The Annals of Applied Statistics
2021, Vol. 15, No. 2, 831–855
https://doi.org/10.1214/20-AOAS1413

INFERENCE OF LARGE MODIFIED POISSON-TYPE GRAPHICAL MODELS: APPLICATION TO RNA-SEQ DATA IN CHILDHOOD ATOPIC ASTHMA STUDIES

BY RONG ZHANG[1,*], ZHAO REN[1,†], JUAN C. CELEDÓN[2,‡] AND WEI CHEN[2,§]

[1]*Department of Statistics, University of Pittsburgh,* [*]*roz16@pitt.edu;* [†]*zren@pitt.edu*
[2]*Department of Pediatrics, UPMC Children's Hospital of Pittsburgh, University of Pittsburgh,* [‡]*juan.celedon@chp.edu;* [§]*wei.chen@chp.edu*

Recent advances in next-generation sequencing technology have yielded huge amounts of transcriptomic data. The discreteness and the high dimensions of RNA-seq data have posed great challenges in biological network analysis. Although estimation theories for high-dimensional modified Poisson-type graphical models have been proposed for the network analysis of count-valued data, the statistical inference of these models is still largely unknown. We herein propose a two-step procedure in both edgewise and global statistical inference of these modified Poisson-type graphical models using a cutting-edge generalized low-dimensional projection approach for bias correction. Extensive simulations and a real example with ground truth illustrate asymptotic normality of edgewise inference and more accurate inferential results in multiple testing compared to the sole estimation and the inferential method under normal assumption. Furthermore, the application of our method to novel RNA-seq data of childhood atopic asthma in Puerto Ricans demonstrates more biologically meaningful results compared to the sole estimation and the inferential methods based on Gaussian and nonparanormal graphical models.

1. Introduction. Recent developments of high-throughput sequencing technologies have generated unprecedented amounts of RNA-seq data for transcriptomics. Network studies of conditional dependency among genes provide new insights to understand a complex biological process or disease.

Gaussian graphical model (GGM) has been widely used in characterizing the conditional relationships among genes in a biological network. However, discrete omics data sets from the next generation sequencing technology are common because the count values are usually used to quantify the genetic or genomic information. One typical example is the bulk RNA-seq data which summarizes the expression of each gene using the number of counts mapped to it. Another example is the droplet-based single-cell RNA-seq data which quantifies the cell-level gene expression with unique molecular identifiers (UMIs) (Islam et al. (2014)), a direct counting of transcript copies. Therefore, the use of GGM on those non-Gaussian discrete-type data requires a continuous transformation, for example, using the fragments per kilo base of transcript per million (FPKM) or a log-transformation on the count values. Converting count values into continuous values tends to alter their biological meanings with the straightforward interpretation and, sometimes, can be inappropriate (Zwiener, Frisch and Binder (2014)). Poisson distribution, however, is a popular choice and has been shown more reasonable than using FPKM in modeling the count data (Anders and Huber (2010)). To describe the conditional dependency among genes from count-valued omics data, Besag (1974) proposed a natural extension of the univariate Poisson model to a multivariate case, and Yang

Received October 2019; revised October 2020.
Key words and phrases. Poisson graphical model, asthma genomics, RNA-seq data, asymptotic normality, bias correction, multiple testing.

et al. (2015) further extended this to a general graphical model setting called the Poisson graphical model (PGM). Moreover, three modified Poisson-type graphical models: the truncated PGM (TPGM), the sublinear PGM (SPGM) (Yang et al. (2013)) and the square-root PGM (SqrtPGM) (Inouye, Ravikumar and Dhillon (2016)) were proposed to overcome the major drawback of PGM for count data modeling (see Section 2 for more details).

On the other hand, omics data sets are usually large scale with the number of genes p allowed to be far larger than the sample size n. To provide reliable estimation for pairwise conditional dependency with its confidence interval and p-value under such settings, statistical inference of high-dimensional GGM has been well developed within the recent six years; see Liu (2013), Ren et al. (2015), Janková and van de Geer (2015), Janková and van de Geer (2017). Recently, attention has started being paid attention to inference of large non-Gaussian graphical models; see Li et al. (2016) and Cai et al. (2019) for Ising graphical model (IGM). Unfortunately, all current methods based on the three aforementioned high-dimensional modified Poisson-type graphical models only involve estimation, and a unified framework for their statistical inference is still largely unknown.

In this paper we intend to propose a new inferential procedure that particularly tailors to the analysis of nonnegative, discrete and high-dimensional transcriptomic data based on the modified Poisson-type graphical models. Our motivation comes from the novel RNA-seq gene expression data from the study of the Epigenetic Variation and Childhood Asthma in Puerto Ricans (EVA-PR) aged nine to 20 years (Forno et al. (2019)). To our knowledge, it is the first study of atopic asthma in nasal epithelium of a large sample of Hispanic children. Further details of the data are deferred to Section 1 of the Supplement (Zhang et al. (2020)).

Atopic asthma is one of the most prevalent diseases affecting all ages, but efficient methods for its accurate diagnosis are still under development. Clinicians have recently considered using nasal epithelial samples which are much easier to extract and more disease-relevant to replace white blood cell samples in study of the pathogenesis of atopic asthma. According to Forno et al. (2019), studies in nasal epithelial samples provide promising results in identifying epigenetic variants of childhood atopic asthma in Puerto Ricans. Besides, Pandey et al. (2018) has illustrated differentially expressed genes from transcriptomic profiles that are more closely related to the mechanism of asthma using adult nasal epithelial samples. However, conditional dependence among genes underlying atopic asthma from nasal epithelium is largely unknown, a knowledge of which will no doubt facilitate its accurate diagnosis and the development of its precision medicine.

Inspired by the cutting-edge low-dimensional projection estimator (LDPE) approach in inference of high-dimensional linear regression (Zhang and Zhang (2014)) and the recent developments in statistical inference of large IGM, we have developed a novel two-step procedure in inference of pairwise conditional dependency from large modified Poisson-type graphical models. The first step involves ℓ_1-penalized nodewise regressions, and the second step is based on a likelihood-based nonlinear projection which relies on the graph structure itself and is intrinsically different from the essentially linear projection approach considered in van de Geer et al. (2014) for generalized linear models. For further details, please refer to Section 3. From the computational perspective our method only requires $O(p)$ ℓ_1-penalized regressions, due to the novelty of our second step, and is computationally less intensive than the composite likelihood approach and the score matching method proposed in Wang and Kolar (2016) and Yu, Kolar and Gupta (2016), respectively, targeting on exponential family graphical model inference.

In Section 2, we briefly review the properties of three typical modified Poisson-type graphical models. We formally propose a general framework of our procedure with its application to the three modified Poisson-type models in Section 3. Section 4 includes implementations with selection of tuning parameters. Then, we demonstrate the validity and

advantages of our procedure through simulations and a real example with ground truth in Section 5 and an application to the motivating RNA-seq data of childhood allergic asthma in Section 6. We finally conclude with discussion in Section 7. Our approach is implemented in a publicly available `R` package `ModPGMInference` on the GitHub repository: https://github.com/zhangr100/ModPGMInference.

2. The modified Poisson-type graphical models. Let $X = (X_1, X_2, \ldots, X_p)^\intercal$ be a sequence of genes with each $X_i \in \{0, 1, 2, \ldots\}$ for $i = 1, 2, \ldots, p$. An undirected Poisson graph $G = (V, E)$ associated with X consists of the node set $V = \{X_1, X_2, \ldots, X_p\}$ and the edge set $E = \{\text{pairs of } (i, j) \text{ if there is an undirected edge between } X_i \text{ and } X_j\}$. X_i and X_j are conditionally dependent, given all the other genes $\{X_r, r \neq i, j\}$, if and only if there is an edge between the two nodes. More formally speaking, the joint distribution of PGM (Yang et al. (2015)) is defined as $\mathbb{P}_{\psi,\mathbf{\Theta}}(X) = \exp(\sum_{1\le i<j\le p} \theta_{ij} X_i X_j + \sum_{i=1}^{p}(\psi_i X_i - \log(X_i!)) - A(\psi, \mathbf{\Theta}))$, where $A(\psi, \mathbf{\Theta})$ is the log-normalization constant. The parameter θ_{ij} represents the pairwise strength between nodes X_i and X_j and is encoded in a parameter set $\mathbf{\Theta}$. It is easy to see that X_i and X_j are conditionally independent if and only if $\theta_{ij} = 0$. Therefore, if the two nodes are connected in a graph, we set $\theta_{ij} \neq 0$; otherwise, $\theta_{ij} = 0$. However, PGM can only model negative pairwise dependency (or $\theta_{ij} \le 0$) if $A(\psi, \mathbf{\Theta}) < +\infty$ is achieved. This fact is due to $x^2/\log(x!) \to +\infty$ as $x \to +\infty$ which can be shown by the Stirling's approximation. To overcome the major constraint of PGM, three modified Poisson-type graphical models are proposed in the literature to allow for both positive and negative dependencies between pairwise nodes.

2.1. *TPGM.* Since the domain of PGM is $\{0, 1, \ldots\}^p$, the quadratic terms dominate the distribution when count values are very large which leads to negative dependency. Therefore, a natural remedy is to truncate the domain of each node to a finite level so as to capture both positive and negative dependencies. We can make a reasonable assumption that each node X_i is bounded by a finite number D_i with $i = 1, 2, \ldots, p$. The joint distribution of TPGM (Yang et al. (2013)) is defined as

$$\mathbb{P}_{\psi,\mathbf{\Theta}}(X) \propto \exp\left(\sum_{i=1}^{p} \psi_i X_i + \sum_{1\le i<j\le p} \theta_{ij} X_i X_j - \sum_{i=1}^{p} \log(X_i!)\right) \tag{2.1}$$

which has the same format as PGM but with a different log-normalization constant due to the domain $X_i \in \{0, 1, \ldots, D_i\}$ for $i = 1, 2, \ldots, p$. We mention that the Ising graphical model (IGM), studied in Ravikumar, Wainwright and Lafferty (2010), Li et al. (2016) and Cai et al. (2019), is a special case of TPGM when $D_i = 1$ for all $i = 1, 2, \ldots, p$.

2.2. *SPGM.* Unlike TPGM, Yang et al. (2013) also proposed sublinear PGM (SPGM), an alternative to modify the original PGM without a change on the domain of each node. Specifically, by replacing the linear statistic of each node in $\psi_i X_i$ and $\theta_{ij} X_i X_j$ in (2.1) with a newly-constructed statistic that increases even slower than a linear term, both positive and negative dependencies are allowed in the modified distribution without a domination of quadratic terms when the value of each node goes to $+\infty$. Therefore, a modified statistic for each node X_i with $i = 1, 2, \ldots, p$ is defined as

$$S(X_i) = \begin{cases} X_i & \text{if } X_i \le D_{i0}, \\ -\dfrac{1}{2(D_{i1} - D_{i0})} X_i^2 + \dfrac{D_{i1}}{D_{i1} - D_{i0}} X_i - \dfrac{D_{i0}^2}{2(D_{i1} - D_{i0})} & \text{if } D_{i0} < X_i \le D_{i1}, \\ \dfrac{D_{i0} + D_{i1}}{2} & \text{if } X_i \ge D_{i1}, \end{cases}$$

where D_{i0} and D_{i1} are predefined thresholds. The joint distribution of SPGM is thus defined as

$$\mathbb{P}_{\psi,\mathbf{\Theta}}(X) \propto \exp\left(\sum_{i=1}^{p} \psi_i S(X_i) + \sum_{1\le i<j\le p} \theta_{ij} S(X_i)S(X_j) - \sum_{i=1}^{p} \log(X_i!)\right).$$

SPGM will be close to the original PGM as the upper threshold $D_{i1} \to +\infty$. In particular, SPGM still has a relatively thick tail which is approachable to the Poisson case.

2.3. *SqrtPGM*. In addition to the aforementioned two models, Inouye, Ravikumar and Dhillon (2016) proposed a new class of parametric graphical model called Square Root Graphical Model that allows both positive and negative dependencies. In the Poisson case, SqrtPGM essentially uses square root to replace the linear statistic of each node in $\psi_i X_i$ and $\theta_{ij} X_i X_j$ in (2.1), so the interaction terms become linear to avoid the problem that the quadratic terms dominate the distribution when the value of each node goes to $+\infty$. The joint distribution of SqrtPGM is thus defined as

$$\mathbb{P}_{\psi,\mathbf{\Theta}}(X) \propto \exp\left(\sum_{i=1}^{p} \psi_i \sqrt{X_i} + \sum_{1\le i<j\le p} \theta_{ij} \sqrt{X_i}\sqrt{X_j} - \sum_{i=1}^{p} \log(X_i!)\right).$$

2.4. *A unified representation*. Let $T(X)$ and $B(X)$ be the sufficient statistic and the base measure, respectively. All three modified Poisson-type graphical models can be described in the following generalized joint distribution:

$$\mathbb{P}_{\psi,\mathbf{\Theta}}(X) \propto \exp\left(\sum_{i=1}^{p} \psi_i T(X_i) + \sum_{1\le i<j\le p} \theta_{ij} T(X_i)T(X_j) + \sum_{i=1}^{p} B(X_i)\right). \tag{2.2}$$

The corresponding sufficient statistic, base measure and domain of X_i for each model are summarized in Table 1. Although so far we only define those θ_{ij} for which $i < j$, we set $\theta_{ij} = \theta_{ji}$ to ease our notation whenever $\theta_{ij}, i > j$ is used hereafter.

3. Statistical inference of modified Poisson-type graphical models. We first introduce a general two-step procedure to obtain each debiased estimator $\tilde{\theta}_{ij}$ of conditional dependency between variables X_i and X_j, with applications to three modified Poisson-type graphical models specified later. The goal is to achieve the desired asymptotic normality $(nF_{ij})^{1/2}(\tilde{\theta}_{ij} - \theta_{ij}) \to \mathcal{N}(0, 1)$ with a bounded variance $(F_{ij})^{-1}$ as $(n, p) \to +\infty$ under certain sparsity condition of the graph. In addition, we also introduce a global test to discover the entire graph structure.

3.1. *The general framework*. The first step is to provide a globally good initial estimator $\hat{\theta}_{ij}$ of θ_{ij}, and the second step is to correct the potential bias of $\hat{\theta}_{ij}$, via a variant of LDPE approach (Zhang and Zhang (2014)), to obtain the final estimator $\tilde{\theta}_{ij}$.

TABLE 1
Sufficient statistics, base measures and domain of X_i in the three models

Model	$T(X)$	$B(X)$	Domain of X_i
TPGM	X	$-\log(X!)$	$\{0, 1, \ldots, D_i\}$
SPGM	$S(X)$	$-\log(X!)$	$\{0, 1, \ldots\}$
SqrtPGM	$\sqrt{X}$	$-\log(X!)$	$\{0, 1, \ldots\}$

Step 1 (Initialization): From the joint distribution (2.2) we can obtain that the conditional distribution of the random variable X_i, given all other random variables $X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)^\intercal$, belongs to the univariate exponential family. More specifically, the log-likelihood function $\log(\mathbb{P}_{\eta_i}(X_i|X_{-i}))$ can be written as $T(X_i)\mu_i + B(X_i) - f(\mu_i)$ with the parameters $\eta_i = (\psi_i, \theta_i) = (\psi_i, \theta_{i1}, \theta_{i2}, \ldots, \theta_{i(i-1)}, \theta_{i(i+1)}, \ldots, \theta_{ip})^\intercal \in \mathbb{R}^p$ and the sufficient statistic $T(X_i)$. In the above equation we have $\mu_i = \psi_i + \sum_{j\neq i} \theta_{ij} T(X_j)$, and $f(\mu_i)$ is the log-normalization term. To ease notations, we introduce $X^* = (1, X^\intercal)^\intercal$, and $X^*_{-i} = (1, X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)^\intercal$ denotes the subvector of X^* with X_i removed. Similarly we denote $T(X^*_{-i}) = (1, T(X_1), T(X_2), \ldots, T(X_{i-1}), T(X_{i+1}), \ldots, T(X_p))^\intercal$. Therefore, we have a simple notation of $\mu_i = T(X^*_{-i})^\intercal \eta_i$.

Due to the sparse structure of a biological network, the whole parameter set $\boldsymbol{\Theta}$ is commonly assumed sparse in the sense that $\theta_{ij} = 0$ for most pairs of (i, j). Thus, it is natural to estimate $\boldsymbol{\Theta}$ by solving p ℓ_1-penalized nodewise regressions with $i = 1, 2, \ldots, p$ based on the conditional distribution $\mathbb{P}_{\eta_i}(X_i|X_{-i})$. Suppose that $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ are denoted as n i.i.d. samples from the joint distribution $\mathbb{P}_{\psi,\boldsymbol{\Theta}}(X)$. $\theta_i \in \mathbb{R}^{p-1}$ can be estimated by solving the following convex optimization problem:

$$\hat{\eta}_i = (\hat{\psi}_i, \hat{\theta}_i) = \arg\min_{\eta_i} \big\{ l\big(\eta_i; \{X^{(k)}\}_{k=1}^n\big) + \lambda_i \|\theta_i\|_1 \big\}, \tag{3.1}$$

where λ_i is a tuning parameter and the negative joint log-likelihood function $l(\eta_i; \{X^{(k)}\}_{k=1}^n) = -\sum_{k=1}^n \log(\mathbb{P}_{\eta_i}(X_i^{(k)}|X_{-i}^{(k)}))$ takes the form with $\mu_i^{(k)} = T(X_{-i}^{*(k)})^\intercal \eta_i$

$$l\big(\eta_i; \{X^{(k)}\}_{k=1}^n\big) = -\sum_{k=1}^n \big(T\big(X_i^{(k)}\big)\mu_i^{(k)} + B\big(X_i^{(k)}\big) - f\big(\mu_i^{(k)}\big)\big).$$

Of note, we only penalize θ_i instead of entire η_i. If one has certain prior knowledge of the biological network such as group or order structure, then the generic ℓ_1 penalty can be replaced by group Lasso or fused Lasso. To demonstrate the general purpose, we only use generic ℓ_1 in our algorithm.

High-dimensional generalized linear model theory suggests that the estimator $\hat{\theta}_i$ has good statistical properties in a global sense under certain regularity conditions. Indeed, the existing method for estimation of entire graph took this approach with theoretical justifications (Yang et al. (2013), Inouye, Ravikumar and Dhillon (2016)). However, this step itself is not sufficient for our inference purpose due to the bias incurred from the ℓ_1 penalty.

Step 2 (Likelihood-based bias correction): In this step we take a variant of LDPE approach to correct the bias of $\hat{\theta}_{ij}$ obtained from (3.1) for each pair (i, j) with $i < j$.

The original LDPE (Zhang and Zhang (2014)) can be seen as an extension of the least squares estimator in the classical theory of linear model to the high-dimensional settings. We first briefly review the intuition before LDPE. For a low-dimensional linear model with $n < p$, $Y = \mathbf{Z}\beta + \epsilon \in \mathbb{R}^n$, where $Y = (Y^{(1)}, \ldots, Y^{(n)})^\intercal$, $\epsilon = (\epsilon^{(1)}, \ldots, \epsilon^{(n)})^\intercal$, $\beta = (\beta_1, \ldots, \beta_p)^\intercal$ and the jth column of $\mathbf{Z}$ is $Z_j = (Z_j^{(1)}, \ldots, Z_j^{(n)})^\intercal$, the least squares estimator of β_j can be written as as a linear projection of Y onto the orthogonal complement of the column space of $\mathbf{Z}_{-j}$. In order words, with a score vector $V = (v_1, v_2, \ldots v_n)^\intercal$, we have

$$\tilde{\beta}_j^{\text{proj}} = \frac{\sum_{k=1}^n v_k Y^{(k)}}{\sum_{k=1}^n v_k Z_j^{(k)}} = \beta_j + \frac{\sum_{k=1}^n v_k \epsilon^{(k)}}{\sum_{k=1}^n v_k Z_j^{(k)}} + \sum_{l\neq j} \frac{\sum_{k=1}^n v_k Z_l^{(k)} \beta_l}{\sum_{k=1}^n v_k Z_j^{(k)}},$$

and when $V = Z_j^\perp$, the third term vanishes, resulting in the desired least squares estimator $\tilde{\beta}_j^{\text{proj}} = \beta_j + \sum_{k=1}^n v_k \epsilon^{(k)} / (\sum_{k=1}^n v_k Z_j^{(k)})$. However, in high-dimensional cases with $p > n$ and $\mathbf{Z}$ in general position, the orthogonal complement of the column space of $\mathbf{Z}_{-j}$ vanishes, and thus the ideal score vector is undefined as $Z_j^\perp = 0$. Following the linear-based projection

idea but with a general nonzero score vector V, the third term in the decomposition above presents a nonzero bias. Although we do not know the exact bias term, as β is unknown, this analysis of the linear estimator suggests a one-step bias correction with an initial estimator $\hat{\beta}$,

$$\tilde{\beta}_j = \tilde{\beta}_j^{\text{proj}} - \sum_{l \neq j} \frac{\sum_{k=1}^n v_k Z_l^{(k)} \hat{\beta}_l}{\sum_{k=1}^n v_k Z_j^{(k)}} = \hat{\beta}_j + \frac{\sum_{k=1}^n v_i (Y^{(k)} - Z^{(k)\intercal} \hat{\beta})}{\sum_{k=1}^n v_k Z_j^{(k)}}.$$

Therefore, with a globally good initial estimator $\hat{\beta}$ and a well-chosen score vector V, it is expected that the bias due to the third term becomes negligible, resulting in an asymptotically normal estimator $\tilde{\beta}_j$.

Since LDPE was originally introduced in linear model, for our model we first linearize the nodewise regression using initial estimators. The parameter of interest θ_i is encoded in μ_i which corresponds to the sufficient statistic $T(X_i)$. For this reason we expand the conditional expectation of $T(X_i)$, given X_{-i}, which equals the first derivative of $f(\mu_i)$. To further ease our notations, we denote the first and second derivatives of $f(\cdot)$ by $\dot{f}(\cdot)$ and $\ddot{f}(\cdot)$, respectively. Then, at the population level we have the following decomposition:

$$T(X_i) = \mathbb{E}_{\eta_i}\big(T(X_i)|X_{-i}\big) + \epsilon_i = \dot{f}(\mu_i) + \epsilon_i = \dot{f}\big(T\big(X_{-i}^*\big)^\intercal \eta_i\big) + \epsilon_i, \tag{3.2}$$

where ϵ_i has zero mean, given X_{-i}. Since $\hat{\eta}_i$ is a globally good estimator of η_i obtained in (3.1), we may take a local Taylor expansion of $\dot{f}(\mu_i)$ about $\hat{\mu}_i$ with $\hat{\mu}_i = T(X_{-i}^*)^\intercal \hat{\eta}_i$, that is, $\dot{f}(\mu_i) = \dot{f}(\hat{\mu}_i) + \ddot{f}(\hat{\mu}_i) T(X_{-i}^*)^\intercal (\eta_i - \hat{\eta}_i) + Re$, where Re denotes the remainder term. By rearranging terms in the above equation, we have the following linearized version of (3.2):

$$T(X_i) - \dot{f}(\hat{\mu}_i) + \ddot{f}(\hat{\mu}_i) T\big(X_{-i}^*\big)^\intercal \hat{\eta}_i = \ddot{f}(\hat{\mu}_i) T\big(X_{-i}^*\big)^\intercal \eta_i + (Re + \epsilon_i).$$

We are in the position to apply the projection-based idea to the regression above with i.i.d. observations. Specifically, to obtain a better estimator of θ_{ij} $(i < j)$, given some initial estimator $\hat{\eta}_i = (\hat{\psi}_i, \hat{\theta}_i)$, one needs to find an appropriate score vector $V = (v_1, v_2, \ldots, v_n)^\intercal \in \mathbb{R}^n$ and apply a one-step bias correction from $\hat{\eta}_i$ as follows:

$$\tilde{\theta}_{ij} = \hat{\theta}_{ij} + \frac{\sum_{k=1}^n v_k (T(X_i^{(k)}) - \dot{f}(\hat{\mu}_i^{(k)}))}{\sum_{k=1}^n v_k \ddot{f}(\hat{\mu}_i^{(k)}) T(X_j^{(k)})} \quad (1 \leq i < j \leq p). \tag{3.3}$$

With some algebra it is easy to see that the decomposition of the estimation error for $\tilde{\theta}_{ij}$ becomes

$$\begin{aligned} \tilde{\theta}_{ij} - \theta_{ij} = {} & \frac{\frac{1}{n}\sum_{k=1}^n v_k \epsilon_i^{(k)}}{\frac{1}{n}\sum_{k=1}^n v_k \ddot{f}(\hat{\mu}_i^{(k)}) T(X_j^{(k)})} + \frac{\frac{1}{n}\sum_{k=1}^n v_k Re^{(k)}}{\frac{1}{n}\sum_{k=1}^n v_k \ddot{f}(\hat{\mu}_i^{(k)}) T(X_j^{(k)})} \\ & + \frac{\frac{1}{n}\sum_{k=1}^n v_k \ddot{f}(\hat{\mu}_i^{(k)}) T(X_{-\{i,j\}}^{*(k)})^\intercal (\eta_{i,-j} - \hat{\eta}_{i,-j})}{\frac{1}{n}\sum_{k=1}^n v_k \ddot{f}(\hat{\mu}_i^{(k)}) T(X_j^{(k)})}. \end{aligned} \tag{3.4}$$

The first term in the right-hand side of (3.4) is denoted as the error term, and the second and the third terms can be regarded as the bias terms. Intuitively, to achieve the inference purpose, we need to pick an V such that the bias terms are asymptotically negligible with respect to the error term while the error term has asymptotic normality with root-n consistency.

To achieve our goal discussed in last paragraph, we look for a population version of V first. Denote $\langle a, b \rangle = \mathbb{E}(a \ddot{f}(\mu_i) b)$. To have a centered asymptotic normality for the first (error) term, it suffices to pick V, as any function of X_{-i} as ϵ_i has mean zero given X_{-i}. Indeed, for such a choice we have $\mathbb{E}(V\epsilon_i) = 0$ with variance $\text{Var}(V\epsilon_i) = \langle V, V \rangle$. Consequently, the entire first term has the desired asymptotic normality. We leave the mathematical derivation

of these facts in the Supplementary Material (Zhang et al. (2020)). Besides, for the second (bias) term we expect that, with a reasonable choice of V, this term itself is small since it contains a remainder term Re from the second order Taylor expansion. It remains to find a specific V under this constraint (i.e., V is a measurable function of X_{-i}) so that the third (bias) term is negligible. To this end, ideally one needs that $\langle V, T(X_{-\{i,j\}})\rangle$ is a zero vector, where $X_{-\{i,j\}} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)^\intercal$ and $T(X_{-\{i,j\}}) \in \mathbb{R}^{p-2}$ is defined accordingly. Then it is reasonable to expect the third term is small, given that $\hat{\eta}_i$ is a globally good estimator.

The major novelty of our method is on the choice of score vector V. Intrinsic to the graphical model joint distribution (2.2), we propose to choose the population V based on the conditional expectation of X_j, given $X_{-\{i,j\}}$, with respect to the inner product $\langle a, b\rangle$ as follows:

$$
\begin{aligned}
V &= T(X_j) - \frac{\mathbb{E}_{\eta_i,\eta_j}(T(X_j)\ddot{f}(\mu_i)|T(X_{-\{i,j\}}))}{\mathbb{E}_{\eta_i,\eta_j}(\ddot{f}(\mu_i)|T(X_{-\{i,j\}}))} \\
&:= T(X_j) - g\big(T(X_{-\{i,j\}}), \eta_i, \eta_j\big).
\end{aligned} \tag{3.5}
$$

It is worthwhile to point out that the conditional expectation function $g(\cdot)$ depends on unknown parameters only through η_i and η_j. In particular, μ_i is known given η_i. By our choice, one can check that $\langle V, m(T(X_{-\{i,j\}}))\rangle = 0$ for any measurable function $m(\cdot)$. Thus, we have achieved that $\langle V, T(X_{-\{i,j\}})\rangle$ is a zero vector, and, at the population level, the third (bias) term in (3.4) becomes zero.

REMARK 1. Our choice of the score V is new and intrinsic to the joint likelihood of the specific graphical model. Other methods of bias correction for GLMs were discussed in literature, for example, van de Geer et al. (2014). The difference is that our construction of V relies on the explicit knowledge of joint conditional distribution of $T(X_j)$, given all other covariates $T(X_{-\{i,j\}})$ in which the conditional expectation of $T(X_j)$ is a nonlinear function of $T(X_{-\{i,j\}})$. In contrast, the method proposed in van de Geer et al. (2014) does not impose the specific conditional likelihood pattern but, essentially, assumes certain linear sparsity structure among all covariates, and thus the proposed score vector is linear. We emphasize that this linear sparsity structure is invalid in general in our graphical model settings. For the reasons above we call this step of our method the likelihood-based bias correction.

In the end, given the population expression of V in (3.5), we need to represent the empirical element v_k in the score vector V. Denote the oracle score of the kth observation as $v_k^{(o)} = T(X_j^{(k)}) - g(T(X_{-\{i,j\}}^{(k)}), \eta_i, \eta_j)$. Here, we call $v_k^{(o)}$ the oracle score since η_i, η_j are unknown to us. Those points where $T(X_{-\{i,j\}}^{(k)})$ has explained most variability of $T(X_j^{(k)})$ would receive scores with a small magnitude and thus play a less significant role in our method. Intuitively, we expect that the first term in the right-hand side of (3.4) dominates $\tilde{\theta}_{ij} - \theta_{ij}$ with our choice of V. One can show that if ignoring the minor difference between $\ddot{f}(\mu_i)$ and $\ddot{f}(\hat{\mu}_i)$, then the asymptotic variance of this first term is F_{ij}^{-1}, where

$$
F_{ij} = \mathbb{E}_{\eta_i,\eta_j}\big(\big(T(X_j) - g\big(T(X_{-\{i,j\}}), \eta_i, \eta_j\big)\big)^2 \ddot{f}(\mu_i)\big) = \langle V, V\rangle.
$$

We leave its mathematical derivation in the Supplementary Material (Zhang et al. (2020)). Thanks to the globally good estimators $\hat{\eta}_i$ and $\hat{\eta}_j$ obtained from Step 1, it is natural for us to finally use the plugged-in estimator of the oracle, $v_k = T(X_j^{(k)}) - g(T(X_{-\{i,j\}}^{(k)}), \hat{\eta}_i, \hat{\eta}_j)$ in the bias correction step (3.3). We defer the specification of complete implementations in Section 4.

Intuitively, we expect that our choice of the nonlinear score vector leads to the following asymptotic normality under some regularity conditions:

$$\sqrt{nF_{ij}}(\tilde{\theta}_{ij} - \theta_{ij}) \to \mathcal{N}(0, 1).$$

While we do not have access to F_{ij}, due to its dependence on unknown parameters, it is natural to replace it by the empirical estimator $\frac{1}{n}\sum_{k=1}^{n} v_k^2 \ddot{f}(\hat{\mu}_i^{(k)})$. Therefore, we expect the following asymptotic normality result:

$$\left(\sum_{k=1}^{n} v_k^2 \ddot{f}\big(\hat{\mu}_i^{(k)}\big)\right)^{1/2} (\tilde{\theta}_{ij} - \theta_{ij}) \to \mathcal{N}(0, 1). \tag{3.6}$$

3.2. *Applications to three modified Poisson-type graphical models*. We apply the proposed general framework of statistical inference to the three modified Poisson-type graphical models described in Section 2. Our current method for modified-Poisson graphical models is an extension of Li et al. (2016), which only considered Ising graphical model, a special case of TPGM.

Each nodewise regression in Step 1 for all three models relies on the conditional distribution $\mathbb{P}_{\eta_i}(X_i|X_{-i})$. A more specific representation is provided as

$$\mathbb{P}_{\eta_i}(X_i|X_{-i}) = \frac{\exp[T(X_i)(\psi_i + \sum_{j\neq i}\theta_{ij}T(X_j)) + B(X_i))]}{\sum_{m=0}^{D_i}\exp[T(m)(\psi_i + \sum_{j\neq i}\theta_{ij}T(X_j)) + B(m)]}, \tag{3.7}$$

where the corresponding sufficient statistic and the base measure for each model are referred to Table 1. The threshold D_i for each X_i is finite in TPGM, while its value becomes $+\infty$ in the other two models.

The bias correction in Step 2 needs the knowledge of $f(\mu_i)$ which is the denominator of the right-hand side of (3.7) for the three models. Specifically, the expression of $f(\mu_i)$ for each of three models is shown in Table 2, and the details of corresponding $\dot{f}(\mu_i)$ and $\ddot{f}(\mu_i)$ are referred to Tables 3 and 4 in the Supplementary Material (Zhang et al. (2020)). Moreover, the expression of v_k in each model is based on the function $g(T(X_{-\{i,j\}}), \eta_i, \eta_j)$. In general, the expression of $g(T(X_{-\{i,j\}}), \eta_i, \eta_j)$ can be presented as

$$\begin{aligned} g\big(T(X_{-\{i,j\}}), \eta_i, \eta_j\big) &= \frac{\mathbb{E}_{\eta_i,\eta_j}[T(X_j)\ddot{f}(\mu_i)|T(X_{-\{i,j\}})]}{\mathbb{E}_{\eta_i,\eta_j}[\ddot{f}(\mu_i)|T(X_{-\{i,j\}})]} \\ &= \frac{\sum_{k_2=0}^{D_j}(T(k_2)\cdot \ddot{f}(\theta_{ij}T(k_2) + T(X^*_{-\{i,j\}})^{\intercal}\eta_{i,-j})\cdot Q)}{\sum_{k_2=0}^{D_j}(\ddot{f}(\theta_{ij}T(k_2) + T(X^*_{-\{i,j\}})^{\intercal}\eta_{i,-j})\cdot Q)} \end{aligned}$$

TABLE 2
Details of $f(\mu_i)$ in the three models

Model	$f(\mu_i)$
TPGM	$\log(\sum_{m=0}^{D_i}\exp(m\mu_i - \log(m!)))$
SPGM	$\log(\sum_{m=0}^{+\infty}\exp(S(m)\mu_i - \log(m!)))$
SqrtPGM	$\log(\sum_{m=0}^{+\infty}\exp(\sqrt{m}\mu_i - \log(m!)))$

TABLE 3
Details of $g(T(X_{-\{i,j\}}), \eta_i, \eta_j)$ in the three models

Model	$g(T(X_{-\{i,j\}}), \eta_i, \eta_j)$
TPGM	$\frac{\sum_{k_2=0}^{D_j}(k_2 \cdot \ddot{f}(\theta_{ij}k_2 + X_{-\{i,j\}}^{*\mathsf{T}}\eta_{i,-j}) \cdot Q)}{\sum_{k_2=0}^{D_j}(\ddot{f}(\theta_{ij}k_2 + X_{-\{i,j\}}^{*\mathsf{T}}\eta_{i,-j}) \cdot Q)}$
SPGM	$\frac{\sum_{k_2=0}^{+\infty}(S(k_2) \cdot \ddot{f}(\theta_{ij}S(k_2) + S(X_{\{-i,j\}}^{*})^{\mathsf{T}}\eta_{i,-j}) \cdot Q)}{\sum_{k_2=0}^{+\infty}(\ddot{f}(\theta_{ij}S(k_2) + S(X_{-\{i,j\}}^{*})^{\mathsf{T}}\eta_{i,-j}) \cdot Q)}$
SqrtPGM	$\frac{\sum_{k_2=0}^{+\infty}(\sqrt{k_2} \cdot \ddot{f}(\theta_{ij}\sqrt{k_2} + \sqrt{X_{-\{i,j\}}^{*}}^{\mathsf{T}}\eta_{i,-j}) \cdot Q)}{\sum_{k_2=0}^{+\infty}(\ddot{f}(\theta_{ij}\sqrt{k_2} + \sqrt{X_{-\{i,j\}}^{*}}^{\mathsf{T}}\eta_{i,-j}) \cdot Q)}$

with

$$Q = \sum_{k_1=0}^{D_i} \exp\big(T(k_1)T\big(X_{-\{i,j\}}^{*}\big)^{\mathsf{T}}\eta_{i,-j} + T(k_2)T\big(X_{-\{i,j\}}^{*}\big)^{\mathsf{T}}\eta_{j,-i} + B(k_1) + B(k_2) + \theta_{ij}T(k_1)T(k_2)\big),$$

where $\eta_{i,-j}$ is the subvector of η_i with θ_{ij} removed and $\eta_{j,-i}$ is the subvector of η_j with θ_{ji} removed. Specific $g(T(X_{-\{i,j\}}), \eta_i, \eta_j)$ for each model is summarized in Table 3, and the details of corresponding Q are shown in Table 5 in the Supplementary Material (Zhang et al. (2020)).

3.3. *Multiple testing with false discovery rate control.* If the structure of an overall graph is paid attention to, then there involves a multiple testing problem for all θ_{ij}'s

$$H_0 : \theta_{ij} = 0 \quad \text{vs.} \quad H_1 : \theta_{ij} \neq 0 \quad (1 \le i < j \le p) \tag{3.8}$$

that tests all pairs simultaneously. One of the most popular large-scale multiple testing procedures is the false discovery rate (FDR) analysis (Benjamini and Hochberg (1995)). It is well known that the false discovery rate $\mathrm{FDR}(t) = \mathbb{E}(\mathrm{FDP}(t))$ is the expectation of false discovery proportion (FDP), which is defined as

$$\mathrm{FDP}(t) = \frac{\sum_{(i,j)\in\mathcal{H}_0} I\{|\hat{T}_{ij}| \ge t\}}{\max\{\sum_{1\le i<j\le p} I\{|\hat{T}_{ij}| \ge t\}, 1\}}, \tag{3.9}$$

where $\hat{T}_{ij}$ is some generic test statistic for each individual hypothesis with a given threshold level t, $\mathcal{H}_0 = \{(i,j) : i < j, \theta_{ij} = 0\}$ denotes the set of true nulls (i.e., the edge set E), the numerator is the total number of false positives and the denominator is the total number of rejections. The numerator in (3.9) is generally unknown, but, under certain mild sparsity assumption of the underlying graph, one can estimate it by $2(1-\Phi(t))(p^2-p)/2$, as suggested in Liu (2013), where $\Phi(\cdot)$ is a standard normal CDF.

The test statistic in our case is $\hat{T}_{ij} = (\sum_{k=1}^{n} v_k^2 \ddot{f}(\hat{\mu}_i^{(k)}))^{\frac{1}{2}}\tilde{\theta}_{ij}$, a standardized version of $\tilde{\theta}_{ij}$. Following the idea in Liu (2013), we set a predefined level of FDR as $0 < \alpha < 1$ and choose the threshold of the test statistic as

$$\hat{t} = \inf\left\{0 \le t \le 2\sqrt{\log p} : \frac{2(1-\Phi(t))(p^2-p)/2}{\max\{\sum_{1\le i<j\le p} I\{|\hat{T}_{ij}| \ge t\}, 1\}} \le \alpha\right\}. \tag{3.10}$$

We reject H_0 in (3.8) if $|\hat{T}_{ij}| \ge \hat{t}$. If no $\hat{t}$ is chosen, we set $\hat{t} = 2\sqrt{\log p}$ as under null; the distribution of each $\hat{T}_{ij}$ is expected to be close to a standard normal such that the largest magnitude of $(p^2-p)/2$ statistics is no larger than $2\sqrt{\log p}$ with probability going to 1. Although

we do not provide any theory, we comment that with the constraint $t \le 2\sqrt{\log p}$ in (3.10), the weak dependency among all $\hat{T}_{ij}$'s will not influence the FDR control asymptotically. For further theoretical justification, please refer to Liu (2013).

4. Implementations for graph inference.

4.1. *Algorithm.* Step 1 involves a nodewise ℓ_1-penalized regression for each node X_i on all other nodes X_{-i}; see Ravikumar, Wainwright and Lafferty (2010), Yang et al. (2013). Here, the intercept ψ_i is excluded from the penalization. The total computational complexity of Step 1 is essentially equivalent to solving $O(p)$ ℓ_1-penalized regression problems. Each problem can be solved efficiently using the proximal gradient descent. Set $T(\boldsymbol{X}^*)$ as the $n \times (p+1)$ matrix with the kth row being $T(X^{*(k)})^\intercal$ for $k = 1, \ldots, n$. In addition, all the regressions rely on a single matrix with the ijth element being the inner product between the ith and jth columns of $T(\boldsymbol{X}^*)$ which includes $O(np^2)$ operations. The precalculation of this matrix can help avoid its repetitive calculation.

The bias correction for the parameter set $\boldsymbol{\Theta}$ in Step 2 has a total of $O(p^2)$ loops, and each loop for $\hat{\theta}_{ij}$ involves the calculation of inner products $\sum_{r \ne \{i,j\}} \hat{\theta}_{ir} T(X_r^{(k)})$ and $\sum_{r \ne \{i,j\}} \hat{\theta}_{jr} T(X_r^{(k)})$ for all $k = 1, \ldots, n$. The naïve matrix calculation tends to increase the computational complexity to $O(np^3)$. To simplify the computational steps, we precalculate the inner product between each $T(X_{-i}^{*(k)})$ and the initial estimators $\hat{\eta}_i$ and save the value in a prediction matrix which can be repetitively used for the inner product calculation within each loop. It can be seen that the precalculation of the prediction matrix helps reduce the computational complexity of these inner products to $O(np^2)$.

Besides the aforementioned implementations with high computational convenience, all the algorithms are achieved with the `Rcpp` library. Due to the lack of closed-form expressions for the normalization terms in the conditional distributions of SPGM and SqrtPGM, the numerical approximations that require a summation from zero to a large number are highly involved with many loop operations. The usage of `Rcpp` library, which incorporates the efficient `C++` code under the `R` environment, helps lower the computational burden for loops. In the end, we summarize all the steps of our two-step inference method in Algorithm 1.

4.2. *Selection of tuning parameters.* The tuning parameter λ_i in (3.1) controls the neighborhood sparsity of each node X_i or the number of edges extending out from X_i, so we need to select a sequence of λ_i with $i = 1, 2, \ldots, p$ for initial estimators in Step 1. According to the different purpose of inference, we provide two ways for selection of tuning parameters.

We at first focus on the inference of each individual θ_{ij}. The extended BIC (EBIC) criterion has been well studied under the regime of high-dimensional graphical models (Barber and Drton (2015)). We write EBIC for each regression as follows:

$$\text{EBIC}_\gamma(J) = 2l\big(\eta_i; \{X^{(k)}\}_{k=1}^n\big) + |J|\big(\log(n) + 2\gamma \log(p-1)\big), \tag{4.1}$$

where $l(\eta_i; \{X^{(k)}\}_{k=1}^n) = -\sum_{k=1}^n \log(\mathbb{P}_{\eta_i}(X_i^{(k)} | X_{-i}^{(k)}))$, $|J|$ is the cardinality of $J = \{j : j \ne i \text{ and } \hat{\theta}_{ij} \ne 0\}$ and some universal $\gamma \ge 0$. Following the suggestion in Barber and Drton (2015), we set $\gamma = 0.5$ as the default value in real implementations and select the tuning parameters that minimize (4.1).

For multiple testing, the tuning parameters are chosen as in Liu (2013) to guarantee $2(1 - \Phi(t))(p^2 - p)/2$ as close to $\sum_{(i,j) \in \mathcal{H}_0} I\{|\hat{T}_{ij}| \ge t\}$ as possible. We leave further details in Section 5 of the Supplementary Material (Zhang et al. (2020)).

To ensure the validity of EBIC to select tuning parameters, we further performed a comprehensive study of hyperparameter selection. We compared inferred networks between the

Algorithm 1 Statistical inference of the modified Poisson-type graphical models

- *Step 1: Initialization*
  1. Precalculate and save the inner product matrix, where the ijth element denotes the inner product between the column i and column j of $T(\boldsymbol{X}^*)$.
  2. For each X_i, $i = 1, 2, \ldots, p$, do nodewise ℓ_1-penalized regression (3.1).
  3. Obtain each initial estimator $\hat{\theta}_{ij}$ for Step 2.
- *Step 2: Likelihood-based Bias Correction*
  1. Precalculate and save the $n \times p$ prediction matrix $\boldsymbol{M}$, where each element $\hat{\mu}_i^{(k)}$ denotes the inner product of $T(X_{-i}^{*(k)})$ and $\hat{\eta}_i$ with $k = 1, 2, \ldots, n$ and $i = 1, 2, \ldots, p$.
  2. For each $i = 1, 2, \ldots, p-1$, do:
     (a) Calculate $\dot{f}(\hat{\mu}_i^{(k)})$ and $\ddot{f}(\hat{\mu}_i^{(k)})$ in (3.3) with $k = 1, \ldots, n$.
     (b) With each fixed i, for each $j = i+1, i+2, \ldots, p$, do:
        i. Calculate $q_1^{(k)} = \hat{\mu}_i^{(k)} - \hat{\theta}_{ij} T(X_j^{(k)})$ and $q_2^{(k)} = \hat{\mu}_j^{(k)} - \hat{\theta}_{ji} T(X_i^{(k)})$ with $k = 1, \ldots, n$.
        ii. Plug $q_1^{(k)}$ and $q_2^{(k)}$ into (3.5) to obtain the score vector V.
        iii. Generate the final estimator $\tilde{\theta}_{ij}$ in (3.3).
  3. Estimate the standard deviation, the 95% confidence interval, the p-value and the z-score for each $\tilde{\theta}_{ij}$.

proposed method and the sole estimation procedure with only nodewise ℓ_1-penalized regressions using EBIC and cross-validation under both simulation settings and the real data application in Sections 5–6. It is intriguing to notice that the proposed method is robust to different hyperparameter selection methods, while the sole estimation is very sensitive to various model selection criteria. Moreover, the proposed method can reach a better balance between false and true discoveries than the sole estimation based on different hyperparameter selection methods. More details are left in Section 6 of the Supplementary Material (Zhang et al. (2020)).

5. Simulations and a real example with ground truth. To show the validity of the proposed two-step procedure, we evaluated its performance from two-folds: 1. Asymptotic normality; 2. False discovery rate control for multiple testing. We considered four different graph settings: (a) the chain graph with two consecutive nodes arranged to be connected, (b) the grid graph (four-nearest neighbor graph) with nodes arranged to a lattice with maximal degree $d = 4$, (c) the Erdős–Rényi (E-R) random graph with average node degree $d = 4$ and (d) the scale-free network (Barabási and Albert (1999)). We generated random samples from the three modified Poisson-type models via Gibbs sampling (Zhang, Ouyang and Zhao (2017)). The first 5000 draws were discarded in the burn-in period. Then, we took one sample every 100 draws to guarantee independence. In addition, we also compared the proposed method to the popular Gaussian graphical model estimation with FDR control using Lasso (GFC_L) (Liu (2013)) by evaluating their performance on simulated RNA-seq data and a real example with ground truth.

5.1. *Asymptotic normality.* The lattice size of Grid graph is $\sqrt{p} \times \sqrt{p}$ here. For each given graph we generated 100 data sets with $n = 300$, $p = 100$ and 400, respectively, from the three models with each of nonzero entries drawn randomly from the set $(-0.4, -0.3, -0.2, -0.1, 0.1, 0.2, 0.3, 0.4)$. Other parameter details in the three models are described as below:

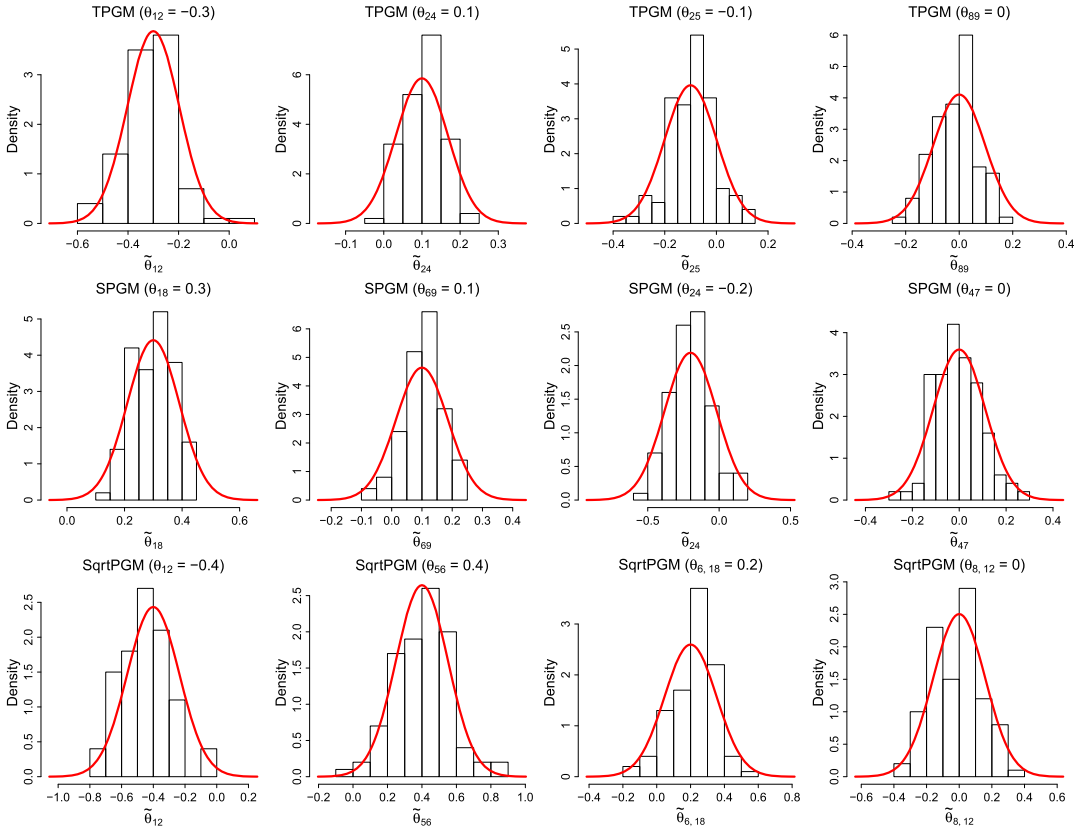

FIG. 1. *Histograms of the estimated pairwise entries for* $p = 400$ *from the three models in scale-free graph.*

- TPGM: The intercept term $\psi_i = 0$, and the threshold value $D_i = 3$.
- SPGM: The intercept term $\psi_i = -0.5$, and two threshold values $D_{i0} = 3$ and $D_{i1} = 6$.
- SqrtPGM: The intercept term $\psi_i = 0$.

The proposed estimates of each pairwise parameter were obtained based on Algorithm 1 with EBIC criterion for selection of tuning parameters in all the graph settings. Figure 1 shows the histograms of randomly selected entries that cover all possible values of true parameters from the three modified Poisson-type graphical models under high-dimensional settings with $p = 400$ for Scale-free graph. Each red curve is denoted as the approximate Gaussian density of a particular entry. It can be seen that the histograms of each entry match with the corresponding normal distribution very well. The histograms for scale-free graph with $p = 100$ and the other three graph settings are referred to Figures 5 to 10 in the Supplementary Material (Zhang et al. (2020)). Similarly, all the histograms of estimated entries are also in good accordance with their corresponding normal distributions.

The $(1 - \alpha)$ confidence interval for each θ_{ij} can be derived straightforwardly from the asymptotic normality in (3.6),

$$\left(\tilde{\theta}_{ij} - z_{\alpha/2}\left(\sum_{k=1}^{n} v_k^2 \ddot{f}(\hat{\mu}_i^{(k)})\right)^{-1/2}, \quad \tilde{\theta}_{ij} + z_{\alpha/2}\left(\sum_{k=1}^{n} v_k^2 \ddot{f}(\hat{\mu}_i^{(k)})\right)^{-1/2}\right),$$

where $z_{\alpha/2}$ is the z-score with the right tail probability equal to $\alpha/2$, that is, $P(\mathcal{N}(0,1) > z_{\alpha/2}) = \alpha/2$.

In addition, we also evaluated the performance of empirical coverage probabilities of the 95% confidence intervals of θ_{ij}'s to demonstrate the validity of our inference results. Considering the sparse structures of both graph settings, we separated all θ_{ij}'s into two sets: the

TABLE 4
Medians (standard deviations) of empirical coverage probabilities of the 95% confidence intervals in S_0 and S_0^c with $p = 400$

	S_0				S_0^c			
	Chain	Grid	E-R	Scale-free	Chain	Grid	E-R	Scale-free
				$n = 300$, $p = 400$				
TPGM	0.9436	0.9352	0.9284	0.9311	0.9511	0.9508	0.9497	0.9501
	(0.0118)	(0.0087)	(0.0081)	(0.0117)	(0.0010)	(0.0010)	(0.0012)	(0.0009)
SPGM	0.9499	0.9153	0.8910	0.9135	0.9528	0.9536	0.9519	0.9505
	(0.0118)	(0.0087)	(0.0100)	(0.0109)	(0.0010)	(0.0011)	(0.0011)	(0.0009)
SqrtPGM	0.9524	0.9512	0.9512	0.9549	0.9512	0.9501	0.9501	0.9510
	(0.0110)	(0.0087)	(0.0087)	(0.0108)	(0.0009)	(0.0009)	(0.0010)	(0.0008)

edge S_0 and nonedge S_0^c:

$$S_0 = \{(i, j) : \theta_{ij} \neq 0\}, \qquad S_0^c = \{(i, j) : \theta_{ij} = 0\}.$$

Then, based on all the estimates $\tilde{\theta}_{ij}$'s, the average empirical coverage probabilities of the 95% confidence intervals were evaluated in S_0 and S_0^c, respectively. Table 4 reports the medians (standard deviations) of average empirical coverage probabilities of the 95% confidence intervals over 100 replications for $p = 400$. As we can see, all results are close to 0.95, the target confidence level. Additional results towards individual inference with $p = 100$, $n = 150$ and 100 and, using settings in Section 5.2, are summarized in Tables 11–14 in the Supplementary Material (Zhang et al. (2020)).

5.2. *False discovery rate control for multiple testing*. To evaluate the performance of our estimates for multiple testing with false discovery rate (FDR) control, we considered the four graphs with a two-block structure. More specifically, the first half of nodes form one block, leaving the remaining nodes as another block. Two cases were evaluated: $p = 200$ and 400. The detailed parameter settings are described as below:

- TPGM: Each of nonzero entries is randomly drawn: (i) either -0.3 or 0.3 in Block 1; (ii) either -0.4 or 0.4 in Block 2. For both blocks, each intercept term $\psi_i = -0.5$, and each threshold value $D_i = 3$.
- SPGM: Each of nonzero entries is randomly drawn: (i) either -0.3 or 0.3 in Block 1; (ii) either -0.4 or 0.4 in Block 2. For two blocks the intercept term ψ_i is: (i) -0.5 for chain and scale-free graphs; (ii) -1 for grid and E-R graphs. Two threshold values $D_{i0} = 2$ and $D_{i1} = 5$.
- SqrtPGM: Each of nonzero entries is randomly drawn: (i) either -0.6 or 0.6 in Block 1; (ii) either -0.9 or 0.9 in Block 2. The intercept term $\psi_i = 0$.

For each of the two cases, we generated 100 data sets with $n = 400$. We investigated the performance of our procedure by evaluating true positive rate (TPR) and false positive rate (FPR) over a range of FDR control levels. Here, we used the tuning selection scheme, described in Section 5 of the Supplementary Material (Zhang et al. (2020)), for multiple testing. To compare, we also applied the sole estimation procedure with nodewise ℓ_1-penalized regressions in each same data set through a range of regularization parameters. The medians of TPRs and FPRs at each cut-off over 100 replications from the two procedures are presented in the receiver operating characteristic (ROC) curves for $p = 400$, as shown in Figure 2. It can be seen that all curves from the proposed inferential procedure lie above the ones from

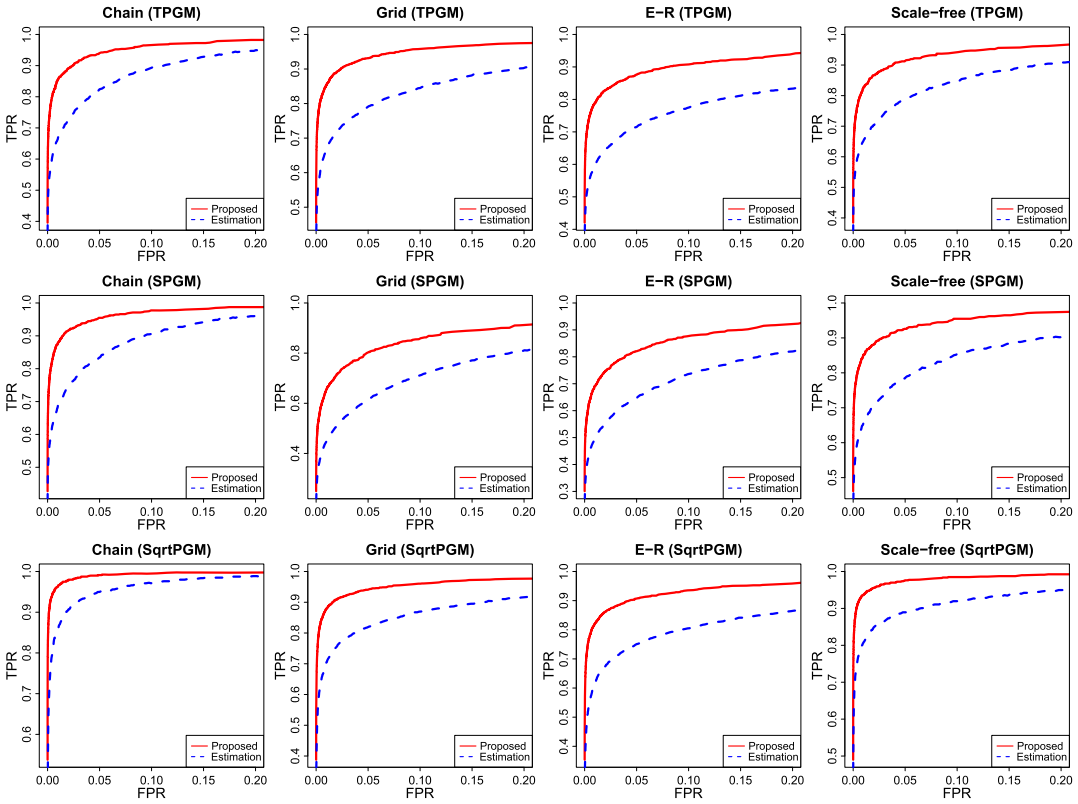

FIG. 2. *ROC curves based on TPRs and FPRs for the proposed inferential procedure and the sole estimation in the case of* $p = 400$.

the sole estimation and show noticeably better performance in detecting true conditional dependency while simultaneously maintaining false discovers at a low level. ROC curves for $p = 200$ which share similar patterns are shown in Figure 11 in the Supplementary Material (Zhang et al. (2020)).

Furthermore, we report the medians (standard deviations) of empirical FDRs with prespecified levels 0.1 and 0.2 for both $p = 200$ and 400 in Table 5. The medians (standard deviations) of their corresponding power values are shown in Table 6. The empirical FDRs are well controlled at the desired levels with a relatively good performance of power. Additional simulation results toward global inference with $n = 150$ are summarized in Tables 15–16 in the Supplementary Material (Zhang et al. (2020)).

5.3. *Evaluation on simulated RNA-seq data.* We further evaluated the performance of the proposed method by comparing it to GFC_L on simulated RNA-seq data. RNA-seq data was simulated based on the following steps:

(a) *(Incorporation of conditional dependence)* A simulated normalized count-valued RNA-seq data set X with n samples and p genes is generated via Gibbs sampling from SqrtPGM with scale-free graph.

(b) *(Pseudo-random number addition)* A randomly generated number from uniform distribution between zero and one is added to each element of the simulated count data to ensure randomness.

(c) *(Inverse power transform)* Inverse power transform is performed on X^β for $0 < \beta < 1$ with the value of β from real data applications.

TABLE 5
Medians (standard deviations) of empirical false discovery rates

	$\alpha = 0.1$				$\alpha = 0.2$			
	Chain	Grid	E-R	Scale-free	Chain	Grid	E-R	Scale-free
				$n = 400,\ p = 200$				
TPGM	0.0892	0.0948	0.0939	0.0939	0.1777	0.1794	0.1792	0.1856
	(0.0277)	(0.0188)	(0.0189)	(0.0246)	(0.0397)	(0.0260)	(0.0277)	(0.0338)
SPGM	0.0858	0.0840	0.0964	0.0938	0.1744	0.1623	0.1760	0.1895
	(0.0259)	(0.0209)	(0.0222)	(0.0253)	(0.0327)	(0.0261)	(0.0299)	(0.0361)
SqrtPGM	0.0884	0.0903	0.0955	0.0956	0.1762	0.1784	0.1721	0.1810
	(0.0237)	(0.0323)	(0.0277)	(0.0262)	(0.0327)	(0.0318)	(0.0368)	(0.0346)
				$n = 400,\ p = 400$				
TPGM	0.0940	0.1007	0.1120	0.0937	0.1866	0.1931	0.2052	0.1939
	(0.0142)	(0.0128)	(0.0187)	(0.0225)	(0.0205)	(0.0204)	(0.0213)	(0.0251)
SPGM	0.0998	0.1154	0.1159	0.1054	0.1852	0.2032	0.2145	0.2092
	(0.0212)	(0.0173)	(0.0141)	(0.0242)	(0.0218)	(0.0231)	(0.0201)	(0.0249)
SqrtPGM	0.0986	0.0907	0.0997	0.0976	0.2016	0.1818	0.1885	0.2021
	(0.0197)	(0.0117)	(0.0123)	(0.0176)	(0.0280)	(0.0174)	(0.0152)	(0.0254)

(d) *(Final count generation)* Elements are rounded down to its nearest integer to obtain the final simulated RNA-seq data set.

Here, we considered the two-block scale-free graph in Section 5.2. The step (c) was motivated by the preprocessing steps in Allen and Liu (2013) on original RNA-seq data. The preprocessing on our motivating count-valued RNA-seq data of childhood atopic asthma returned $\beta = 0.2517$.

We used the proposed procedure to generate 100 simulated RNA-seq data sets with $n = 300$ and $p = 400$. Figures 3(A) and 3(B) demonstrate the histograms of count values from the real data set of childhood atopic asthma and a typical simulated RNA-seq data set. As it can be seen, their distribution shapes are quite close to each other with decaying proportions of large count values. Before implementing the proposed method, we took a power

TABLE 6
Medians (standard deviations) of power values for corresponding FDR control levels

	$\alpha = 0.1$				$\alpha = 0.2$			
	Chain	Grid	E-R	Scale-free	Chain	Grid	E-R	Scale-free
				$n = 400,\ p = 200$				
TPGM	0.7222	0.7116	0.7867	0.7727	0.7828	0.7619	0.8329	0.8131
	(0.0236)	(0.0215)	(0.0197)	(0.0230)	(0.0234)	(0.0204)	(0.0161)	(0.0215)
SPGM	0.7172	0.4881	0.4746	0.7449	0.7677	0.5556	0.5278	0.7879
	(0.0226)	(0.0256)	(0.0228)	(0.0247)	(0.0241)	(0.0238)	(0.0241)	(0.0238)
SqrtPGM	0.8838	0.7460	0.6613	0.8939	0.9192	0.8069	0.7295	0.9242
	(0.0238)	(0.0227)	(0.0336)	(0.0198)	(0.0197)	(0.0218)	(0.0327)	(0.0167)
				$n = 400,\ p = 400$				
TPGM	0.6357	0.7131	0.6463	0.6131	0.6910	0.7592	0.6920	0.6633
	(0.0177)	(0.0114)	(0.0123)	(0.0164)	(0.0183)	(0.0113)	(0.0122)	(0.0168)
SPGM	0.6646	0.4347	0.5087	0.6796	0.7186	0.4908	0.5538	0.7236
	(0.0163)	(0.0161)	(0.0157)	(0.0166)	(0.0188)	(0.0150)	(0.0154)	(0.0179)
SqrtPGM	0.8492	0.6966	0.6467	0.8065	0.8920	0.7652	0.7107	0.8492
	(0.0185)	(0.0161)	(0.0141)	(0.0185)	(0.0157)	(0.0146)	(0.0138)	(0.0185)

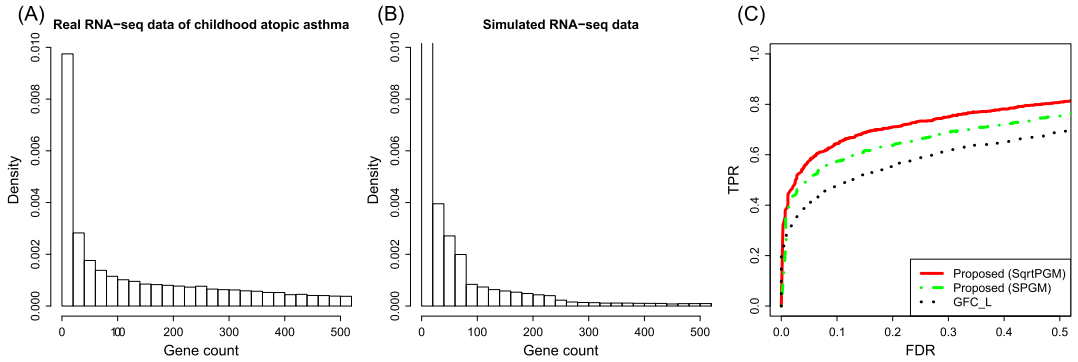

FIG. 3. *(A) Histogram of real RNA-seq data of childhood atopic asthma. (B) Histogram of typical simulated RNA-seq data. (C) ROC-type curves based on TPRs and FDRs for the proposed inferential procedure (SqrtPGM and SPGM) and GFC_L on simulated RNA-seq data.*

transformation on the simulated RNA-seq data with $\beta = 0.2517$ and rounded down each element of data matrices to its nearest integer. Before implementing GFC_L, we performed a log and nonparanormal transformation (Liu, Lafferty and Wasserman (2009)) to continue and Gaussianize the simulated data sets (Jia et al. (2017)) using the R package `huge` (Zhao et al. (2012)). We implemented our proposed method using SqrtPGM and SPGM because they are more general Poisson-type distributions than TPGM. GFC_L was implemented with the `R` package `SILGGM` (Zhang, Ren and Chen (2018)).

The evaluation of methods depends on the performance of TPR over a range of varying FDR control levels. The medians of TPRs and FDRs over 100 replications from our approach (SqrtPGM and SPGM) and GFC_L are illustrated in the ROC-type curves in Figure 3(C). Both curves from our approach with SqrtPGM and SPGM lie above the one from GFC_L which indicates that our proposed method is noticeably more capable of capturing built-in features than GFC_L while controlling FDRs around same levels. Furthermore, we also reported all the medians (standard deviations) of empirical FDRs and corresponding power values with prespecified levels $\alpha = 0.1$ and 0.2 in Table 7. The power values from the proposed method with SqrtPGM and SPGM are both greater than those from GFC_L while all the empirical FDRs are very similar. Additional results with $n = 150$ are shown in Table 17 in the Supplementary Material (Zhang et al. (2020)).

5.4. *A real example with ground truth*: *Liver cytochrome P*450*s*. We also evaluated the validity of our proposed approach by comparing with GFC_L on a real example with established ground truth. It is a count-valued RNA-seq data set for a liver cytochrome P450s

TABLE 7
Medians (standard deviations) of empirical FDRs and power values from our proposed method (SqrtPGM and SPGM) and GFC_L on simulated RNA-seq data with FDR controlled at levels $\alpha = 0.1$ *and* 0.2

Proposed (SqrtPGM)		Proposed (SPGM)		GFC_L	
FDR	Power	FDR	Power	FDR	Power
		$\alpha = 0.1$			
0.1023	0.6445	0.1143	0.5842	0.0993	0.4786
(0.0181)	(0.0259)	(0.0191)	(0.0224)	(0.0250)	(0.0237)
		$\alpha = 0.2$			
0.1965	0.7085	0.2295	0.6508	0.2018	0.5553
(0.0234)	(0.0240)	(0.0208)	(0.0224)	(0.0248)	(0.0236)

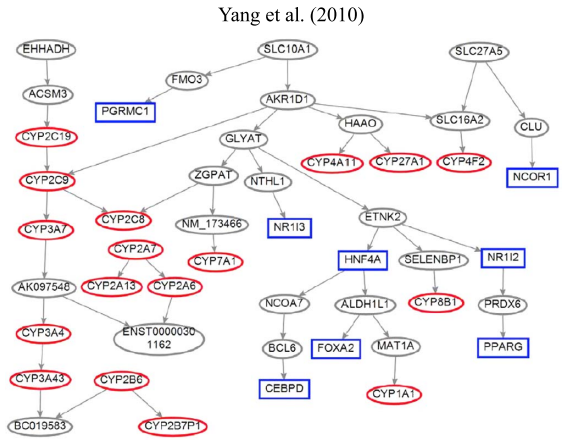

FIG. 4. *The subnetwork of P450 regulatory system from Yang et al. (2010) with the known regulators and P450 genes shown in blue rectangles and red ovals, respectively.*

subnetwork from humans with $n = 100$ samples and $p = 44$ genes, and we downloaded the data from the Supplementary Material of Jia et al. (2017). Through experimental work, Yang et al. (2010) uncovered a subnetwork of P450 regulatory system for human liver shown in Figure 4.

At first, we evaluated the scale-free (or power-law) topology of inferred networks because a biological network generally has a scale-free pattern (Almaas and Barabási (2006)). After implementing the preprocessing steps in Allen and Liu (2013), we performed the proposed approach using TPGM, SPGM and SqrtPGM with FDR control at levels 0.001, 0.005, 0.01, 0.05, 0.1 and 0.15. We implemented GFC_L on the data after a log and nonparanormal transformation on the original count values at same levels of FDR control. The power law can be described as $p(\lambda) \propto \lambda^{-\alpha}$, where λ and $p(\lambda)$ are denoted as node degree and its corresponding probability, and α is a positive number. A correlation value between the $\log 2$ of node degree and the $\log 2$ of its corresponding probability closer to -1 indicates a better conformation to the power law. Speaking overall, our proposed approach generates networks that are more consistently follow scale-free topology than GFC_L, as shown in Table 8. The correlation values from our approach with TPGM and SqrtPGM are consistently around -0.8 or -0.9. When prespecified levels become 0.1 and 0.15, the inferred networks from GFC_L are highly deviated from a scale-free pattern with correlation values of -0.4322 and 0.1344.

We then evaluated the identified gene interactions from inferred networks of GFC_L and the proposed approach with FDR control at level 0.001 because all of them follow scale-free topology well with negative correlation values stronger than -0.9. In terms of the proposed approach, we focused on the results with SPGM, due to its slightly better performance

TABLE 8
Correlations between the $\log 2$ *of node degree and the* $\log 2$ *of its corresponding probability of inferred networks*

	FDR level					
	0.001	0.005	0.01	0.05	0.10	0.15
GFC_L	−0.9304	−0.9304	−0.9304	−0.7734	−0.4322	0.1344
Proposed (TPGM)	−0.9395	−0.9000	−0.9000	−0.8788	−0.8192	−0.8041
Proposed (SPGM)	−0.9105	−0.8216	−0.7007	−0.6336	−0.5887	−0.5999
Proposed (SqrtPGM)	−0.9398	−0.9393	−0.9278	−0.9151	−0.9342	−0.8439

TABLE 9
The identified gene interactions that overlap the subnetwork from Yang et al. (2010) by GFC_L and the proposed approach with SPGM

GFC_L	Proposed (SPGM)
CYP3A4 —— CYP3A43	CYP3A4 —— CYP3A43
CYP2A7 —— CYP2A13	CYP2A7 —— CYP2A13
AKR1D1 —— GLYAT	CYP2C9 —— CYP2C8
NCOA7 —— BCL6	NCOA7 —— BCL6
ETNK2 —— NR1I2	CYP2B6 —— CYP2B7P1
	CYP2A7 —— CYP2A6
	FMO3 —— SLC10A1
	SLC10A1 —— AKR1D1

than TPGM and SqrtPGM. We listed the identified gene interactions that overlap the subnetwork from Yang et al. (2010) in Figure 4 from our proposed approach with SPGM and GFC_L in Table 9. As it can be seen, our proposed approach with SPGM can capture most of these interactions identified from GFC_L, for example, CYP3A4 and CYP3A43, BCL6 and NCOA7, and CYP2A13 and CYP2A7. Moreover, the proposed approach with SPGM can recover more functionally important interactions between the known cytochrome P450 genes shown in red ovals in Figure 4 than GFC_L, for example, CYP2C8 and CYP2C9, CYP2B6 and CYP2B7P1, and CYP2A6 and CYP2A7.

6. Application to RNA-seq data of childhood allergic asthma. We applied our proposed approach to the motivating RNA-seq gene expression data that illustrates the count-valued transcripts of genes from the nasal epithelial cells of $n = 157$ children (62 females and 95 males) with allergic asthma in Puerto Ricans. These children have an average age of 15.3 years with a median total IgE (Immunoglobulin E) of 372 IU/mL. More detailed demographic information of these children is deferred to Table 1 in the Supplementary Material (Zhang et al. (2020)). Before using the proposed approach, we normalized the RNA-seq data following the preprocessing steps described in Allen and Liu (2013). After preprocessing, the normalized data includes $p = 500$ genes and is more approachable to a Poisson-type distribution than the original one; see Figure 12 in the Supplementary Material (Zhang et al. (2020)).

We inferred gene network using the proposed method in the three models with FDR control at level 0.001. As comparison studies, we also constructed gene network using only the estimation results from Step 1 of the procedure. To evaluate the overall network structure with scale-free (or power-law) topology, Figure 5 illustrates the $\log 2$-$\log 2$ plots of node degree distribution for inferred networks and their corresponding correlation measurements. As it can be seen, the correlation values based on the proposed inferential procedure are all around -0.9 and much closer to -1 in TPGM and SPGM, while the values are comparable with the sole estimation in SqrtPGM. Although the correlation values are still good, the sole estimation generally leads to a much sparser network and fails to capture complex coexpression structures, particularly for TPGM which has a maximum node degree of 1 and barely demonstrates any informative interactions. Additional evaluations of the overall network structure based on the two graphical model methods under normal assumption are shown in Figure 13 in the Supplementary Material (Zhang et al. (2020)). Due to their failure to conform the power law, we did not include them for further analysis.

In addition to evaluating the overall network structure, we also studied community structure of all the inferred networks, using the eigenspectrum of the modularity matrix (Newman

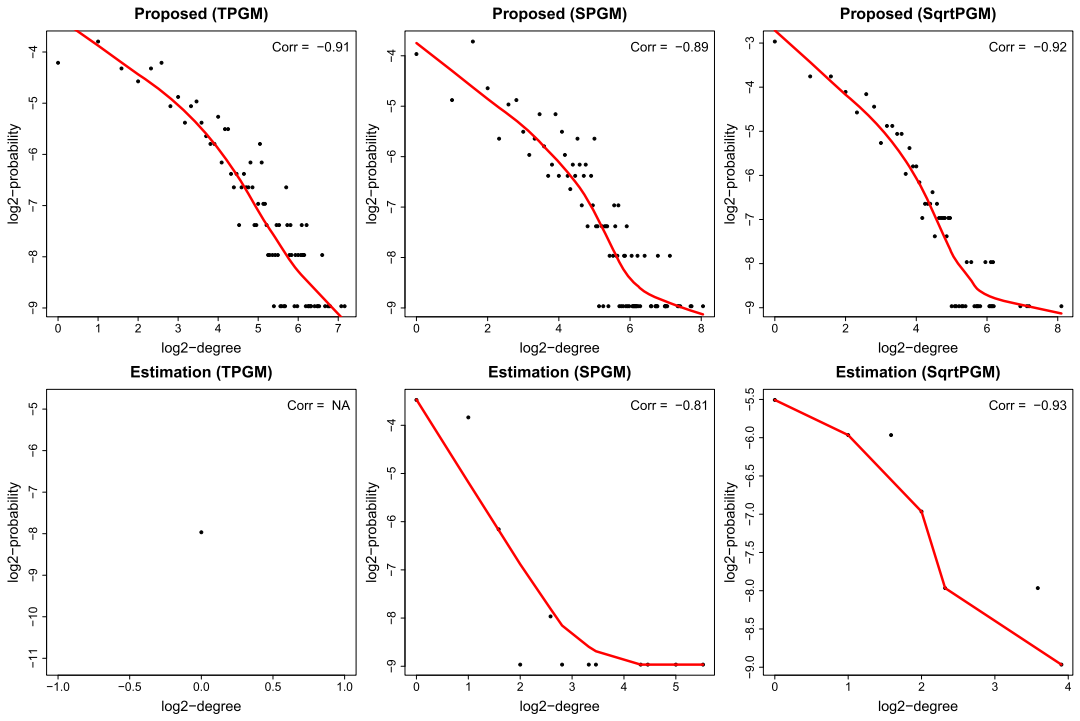

FIG. 5. *The* log 2-log 2 *plots of degree distribution for the inferred networks* (*NA*: *Not available*).

(2006)), to explore important gene pathways within the identified gene modules to atopic asthma. Besides the aforementioned methods with three modified Poisson-type models, we included GFC_L and the nonparanormal SKEPTIC estimator (Liu et al. (2012)) as comparison studies. To ensure the fairness in comparison, we extracted the same 500 genes from the original data and made a log and nonparanormal transformation to the count values before the use of GFC_L with FDR control at level 0.001. For nonparanormal SKEPTIC we obtained Spearman's rho statistics from the original count-valued data with 500 genes and implemented the graphical Lasso to estimate networks. The resulting estimated graph was finally selected by the EBIC criterion. Table 10 demonstrates the identified big gene modules with a size of at least 30 genes from the inferred gene networks. It can be seen that the proposed method successfully detects two to four big gene modules among three models, while the sole estimation in TPGM and SqrtPGM, GFC_L and nonparanormal SKEPTIC fail to identify informative ones.

TABLE 10
The big gene modules identified by different approaches (*NA*: *No modules with a size of at least* 30 *genes available*)

Method	Size of big modules	Number of big modules
Proposed (TPGM)	312, 169	2
Proposed (SPGM)	229, 75, 120	3
Proposed (SqrtPGM)	48, 164, 120, 114	4
Estimation (TPGM)	NA	0
Estimation (SPGM)	49, 32	2
Estimation (SqrtPGM)	NA	0
GFC_L	NA	0
Nonparanormal SKEPTIC	NA	0

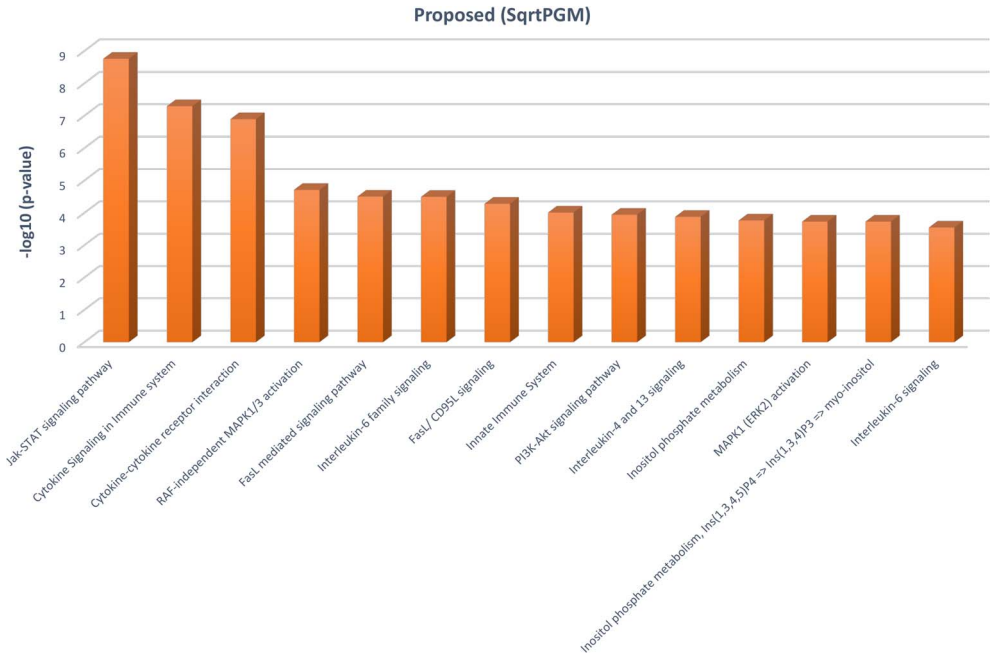

FIG. 6. *Some enriched pathways from the proposed inferential procedure in SqrtPGM.*

We further performed gene pathway enrichment analysis on the identified big modules in Table 10 using `ToppGene Suite` (Chen et al. (2009)) with FDR control at level 0.05; see Table 19 in the Supplementary Material (Zhang et al. (2020)) for complete results. From those modules identified by the proposed inferential procedure, we found some pathways that are shared within three models and critical to atopic asthma, for example, metal sequestration by antimicrobial proteins and FasL/CD95L signaling. The antimicrobial activity of S100A8/A9 proteins can induce a metal-witholding response by starving pathogens with metal nutrients in inflamed upper airway due to the chronic autoimmune diseases like asthma, according to Van Crombruggen et al. (2016). The potential role of Fas and its ligand (FasL) signaling pathway in T helper type 2 (Th2) cells for asthma has been intensively studied (Potapinska and Demkow (2009), Williams et al. (2018)). More interestingly, we also noticed some unique pathways enriched from the different modules in the three models, as shown in Figure 6 for SqrtPGM and Figure 14 in the Supplementary Material (Zhang et al. (2020)) for TPGM and SPGM. CLEC7A/inflammasome pathway (Hadebe, Brombacher and Brown (2018)) from TPGM and TRAIL signaling pathway (Braithwaite, Marriott and Lawrie (2018)) from SPGM are important in regulating immune responses to atopic asthma. More unique pathways were enriched from SqrtPGM, such as Jak-STAT signaling pathway and Interleukin-4 and 13 (IL-4/IL-13) signaling pathway. Jak-STAT signaling pathway has been shown to play an important role in the development of atopic asthma by differentiating Th2 cells from naïve T cells (Vale (2016)) and regulating the level of IgE (Zhang et al. (2018)). IL-4/IL-13 signaling pathway is central for IgE regulation, and genetic alterations in this pathway reveals its significance to the development of childhood atopic asthma (Kabesch et al. (2006)). However, the identified gene modules from the sole estimation are not capable of reflecting critical gene pathways about allergic asthma compared with the proposed inferential procedure. The gene interactions of the module in SqrtPGM with enriched JAK-STAT signaling pathway as well as their corresponding interactions in TPGM and SPGM are further presented in Figure 15 in the Supplementary Material (Zhang et al. (2020)).

Then, we investigated interactions among the 12 genes included in the Jak-STAT signaling pathway which is the most significant enriched pathway from SqrtPGM and also the one

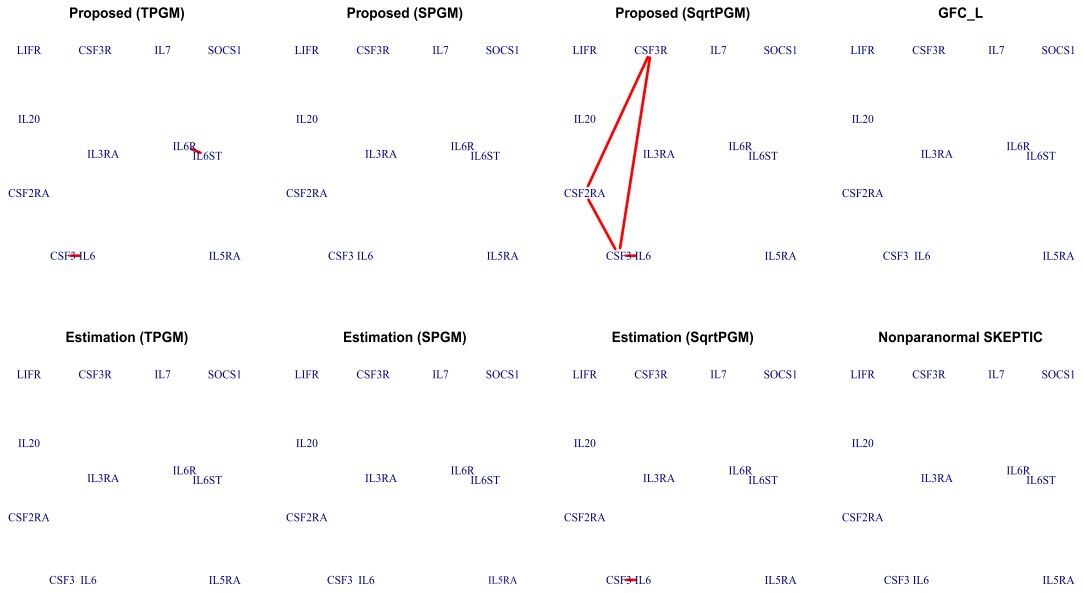

FIG. 7. *The inferred interactions of genes within the Jak-STAT signaling pathway.*

enriched using a total of 500 genes with FDR control at the 0.05 level; see Table 20 in the Supplementary Material (Zhang et al. (2020)). Targeting this pathway will be therapeutically effective on asthma pathology (Vale (2016)). The inferred gene interactions from our procedure, the sole estimation, GFC_L and nonparanormal SKEPTIC are demonstrated at a fixed panel in Figure 7. As we can see, the sole estimation, GFC_L and nonparanormal SKEPTIC can barely detect any informative interactions, except the one between IL6 and CSF3. Conversely, our procedure is capable of identifying more meaningful interactions related to atopic asthma in addition to the one between IL6 and CSF3, for example, IL6R and IL6ST from TPGM, and CSF3 and CSF3R from SqrtPGM. The activation of IL6R requires an association with IL6ST so as to regulate the immune response. CSF3R, which is associated with asthma, is known as the receptor for CSF3 and should be involved in granulopoiesis during the inflammatory process. According to a more recent study in Wang et al. (2019), CSF3 is identified as a major effector that promotes infection-dependent transition to severe asthma, and inhibition of CSF3R can be a potential strategy for preventing the pathological inflammation. These suggest that the sole estimation, GFC_L and nonparanormal SKEPTIC may neglect important functional relationships between genes closely related to atopic asthma.

Last, but not least, the inferred networks from our approach can capture important hub genes that are closely associated with asthma and allergy, for example, NTRK2 (21 and 50 connections to other genes in SPGM and SqrtPGM, respectively) and GSN (27, 23 and 71 connections to other genes in TPGM, SPGM and SqrtPGM, respectively). The two genes are also listed as the top differentially expressed genes in Forno et al. (2020) and are well replicated by the two external cohorts from Giovannini-Chami et al. (2012) and Yang et al. (2017). However, the sole estimation, GFC_L and nonparanormal SKEPTIC fail to identify these important hub genes.

In summary, our refined inference is more useful compared to the sole estimation, GFC_L and nonparanormal SKEPTIC. It not only reveals more significant pathways related to atopic asthma but also captures more complex gene coexpression structures and important hub genes. The sole estimation and the methods based on Gaussian and nonparanormal graphical models may lead to information loss and are less powerful to obtain informative disease-relevant results. Therefore, our procedure can be potentially useful for new treatment development in atopic asthma. To further demonstrate the advantages of our proposed method,

we performed additional analysis of GFC_L on transcript per million (TPM) values from RNA-seq data; see Section 8.5 of the Supplementary Material (Zhang et al. (2020)) for more details.

7. Conclusion and discussion. We have developed a novel procedure for statistical inference of three modified Poisson-type graphical models which provides reliable confidence intervals and p-values of pairwise edge and desirable false discovery rate control of multiple edges to tailor the network analysis of nonnegative, discrete and high-dimensional data. The procedure essentially relies on the intrinsic property of graphical models and is different from the existing regression-based bias correction. Compared to the sole estimation approach, the proposed method is robust to different hyperparameter selection criteria, which results in its noticeably better performance in inferring a more biologically meaningful network by identifying more true signals while simultaneously controlling false discovers at a reasonably low level. Compared to the application of graphical model methods under normal and nonparanormal assumptions, the proposed method tends to reveal more biological meaningful networks and is more capable of capturing important gene interactions with less information loss. From Yang et al. (2013) they mentioned another modified Poisson-type model called quadratic Poisson graphical model (QPGM). However, unlike the desired Poisson tail, QPGM is more like Gaussian distribution and has Gaussian-esque thin tail. Due to this major drawback, we do not consider QPGM here.

The proposed method can be applied to more different types of omics data even though it is mainly motivated by the count-valued RNA-seq, for example, DNA copy number variation (CNV) data and single nucleotide polymorphism (SNP) data for genomics. In practice, TPGM is more suitable for the context with a relatively small range of discrete values, such as CNV or SNP data. For RNA-seq data, which generally has much larger discrete values, we recommend to explore SPGM or SqrtPGM first because they are more general Poisson-type distributions and allow a broader set of feasible parameters for pairwise conditional dependence than TPGM. Indeed, when the upper bound D_i of TPGM becomes larger, the behavior of TPGM tends to be closer to original PGM which suffers from the limitation of negative pairwise dependency. With sufficient computational resource we suggest to explore all three models by comparing the results of overall network inference, gene modularity detection and gene network construction for important pathways, according to different purpose of each study.

There are several limitations of our proposed method which need future study as well. On the one hand, the proposed method is not symmetric between i and j in estimating each θ_{ij} and, generally, depends on the ordering of variables. One can naively apply a sample splitting scheme for symmetrization. More specifically, we randomly split the data into two halves. Then, for each fixed pair $i < j$, we fit the first half of the data into our method to obtain estimator $\tilde{\theta}_{ij}$ and then apply the second half to our method with i and j switched to obtain $\tilde{\theta}_{ji}$. The final asymptotically normal estimator is the average of these two independent estimators $\tilde{\theta}_{ij}^{\mathrm{sym}} = (\tilde{\theta}_{ij} + \tilde{\theta}_{ji})/2$. However, both that sample splitting scheme only uses part of the data for inference and that the result depends on the random split of the data make it less preferred in practice. Some preliminary analysis suggests that sample splitting is not necessary for asymptotic normality of $\tilde{\theta}_{ij}^{\mathrm{sym}}$, but the dependency between $\tilde{\theta}_{ij}$ and $\tilde{\theta}_{ji}$, obtained with the same entire samples, requires a refined theoretical analysis. We thus leave it as a future study. On the other hand, our method allows only a single discrete-type data set as an input. Due to the increasing popularity of multiomics study, the integrative network analysis of multilayered data sets with both continuous and discrete values is a promising future direction. To this end, we will further expand our procedure to more generalized or mixed-type exponential family graphical models as a future work.

Acknowledgments. The authors are grateful to the four anonymous referees, an Associate Editor and the Editor for their highly valuable comments that improved the quality of this paper.

Funding. The first and second authors were supported in part by NSF Grant DMS-1812030.

The third and fourth authors were supported by NIH Grants HL079966, HL117191, and MD011764.

SUPPLEMENTARY MATERIAL

Supplementary materialary material to "Inference of large modified Poisson-type graphical models: Application to RNA-seq data in childhood atopic asthma studies" (DOI: 10.1214/20-AOAS1413SUPPA; .pdf). In this supplement, we provide more technical details for our inference procedure, and additional results for simulations and the real application.

Code (DOI: 10.1214/20-AOAS1413SUPPB; .zip). `R` package `ModPGMInference`, the package manual, code and data for simulations and real data applications contained in this paper. In the future, `ModPGMInference` will be maintained and updated on the GitHub repository: https://github.com/zhangr100/ModPGMInference.

REFERENCES

ALLEN, G. I. and LIU, Z. (2013). A local Poisson graphical model for inferring networks from sequencing data. *IEEE Trans. Nanobiosci.* **12** 189–198.

ALMAAS, E. and BARABÁSI, A.-L. (2006). Power laws in biological networks. In *Power Laws, Scale-Free Networks and Genome Biology* 1–11. Springer, Berlin.

ANDERS, S. and HUBER, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* **11** R106. https://doi.org/10.1186/gb-2010-11-10-r106

BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. MR2091634 https://doi.org/10.1126/science.286.5439.509

BARBER, R. F. and DRTON, M. (2015). High-dimensional Ising model selection with Bayesian information criteria. *Electron. J. Stat.* **9** 567–607. MR3326135 https://doi.org/10.1214/15-EJS1012

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236. MR0373208

BRAITHWAITE, A. T., MARRIOTT, H. M. and LAWRIE, A. (2018). Divergent roles for TRAIL in lung diseases. *Front. Med.* **5** 212. https://doi.org/10.3389/fmed.2018.00212

CAI, T. T., LI, H., MA, J. and XIA, Y. (2019). Differential Markov random field analysis with an application to detecting differential microbial community networks. *Biometrika* **106** 401–416. MR3949311 https://doi.org/10.1093/biomet/asz012

CHEN, J., BARDES, E. E., ARONOW, B. J. and JEGGA, A. G. (2009). ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* **37** W305–W311. https://doi.org/10.1093/nar/gkp427

FORNO, E., WANG, T., QI, C., YAN, Q., XU, C.-J., BOUTAOUI, N., HAN, Y.-Y., WEEKS, D. E., JIANG, Y. et al. (2019). DNA methylation in nasal epithelium, atopy, and atopic asthma in children: A genome-wide study. *Lancet Respir. Med.* **7** 336–346.

FORNO, E., ZHANG, R., JIANG, Y., KIM, S., YAN, Q., REN, Z., HAN, Y.-Y., BOUTAOUI, N., ROSSER, F. et al. (2020). Transcriptome-wide and differential expression network analyses of childhood asthma in nasal epithelium. *J. Allergy Clin. Immunol.* **146** 671–675.

GIOVANNINI-CHAMI, L., MARCET, B., MOREILHON, C., CHEVALIER, B., ILLIE, M. I., LEBRIGAND, K., ROBBE-SERMESANT, K., BOURRIER, T., MICHIELS, J.-F. et al. (2012). Distinct epithelial gene expression phenotypes in childhood respiratory allergy. *Eur. Respir. J.* **39** 1197–1205.

HADEBE, S., BROMBACHER, F. and BROWN, G. D. (2018). C-type lectins receptors in asthma. *Front. Immunol.* **9** 733.

INOUYE, D., RAVIKUMAR, P. and DHILLON, I. (2016). Square root graphical models: Multivariate generalizations of univariate exponential families that permit positive dependencies. In *International Conference on Machine Learning* 2445–2453.

ISLAM, S., ZEISEL, A., JOOST, S., LA MANNO, G., ZAJAC, P., KASPER, M., LÖNNERBERG, P. and LINNARSSON, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11** 163–166.

JANKOVÁ, J. and VAN DE GEER, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Stat.* **9** 1205–1229. MR3354336 https://doi.org/10.1214/15-EJS1031

JANKOVÁ, J. and VAN DE GEER, S. (2017). Honest confidence regions and optimality in high-dimensional precision matrix estimation. *TEST* **26** 143–162. MR3613609 https://doi.org/10.1007/s11749-016-0503-5

JIA, B., XU, S., XIAO, G., LAMBA, V. and LIANG, F. (2017). Learning game regulatory networks from next generation sequencing data. *Biometrics* **73** 1221–1230. MR3744536 https://doi.org/10.1111/biom.12682

KABESCH, M., SCHEDEL, M., CARR, D., WOITSCH, B., FRITZSCH, C., WEILAND, S. K. and VON MUTIUS, E. (2006). IL-4/IL-13 pathway genetics strongly influence serum IgE levels and childhood asthma. *J. Allergy Clin. Immunol.* **117** 269–274.

LI, S., REN, Z., ZHANG, C.-H. and ZHOU, H. H. (2016). Asymptotic normality in estimation of large Ising graphical models. Unpublished Manuscript.

LIU, W. (2013). Gaussian graphical model estimation with false discovery rate control. *Ann. Statist.* **41** 2948–2978. MR3161453 https://doi.org/10.1214/13-AOS1169

LIU, H., LAFFERTY, J. and WASSERMAN, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.* **10** 2295–2328. MR2563983

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. and WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40** 2293–2326. MR3059084 https://doi.org/10.1214/12-AOS1037

NEWMAN, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E* (3) **74** 036104. MR2282139 https://doi.org/10.1103/PhysRevE.74.036104

PANDEY, G., PANDEY, O. P., ROGERS, A. J., AHSEN, M. E., HOFFMAN, G. E., RABY, B. A., WEISS, S. T., SCHADT, E. E. and BUNYAVANICH, S. (2018). A nasal brush-based classifier of asthma identified by machine learning analysis of nasal RNA sequence data. *Sci. Rep.* **8** 8826. https://doi.org/10.1038/s41598-018-27189-4

POTAPINSKA, O. and DEMKOW, U. (2009). T lymphocyte apoptosis in asthma. *Eur. J. Med. Res.* **14** 192–195.

RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343 https://doi.org/10.1214/09-AOS691

REN, Z., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43** 991–1026. MR3346695 https://doi.org/10.1214/14-AOS1286

VALE, K. (2016). Targeting the JAK-STAT pathway in the treatment of Th2-high severe asthma. *Future Med. Chem.* **8** 405–419.

VAN CROMBRUGGEN, K., VOGL, T., PÉREZ-NOVO, C., HOLTAPPELS, G. and BACHERT, C. (2016). Differential release and deposition of S100A8/A9 proteins in inflamed upper airway tissue. *Eur. Respir. J.* **47** 264–274.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

WANG, J. and KOLAR, M. (2016). Inference for high-dimensional exponential family graphical models. In *Artificial Intelligence and Statistics* 1042–1050.

WANG, H., FITZPATRICK, M., WILSON, N. J., ANTHONY, D., READING, P. C., SATZKE, C., DUNNE, E. M., LICCIARDI, P. V., SEOW, H. J. et al. (2019). CSF3R/CD114 mediates infection-dependent transition to severe asthma. *J. Allergy Clin. Immunol.* **143** 785–788.

WILLIAMS, J. W., FERREIRA, C. M., BLAINE, K. M., RAYON, C., VELÁZQUEZ, F., TONG, J., PETER, M. E. and SPERLING, A. I. (2018). Non-apoptotic Fas (CD95) signaling on T cells regulates the resolution of Th2-mediated inflammation. *Front. Immunol.* **9** 2521.

YANG, X., ZHANG, B., MOLONY, C., CHUDIN, E., HAO, K., ZHU, J., GAEDIGK, A., SUVER, C., ZHONG, H. et al. (2010). Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver. *Genome Res.* **20** 1020–1036.

YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2013). On Poisson graphical models. In *Advances in Neural Information Processing Systems* 1718–1726.

YANG, E., RAVIKUMAR, P., ALLEN, G. I. and LIU, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16** 3813–3847. MR3450553

YANG, I. V., PEDERSEN, B. S., LIU, A. H., O'CONNOR, G. T., PILLAI, D., KATTAN, M., MISIAK, R. T., GRUCHALLA, R., SZEFLER, S. J. et al. (2017). The nasal methylome and childhood atopic asthma. *J. Allergy Clin. Immunol.* **139** 1478–1488.

YU, M., KOLAR, M. and GUPTA, V. (2016). Statistical inference for pairwise graphical models using score matching. In *Advances in Neural Information Processing Systems* 2829–2837.

ZHANG, Y., OUYANG, Z. and ZHAO, H. (2017). A statistical framework for data integration through graphical models with application to cancer genomics. *Ann. Appl. Stat.* **11** 161–184. MR3634319 https://doi.org/10.1214/16-AOAS998

ZHANG, R., REN, Z. and CHEN, W. (2018). SILGGM: An extensive R package for efficient statistical inference in large-scale gene networks. *PLoS Comput. Biol.* **14** e1006369.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

ZHANG, N.-Z., CHEN, X.-J., MU, Y.-H. and WANG, H. (2018). Identification of differentially expressed genes in childhood asthma. *Medicine* **97** e10861.

ZHANG, R., REN, Z., CELEDÓN, J. C and CHEN, W. (2021). Supplement to "Inference of large modified poisson-type graphical models: Application to RNA-seq data in childhood atopic asthma studies." https://doi.org/10.1214/20-AOAS1413SUPPA, https://doi.org/10.1214/20-AOAS1413SUPPB

ZHAO, T., LIU, H., ROEDER, K., LAFFERTY, J. and WASSERMAN, L. (2012). The `huge` package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13** 1059–1062. MR2930633

ZWIENER, I., FRISCH, B. and BINDER, H. (2014). Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS ONE* **9** e85150. https://doi.org/10.1371/journal.pone.0085150