensional semiparametric gaussian copula graph-
635 ical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.
- 636 Yanyuan Ma, Marc G Genton, and Anastasios A Tsiatis. Locally efficient semiparametric estimators for generalized skew-
637 elliptical distributions. *Journal of the American Statistical Association*, 100(471):980–989, 2005.
- 638 Nicolai Meinshausen, Peter Bühlmann, et al. High-dimensional graphs and variable selection with the lasso. *The Annals of*
639 *Statistics*, 34(3):1436–1462, 2006.

- 640 Adria Caballe Mestres, Natalia Bochkina, and Claus Mayer. Selection of the regularization parameter in graphical models using
641 network characteristics. *Journal of Computational and Graphical Statistics*, 27(2):323–333, 2018.
- 642 Ricardo Pio Monti, Christoforos Anagnostopoulos, Giovanni Montana, et al. Learning population and subject-specific brain
643 connectivity networks via mixed neighborhood selection. *The Annals of Applied Statistics*, 11(4):2142–2164, 2017.
- 644 Rebecca Morrison, Ricardo Baptista, and Youssef Marzouk. Beyond normality: Learning sparse probabilistic graphical models
645 in the non-Gaussian setting. In *Advances in Neural Information Processing Systems*, pages 2359–2369, 2017.
- 646 Cornelis J Potgieter and Marc G Genton. Characteristic function-based semiparametric inference for skew-symmetric models.
647 *Scandinavian Journal of Statistics*, 40(3):471–490, 2013.
- 648 Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications
649 for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.
- 650 Aritra Sengupta, Noel Cressie, Brian H Kahn, and Richard Frey. Predictive inference for big, spatial, non-gaussian data: Modis
651 cloud data and its change-of-support. *Australian & New Zealand Journal of Statistics*, 58(1):15–45, 2016.
- 652 Valentin Todorov and Peter Filzmoser. An object-oriented framework for robust multivariate analysis. *Journal of Statistical
653 Software*, 32(3):1–47, 2009. URL <https://www.jstatsoft.org/article/view/v032i03/>.
- 654 Dietrich von Rosen. Moments for the inverted Wishart distribution. *Scandinavian Journal of Statistics*, 15:97–109, 1988.
- 655 Arend Voorman. *spacejam: Sparse conditional graph estimation with joint additive models.*, 2013. URL [https://CRAN.R-](https://CRAN.R-project.org/package=spacejam)
656 [project.org/package=spacejam](https://CRAN.R-project.org/package=spacejam). R package version 1.1.
- 657 Arend Voorman, Ali Shojaie, and Daniela Witten. Graph estimation with joint additive models. *Biometrika*, 101(1):85–101,
658 2014.
- 659 Jiuzhou Wang, Joseph Boyer, and Marc G Genton. A skew-symmetric representation of multivariate distributions. *Statistica
660 Sinica*, pages 1259–1270, 2004.
- 661 Adriano V Werhli, Marco Grzegorzczak, and Dirk Husmeier. Comparative evaluation of reverse engineering gene regulatory
662 networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*, 22(20):2523–2531,
663 2006.
- 664 Yingcun Xia, Howell Tong, and Wai Keung Li. On extended partially linear single-index models. *Biometrika*, 86(4):831–842,
665 1999.
- 666 Lingzhou Xue, Hui Zou, et al. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The
667 Annals of Statistics*, 40(5):2541–2571, 2012.
- 668 Guan Yu and Yufeng Liu. Sparse regression incorporating graphical structure among predictors. *Journal of the American Statis-
669 tical Association*, 111(514):707–720, 2016.
- 670 Shiqing Yu, Mathias Drton, and Ali Shojaie. Graphical models for non-negative data using generalized score matching. *arXiv
671 preprint arXiv:1802.06340*, 2018.
- 672 Xiaotong Yuan, Ping Li, Tong Zhang, Qingshan Liu, and Guangcan Liu. Learning additive exponential family graphical models
673 via $\ell_{2,1}$ -norm regularized m-estimation. In *Advances in Neural Information Processing Systems*, pages 4367–4375, 2016.
- 674 Hamid Zareifard, Håvard Rue, Majid Jafari Khaledi, and Finn Lindgren. A skew gaussian decomposable graphical model. *Journal
675 of Multivariate Analysis*, 145:58–72, 2016.
- 676 Rui Zhuang, Noah Simon, and Johannes Lederer. Graphical models for discrete and continuous data. *arXiv preprint
677 arXiv:1609.05551*, 2016.

9 | APPENDIX: PROOFS

9.1 | Proof of Theorem 1

Let \mathcal{H} denote the generalized precision matrix \mathcal{H} with element $h_{jk} = \partial^2 \log(f) / (\partial Y_j \partial Y_k)$. Morrison et al. (2017) showed that Y_j and Y_k are conditionally independent if and only if $h_{jk} = 0$ at all \mathbf{Y} . Using this result, we have $\log f(\mathbf{Y}) = \log \phi_p(\mathbf{Y}; \boldsymbol{\xi}, \boldsymbol{\Sigma}) + \log \{G(\boldsymbol{\alpha}^\top \mathbf{D}^{-1}(\mathbf{Y} - \boldsymbol{\xi}))\} + C$, where C is a constant that does not depend on \mathbf{Y} . The first term is the log likelihood of the multivariate normal distribution, so

$$\frac{\partial^2}{\partial Y_j \partial Y_k} \phi_p(\mathbf{Y}^*; \boldsymbol{\xi}, \boldsymbol{\Sigma}) = \omega_{jk}.$$

For the second term, we have

$$\frac{\partial}{\partial Y_j} \log G \{ \boldsymbol{\alpha}^\top \mathbf{D}^{-1}(\mathbf{Y} - \boldsymbol{\xi}) \} = \frac{G' \{ \boldsymbol{\alpha}^\top \mathbf{D}^{-1}(\mathbf{Y} - \boldsymbol{\xi}) \}}{G \{ \boldsymbol{\alpha}^\top \mathbf{D}^{-1}(\mathbf{Y} - \boldsymbol{\xi}) \}} \sigma_j^{-1} \alpha_j,$$

and

$$\frac{\partial^2}{\partial Y_j \partial Y_k} \log G \{ \boldsymbol{\alpha}^\top \mathbf{D}^{-1}(\mathbf{Y} - \boldsymbol{\xi}) \} = \left\{ \frac{G'' \{ \boldsymbol{\alpha}^\top \mathbf{D}^{-1}(\mathbf{Y} - \boldsymbol{\xi}) \}}{G \{ \boldsymbol{\alpha}^\top \mathbf{D}^{-1}(\mathbf{Y} - \boldsymbol{\xi}) \}} - \left[\frac{G' \{ \boldsymbol{\alpha}^\top \mathbf{D}^{-1}(\mathbf{Y} - \boldsymbol{\xi}) \}}{G \{ \boldsymbol{\alpha}^\top \mathbf{D}^{-1}(\mathbf{Y} - \boldsymbol{\xi}) \}} \right]^2 \right\} \sigma_j^{-1} \sigma_k^{-1} \alpha_j \alpha_k.$$

Since $\sigma_j \neq 0$ for all $j = 1, \dots, p$, the quantity $h_{jk} = 0$ for all \mathbf{Y} if and only if $\Omega_{jk} = 0$ and $\alpha_j \alpha_k = 0$, as claimed by the theorem.

9.2 | Proof of Theorem 2

Throughout the proof, we will use $\phi(x; \mu, \sigma^2)$ to denote the pdf function of the univariate normal random variable with mean μ and variance σ^2 , and $\phi(x)$ the pdf of the standard normal distribution. Let $\pi(Z_j | \mathbf{Z}_{-j}) = \phi(Z_j; \mathbf{Z}_{-j}^\top \boldsymbol{\beta}_j, \tilde{\sigma}_{jj, -j})$ be the conditional density function of Z_j given \mathbf{Z}_{-j} when $\mathbf{Z} \sim N_p(\mathbf{0}, \tilde{\boldsymbol{\Sigma}})$. Also, let $\pi(\mathbf{Z}_{-j}) = \phi(\mathbf{Z}_{-j}; \mathbf{0}, \tilde{\boldsymbol{\Sigma}}_{-j}^{(-j)})$ denote the marginal distribution of \mathbf{Z}_{-j}

697 in this case, First, we compute the marginal distribution of \mathbf{Z}_{-j} as

$$\begin{aligned}
 698 \quad f(\mathbf{Z}_{-j}) &= 2 \int \phi_p(\mathbf{z}^*; \mathbf{0}, \Sigma) G(\alpha_j z_j + \mathbf{Z}_{-j}^\top \alpha_{-j}) dz_j \\
 699 \quad &= 2\pi(\mathbf{Z}_{-j}) \int \pi(Z_j | \mathbf{Z}_{-j}) G(\alpha_j z_j + \mathbf{Z}_{-j}^\top \alpha_{-j}) dz_j \\
 700 \quad &= 2\pi(\mathbf{Z}_{-j}) \int (\tilde{\sigma}_{jj,-j}^{-1/2}) \phi\left(\frac{z_j - \mathbf{Z}_{-j}^\top \beta_j}{\sqrt{\tilde{\sigma}_{jj,-j}}}\right) G(\alpha_j z_j + \mathbf{Z}_{-j}^\top \alpha_{-j}) dz_j. \\
 701
 \end{aligned}$$

702 Now, we apply the change of variable to $z = \tilde{\sigma}_{jj,-j}^{-1/2} (z_j - \mathbf{Z}_{-j}^\top \beta_j)$ to obtain

$$703 \quad f(\mathbf{Z}_{-j}) = 2\pi(\mathbf{Z}_{-j}) \int \phi(z) G\{z \alpha_j \tilde{\sigma}_{jj,-j}^{-1/2} + \mathbf{Z}_{-j}^\top (\alpha_j \beta_j + \alpha_{-j})\} dz \quad (7)$$

$$704 \quad = 2\pi(\mathbf{Z}_{-j}) E\left[G\left\{\alpha_j \tilde{\sigma}_{jj,-j}^{-1/2} \tilde{Z} + \mathbf{Z}_{-j}^\top (\alpha_j \beta_j + \alpha_{-j})\right\}\right], \quad \tilde{Z} \sim N(0, 1). \quad (8)$$

706 We now prove the three parts of the theorem separately.

707 **(a)** When $\alpha_j = 0$, we have $f(\mathbf{Z}_{-j}) = \pi(\mathbf{Z}_{-j}) G(\mathbf{Z}_{-j}^\top \alpha_{-j})$, so the conditional distribution of Z_j given
 708 \mathbf{Z}_{-j} is

$$709 \quad f(Z_j | \mathbf{Z}_{-j}) = \frac{f(\mathbf{Z})}{f(\mathbf{Z}_{-j})} = \frac{\phi_p(\mathbf{Z}; \mathbf{0}, \Sigma) G(\mathbf{Z}_{-j}^\top \alpha_{-j})}{\pi(\mathbf{Z}_{-j}) G(\mathbf{Z}_{-j}^\top \alpha_{-j})} = N(Z_j; \mu_j, \sigma_{jj,-j}),$$

710 with $\mu_j = \mathbf{Z}_{-j}^\top \beta_j$ and variance $\tilde{\sigma}_{jj,-j} = 1 - \tilde{\sigma}_j^{(-j)} \left(\tilde{\Sigma}_{-j}^{(-j)}\right)^{-1} \tilde{\sigma}_j^{(j)}$. This is similar to the case when
 711 \mathbf{Z} is multivariate Gaussian.

712 **(b)** Next, we compute the conditional expectation

$$\begin{aligned}
 713 \quad E(Z_j | \mathbf{Z}_{-j}) &= \int Z_j f(Z_j | \mathbf{Z}_{-j}^*) dZ_j = \frac{1}{f(\mathbf{Z}_{-j})} \int Z_j f(Z_j, \mathbf{Z}_{-j}) dZ_j \\
 714 \quad &= \frac{\pi(\mathbf{Z}_{-j}) \int Z_j \pi(z_j | \mathbf{Z}_{-j}) G(\alpha_j z_j + \mathbf{Z}_{-j}^\top \alpha_{-j}) dZ_j}{\pi(\mathbf{Z}_{-j}) E\left[G\left\{\alpha_j \tilde{\sigma}_{jj,-j}^{-1/2} \tilde{Z} + \mathbf{Z}_{-j}^\top (\alpha_j \beta_j + \alpha_{-j})\right\}\right]} \\
 715 \quad &= \frac{\int Z_j \pi(Z_j | \mathbf{Z}_{-j}) G(\alpha_j Z_j + \mathbf{Z}_{-j}^\top \alpha_{-j}) dZ_j}{E\left[G\left\{\alpha_j \tilde{\sigma}_{jj,-j}^{-1/2} \tilde{Z} + \mathbf{Z}_{-j}^\top (\alpha_j \beta_j + \alpha_{-j})\right\}\right]}. \quad (9) \\
 716
 \end{aligned}$$

717 Consider the numerator of the last expression. We have

$$\begin{aligned}
 718 \quad & \int Z_j \pi(Z_j | \mathbf{Z}_{-j}) G(\alpha_j Z_j + \mathbf{Z}_{-j}^\top \alpha_{-j}) dZ_j = \int Z_j \tilde{\sigma}_{jj \cdot -j}^{-1/2} \phi\left(\frac{Z_j - \mathbf{Z}_{-j}^\top \beta_j}{\sqrt{\tilde{\sigma}_{jj \cdot -j}}}\right) G(\alpha_j Z_j + \mathbf{Z}_{-j}^\top \alpha_{-j}) dZ_j \\
 719 \quad & = \int (z \tilde{\sigma}_{jj \cdot -j}^{1/2} + \mathbf{Z}_{-j}^\top \beta_j) \phi(z) G\left\{z \alpha_j \tilde{\sigma}_{jj \cdot -j}^{-1/2} + \mathbf{Z}_{-j}^\top (\alpha_j \beta_j + \alpha_{-j})\right\} dz \\
 720 \quad & = \mathbf{Z}_{-j}^\top \beta_j E\left\{G\left\{\tilde{Z} \alpha_j \tilde{\sigma}_{jj \cdot -j}^{-1/2} + \mathbf{Z}_{-j}^\top (\alpha_j \beta_j + \alpha_{-j})\right\}\right\} \\
 721 \quad & + \tilde{\sigma}_{jj \cdot -j}^{1/2} E\left[\tilde{Z} G\left\{\tilde{Z} \alpha_j \tilde{\sigma}_{jj \cdot -j}^{-1/2} + \mathbf{z}_{-j}^\top (\alpha_j \beta_j + \alpha_{-j})\right\}\right], \quad \tilde{Z} \sim N(0, 1). \\
 722
 \end{aligned}$$

723 Substituting back into (9), we obtain $E(Z_j | \mathbf{z}_{-j}) = \mathbf{Z}_{-j}^\top \beta_j + g_j(\mathbf{Z}_{-j}^\top \zeta_j)$ with

$$724 \quad g_j(\mathbf{Z}_{-j}^\top \zeta_j) = \tilde{\sigma}_{jj \cdot -j}^{1/2} \frac{E\left\{\tilde{Z} G\left(\mathbf{Z}_{-j}^\top \alpha_j \tilde{\sigma}_{jj \cdot -j}^{-1/2} + \mathbf{Z}_{-j}^\top \zeta_j\right)\right\}}{E\left\{G\left(\tilde{Z} \alpha_j \tilde{\sigma}_{jj \cdot -j}^{-1/2} + \mathbf{Z}_{-j}^\top \zeta_j\right)\right\}}. \quad (10)$$

725 and $\zeta_j = \alpha_j \beta_j + \alpha_{-j}$.

726 (c) We next prove that $E\left\{h(\mathbf{Z}_{-j}) g_j(\mathbf{Z}_{-j}^\top \zeta_j)\right\} = 0$ for any odd function h of \mathbf{Z}_{-j} . Indeed,

$$\begin{aligned}
 727 \quad & E\left\{h(\mathbf{Z}_{-j}) g_j(\mathbf{Z}_{-j}^\top \zeta_j)\right\} = \int_{\mathbb{R}^{p-1}} h(\mathbf{z}_{-j}^*) g_j(\mathbf{Z}_{-j}^\top \zeta_j) f(\mathbf{z}_{-j}) d\mathbf{z}_{-j} \\
 728 \quad & \propto \int_{\mathbb{R}^{p-1}} h(\mathbf{Z}_{-j}) \pi(\mathbf{Z}_{-j}) \tilde{h}(\mathbf{Z}_{-j}) d\mathbf{Z}_{-j}, \quad (11) \\
 729
 \end{aligned}$$

730 with $\tilde{h}(\mathbf{z}_{-j}) = E\left\{\tilde{Z} G\left(\tilde{Z} \alpha_j \tilde{\sigma}_{jj \cdot -j}^{-1/2} + \mathbf{Z}_{-j}^\top \zeta_j\right)\right\}$. Next we show that $\tilde{h}(\mathbf{Z}_{-j})$ is an even function of
 731 \mathbf{z}_{-j} for any skewing function G ; in other words $h(\mathbf{Z}_{-j}) = \tilde{h}(-\mathbf{Z}_{-j})$. Indeed,

$$\begin{aligned}
 732 \quad & \tilde{h}(\mathbf{Z}_{-j}) = \int_{-\infty}^{\infty} z \phi(z) \left\{G\left(z \alpha_j \tilde{\sigma}_{jj \cdot -j}^{-1/2} + \mathbf{Z}_{-j}^\top \zeta_j\right)\right\} dz \stackrel{(i)}{=} \int_{-\infty}^{\infty} z \phi(z) \left\{1 - G\left(-z \alpha_j \tilde{\sigma}_{jj \cdot -j}^{-1/2} - \mathbf{Z}_{-j}^\top \zeta_j\right)\right\} dz \\
 733 \quad & = E(Z) - \int_{-\infty}^{\infty} z \phi(z) G\left(-z \alpha_j \tilde{\sigma}_{jj \cdot -j}^{-1/2} - \mathbf{Z}_{-j}^\top \zeta_j\right) dz \stackrel{(ii)}{=} \int_{-\infty}^{\infty} u \phi(u) G\left(u \alpha_j \tilde{\sigma}_{jj \cdot -j}^{-1/2} - \mathbf{Z}_{-j}^\top \zeta_j\right) du \\
 734 \quad & = \tilde{h}(-\mathbf{Z}_{-j}), \\
 735
 \end{aligned}$$

736 where step (i) follows from $G(x) = 1 - G(-x)$ and step (ii) follows from $E(\tilde{Z}) = 0$ and the
 737 change of variable $u = -z$. Finally, because $\tilde{h}(\mathbf{z}_{-j})$ is an even function, then the integrand in
 738 (11) is an odd function, and the expectation $E\left\{h(\mathbf{Z}_{-j}) g_j(\mathbf{Z}_{-j}^\top \zeta_j)\right\} = 0$ as claimed.

739 **9.3 | Proof of Theorem 3**

740 Without loss of generality, we assume Σ is a correlation matrix, i.e $\Sigma = \tilde{\Sigma}$ and \mathbf{D} is the identity
 741 matrix of dimension $p \times p$.

742 (1) \iff (2): This is proved in Theorem 1.

743 For the other relationships, first note that because Ω is the inverse of Σ , then we have $\Omega_{-j}^{(j)} =$
 744 $-\tilde{\sigma}_{jj.-j}^{-1} \left(\tilde{\Sigma}_{-j}^{(-j)} \right)^{-1} \tilde{\sigma}_{-j}^{(j)}$, with $\sigma_{jj.-j}$ a scalar defined in Theorem 2. Note that $\tilde{\sigma}_{jj.-j} \neq 0$ in general.
 745 Hence $\beta_j \propto \Omega_{-j}^{(j)}$, so $\mathbf{D}_j^{(k)} = 0$ if and only if $\beta_j^{(k)} = 0$.

746 (2) \implies (3a), (3b), and (3c): Consider the situation when $\omega_{kj} = 0$ and $\alpha_j \alpha_k = 0$. By the above
 747 argument, we have $\beta_j^{(k)} = \beta_k^{(j)} = 0$. Hence, $\zeta_j^{(k)} \zeta_k^{(j)} \propto (\alpha_k + \alpha_j \beta_j^{(k)})(\alpha_j + \alpha_k \beta_k^{(j)}) = 0$. Without loss
 748 of generality, assume $\alpha_j = 0$. In this situation, $f_j = E(\tilde{Z}) = 0$, where $\tilde{Z} \sim N(0, 1)$.

749 (3a) and (3b) \implies (2): First, (3a) implies $\omega_{jk} = \omega_{kj} = 0$. Next, we have $\zeta_j^{(k)} \zeta_k^{(j)} \propto (\alpha_k +$
 750 $\alpha_j \beta_j^{(k)})(\alpha_j + \alpha_k \beta_k^{(j)}) = \alpha_j \alpha_k$, so $\zeta_j^{(k)} \zeta_k^{(j)} = 0$ implies $\alpha_j \alpha_k = 0$.

751 (3a) and (3c) \implies (2): Again, (3a) implies $\omega_{kj} = \omega_{jk} = 0$. Without loss of generality, assume
 752 the function $f_j = 0$ for all \mathbf{z}_{-j} . This implies

753
$$E \left\{ \tilde{Z} G \left(\tilde{Z} \alpha_j \sqrt{\tilde{\sigma}_{jj.-j}} \right) \right\} = 0. \tag{12}$$

754 By definition,

755
$$E \left\{ \tilde{Z} G \left(\tilde{Z} \alpha_j \sqrt{\tilde{\sigma}_{jj.-j}} \right) \right\} = (2\pi)^{-1/2} \int_{-\infty}^{\infty} z G \left(z \alpha_j \sqrt{\tilde{\sigma}_{jj.-j}} \right) e^{-z^2/2} dz$$

756 Writing $a = \alpha_j \sqrt{\tilde{\sigma}_{jj.-j}}$, integration by part gives

757
$$\int_{-\infty}^{\infty} z G(az) e^{-z^2/2} dz = -a e^{-z^2/2} G(az) \Big|_{-\infty}^{\infty} + a \int_{-\infty}^{\infty} e^{-z^2/2} G'(az) dz = a \int_{-\infty}^{\infty} e^{-z^2/2} G'(az) dz.$$

758 The last equality follows from the fact that $e^{-z^2/2} \rightarrow 0$ when $z \rightarrow \pm\infty$, and $G(z) \rightarrow 0$ when
 759 $z \rightarrow -\infty$ and $G(z) \rightarrow 1$ when $z \rightarrow \infty$. Also $G'(z)$ is the pdf of a symmetric random variable, so it
 760 is an even function. Together, (12) implies $\alpha_j = a = 0$, so $\alpha_j \alpha_k = 0$, as claimed.

9.4 | Proof of Theorem 4

1. Unbiasedness: By definition, we have

$$\begin{aligned}
 E \{ \hat{\beta}_j \} &= E \left\{ \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \mathbf{Z}_{-j}^\top \mathbf{z}_j \right\} = E \left[E \left\{ \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \mathbf{Z}_{-j}^\top \mathbf{z}_j \right\} \middle| \mathbf{Z}_{-j} \right] \\
 &= E \left[\left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \mathbf{Z}_{-j}^\top E(\mathbf{z}_j | \mathbf{Z}_{-j}) \right] = E \left[\left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \mathbf{Z}_{-j}^\top (\mathbf{Z}_{-j} \beta_j + \mathbf{g}_j(\mathbf{Z}_{-j} \zeta_j)) \right] \\
 &= \beta_j + E \left[\left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \mathbf{Z}_{-j}^\top \mathbf{g}_j(\mathbf{Z}_{-j} \zeta_j) \right] = \beta_j.
 \end{aligned}$$

The last equality follows from the fact that $E \left[\left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \mathbf{Z}_{-j}^\top \mathbf{g}_j(\mathbf{Z}_{-j} \zeta_j) \right] = 0$ due to part (c) of Theorem 2.

2. Error bound:

$$\begin{aligned}
 E \{ \|\hat{\beta}_j - \beta_j\|_2^2 \} &= E \left\{ \left\| \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \mathbf{Z}_{-j}^\top \mathbf{z}_j - \beta_j \right\|_2^2 \right\} \\
 &= E \left[\left\{ \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \mathbf{Z}_{-j}^\top \mathbf{z}_j - \beta_j \right\}^\top \left\{ \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \mathbf{Z}_{-j}^\top \mathbf{z}_j - \beta_j \right\} \right] \\
 &= E \left[\mathbf{z}_j^\top \mathbf{Z}_{-j} \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-2} \mathbf{Z}_{-j}^\top \mathbf{z}_j - 2 \mathbf{z}_j^\top \mathbf{Z}_{-j} \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \beta_j \right] + \beta_j^\top \beta_j. \quad (13)
 \end{aligned}$$

For the second term, we have

$$\begin{aligned}
 E \left\{ \mathbf{z}_j^\top \mathbf{Z}_{-j} \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \beta_j \right\} &= E \left[E(\mathbf{z}_j^\top | \mathbf{Z}_{-j}) \mathbf{Z}_{-j} \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \beta_j \right] \\
 &= E \left[\left\{ \beta_j^\top \mathbf{Z}_{-j}^\top + \mathbf{g}_j^\top(\mathbf{Z}_{-j} \zeta_j) \right\} \mathbf{Z}_{-j} \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \beta_j \right] \\
 &= \beta_j^\top \beta_j + E \left[\mathbf{g}_j^\top(\mathbf{Z}_{-j} \zeta_j) \mathbf{Z}_{-j} \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \beta_j \right] \\
 &= \beta_j^\top \beta_j, \quad (14)
 \end{aligned}$$

where the last inequality follows from part (c) of Theorem 2. The first term in (13) is more challenging. First, conditional on \mathbf{Z}_{-j} , we have

$$\begin{aligned}
 &E \left\{ \mathbf{z}_j^\top \mathbf{Z}_{-j} \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-2} \mathbf{Z}_{-j}^\top \mathbf{z}_j \middle| \mathbf{Z}_{-j} \right\} \\
 &= E(\mathbf{z}_j | \mathbf{Z}_{-j})^\top \mathbf{Z}_{-j} \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-2} \mathbf{Z}_{-j}^\top E(\mathbf{z}_j | \mathbf{Z}_{-j}) + \text{trace} \left\{ \mathbf{Z}_{-j} \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-2} \mathbf{Z}_{-j}^\top \text{Var}(\mathbf{z}_j | \mathbf{Z}_{-j}) \right\} \\
 &= \beta_j^\top \beta_j + 2 \mathbf{g}_j^\top(\mathbf{Z}_{-j} \zeta_j) \mathbf{Z}_{-j} \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-1} \beta_j + \mathbf{g}_j^\top(\mathbf{Z}_{-j} \zeta_j) \mathbf{Z}_{-j} \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-2} \mathbf{Z}_{-j}^\top \mathbf{g}_j(\mathbf{Z}_{-j} \zeta_j) \\
 &+ \text{trace} \left\{ \mathbf{Z}_{-j} \left(\mathbf{Z}_{-j}^\top \mathbf{Z}_{-j} \right)^{-2} \mathbf{Z}_{-j}^\top \text{Var}(\mathbf{z}_j | \mathbf{Z}_{-j}) \right\} \\
 &= \beta_j^\top \beta_j + I_1 + I_2 + I_3. \quad (15)
 \end{aligned}$$

785 We now split the proof into two cases.

786 **9.4.1 | When $\alpha_j \neq 0$**

787 In this case, the function g_j is not a zero function. To simplify the notation, we let $\mathbf{X} = \mathbf{Z}_{-j}$, so
 788 $\mathbf{x}^{(i)}$ and x_{ij} denote its i th row and (i, j) -th element respectively. Applying part (c) of Theorem
 789 2 again, we have

790
$$E(I_1) = 0. \tag{16}$$

791 Next, consider the term I_2 and note that

792
$$I_2 = \mathbf{g}_j^\top (\mathbf{X}\zeta_j) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-2} \mathbf{X}^\top \mathbf{g}_j (\mathbf{X}\zeta_j) = \left\| (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{g}_j (\mathbf{X}\zeta_j) \right\|_F^2 \tag{17}$$

793
$$\leq \left\| (\mathbf{X}^\top \mathbf{X})^{-1} \right\|_F^2 \left\| \mathbf{X}^\top \mathbf{g}_j (\mathbf{X}\zeta_j) \right\|_2^2. \tag{18}$$

 794

795 Taking expectation with respect to \mathbf{X} on both sides and applying the Cauchy-Schwartz in-
 796 equality, we have

797
$$\{E(I_2)\}^2 \leq E \left\{ \left\| (\mathbf{X}^\top \mathbf{X})^{-1} \right\|_F^4 \right\} E \left\{ \left\| \mathbf{X}^\top \mathbf{g}_j (\mathbf{X}\zeta_j) \right\|_2^4 \right\} = I_{21} I_{22}.$$

798 Consider the term I_{21} . By the invariance property, $\mathbf{X}^\top \mathbf{X}$ follows a multivariate Wishart dis-
 799 tribution with n degree of freedom $W_p(n, \Sigma_{-j}^{(-j)})$. As a result, the matrix $\{\mathbf{X}^\top \mathbf{X}\}^{-1}$ follows
 800 an Inverse Wishart distribution, Inv-Wishart $\left(n, (\Sigma_{-j}^{(-j)})^{-1} \right)$. To simplify notation, let $\mathbf{B} =$
 801 $(\mathbf{X}^\top \mathbf{X})^{-1}$ with elements b_{rq} , $r, q = 1, \dots, p - 1$. Then, the Cauchy-Schwartz inequality gives

802
$$\|\mathbf{B}\|_F^4 = \left(\sum_{r,q} b_{rq}^2 \right)^2 \leq (p - 1)^2 \sum_{r,q} b_{rq}^4.$$

803 Now, we bound each of the b_{rq}^4 term. For any $m \times n$ matrix \mathbf{A} , let $\text{vec}(\mathbf{A})$ denote a mn -
 804 dimensional vector formed by stacking columns of \mathbf{A} together, and denote $\overset{r}{\otimes} \mathbf{A} = \underbrace{\mathbf{A} \otimes \dots \otimes \mathbf{A}}_r$,
 805 with \otimes being the Kronecker product. Note that for any square $p \times p$ matrix \mathbf{A} , the matrix $\overset{r}{\otimes} \mathbf{A}$
 806 has dimension $p^r \times p^r$. When $r = 0$, then $\overset{0}{\otimes} \mathbf{A} = 1$. Applying the results of von Rosen (1988,

equation 4.2), we have for any $r \in \mathbb{N}$,

$$\text{vec} \left[E \left(\begin{smallmatrix} r+1 \\ \otimes \\ \mathbf{B} \end{smallmatrix} \right) \right] = \left[(n - \bar{\rho} - 1) \mathbf{I}_{2r+2} - \sum_{i=0}^{r-1} \{ \mathbf{P}_1(i) + \mathbf{P}_2(i) \} \right]^{-1} \text{vec} \left[E \left(\begin{smallmatrix} r \\ \otimes \\ \mathbf{B} \end{smallmatrix} \right) \otimes (\boldsymbol{\Sigma}_{-j}^{(-j)})^{-1} \right],$$

where $\bar{\rho} = \rho - 1$, $\mathbf{P}_1(i)$ and $\mathbf{P}_2(i)$ are appropriately defined permutation matrices of dimensions p^{2r+2} in von Rosen (1988, section 4). Next, let $\mathbf{V}(r) = (n - \bar{\rho} - 1) \mathbf{I}_{2r+2} - \sum_{i=0}^{r-1} \{ \mathbf{P}_1(i) + \mathbf{P}_2(i) \}$. For any r such that $n - \bar{\rho} - 1 > 2r$, the matrix \mathbf{V}_r is diagonally dominant and hence non-singular, so its inverse exists. In this case, applying the Holder inequality, we obtain

$$\left\| \text{vec} \left[E \left(\begin{smallmatrix} r+1 \\ \otimes \\ \mathbf{B} \end{smallmatrix} \right) \right] \right\|_{\infty} \leq \left[\sigma_{\max} \{ \mathbf{V}(r)^{-1} \} \right] \left\| \text{vec} \left[E \left(\begin{smallmatrix} r \\ \otimes \\ \mathbf{B} \right) \otimes (\boldsymbol{\Sigma}_{-j}^{(-j)})^{-1} \right] \right\|_{\infty}. \quad (19)$$

Note that $\sigma_{\max} \{ \mathbf{V}(r)^{-1} \} = 1 / \sigma_{\min} (\mathbf{V}_r)$. Because $\mathbf{P}_1(i)$ and $\mathbf{P}_2(i)$ are permutation matrices, all their singular values are 1, and so $\sigma_{\max} (\sum_{i=0}^{r-1} \{ \mathbf{P}_1(i) + \mathbf{P}_2(i) \}) = 2r$, hence $\sigma_{\min} (\mathbf{V}_r) = n - \bar{\rho} - 1 - 2r$. Therefore, we have

$$\sigma_{\max} \{ \mathbf{V}(r)^{-1} \} = O \{ (n - \bar{\rho})^{-1} \} = O \{ (n - \rho)^{-1} \}.$$

Under the restricted eigenvalue condition (Condition 2) for $\boldsymbol{\Sigma}$, all elements of $(\boldsymbol{\Sigma}_{-j}^{(-j)})^{-1}$ are $O(1)$. Therefore, using the recursive relation (19), we have

$$\left\| \text{vec} \left[E \left(\begin{smallmatrix} r+1 \\ \otimes \\ \mathbf{B} \end{smallmatrix} \right) \right] \right\|_{\infty} = O \{ (n - \rho)^{-r} \}.$$

Applying the above relationship with $r = 4$ and noting that $E(b_{r,q}^4)$ is a component of $\text{vec} \left[E \left(\begin{smallmatrix} 4 \\ \otimes \\ \mathbf{B} \end{smallmatrix} \right) \right]$, we have

$$I_{21} = E \left(\|\mathbf{B}\|_F^4 \right) \leq (\rho - 1)^2 \sum_{r,q} E(b_{r,q}^4) \leq O \left\{ \frac{(\rho - 1)^4}{(n - \rho)^4} \right\}.$$

Now we consider I_{22} . We have

$$\begin{aligned} \|\mathbf{X}^T \mathbf{g}_j(\mathbf{X}\zeta_j)\|_2^4 &= \left[\text{trace} \{ \mathbf{X}^T \mathbf{g}_j(\mathbf{X}\zeta_j) \mathbf{g}_j^T(\mathbf{X}\zeta_j) \mathbf{X} \} \right]^2 = \left[\sum_{r=1}^{p-1} \left\{ \sum_{i=1}^n x_{ir} \mathbf{g}_j(\mathbf{z}_{-j}^{(i)} \zeta_j) \right\}^2 \right]^2 \\ &\leq (\rho - 1) \sum_{r=1}^{p-1} \left\{ \sum_{i=1}^n x_{ir} \mathbf{g}_j(\mathbf{x}^{(i)} \zeta_j) \right\}^4, \end{aligned}$$

where the last inequality follows from the Cauchy-Schwartz inequality. Hence, we have

$$E(I_{22}) \leq (p-1) \sum_{r=1}^{p-1} E \left\{ \sum_{i=1}^n x_{ir} g_j(\mathbf{x}^{(i)} \zeta_j) \right\}^4.$$

Furthermore, because the rows of \mathbf{X} are independent, then the expectation $E \{x_{ir} g_j(\mathbf{x}^{(i)} \zeta_j)\} = 0$. As a result, for any r , we obtain

$$\begin{aligned} E \left\{ \sum_{i=1}^n x_{ir} g_j(\mathbf{z}_{-j}^{(i)} \zeta_j) \right\}^4 &= \sum_{i=1}^n E \{x_{ir} g_j(\mathbf{z}_{-j}^{(i)} \zeta_j)\}^4 \\ &+ \sum_{i=1}^n \sum_{k=1}^n E \{x_{ir} g_j(\mathbf{x}^{(i)} \zeta_j)\}^2 E \{x_{kr} g_j(\mathbf{x}^{(k)} \zeta_j)\}^2 \\ &\stackrel{(ii)}{=} O(n) + O(n^2) = O(n^2), \end{aligned}$$

where step (ii) follows from Lemma 2 in the Supporting Information. As a result, $E(I_{22}) \leq O\{(p-1)^2 n^2\}$, and together,

$$E(I_2) = O \left(\sqrt{\frac{(p-1)^6 n^2}{(n-p)^4}} \right) = O \left(\frac{n(p-1)^3}{(n-p)^2} \right). \tag{20}$$

Finally, we consider the term $I_3 = \text{trace} \{ \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-2} \mathbf{X}^\top \text{Var}(\mathbf{z}_j | \mathbf{X}) \}$. Because components of the vector \mathbf{z}_j are mutually independent, the covariance matrix $\text{Var}(\mathbf{z}_j | \mathbf{X})$ is diagonal with elements $\text{Var}(z_j^{(i)} | \mathbf{z}_{-j}^{(i)})$, $i = 1, \dots, n$. Lemma 2 in the Supporting Information shows that these elements are $O(1)$, so collectively we have

$$E(I_3) = O(1) E \left[\text{trace} \{ \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-2} \mathbf{X}^\top \} \right] = O(1) E \left[\text{trace} \{ (\mathbf{X}^\top \mathbf{X})^{-1} \} \right]. \tag{21}$$

Again, $(\mathbf{X}^\top \mathbf{X})^{-1}$ follows an Inverse Wishart distribution $\text{Inv-Wishart} \left(n, \left(\boldsymbol{\Sigma}_{-j}^{(-j)} \right)^{-1} \right)$, so

$$E \left[\text{trace} \{ (\mathbf{X}^\top \mathbf{X})^{-1} \} \right] = O \left(\frac{p-1}{n-p} \right).$$

Substituting (14), (15), (16), (20) and (21) in (13), we have

$$E \left(\left\| \hat{\beta}_j - \beta_j \right\|_2^2 \right) = O \left(\frac{n(p-1)^3}{(n-p)^2} + \frac{p-1}{n-p} \right).$$

848 Note that if p is fixed, then the rate of convergence will be $O(n^{-1})$. If p is growing, consistency
849 will be achieved if $p = o(n^{1/3})$.

850 9.4.2 | When $\alpha_j = 0$

851 If $\alpha_j = 0$, then the function $g_j = 0$ everywhere. In this case, the term $I_2 = 0$, so only the term
852 I_3 matters. Hence, it is straightforward to see that

$$853 E \left(\|\hat{\beta}_j - \beta_j\|_2^2 \right) = O \left(\frac{p-1}{n-p} \right).$$

854 9.5 | Proof of Theorem 5

855 In this proof, we let $\mathbf{X} = \mathbf{Z}_{-j}$, and we remove the subscript j to further simplify the notation.

856 First, we prove that when $p^4/n \rightarrow 0$, the conditional expectation $E(\mathbf{r}|\mathbf{X}) \xrightarrow{P} \mathbf{g}(\mathbf{X}\zeta)$ compo-
857 nentwise. We have $\mathbf{r} = \mathbf{z} - \mathbf{X}\hat{\beta} = \mathbf{z} - \mathbf{H}\mathbf{z} = (\mathbf{I} - \mathbf{H})\mathbf{z}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}$. Therefore,

$$858 E(\mathbf{r}|\mathbf{X}) = (\mathbf{I} - \mathbf{H}) \{ \mathbf{X}\beta + \mathbf{g}(\mathbf{X}\zeta) \} = \mathbf{g}(\mathbf{X}\zeta) - \mathbf{H}\mathbf{g}(\mathbf{X}\zeta).$$

859 Next, we show that each element of the vector $\mathbf{a} = \mathbf{H}\mathbf{g}(\mathbf{X}\zeta)$ goes to zero in probability. Consid-
860 ering $a_i = \mathbf{x}^{(i)}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}\mathbf{g}(\mathbf{X}\zeta)$, we will show that both $E(a_i) \rightarrow 0$ and $E(a_i^2) \rightarrow 0$ with probability
861 tending to one. Using the Cauchy-Schwartz inequality and (20) with $I_2 = \|(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{g}(\mathbf{X}\zeta)\|_F^2$,
862 we have

$$863 E(a_i) \leq \sqrt{E(\|\mathbf{x}^{(i)}\|_2^2) E(\|(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}\mathbf{g}(\mathbf{X}\zeta)\|_F^2)} \leq \sqrt{O(p)O\left(\frac{np^3}{(n-p)^2}\right)} = O\left(\frac{\sqrt{np^2}}{|n-p|}\right) \rightarrow 0,$$

864 and

$$865 E(a_i^2) \leq \sqrt{E(\|\mathbf{x}^{(i)}\|_2^4) E(\|(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}\mathbf{g}(\mathbf{X}\zeta)\|_F^4)} \leq \sqrt{O(p^2)O\left(\frac{n^2p^6}{(n-p)^4}\right)} = O\left(\frac{np^4}{(n-p)^2}\right) \rightarrow 0.$$

866 Therefore, when $p^4/n \rightarrow 0$, we have $E\{r^{(i)} | \mathbf{X}\} \xrightarrow{P} \mathbf{g}(\mathbf{x}^{(i)}\zeta)$, $i = 1, \dots, n$. Note that both the
867 left and right hand sides are functions of random variables \mathbf{X} . Taking expectation on both sides
868 with respect to \mathbf{X} , we obtain

$$869 E\{r^{(i)}\} = E\{\mathbf{g}(\mathbf{x}^{(i)}\zeta)\} + o(1), \quad i = 1, \dots, n \quad (22)$$

870 by uniform integrability of $E \{r^{(i)} \mid \mathbf{X}\}$ by Condition 2 in the main paper. Next, recall that $\hat{\mu}_0 =$
 871 $n^{-1} \sum_{k=1}^n r^{(k)}$, so for each $i = 1, \dots, n$, we have

$$\begin{aligned}
 872 \quad \hat{\mu}_0 - E \{g(\mathbf{x}^{(1)}\zeta)\} &= n^{-1} \sum_{k=1}^n r^{(k)} - E \{g(\mathbf{x}^{(i)}\zeta)\} \\
 873 \quad &= \left[n^{-1} \sum_{k=1}^n r^{(k)} - E \{r^{(1)}\} \right] + \left[E \{r^{(1)}\} - E \{g(\mathbf{x}^{(1)}\zeta)\} \right] \quad (23) \\
 874
 \end{aligned}$$

875 The second term in the square bracket of (23) is $o(1)$ by (22). Note that $r^{(1)}, \dots, r^{(n)}$ have the
 876 same marginal distribution, and hence have the same expectation. Although they are corre-
 877 lated, the conditional variance-covariance matrix

$$878 \quad \text{Var}(\mathbf{r} \mid \mathbf{X}) = (\mathbf{I} - \mathbf{H})\text{Var}(\mathbf{y} \mid \mathbf{X})(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})\mathcal{O}(1).$$

879 Since all the elements of the \mathbf{H} matrix is $o(1)$, we have

$$880 \quad \max_{i,k} \{ \text{Cov}(r^{(i)}, r^{(k)} \mid \mathbf{X}) \} = o(1).$$

881 By uniform integrability, for any i and k , we then have $\text{Cov}(r^{(i)}, r^{(k)}) = o(1)$. Hence for any $\varepsilon > 0$,
 882 we can apply the Chebyshev inequality to obtain

$$883 \quad P \left(\left| n^{-1} \sum_{k=1}^n r^{(k)} - E \{r^{(1)}\} \right| \leq \varepsilon \right) \leq \frac{\text{Var}(\sum_{k=1}^n r^{(k)})}{n^2 \varepsilon^2} = \frac{\sum_{i=1}^n \sum_{k=1}^n \text{Cov}(r^{(i)}, r^{(k)})}{n^2 \varepsilon^2} = o(1).$$

884 Therefore, $n^{-1} \sum_{k=1}^n r^{(k)} - E \{r^{(1)}\} = o_p(1)$. Substituting into (23), we get

$$885 \quad \hat{\mu}_0 - E \{g(\mathbf{x}^{(1)}\zeta)\} = o_p(1) + o(1) = o_p(1).$$

886 Next, let $\hat{g}(\mathbf{X}\hat{\zeta}) = \hat{\mu}_0 + \hat{\tau}\hat{\nu}(\mathbf{X}^T \hat{\theta})$ be the estimated function from the one-component projec-
 887 tion pursuit. Using the result from Chen (1991, Theorem 1) and the fact that the function g is
 888 twice continuously differentiable, we then have

$$889 \quad \lim_{n \rightarrow \infty} P \left(n^{-1} \sum_{i=1}^n \{ \hat{g}(\mathbf{x}^{(i)}\hat{\zeta}) - g(\mathbf{x}^{(i)}\zeta) \}^2 \geq Cn^{-4/5} \right) = 0 \quad (24)$$

890 for a constant C that does not depend on either n or p . By the reverse triangular inequality, we

891 then have

$$\begin{aligned}
 892 \quad & n^{-1} \sum_{i=1}^n \left\{ \hat{g}(\mathbf{x}^{(i)} \hat{\zeta}) - g(\mathbf{x}^{(i)} \zeta) \right\}^2 = n^{-1} \sum_{i=1}^n \left[\hat{\tau} \hat{\nu}(\mathbf{x}^{(i)} \hat{\theta}) - \{g(\mathbf{x}^{(i)} \zeta) - \hat{\mu}_0\} \right]^2 \\
 893 \quad & \geq n^{-1} \left| \sqrt{\sum_{i=1}^n \hat{\tau}^2 \hat{\nu}^2(\mathbf{x}^{(i)} \hat{\theta})} - \sqrt{\sum_{i=1}^n \{g(\mathbf{x}^{(i)} \zeta) - \hat{\mu}_0\}^2} \right|^2 \\
 894 \quad & \geq \left| \hat{\tau} - \sqrt{n^{-1} \sum_{i=1}^n [g(\mathbf{x}^{(i)} \zeta) - E\{g(\mathbf{x}^{(1)} \zeta)\} + E\{g(\mathbf{x}^{(1)} \zeta)\} - \hat{\mu}_0]^2} \right|^2 \quad (25) \\
 895
 \end{aligned}$$

896 where the second inequality follows from the fact that $n^{-1} \sum_{i=1}^n \hat{\nu}^2(\mathbf{x}^{(i)} \hat{\theta}) = 1$. For the term
 897 inside the square root of the last expression, we have

$$\begin{aligned}
 898 \quad & n^{-1} \sum_{i=1}^n [g(\mathbf{x}^{(i)} \zeta) - E\{g(\mathbf{x}^{(1)} \zeta)\} + E\{g(\mathbf{x}^{(1)} \zeta)\} - \hat{\mu}_0]^2 \\
 899 \quad & \leq 2n^{-1} \sum_{i=1}^n [g(\mathbf{x}^{(i)} \zeta) - E\{g(\mathbf{x}^{(1)} \zeta)\}]^2 + 2n^{-1} \sum_{i=1}^n [E\{g(\mathbf{x}^{(1)} \zeta)\} - \hat{\mu}_0]^2 \\
 900 \quad & = \tau^2 + o_p(1) + o_p(1) = \tau^2 + o_p(1), \quad (26) \\
 901
 \end{aligned}$$

902 where $\tau^2 = \text{Var}\{g(\mathbf{x}^{(1)} \zeta)\}$. Hence, substituting (26) into (24) and (25), we obtain

$$903 \quad \lim_{n \rightarrow \infty} P(|\hat{\tau} - \tau|^2 \geq Cn^{-4/5}) \leq \lim_{n \rightarrow \infty} P\left(n^{-1} \sum_{i=1}^n \left\{ \hat{g}(\mathbf{x}^{(i)} \hat{\theta}) - g(\mathbf{x}^{(i)} \zeta) \right\}^2 \geq Cn^{-4/5}\right) = 0$$

904 or in other words, $\lim_{n \rightarrow \infty} P(|\hat{\tau}_j - \tau_j| \geq Cn^{-2/5}) = 0$ as claimed.

905 9.6 | Proof of Theorem 6

906 First, consider the matrix $\mathbf{B}^{(1)}$ with element $b_{jk}^{(1)} = \max(|\hat{\beta}_j^{(k)}|, |\hat{\beta}_k^{(j)}|)$. By the Markov inequality,
 907 for any $\delta > 0$, we have

$$908 \quad P(\|\hat{\beta}_j - \beta_j\|_\infty \geq \delta) \leq \frac{E(\|\hat{\beta}_j - \beta_j\|_\infty^2)}{\delta^2} \leq \frac{E(\|\hat{\beta}_j - \beta_j\|_2^2)}{\delta^2}.$$

909 Therefore, by Theorem 4, we obtain

$$910 \quad P\left(\|\hat{\beta}_j - \beta_j\|_\infty \geq \delta\right) \leq \begin{cases} \frac{C_1 p^3}{n\delta^2}, & \alpha_j \neq 0 \\ \frac{C_2 p}{n\delta^2}, & \alpha_j = 0. \end{cases}$$

911 Hence, by the union bound,

$$912 \quad P\left(\max_{j \in S_\alpha} \|\hat{\beta}_j - \beta_j\|_\infty \geq \delta\right) \leq \sum_{j \in S_\alpha} P(\|\hat{\beta}_j - \beta_j\|_\infty \geq \delta) \leq \frac{C_1 s_\alpha p^3}{n\delta^2}, \text{ and}$$

913

$$914 \quad P\left(\max_{j \in S_\alpha^c} \|\hat{\beta}_j - \beta_j\|_\infty \geq \delta\right) \leq \sum_{j \in S_\alpha^c} P(\|\hat{\beta}_j - \beta_j\|_\infty \geq \delta) \leq \frac{C_2(p - s_\alpha)p}{n\delta^2} \leq \frac{C_2 p^2}{n\delta^2}.$$

915 Next, we consider the matrix $\mathbf{B}^{(2)}$ with elements $b_{jk}^{(2)} = \hat{\tau}_j \hat{\tau}_k$, $j, k = 1, \dots, p$. By Theorem 5, as
 916 $n \rightarrow \infty$, we have $|\hat{\tau}_j - \tau_j| < C_3 n^{-2/5}$ with probability tending to one for a sufficiently large constant
 917 C_3 , such that for any pair (j, k) with $k \neq j$, we have $|\hat{\tau}_j \hat{\tau}_k - \tau_j \tau_k| < C_4 n^{-2/5}$ with probability
 918 tending to one for a sufficiently large constant C_4 . Therefore, the union bound gives

$$919 \quad \max_{\substack{j,k=1,\dots,p \\ k \neq j}} |\hat{\tau}_j \hat{\tau}_k - \tau_j \tau_k| \leq C_4 p(p-1) n^{-2/5} \leq C_4 p^2 n^{-2/5}$$

920 with probability tending to one. Finally, since $(j, k) \in E$ corresponds to both $\beta_j^{(k)}$ and $\tau_j \tau_k$ being
 921 zero, then we obtain

$$922 \quad o_{jk} = \left(b_{jk}^{(1)}\right)^2 + \left(b_{jk}^{(2)}\right)^2 \leq \delta^2 + C_4 p^4 n^{-4/5}$$

923 If we choose δ and threshold T as stated in the theorem, then with probability tending to one,
 924 $o_{jk} \leq T$ if $(j, k) \in E$. On the other hand, if $(j, k) \notin E$, $o_{jk} > T$ with probability tending to one as
 925 claimed.