Received: 21 June 2021 | Revised: 23 May 2022 | Accepted: 25 May 2022

DOI: 10.1002/sim.9493

RESEARCH ARTICLE

Statistics in Medicine WILEY

Markov neighborhood regression for statistical inference of high-dimensional generalized linear models

Lizhe Sun | **Faming Liang**

Department of Statistics, Purdue University, West Lafayette, Indiana,

Correspondence
Faming Liang, Department of Statistics, Purdue University, West Lafayette, IN 47907, USA.
Email: fmliang@purdue.edu

Funding information
Division of Mathematical Sciences, Grant/Award Number: DMS-2015498; National Institute of General Medical Sciences, Grant/Award Numbers: R01-GM117597, R01-GM126089

High-dimensional inference is one of fundamental problems in modern biomedical studies. However, the existing methods do not perform satisfactorily. Based on the Markov property of graphical models and the likelihood ratio test, this article provides a simple justification for the Markov neighborhood regression method such that it can be applied to statistical inference for high-dimensional generalized linear models with mixed features. The Markov neighborhood regression method is highly attractive in that it breaks the high-dimensional inference problems into a series of low-dimensional inference problems. The proposed method is applied to the cancer cell line encyclopedia data for identification of the genes and mutations that are sensitive to the response of anti-cancer drugs. The numerical results favor the Markov neighborhood regression method to the existing ones.

KEYWORDS
confidence interval, graphical model, likelihood ratio test, nodewise regression, *P*-value

1 | INTRODUCTION

During the past two decades, dramatic improvements in data collection and acquisition technologies have enabled scientists to collect a great amount of high-dimensional data, for which the dimension p can be much larger than the sample size n (a.k.a. small-n-large-p). The current research on high-dimensional data mainly focuses on variable selection and graphical modeling. The former aims to provide a consistent estimate for the regression model under sparsity constraints. The existing methods include Lasso,[1] SCAD,[2] MCP,[3] elastic net,[4] and rLasso,[5] among others. The latter aims to learn conditional independence relationships for a large set of variables. The existing methods include graphical Lasso,[6,7] nodewise regression,[8] and ψ-learning,[9,10] among others developed. Quite recently, more and more researchers turn their attention to statistical inference, which is to seek for statistical procedures that are able to quantify uncertainty of high-dimensional regression, for example, constructing confidence intervals and assessing *P*-values for a single or subset of regression coefficients. A non-exhaustive list of the existing methods include desparsified Lasso,[11-13] multi sample-splitting,[14] ridge projection,[15] and Markov neighborhood regression (MNR).[16] See Section S1 of the supplementary material and Section 2 of this article for a brief review of these methods. Among the existing methods, the MNR method is a promising one. Based on the Markov property of Gaussian graphical models (GGMs), it successfully breaks the high-dimensional inference problem into a series of low-dimensional inference problems from which the desired confidence interval and *P*-value can be computed as for the conventional low-dimensional regression problems. Compared to the existing methods, the MNR

method tends to produce confidence intervals with more accurate coverage rates. However, based on the theory developed in Reference 16, the MNR method is only applicable to the case that the explanatory variables follow a multivariate Gaussian distribution. This has severely limited the scope of its applications.

This article provides a simple justification, based on the Markov property of graphical models and the likelihood ratio test, for the MNR method such that it can be extended to general high-dimensional inference problems. In particular, it can be applied to statistical inference for high-dimensional generalized linear models (GLMs) with mixed features, where the features can be continuous, discrete or both, and the response variable can be Gaussian, Poisson, multinomial, or even survival time (for Cox regression). This article also provides an algorithm for implementation of the MNR method and proves its validity. The numerical results favor the MNR method to the existing ones.

The remaining part of this article is organized as follows. Section 2 provides a brief review for the MNR method with Gaussian explanatory variables. Section 3 extends the MNR method to general high-dimensional inference problems. Section 4 illustrates the performance of MNR along with comparisons with the desparsified Lasso and ridge projection methods. Section 5 presents the application of the MNR method to the cancer cell line encyclopedia (CCLE) data. Section 6 concludes the article with a brief discussion.

2 | A BRIEF REVIEW OF THE MNR METHOD

Suppose that a set of n independent samples $D_n = \{(Y^{(i)}, \boldsymbol{X}^{(i)})_{i=1}^n\}$ have been collected from the linear regression with a random design:

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \epsilon, \tag{1}$$

where ϵ follows the Gaussian distribution $N(0, \sigma^2)$, and the explanatory variables (also known as features) $\boldsymbol{X} = (X_1, \ldots, X_p)$ follows a multivariate normal distribution $N_p(\boldsymbol{0}, \Sigma)$. Let S_* denote the support of the true model, which is sparse. Suppose that $\boldsymbol{X}$ has been represented by a GGM denoted by $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V} = \{1, 2, \ldots, p\}$ represents the set of p vertices, $\boldsymbol{E} = (e_{ij})$ represents the adjacency matrix, $e_{ij} = 1$ if the (i,j)th entry of the precision matrix $\Theta = \Sigma^{-1}$ is nonzero and 0 otherwise. Let $\boldsymbol{X}_A = \{X_k : k \in A\}$ denote a set of features indexed by $A \subset \boldsymbol{V}$. Let $\xi_j = \{k : e_{jk} = 1\}$ denote the neighboring set of X_j in $\boldsymbol{G}$. It follows from the Markov property of the GGM that $X_j \perp X_i | \boldsymbol{X}_{\xi_j}$ for any $i \in \boldsymbol{V} \setminus \xi_j$, ξ_j is called the minimum Markov neighborhood of X_j in $\boldsymbol{G}$. The minimum Markov neighborhood is also termed as Markov blanket in Bayesian networks or general Markov networks. Any subset S_j is a Markov neighborhood of X_j if $\xi_j \subseteq S_j \subseteq \boldsymbol{V} \setminus \{j\}$.

Without loss of generality, we let $S_1 = \{2, \ldots, d\}$ denote a Markov neighborhood of X_1, let Σ_d denote the covariance matrix of $\{X_1\} \cup \boldsymbol{X}_{S_1}$, and partition Θ as

$$\Theta = \begin{bmatrix} \Theta_d & \Theta_{d,p-d} \\ \Theta_{p-d,d} & \Theta_{p-d} \end{bmatrix}. \tag{2}$$

Following from the well-known property of the GGM,[17] for any variables X_i and X_j,

$$X_i \perp X_j | X_{\boldsymbol{V}\setminus\{i,j\}} \Leftrightarrow \theta_{ij} = 0, \tag{3}$$

where θ_{ij} denotes the (i,j)th entry of Θ. Therefore, the first row of $\Theta_{d,p-d}$ and the first column of $\Theta_{p-d,d}$ in (2) are exactly zero, as $X_1 \perp \boldsymbol{X}_{\boldsymbol{V}\setminus(\{1\}\cup S_1)} | \boldsymbol{X}_{S_1}$ holds. Inverting Θ, we have $\Sigma_d = (\Theta_d - \Theta_{d,p-d}\Theta_{p-d}^{-1}\Theta_{p-d,d})^{-1}$, which is equal to the top $d \times d$-submatrix of $\Sigma = \Theta^{-1}$. Therefore,

$$\Sigma_d^{-1} = \Theta_d - \Theta_{d,p-d}\Theta_{p-d}^{-1}\Theta_{p-d,d}. \tag{4}$$

Since the first row of $\Theta_{d,p-d}$ and the first column of $\Theta_{p-d,d}$ are exactly zero, the $(1,1)$th element of $\Theta_{d,p-d}\Theta_{p-d}^{-1}\Theta_{p-d,d}$ is exactly zero. Therefore, the $(1,1)$th entry of Θ_d (and Θ) equals to the $(1,1)$th entry of Σ_d^{-1}. This suggests that *if* $\{X_1\} \cup \boldsymbol{X}_{S_1} \supset \boldsymbol{X}_{S_*}$ *holds and n is sufficiently large, then the statistical inference for* β_1 *can be made based on the subset regression:*

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_d\beta_d + \epsilon. \tag{5}$$

Since S_1 forms a Markov neighborhood of X_1 in the GGM formed by all features, the subset regression is called a *Markov neighborhood regression*, which breaks the high-dimensional inference problem into a series of low-dimensional inference problems by solving a subset regression for each feature.

Let $\hat{\xi}_j$ denote an estimate of ξ_j, let $\hat{S}_*$ denote an estimate of S_*, and let $D_j = \{j\} \cup \hat{\xi}_j \cup \hat{S}_*$. Reference 16 proved the validity of the MNR method under the following conditions:

$$\hat{S}_* \supseteq S_*, \tag{6}$$

$$\hat{\xi}_j \supseteq \xi_j, \quad \forall j \in \{1, 2, \ldots, p\}, \tag{7}$$

$$|D_j| = |\{j\} \cup \hat{\xi}_j \cup \hat{S}_*| = o\left(\sqrt{n}\right). \tag{8}$$

For each $j \in \boldsymbol{V}$, if the conditions (6) to (8) are satisfied, then $\sqrt{n}(\hat{\beta}_j - \beta_j) \sim N(0, \sigma^2\theta_{jj})$, where θ_{jj} is the (j,j)th entry of the precision matrix Θ. For the case that n is finite, one can use $t(n - |D_j| - 1)$ to approximate the distribution of $\sqrt{n}\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}_n^2\hat{\theta}_{jj}}}$; that is, *the estimate, P-value, and confidence interval of* β_j *can be calculated from a subset regression as in conventional low-dimensional multiple linear regression.*

As implied by the proof of Theorem 1 of Reference 16, the conditions (6) and (8) together ensure the convergence $\sqrt{n}(\hat{\beta}_j - \beta_j) \xrightarrow{d} N(0, \sigma^2\theta_{D_j})$, where $\xrightarrow{d}$ denotes convergence in distribution and θ_{D_j} denotes the diagonal elements of $\Sigma_{D_j}^{-1}$ corresponding to β_j, and Σ_{D_j} denotes the covariance matrix of the features included in D_j; while the condition (7) ensures $\theta_{D_j} = \theta_{jj}$ as explained above (around Equation 4).

Finally, we note that the inference problem addressed MNR is very different from the post-selection inference problem considered in the literature.[18,19] Let $\boldsymbol{X_V}$ denotes the collection of all features of a high-dimensional regression, and let $M \subset \boldsymbol{V}$ denote a subset model. The post-selection inference is to construct a confidence interval C_j^M of β_j^M for any feature $j \in M$, conditioned on the event that the model M is selected (ie, $\hat{M} = M$), such that $P(\beta_j^M \in C_j^M | \hat{M} = M, \boldsymbol{X_V}) \geq 1 - \alpha$ for a prespecified confidence level $1 - \alpha$. Due to the intrinsic correlation between the selected model and the outputs of statistical tests, the theory for post-selection inference is rather intricate. In contrast, the problem addressed by MNR is relatively simple, which is to find a confidence interval C_j for any feature $j \in \boldsymbol{V}$ such that $P(\beta_j \in C_j | \boldsymbol{X_V}) \geq 1 - \alpha$. Without conditioning on the selected model makes the theory, as developed in Reference 16 and the current article, much simpler than that of post-selection inference. Under appropriate sparsity assumptions, our inference procedure is valid as long as the consistency or sure screening properties hold for the variable/structure selection procedures employed in steps (a) and (b) of Algorithm 1 (see Section 3 for the detail).

3 | EXTENSION OF THE MNR METHOD TO GENERAL HIGH-DIMENSIONAL INFERENCE PROBLEMS

From the brief review given in Section 2, we know that the validity of the MNR method depends crucially on the normality assumption for the features $X_1, X_2, \ldots, X_p$. Otherwise, (3) does not hold and the proof cannot go through any more. The normality assumption has severely limited the application scope of the method. In what follows, we provide a new proof for the validity of the MNR method such that it can be used for statistical inference of high-dimensional GLMs with mixed features.

The density function of the GLM is given by

$$f(y|\boldsymbol{x}, \boldsymbol{\beta}, \sigma) = c(y, \sigma) \exp\{[\vartheta y - \varphi(\vartheta)]/d(\sigma)\}, \tag{9}$$

where σ is the dispersion parameter, $\varphi(\cdot)$ is continuously differentiable, and ϑ is the natural parameter relating y to the features via a linear function

$$\vartheta = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p, \tag{10}$$

where the features can be continuous, discrete or both. This class of GLMs includes normal linear regression, Poisson regression, and logistic regression, among others. Further, we assume that the joint distribution of the features

Algorithm 1. Markov neighborhood regression for high-dimensional GLMs

(a) (Variable selection) Conduct variable selection for the model $Y \sim \mathbf{X_V}$ to get a consistent estimate of S_*. Denote the estimate by $\hat{S}_*$.

(b) (*Markov blanket estimation*) For each variable X_j, $j = 1, 2, \ldots, p$, obtain a consistent estimate of its Markov blanket and denote the estimate by $\hat{\xi}_j$.

(c) (Subset regression) For each variable X_j,$j = 1, 2, \ldots, p$,let $D_j = \{j\} \cup \hat{\xi}_j \cup \hat{S}_*$ and conduct a subset GLM:$Y \sim X_j + \mathbf{X}_{\hat{\xi}_j \cup \hat{S}_*}$.Calculate the P-value and construct the confidence interval for the coefficient of X_j as in a low-dimensional GLM.

$X_1, X_2, \ldots, X_p$ can be represented by a graphical model, and the conditional distribution of each feature X_i can be represented by a GLM. We refer to Reference 20 for discussions on compatibility of the joint and conditional distributions. For example, for the case that the features are mixed by Gaussian and Bernoulli random variables, their joint distribution can be found in Reference 21, where it is shown that the conditional distribution of each Gaussian random variable can be represented by a linear regression and that of each binomial random variable can be represented by a logistic regression.

Toward the goal of making statistical inference for such a high-dimensional GLM, evaluating P-value, and constructing confidence interval for the coefficient of each feature, we extend the MNR method as follows. First, let's consider $P(Y, X_j|\boldsymbol{X}_{\boldsymbol{V}\setminus\{j\}})$, the joint distribution of X_j and Y conditioned on all other variables $\boldsymbol{X}_{\boldsymbol{V}\setminus\{j\}}$. Suppose that a Markov network has been constructed for the features $X_1, X_2, \ldots, X_p$ and the Markov blanket ξ_j has been identified for X_j. Here the Markov network can be a Bayesian network or its moral graph, based on which a Markov blanket can be identified for each variable X_j. Recall that the Markov blanket of a node X_j is the minimum subset of $\{X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p\}$ such that X_j is independent of all other variables conditional on it. Following from the property of the Markov blanket, $P(Y, X_j|\boldsymbol{X}_{\boldsymbol{V}\setminus\{j\}})$ can be simplified as follows:

$$\begin{aligned} P(Y, X_j|\boldsymbol{X}_{\boldsymbol{V}\setminus\{j\}}) &= P(Y|X_{\boldsymbol{V}})P(X_j|\boldsymbol{X}_{\boldsymbol{V}\setminus\{j\}}) = P(Y|\boldsymbol{X}_{S_*})P(X_j|\boldsymbol{X}_{\xi_j}) \\ &= P(Y, X_j|\boldsymbol{X}_{S_*\cup\xi_j}) = P(Y|\boldsymbol{X}_{S_*\cup\xi_j\cup\{j\}})P(X_j|\boldsymbol{X}_{S_*\cup\xi_j}), \end{aligned} \tag{11}$$

where $P(Y|\boldsymbol{X}_{S_*\cup\xi_j\cup\{j\}})$ can be modeled by a subset GLM with the natural parameter given by

$$\vartheta = \beta_0 + X_j\beta_j + \boldsymbol{X}_{S_*\cup\xi_j}\boldsymbol{\beta}_{S_*\cup\xi_j}, \tag{12}$$

where $\boldsymbol{\beta}_{S_*\cup\xi_j}$ denotes the regression coefficients corresponding to the features $\boldsymbol{X}_{S_*\cup\xi_j}$. Suppose that the likelihood ratio test method is used to test the hypothesis $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$ with respect to the GLM (9), which characterizes the relationship between Y and X_j conditioned on $\boldsymbol{X}_{\boldsymbol{V}\setminus\{j\}}$. By (11), this test is reduced to the likelihood ratio test for the hypothesis with respect to a subset GLM with the natural parameter given by (12). *In summary, the MNR method can be described in Algorithm 1, which breaks the high-dimensional inference problem into a series of low-dimensional inference problems by solving a subset GLM for each feature.*

The likelihood ratio test has been well studied for GLMs. For linear regression, the likelihood ratio test and the Wald test give the same results in the case that n is finite and $n > |D_j|$ holds. For example, the square of the Wald t-test for a single coefficient is numerically identical to the likelihood ratio F-test for the same coefficient. For other GLMs such as logistic regression, this equality does not hold. However, the two tests are asymptotically equivalent in the sense that they will produce the same inference when n is large and $|D_j| = o(\sqrt{n})$ holds, following from the asymptotic theory established in Reference 22 where the number of parameters is allowed to grow with n. That is, if the conditions (6) to (8) are satisfied, the inference for each of the subset GLM in Algorithm 1 can be done as for the conventional low-dimensional GLMs.

The conditions (6) to (8) imply that the MNR method can be implemented in many different ways. The condition (6) is the so-called screening property, which is known to hold for many high-dimensional variable selection algorithms, such as SCAD,[2] MCP,[3] elastic net,[4] and adaptive Lasso.[23] Lasso also satisfies this condition under appropriate conditions of the design matrix, see Reference 24 for a review. For Markov blanket estimation, there are at least two methods we can use. The first one is p-learning,[10] which provides a consistent estimate of the moral graph for mixed data. The second one is nodewise regression, which was first proposed in Reference 8 for learning GGMs with an ℓ_1-penalty, and then extended in Reference 25 for learning graphical models for binary data and in Reference 20 for learning graphical models

for mixed data. This article extends the method further. In Appendix A, we show that the nodewise regression method can be applied to learn mixed graphical models with an amenable penalty,[26] which includes the Lasso, SCAD, and MCP penalties as special cases. Further, the condition (8) can be easily satisfied by a slight twist of the sparsity constraints imposed on the variable selection and Markov blanket estimation algorithms. Theorem 1 provides a formal justification for the validity of Algorithm 1, whose proof is given in Appendix A.

Theorem 1 (Validity of Algorithm 1). *Consider a GLM given in (9) with sample size n and dimension p, where the features form a mixed graphical model with compatible joint and conditional distributions. Suppose that the GLM is sparse such that the true model size* $|S_*| \prec \min\{n/\log(p), \sqrt{n}\}$, *the mixed graphical model is sparse such that the maximum neighborhood size* $k \prec \min\{n/\log(p), \sqrt{n}\}$, *and the conditions 1 to 11 (given in Appendix A) are satisfied. If a regularization method with an amenable penalty function is used for variable selection in step (a), and the nodewise regression method with an amenable penalty function is used for Markov blanket estimation in step (b), then Algorithm 1 is valid for statistical inference of high-dimensional GLMs.*

Regarding Algorithm 1 and the proof for its validity, we have two remarks.

Remark 1 (Joint inference). Algorithm 1 can be easily extended to joint inference for a finite number of variables. For example, we want to test a linear hypothesis $H_0 : \boldsymbol{\alpha}'\boldsymbol{\beta} = 0$ vs $H_1 : \boldsymbol{\alpha}'\boldsymbol{\beta} \neq 0$, where $\boldsymbol{\alpha} = (a_1, a_2, \ldots, a_p)$ denotes a p-vector with r nonzero elements and r is finite. For this hypothesis, the subset GLM can be simply constructed as $Y \sim \beta_0 + \boldsymbol{X}_D\boldsymbol{\beta}_D$, where $D = R \cup (\cup_{j\in R}\hat{\xi}_j) \cup \hat{S}_*$ and $R = \{j : a_j \neq 0, j = 1, 2, \ldots, p\}$.

Remark 2 (Accelerating computing by screening). For the subset GLM, if we do not stick to the equivalent Wald test, but directly perform the likelihood ratio test, then the condition (8) can be much relaxed. For linear regression, by Theorem 1 of Reference 27, the condition (8) can be relaxed as $|D_j| = o(n)$, which is actually a sufficient and necessary condition for the Chi-square approximation of the likelihood ratio test. For the logistic regression, the above condition can be relaxed further. Reference 28 showed that if $|D_j|$ and n grows large in such a way that $|D_j|/n \to \kappa$ for some $\kappa < 1/2$, then the likelihood ratio test can be approximated by a rescaled Chi-square. Based on these results, the sparsity conditions in Theorem 1 can be relaxed as $|S_*| \prec n/\log(p)$ and $k \prec n/\log(p)$. More importantly, in this case, Algorithm 1 can be much accelerated by replacing the variable selection procedure, performed in step (a) as well step (b) for each node, by a sure independence screening procedure.[29,30]

4 | SIMULATION STUDIES

4.1 | Linear regression

We generated 500 datasets from the linear regression $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where the sample size $n = 300$, the dimension $p = 500$, and the random error $\boldsymbol{\epsilon} \sim N(0, I_n)$. The features were generated in the following procedure according to a Bayesian network with the adjacency matrix $\boldsymbol{E}$ given by

$$E_{i,j} = \begin{cases} 1, & \text{if } j - i = 1,\ i = 1, \ldots, (j-1), \\ 1, & \text{if } j - i = 2,\ i = 1, \ldots, (j-2), \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

First, we ordered the variables as $X_1, X_2, \ldots, X_p$, and randomly marked half of them as continuous and half as binary. Next, we generated $Z_1 \sim \mathcal{N}(0, 1)$, and set $X_1 = Z_1$ if X_1 is continuous, and $X_1 \sim \text{Binomial}(1, 1/(1 + \exp\{-Z_1\}))$ otherwise; and generated the variables $X_j, j = 2, 3, \ldots, p$ sequentially by setting

$$Z_j = \sum_{i=1}^{j-1} \rho E_{ij} X_i \quad \text{and} \quad X_j = \begin{cases} Z_j + \epsilon, & \text{if } X_j \text{ is continuous,} \\ \text{Binomial}\left(1, \frac{\exp(Z_j)}{1+\exp(Z_j)}\right) & \text{if } X_j \text{ is binary,} \end{cases} \tag{14}$$

where we set $\rho = 0.5$. The true regression coefficients were given by $(\beta_0^*, \beta_1^*, \beta_2^*, \ldots, \beta_p^*) = (1, 2, 2.5, 3, 3.5, 4, 0, \ldots, 0)$.

TABLE 1 Coverage rates and widths of the 95% confidence intervals produced by MNR, desparsified Lasso, and ridge projection for simulated examples, where "signal" and "noise" denote nonzero and zero regression coefficients, respectively

	Measure		Desparsified-Lasso	Ridge projection	MNR
Linear	Coverage	Signal	0.880 (0.015)	0.975 (0.007)	0.955 (0.010)
		Noise	0.953 (0.010)	0.981 (0.006)	0.950 (0.010)
	Width	Signal	0.374 (0.006)	0.682(0.010)	0.379 (0.005)
		Noise	0.377 (0.007)	0.693(0.012)	0.387 (0.006)
	CPU(s)		390.9	2.393	224.6
Logistic	Coverage	Signal	0.135 (0.015)	0.199 (0.018)	0.940 (0.011)
		Noise	0.990 (0.005)	1.000 (0.0002)	0.948 (0.010)
	Width	Signal	0.831 (0.011)	1.693 (0.025)	1.497 (0.016)
		Noise	0.784 (0.014)	1.677 (0.030)	1.059 (0.017)
	CPU(s)		2036	7.765	532.9
Survival	Coverage	Signal	-	-	0.939 (0.011)
		Noise	-	-	0.945 (0.010)
	Width	Signal	-	-	0.395 (0.004)
		Noise	-	-	0.370 (0.006)
	CPU(s)		-	-	335.4[a]

Note: The CPU time (in seconds) was recorded for a single dataset with the method running in serial on a personal computer of i9-10900k CPU@3.6 GHz with 128 GB memory.

[a]We set the SIS iteration number to 1 in step (a) of Algorithm 1.

Algorithm 1 was applied to this example, where variable selection was conducted using the SIS-MCP method implemented in the R package *SIS*,[31] the Markov blanket was estimated by nodewise regression with SIS-MCP used in regressing for each node. For comparison, the desparsified-Lasso and ridge projection method were also applied to this example. Both methods have been implemented in the R package *hdi*.[32] For all other examples of this article, the algorithms were implemented in the same way.

Table 1 summarizes the coverage rates and widths of the 95% confidence intervals produced by these methods for each regression coefficient. For the nonzero regression coefficients (denoted by "signal"), the mean coverage rate and its standard deviation are calculated by

$$\overline{p}_{\text{cover}} = \sum_{j=1}^{500}\sum_{i\in S_*}\hat{p}_i^{(j)}/(500\cdot|S_*|), \quad \sigma(\overline{p}_{\text{cover}}) = \sqrt{\text{Var}\{\hat{p}_i^{(j)} : i\in S_*, j=1,2,\ldots,500\}/500}, \tag{15}$$

where $\hat{p}_i^{(j)} \in \{0,1\}$ indicates the coverage of β_i by the confidence interval, and $\text{Var}\{\cdot\}$ denotes the variance. By dividing by 500 in its calculation, $\sigma(\overline{p}_{\text{cover}})$ represents the variability of the mean value (averaged over 500 independent datasets) for a single regression coefficient. For the width of the confidence interval, the mean and standard deviation were calculated similarly. For the zero regression coefficients (denoted by "noise"), the mean coverage rate, the mean width, and their standard deviations were also calculated similarly. The comparison indicates that MNR significantly outperforms the existing methods: For both the nonzero and zero regression coefficients, the mean coverage rates produced by MNR are much closer to their nominal level. The reason why desparsified Lasso is coverage deficient has been explained in Reference 16: Desparsified Lasso centers its confidence interval at a bias-corrected Lasso estimator which, unfortunately, is still biased, although its bias has been much smaller than the original Lasso estimator.

4.2 | Logistic regression

We simulated 500 datasets from a logistic regression, where we set $n = 600$, $p = 800$, and the true regression coefficients $(\beta_0^*, \beta_1^*, \beta_2^*, \ldots, \beta_p^*) = (1, 2, 2.5, -3, 3.5, -4, 0, \ldots, 0)$. We set $\rho = 0.5$. The features were generated according to (13) and

(14) as in Section 4.1. The MNR, desparsified Lasso and ridge projection methods were applied to this example with the results summarized in the lower part of Table 1. For MNR, SIS-MCP was used for both procedures of variable selection and Markov blanket construction. The comparison indicates again that MNR significantly outperforms desparsified Lasso and ridge projection in confidence interval construction for GLMs. The desparsified Lasso and ridge regression essentially fail for the example.

4.3 | Cox regression

To valid the MNR method in more general cases, we consider cox regression. We let $\lambda(t)$ denote the hazard rate at time t and let $\lambda_0(t)$ denote the baseline hazard rate. The Cox regression is given by

$$\lambda(t) = \lambda_0(t)\exp\{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p, X_p\},$$

from which we simulated 500 datasets with $n = 400$, $p = 600$. We set the true regression coefficients $(\beta_1^*, \beta_2^*, \ldots, \beta_p^*) = (1, 1, 1, 1, 1, 0, 0, \ldots, 0)$, set the baseline hazard rate $\lambda_0(t) \equiv 0.1$, and set the censoring hazard rate $\lambda_c = 1$. We generated the predictors by using Equations (13) and (14) with $\rho = 0.5$, generated the event time from the Weibull distribution with a shape parameter of 1 and a scale parameter of $\lambda_0 \exp(-\sum_{i=1}^{p} X_i\beta_i)$, generated the censoring time from the Weibull distribution with a shape parameter of 1 and a scale parameter of λ_c, and set the observed survival time as the minimum of the event time and the censoring time for each subject.

The MNR method was applied to the datasets, where SIS-Lasso was applied for variable selection and SIS-MCP was used for Markov blanket estimation. The results were summarized in Table 1. Unfortunately, the desparsified Lasso and ridge regression were not available for this model and thus could not be used for comparison.

4.4 | Variable selection by MNR

This section explores the potential of MNR in variable selection. As discussed in Reference 16, MNR converts the variable selection problem to a multiple hypothesis testing problem. By computing and sorting adjusted P-values[33] or q-values,[34] we can select important variables at a prespecified false discovery rate (FDR). In this study, we generated 20 datasets from a linear regression model under each of the following settings: (a) $n = 200, 300$, $p = 1000$, $\sigma^2 = 1$, true regression coefficients $(\beta_0^*, \beta_1^*, \beta_2^*, \ldots, \beta_p^*) = (1, 2, 2.5, 3, 3.5, 4, 0, \ldots, 0) * \gamma$; (b) $n = 300, 500$, $p = 10\,000$, $\sigma^2 = 1$, true regression coefficients $(\beta_0^*, \beta_1^*, \beta_2^*, \ldots, \beta_p^*) = (1, 3, -3, 3.5, 4, -4, 4.5, 5, -5, 5.5, 6, 0, \ldots, 0) * \gamma$; where the value of $\gamma \in (0, 1]$ is varied for tuning the strength of signal. The explanatory variables were generated as in Section 4.1, but different values of ρ, including $\rho = 0.1, 0.3$ and 0.5, were used in equation (14). The proportion of binary predictors was set to 10%, that is, each dataset consists of 100 and 1000 binary predictors under the settings (a) and (b), respectively.

By step 3 of Algorithm 1, a subset regression is implemented for each predictor, and thus relevant variables can be selected based on the multiple hypothesis tests:

$$H_0 : \beta_j = 0, \quad H_\alpha : \beta_j \neq 0, \qquad j = 1, 2, \ldots p. \tag{16}$$

Given the P-values of the subset regressions, we conduct the multiple hypothesis tests using the empirical Bayesian method developed in Reference 35. As shown in Tables 2 and 3, MNR can exactly identify the true predictors for each dataset at a FDR level of $q = 0.001$ or $q = 0.0001$, where the q-value is as defined in Reference 34. Tables 2 and 3 report these results in terms of false selection rate (FSR) and negative selection rate (NSR), which are defined by:

$$\text{FSR} = \frac{\sum_{j=1}^{20} |\hat{\mathbf{S}}_j / \mathbf{S}_*|}{\sum_{j=1}^{20} |\hat{\mathbf{S}}_j|}, \quad \text{NSR} = \frac{\sum_{j=1}^{20} |\mathbf{S}_* / \hat{\mathbf{S}}_j|}{\sum_{j=1}^{20} |\mathbf{S}_*|}, \tag{17}$$

where $\mathbf{S}_*$ is the set of true variables and $\hat{\mathbf{S}}_j$ is the set of selected variables for dataset j.

For comparison, we applied the popular likelihood regularization methods, including SIS-SCAD, SIS-MCP, SIS-Lasso, and SIS-Elastic-Net, to these datasets for performing variable selection under their default settings in the package SIS. It

TABLE 2 Variable selection results by MNR, SIS-SCAD, SIS-MCP, SIS-Lasso, and SIS-Elastic-Net for linear regression datasets simulated with $n = 200,300$, $p = 1000$, $\gamma = 1, 1/3$, $\sigma^2 = 1$, and $\rho = 0.1$, 0.3, and 0.5

	MNR							SIS-Elastic-Net	
Measure	$q = 0.0001$	$q = 0.001$	$q = 0.01$	$q = 0.05$	**SIS-SCAD**	**SIS-MCP**	**SIS-Lasso**	$\alpha = 0.1$	$\alpha = 0.2$
	$\rho = 0.1, n = 300, \gamma = 1$								
FSR	0	0	0	0.029	0.010	0.010	0.320	0.875	0.812
NSR	0	0	0	0	0	0	0	0.02	0.02
	$\rho = 0.3, n = 300, \gamma = 1$								
FSR	0	0	0.010	0.057	0.010	0.091	0.281	0.829	0.699
NSR	0	0	0	0	0	0	0	0.01	0.01
	$\rho = 0.5, n = 300, \gamma = 1$								
FSR	0	0	0.010	0.057	0.038	0.057	0.254	0.701	0.554
NSR	0	0	0	0	0	0	0	0	0
	$\rho = 0.5, n = 200, \gamma = 1/3$								
FSR	0.010	0.010	0.030	0.117	0.546	0.560	0.429	0.674	0.579
NSR	0.05	0.04	0.02	0.02	0.02	0.01	0	0.01	0.01

Note: For the elastic-net penalty, we tried the setting $\alpha = 0.1, 0.2$.

is known that the likelihood regularization methods tend to select more false predictors to compensate their shrinkage effects on regression coefficients, and this over-selection issue can become worse as the ratio p/n increases due to the increasing likelihood of *spurious correlation*. In statistics, spurious correlation refers to that two or more variables are associated but not causally related due to either coincidence or the presence of unseen confounding factors. The MNR method, as a multiple hypothesis test-based method, provides a promising way for addressing the spurious correlation issue encountered in variable selection by controlling the FDR at a low level. As shown in Tables 2 and 3, MNR can significantly outperform the likelihood regularization methods in high-dimensional variable selection; in particular, MNR tends to have a smaller FSR value and is more robust to the strength of signal than the regularization methods. More discussions on the properties of MNR in variable selection can be found in Sections 5 and 6.

5 | IDENTIFICATION OF DRUG SENSITIVE GENES AND MUTATIONS

Disease heterogeneity is often observed in complex diseases such as cancer. For example, molecularly targeted cancer drugs are only effective for patients with tumors expressing targets.[36,37] The disease heterogeneity has directly motivated the development of precision medicine, aiming to improve patient care by tailoring optimal therapies to an individual patient according to his/her molecular profile and clinical characteristics. Identifying sensitive genes and mutations to different drugs is an important step toward the goal of precision medicine.

In this study, we considered the CCLE dataset, which is publicly available at https://github.com/alexisbellot/GCIT/tree/master/CCLE%20Experiments. The dataset consists of 8-point dose-response curves for 24 drugs (or chemical compounds) across over 400 cell lines. For different drugs, the numbers of cell lines are slightly different. For each cell line, it consists of the expression data of 18 988 genes and 1638 mutations, which bring the dimension of the full dataset to $p = 20\,626$. We used the area under the dose-response curve, which was termed as activity area in Reference 38, to measure the sensitivity of a drug to each cell line. Compared to other measurements, such as IC_{50} and EC_{50}, the activity area could capture the efficacy and potency of the drug simultaneously. Since for each drug, the number of experimented cell lines is small, while the number of genes and mutations is large, accurate identification of the drug sensitive genes/mutations has posed a great challenge on the existing statistical methods. It is known that the regularization methods, such as Lasso, SCAD, and MCP, tend to select more false predictors to compensate their shrinkage effects on regression coefficients. In addition, they tend to select spuriously correlated variables due to their likelihood optimization nature. Spurious correlation often occurs in small-n-large-p regression due to randomness or unknown confounding factors. When spuriously correlated variables exist, they tend to be selected by likelihood-based methods. For a dataset with a small number

TABLE 3 Variable selection results by MNR, SIS-SCAD, SIS-MCP, SIS-Lasso, and SIS-Elastic-Net for linear regression datasets simulated with $n = 300,500$, $p = 10\,000$, $\gamma = 1, 1/3, 1/5$, $\sigma^2 = 1$, and $\rho = 0.5$

	MNR							
Measure	$q = 0.0001$	$q = 0.001$	$q = 0.01$	$q = 0.05$	**SIS-SCAD**	**SIS-MCP**	**SIS-Lasso**	**SIS-Elastic-Net**
	$n = 500, \gamma = 1$							
FSR	0	0	0.005	0.024	0.010	0.476	0.817	0.708
NSR	0	0	0	0	0	0	0	0
	$n = 500, \gamma = 1/3$							
FSR	0	0	0.005	0.024	0.206	0.541	0.845	0.715
NSR	0	0	0	0	0	0	0	0
	$n = 300, \gamma = 1/3$							
FSR	0	0	0.015	0.107	0.631	0.650	0.779	0.668
NSR	0	0	0	0	0	0	0	0.01
	$n = 300, \gamma = 1/5$							
FSR	0	0	0.010	0.099	0.752	0.716	0.771	0.655
NSR	0.025	0.01	0.005	0	0	0	0	0.01
	$n = 300, \gamma = 1/6$							
FSR	0	0	0.010	0.1	0.762	0.739	0.783	0.644
NSR	0.060	0.05	0.035	0.01	0.005	0.005	0	0.01

Note: For the elastic-net penalty, we set $\alpha = 0.2$.

of observations, the spuriously correlated variables often reduce not only the fitting error but also the prediction error in cross-validation. MNR, as a conditional independence test-based method, provides a promising way for limiting the selection of spuriously correlated variables by controlling the FDR at a reasonable level.

Algorithm 1 was applied to the dataset collected for each drug to select the drug-sensitive genes and mutations. The selection was based on the adjusted *P*-values[33] of the conditional independence tests for each single gene/mutation. We set the significance level of the multiple hypothesis test at .05. If there were no genes/mutations selected at this significance level, we just reported one gene/mutation with the smallest adjusted *P*-value. For comparison, the existing methods, including desparsified Lasso[11-13] and ridge projection, were applied to this example. For each drug, desparsified Lasso is simply inapplicable due to the ultra-high dimensionality of the dataset; the package *hdi*[32] aborted due to the excess of memory limit. However, the ridge projection method still performed reasonably well. For this method, we also selected the genes/mutations with the adjusted *P*-values less than .05 as significant, or reported one gene/mutation with the smallest adjusted *P*-value if no gene/mutation was significant at the level .05.

Table 4 summarizes the results produced by the above methods. It shows that MNR and ridge projection can produce similar or overlapped results for many drugs, while the confidence intervals produced by MNR tend to be narrower than those by ridge projection for the genes/mutations selected by both methods. For example, for the drugs Topotecan and Irinotecan, both methods selected the gene SLFN11 as a drug sensitive gene, and the confidence intervals by MNR are narrower than those by ridge projection. In the literature, References 38 and 39 reported that SLFN11 is predictive of treatment response for Topotecan and Irinotecan. For the drug 17-AAG, both methods selected NQO1 as a drug sensitive gene. References 38 and 40 reported NQO1 as the top predictive biomarker for 17-AAG. Other examples include the drug Nilotinib for which both methods selected APOL4, the drug PF2341066 for which both methods selected the mutation SCD5, the drug PLX4720 for which both methods selected the mutation BRAFV600E, and the drug Erlotinib for which both methods selected the mutation EGFR. It is known that EGFR is the target gene of the drug Erlotinib, and this target gene has been correctly identified by MNR.

For a thorough study for the performance of MNR, we have also compared it with some popular high-dimensional variable selection methods such as SIS-SCAD, SIS-MCP, and SIS-Lasso, which all fall into the class of likelihood regularization methods. We compared the performance of these methods in variable selection, goodness-of-fit and prediction. For this purpose, a 5-fold cross validation experiment was conducted for each drug. Table 5 reports results for three selected drugs, 17-AAG, Irinotecan, and PLX4720. More results are presented in Table S3 of the supplementary material.

TABLE 4 Comparison of drug sensitive genes/mutations selected by desparsified Lasso, ridge projection, and MNR for 24 anti-cancer drugs, where "*" indicates that this gene was significantly selected and the number in the parentheses denotes the width of the 95% confidence interval, and "-MUT" indicates a mutation

Drug	Desparsified-Lasso	Ridge	MNR
17-AAG	-	NQO1(0.194)	NQO1(0.247)
AEW541	-	NFE2L3(0.327)	GPATCH3(0.245)
AZD0530	-	STK39(0.331)	PYY(0.208)
AZD6244	-	SPRY2(0.303)	NRAS-MUT*(0.548)
Erlotinib	-	EGFR-MUT(1.498)	EGFR-MUT*(0.814)
	-		CLK3-MUT*(1.506)
	-		EGFR*(0.261)
Irinotecan	-	SLFN11*(0.337)	SLFN11*(0.2)
L-685458	-	SELPLG(0.473)	WDR86*(0.203)
Lapatinib	-	ERBB2(0.561)	SCO1(0.303)
LBW242	-	SET-MUT(10.27)	SET-MUT*(5.075)
Nilotinib	-	APOL4*(0.474)	CAMK2A-MUT*(2.017)
			NCF4*(0.349)
			CCL23*(0.352)
			TRDC*(0.211)
			RNASE2*(0.437)
			APOL4*(0.277)
Nutlin-3	-	SPIC(0.398)	ASB16*(0.231)
Paclitaxel	-	ABCB1(0.326)	TM2D2*(0.280)
Panobinostat	-	LOC100652995(0.250)	SVIP*(0.201)
PD-0325901	-	SPRY2(0.324)	THRSP-MUT(2.696)
PD-0332991	-	TMTC2(0.346)	NFE2L3*(0.223)
PF2341066	-	SCD5-MUT(8.433)	SCD5-MUT*(3.239)
			ANKRD22*(0.251)
			WDFY4*(0.314)
PHA-665752	-	GCFC2(0.387)	PDPK1-MUT(3.429)
PLX4720	-	BRAFV600E-MUT*(1.830)	BRAFV600E-MUT*(0.899)
	-		PLEKHH3*(0.19)
	-		IRAK1-MUT*(1.66)
RAF265	-	GNPTAB(0.354)	FAM89B*(0.255)
Sorafenib	-	PROSER1(0.523)	DNAJC5B*(0.284)
	-		THAP10*(0.261)
TAE684	-	SELPLG(0.457)	PPFIA1*(0.292)
TKI258	-	WDFY4(0.464)	THEMIS*(0.304)
Topotecan	-	SLFN11*(0.278)	SLFN11(0.17)
ZD-6474	-	APOL4(0.417)	PGBD2*(0.206)

Note: For each dataset, ridge regression cost 2.6 minutes CPU time with a single thread running in serial, and MNR cost 46.5 minutes CPU time with 10 threads running in parallel. All methods were run on the same personal computer with i9-10900k CPU@3.6GHz and 128 GB memory.

TABLE 5 Comparison of MNR with SIS-SCAD, SIS-MCP, and SIS-Lasso for model prediction and variable selection on three selected drugs, 17-AAG, Irinotecan, and PLX4720, via 5-fold cross-validation experiments: "MSFE" denotes the mean squared fitting error, "MSPE" denotes the mean squared prediction error, and "Size" denotes the number of selected gene/mutations, which are reported as the average over 5-fold results with the standard deviation given in the parentheses; "selected Genes/mutations" shows the genes and mutations selected in the 5-fold experiments, where the number in the parentheses represents the selection frequency of each selected gene/mutation

Drug	Methods	MSFE	MSPE	Size	Selected genes/mutations
17-AAG	SIS-SCAD	0.62(0.21)	0.88(0.16)	20.0(11.5)	NQO1(4),CDH6(3),MMP24(3),ZNF610(3), ZFP30(3),ZNF14(3)
	SIS-MCP	0.54(0.02)	0.89(0.14)	16.2(3.5)	NQO1(5),CDH6(3),MMP24(3), ZFP30(3),CBFB(3)
	SIS-Lasso	0.77(0.17)	0.99(0.10)	7.8(11.0)	MMP24(4),NQO1(2),ZFP30(2),CTDSP1(2)
	MNR	0.93(0.04)	0.98(0.11)	1.2(0.5)	NQO1(4)
Irinotecan	SIS-SCAD	0.44(0.05)	0.55(0.08)	6.6(0.9)	ARHGAP19(5),SLFN11(4)
	SIS-MCP	0.46(0.05)	0.56(0.09)	3.8(0.8)	ARHGAP19(5),SLFN11(4)
	SIS-Lasso	0.43(0.06)	0.54(0.09)	9.8(3.0)	ARHGAP19(5),CPSF6(5), SLFN11(4),CD63(3)
	MNR	0.74(0.02)	0.75(0.07)	1.0(0.0)	SLFN11(5)
PLX4720	SIS-SCAD	0.59(0.05)	0.91(0.28)	9.8(5.1)	GAPDHS(3),MAD1L1(3),RXRG(2), LPL(2),ART3(2),ZFP106(2)
	SIS-MCP	0.61(0.04)	0.89(0.27)	5.4(2.8)	GAPDHS(3),ZFP106(2),ZEB2(2)
	SIS-Lasso	0.60(0.06)	0.87(0.23)	10.2(5.6)	SPRYD5(5),GAPDHS(4),RXRG(3)
	MNR	0.52(0.05)	0.65(0.12)	3.2(3.8)	BRAF.V600E-MUT(5), IRAK1-MUT(2)

As expected, MNR tends to select much less numbers of genes/mutations and have slightly larger prediction errors than the likelihood regularization methods.

As mentioned previously, this phenomenon can possibly be explained by *spurious correlation*, which often causes the likelihood-based methods to a high FDR. In contrast, MNR selects variables based on multiple hypothesis tests and it can, as demonstrated by our previous simulation examples, effectively limit the effect of spurious correlation by controlling the FSR at a low level. Further, we note that the genes/mutations selected by MNR for the three drugs in Table 5 have been verified in the literature as described above. Finally, we note that for the drug PLX4720, MNR did not only select a smaller number of genes/mutations, but also predicted more accurately. This is because it selected the right mutation BRAF.V600E, while the likelihood regularization methods failed to do so.

In summary, MNR tends to select a more parsimonious but trustful model than the likelihood regularization methods for high-dimensional regression problems.

6 | DISCUSSION

Based on the Markov property of graphical models and the likelihood ratio test, this article provides a simple justification for the MNR method such that it can be applied to statistical inference for high-dimensional GLMs with mixed features. The MNR method has been tested on both simulated and real data problems. The numerical results indicate its superiority over the existing methods. Compared to desparsified Lasso, MNR does not only produce more accurate confidence intervals, but also is computationally more efficient. Both methods involve nodewise regression, but MNR avoids calculation of the precision matrix Θ required by desparsified Lasso, which is costly when the dimension p is high.

The MNR method is highly attractive in that it has a high-dimensional inference problem reduced to a series of low-dimensional inference problems. Consequently, the MNR method possesses an embarrassingly parallel structure, and its computation can be much accelerated (than reported in the article) if running in parallel on a multi-core computer. Other than parallel implementation, as mentioned in Remark 2, the computation of the MNR method can be further accelerated by replacing the variable selection procedures involved in the method by some sure independent screening procedures. This is worth a further investigation.

As shown in this article, as a by-product, the MNR method can also be used for variable selection for high-dimensional GLMs. Due to its use of the dependence structure among the predictors, the MNR method tends to outperform the existing variable selection methods. A similar finding has been reported in Reference 41 that use of the correlation structure among the predictors can often improve the performance of a variable selection method. In addition, due to its multiple hypothesis testing nature, the MNR method can effectively limit the effect of spurious correlation that has bothered the likelihood regularization methods under the small-n-large-p scenario.

ACKNOWLEDGEMENTS

The authors thank the editor, associate editor, and two referees for their insightful and constructive comments which have led to significant improvement of this article. Liang's research is supported in part by the NSF grant DMS-2015498 and the NIH grants R01-GM117597 and R01-GM126089.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at https://github.com/alexisbellot/GCIT/tree/master/CCLE%20Experiments.

ORCID

Lizhe Sun https://orcid.org/0000-0003-4860-4957
Faming Liang https://orcid.org/0000-0002-1177-5501

REFERENCES

1. Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc Ser B*. 1996;58:267-288.
2. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96:1348-1360.
3. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat*. 2010;38:894-942.
4. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Royal Stat Soc Ser B (Stat Methodol)*. 2005;67:301-320.
5. Song Q, Liang F. High dimensional variable selection with reciprocal L1-regularization. *J Am Stat Assoc*. 2015;110:1607-1620.
6. Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. *Biometrika*. 2007;95:19-35.
7. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. 2008;9:432-441.
8. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. *Ann Stat*. 2006;34:1436-1462.
9. Liang F, Song Q, Qiu P. An equivalent measure of partial correlation coefficients for high dimensional Gaussian graphical models. *J Am Stat Assoc*. 2015;110:1248-1265.
10. Xu S, Jia B, Liang F. Learning moral graphs in construction of high-dimensional bayesian networks for mixed data. *Neural Comput*. 2019;31:1183-1214.
11. van de Geer S, Bühlmann P, Ritov Y, Dezeure R. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann Stat*. 2014;42:1166-1202.
12. Zhang CH, Zhang SS. Confidence intervals for low dimensional parameters in high dimensional linear models. *J Royal Stat Soc Ser B (Stat Methodol)*. 2014;76:217-242.
13. Javanmard A, Montanari A. Confidence intervals and hypothesis testing for high-dimensional regression. *J Mach Learn Res*. 2014;15:2869-2909.
14. Meinshausen N, Meier L, Bühlmann P. p-values for high-dimensional regression. *J Am Stat Assoc*. 2009;104:1671-1681.
15. Bühlmann P. Statistical significance in high-dimensional linear models. *Bernoulli*. 2013;19:1212-1242.
16. Liang F, Xue J, Jia B. Markov neighborhood regression for high-dimensional inference. *J Am Stat Assoc*. 2021;0:1-15.
17. Lauritzen S. *Graphical Models*. Oxford: Oxford University Press; 1996.
18. Berk RA, Brown LD, Buja A, Zhang K, Zhao LH. Valid post-selection inference. *Ann Stat*. 2013;41:802-837.
19. Lee JD, Sun DL, Sun Y, Taylor JE. Exact post-selection inference, with application to the lasso. *Ann Stat*. 2016;44:907-927.
20. Chen S, Witten DM, Shojaie A. Selection and estimation for mixed graphical models. *Biometrika*. 2015;102:47-64.
21. Lee JD, Hastie T. Learning the structure of mixed graphical models. *J Comput Graph Stat*. 2015;24:230-253.
22. Portnoy S. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tend to infinity. *Ann Stat*. 1989;16:356-366.
23. Zou H. The adaptive lasso and its oracle properties. *Ann Stat*. 2006;38:894-942.
24. Hastie T, Tibshirani R, Wainwright M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton: Chapman & Hall; 2015.
25. Ravikumar P, Wainwright MJ, Lafferty JD. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann Stat*. 2010;38:1287-1319.
26. Loh PL, Wainwright MJ. Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. *J Mach Learn Res*. 2015;16:559-616.
27. He Y, Jiang T, Wen J, Xu G. Likelihood ratio test in multivariate linear regression: from low to high dimension. *Stat Sin*. 2021;0:1-21.

28. Sur P, Chen Y, Candès EJ. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled Chi-square. *Probab Theory Relat Fields*. 2019;175:487-558.
29. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space (with discussion). *J R Stat Soc Ser B*. 2008;70:849-911.
30. Fan J, Song R. Sure independence screening in generalized linear model with NP-dimensionality. *Ann Stat*. 2010;38:3567-3604.
31. Saldana DF, Feng Y. SIS: an R package for sure independence screening in ultrahigh-dimensional statistical models. *J Stat Softw*. 2018;83:1-25.
32. Dezeure R, Bühlmann P, Meier L, Meinshausen N. High-dimensional inference: confidence intervals, p-values and r-software HDI. *Stat Sci*. 2015;30:533-558.
33. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6:65-70.
34. Storey JD. A direct approach to false discovery rates. *J R Stat Soc Ser B*. 2002;64:479-498.
35. Liang F, Zhang J. Estimating FDR under general dependence using stochastic approximation. *Biometrika*. 2008;95:961-977.
36. Grünwald V, Hidalgo M. Developing inhibitors of the epidermal growth factor receptor for cancer treatment. *J Natl Cancer Inst*. 2003;95:851-867.
37. Buzdar AU. Role of biologic therapy and chemotherapy in hormone receptor and HER2-positive breast cancer. *Ann Oncol*. 2009;20:993-999.
38. Barretina J, Caponigro G, Stransky N, et al. The cancer cell line encyclopedia enables predictive modeling of anticancer drug sensitivity. *Nature*. 2012;483:603-607.
39. Zoppoli G, Regairaz M, Leo E, et al. Putative DNA/RNA helicase schlafenl1 (SLFN11) sensitizes cancer cells to DNA-damaging agents. *Proc Nat Acad Sci*. 2012;109:15030-15035.
40. Hadley KE, Hendricks DT. Use of NQO1 status as a selective biomarker for oesophageal squamous cell carcinomas with greater sensitivity to 17-AAG. *BMC Cancer*. 2014;14:1-8.
41. Fan Y, Qin G, Zhu Z. Variable selection in robust regression models for longitudinal data. *J Multivar Anal*. 2012;109:156-167.
42. Wainwright MJ, Jordan MI. Graphical models, exponential families, and variational inference. *Found Trends Mach Learn*. 2008;1:1-305.
43. Loh PL, Wainwright MJ. Support recovery without incoherence: a case for nonconvex regularization. *Ann Stat*. 2017;45:2455-2482.
44. Hwang SG. Cauchy's interlace theorem for eigenvalues of Hermitian matrices. *Am Math Mon*. 2004;111:157-159.
45. Yang E, Ravikumar P, Allen GI, Liu Z. Graphical models via univariate exponential family distributions. *J Mach Learn Res*. 2015;16:3813-3847.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Sun L, Liang F. Markov neighborhood regression for statistical inference of high-dimensional generalized linear models. *Statistics in Medicine*. 2022;41(20):4057-4078. doi: 10.1002/sim.9493

APPENDIX A

A.1 Formulation of undirected graphical models

The graphical model is often represented as a graph $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V} = \{1, 2, \ldots, p\}$ denotes the set of nodes and $\boldsymbol{E}$ denotes the set of edges. This article focuses on the case that the graph is undirected, for which $\boldsymbol{E}$ is symmetric. Let $X_1, X_2, \ldots, X_p$ denote the variables associated with the p nodes of the graph. Consider a general pairwise graphical model,[42] whose joint distribution takes the form

$$f(\boldsymbol{x}) \propto \exp\left\{\sum_{r=1}^{p} f_r(x_r) + \sum_{(u,v)\in E} f_{uv}(x_u, x_v)\right\},$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ and $f_{uv} = 0$ for $\{u, v\} \notin \boldsymbol{E}$. In this model, $f_r(x_r)$ is the node potential function and $f_{uv}(x_u, x_v)$ is the edge potential function. Furthermore, the pairwise interactions can be simplified by assuming that $f_{uv}(x_u, x_v) = \theta_{uv} x_u x_v = \theta_{vu} x_v x_u$. In this simplified case, the joint distribution of the graphical model can be expressed as

$$f(\boldsymbol{x}) = \exp\left\{\sum_{r=1}^{p} f_r(x_r) + \sum_{(u,v)\in E} \theta_{uv} x_u x_v - A(\theta)\right\}, \tag{A1}$$

where $A(\theta)$ is the log-partition function.

Let $\boldsymbol{V}_r = \boldsymbol{V} \setminus \{r\}$, and let $\boldsymbol{X}_{V_r} = (X_1, X_2, \ldots, X_{r-1}, X_{r+1}, \ldots, X_p)$. The distribution of X_r conditioned on $\boldsymbol{X}_{\boldsymbol{V}_r}$ is given by

$$f(x_r|\boldsymbol{x}_{\boldsymbol{V}_r}) = \exp\left\{ f_r(x_r) + x_r\Big(\sum_{v\neq r}\theta_{r,v}x_v\Big) - D_r(\eta_r) \right\},$$

where $\eta_r = \eta_r(\boldsymbol{\theta}_r, \boldsymbol{x}_{\boldsymbol{V}_r})$ is a function of $\boldsymbol{\theta}_r = (\theta_{r,1}, \ldots, \theta_{r,r-1}, \theta_{r,r+1}, \ldots, \theta_{r,p})$ and $\boldsymbol{x}_{\boldsymbol{V}_r}$.

A.2 Proof of validity of the MNR method for high-dimensional GLMs

Consider a GLM with the density function given by

$$f(y|\boldsymbol{x}, \boldsymbol{\beta}, \sigma) = c(y, \sigma)\exp\left\{ \frac{y\boldsymbol{x}^T\boldsymbol{\beta} - \varphi(\boldsymbol{x}^T\boldsymbol{\beta})}{d(\sigma)} \right\}, \tag{A2}$$

where σ is the dispersion parameter, $\varphi(\cdot)$ is continuously differentiable, and $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ is a p-vector of features. This class of GLMs includes normal linear regression, Poisson regression, and logistic regression, among others.

Suppose that the features of the GLM (A2) can be modeled by a mixed graphical model, for which the joint distribution can be represented in the form (A1), and the conditional distribution of each variable X_r can be represented by a GLM in the form

$$f(x_r|\boldsymbol{x}_{\boldsymbol{V}_r}, \boldsymbol{\theta}_r, \tau) = c(x_r, \tau)\exp\left\{ \frac{x_r\boldsymbol{x}_{\boldsymbol{V}_r}\boldsymbol{\theta}_r - \psi(\boldsymbol{x}_{\boldsymbol{V}_r}\boldsymbol{\theta}_r)}{d(\tau)} \right\}, \tag{A3}$$

where $\boldsymbol{V} = \{1, 2, \ldots, p\}$ is the index set of features, $\boldsymbol{V}_r = \boldsymbol{V} \setminus \{r\}$, τ is the dispersion parameter, and $\psi(\cdot)$ is the cumulant function. Moreover, the connectivity of the graphical model is determined by $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p$; that is, for any pair of nodes (i, j), the edge exists if and only if both $\theta_{i,j}$ and $\theta_{j,i}$ are nonzero. We refer to Reference 20 for discussions on compatibility of the conditional and joint distributions. Taking the mixed graphical model by Gaussian and binomial random variables as an example, for which the joint distribution is given in Reference 21, the conditional distribution of each Gaussian random variable can be represented as a linear regression, and the conditional distribution of each binomial random variable can be represented as a logistic regression.

To conduct variable selection for the GLM (A2), a regularization method is used, which is to solve the minimization problem

$$\arg\min_{\|\boldsymbol{\beta}\|\leq R} \mathcal{L}_n(\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta}), \tag{A4}$$

where $\mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \log f(y^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{\beta}, \sigma)$ with the super-index i indexing the observations, $\rho_\lambda(\cdot)$ is the regularized function, and R is a scalar.

To determine the structure of the mixed graphical model formed by the features, the nodewise regression method is used, which employs a regularization method to estimate the parameters of the conditional distribution for each node. That is, for each node r, it is to estimate $\boldsymbol{\theta}_r^*$, the true value of $\boldsymbol{\theta}_r$, by solving the minimization problem

$$\arg\min_{\|\boldsymbol{\theta}_r\|\leq R} \mathcal{L}_n(\boldsymbol{\theta}_r) + \rho_\lambda(\boldsymbol{\theta}_r), \tag{A5}$$

where $\mathcal{L}_n(\boldsymbol{\theta}_r) = \frac{1}{n}\sum_{i=1}^n \log f(x_r^{(i)}|\boldsymbol{x}_{\boldsymbol{V}_r}^{(i)}, \boldsymbol{\theta}_r^*, \tau)$ with the super-index i indexing the observations, $\rho_\lambda(\cdot)$ is the regularized function, and R is a scalar. As stated in Reference 8, the support of $\boldsymbol{\theta}_r$ can be used to estimate the neighborhood of the node X_r.

The remaining part of this section is organized as follows. Section A.2.1 gives some definitions on the penalty and loss functions. Section A.2.2 proves the consistency of Markov blanket estimation for each feature X_r. Section A.2.3 proves the consistency of variable selection for the model (9). Finally, Section A.2.4 proves Theorem 1 of the main text.

A.2.1 Definitions on the penalty and loss functions

This section defines two terms, namely, amenable penalty and restricted strong convexity (RSC), which have been discussed in References 26 and 43.

Definition 1 (Amenable regularizer). A penalty function $\rho_\lambda : \mathbb{R} \to \mathbb{R}$ is said μ-amenable for some constant $\mu \geq 0$ if the following conditions are satisfied:

(i) The function $\rho_\lambda(\cdot)$ is symmetric around 0 [ie, $\rho_\lambda(t) = \rho_\lambda(-t)$ for all t] and $\rho_\lambda(0) = 0$.
(ii) The function $\rho_\lambda(\cdot)$ is nondecreasing on $\mathbb{R}^+$.
(iii) The function $t \to \frac{\rho_{\lambda(t)}}{t}$ is nonincreasing on $\mathbb{R}^+$.
(iv) The function $\rho_\lambda(\cdot)$ is differential, for $t \neq 0$.
(v) The function $\rho_\lambda(t) + \frac{\mu}{2}t^2$ is convex, for some $\mu > 0$.
(vi) $\lim_{t\to 0^+} \rho'_\lambda(t) = \lambda$.

The amenable penalty is very general, which includes the Lasso penalty,[1] SCAD penalty,[2] and MCP penalty[3] as special cases. The following lemma on amenable regularizers was taken from Reference 43.

Lemma 1. *If $\rho_\lambda(\cdot)$ is a μ-amenable regularizer, then the following holds:*

(i) For all $t \neq 0$, $|\rho'_\lambda(t)| \leq \lambda$.
(ii) The function $q_\lambda(t) - \frac{\mu}{2}t^2$ is concave and differentiable everywhere.

Definition 2 (RSC). An empirical loss function $\mathcal{L}_n$ is said satisfying an (α, γ)-RSC condition if there exist constants $\alpha_1, \alpha_2 > 0$ and $\gamma_1, \gamma_2 \geq 0$ such that for any pair $\boldsymbol{\theta}, \Delta \in \mathbb{R}^p$,

$$\langle \nabla \mathcal{L}_n(\boldsymbol{\theta} + \Delta) - \nabla \mathcal{L}_n(\boldsymbol{\theta}), \Delta \rangle \geq \begin{cases} \alpha_1 \|\Delta\|_2^2 - \gamma_1 \frac{\log p}{n} \|\Delta\|_1^2, & \forall\ \|\Delta\|_2 \leq 1, \\ \alpha_2 \|\Delta\|_2 - \gamma_2 \sqrt{\frac{\log p}{n}} \|\Delta\|_1, & \forall\ \|\Delta\|_2 \geq 1. \end{cases} \tag{A6}$$

The following lemma was taken from Reference 26, which states that if $\mathcal{L}_n$ is convex, given α_1 and γ_1 in (A6), then $\alpha_2 = \alpha_1$ and $\gamma_2 = 1$ hold.

Lemma 2. *Given the empirical loss function $\mathcal{L}_n$ as defined in (A5), $\|\boldsymbol{\theta}\|_1 \leq R$, R is a scalar, and $\mathcal{L}_n$ is convex. If the first equation in (A6) holds and $n \geq 4R^2\gamma_1^2 \log p$, then*

$$\langle \nabla \mathcal{L}_n(\boldsymbol{\theta} + \Delta) - \nabla \mathcal{L}_n(\boldsymbol{\theta}), \Delta \rangle \geq \alpha_1 \|\Delta\|_2 - \sqrt{\frac{\log p}{n}} \|\Delta\|_1,\ \forall\ \ \|\Delta\|_2 \geq 1.$$

A.2.2 Consistency of Markov blanket estimation

This section provides a theoretical guarantee for the consistency of Markov blanket estimation when the nodewise regression method with a nonconvex penalty function is used for estimation. Refer to the GLM (A3), we let $\boldsymbol{\theta}_r$ denote a p-dimensional parameter vector of the regression for node r, which includes an intercept term stored as the rth element $\theta_{r,r}$. In general, we let $\theta_{r,i}$ denote the ith element of $\boldsymbol{\theta}_r$. Let $\boldsymbol{\theta}_r^*$ denote the true regression parameter vector for node r, and let $S_r = \{i : \theta_{r,i}^* \neq 0, i \neq r\}$ denote the support of $\boldsymbol{\theta}_r^*$. Let $k = \max_{r=1}^p |S_r|$. Formally, the following assumptions are made for the proof.

Assumption 1. For each node r, the empirical loss function $\mathcal{L}_n(\boldsymbol{\theta}_r)$ is convex and satisfies the RSC-condition with (α_1, γ_1); in addition, the penalty function ρ_λ is μ-amenable with $\mu < \frac{4}{3}\alpha_1$, and there exist constants λ and R such that

$$4 \max\left\{ \|\nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)\|_\infty, \alpha_2 \sqrt{\frac{\log k}{n}} \right\} \leq \lambda \leq \sqrt{\frac{(4\alpha_1 - 3\mu)\alpha_2}{384k}},$$
$$\max\left\{ 2\|\boldsymbol{\theta}_r^*\|_1, \frac{48k\lambda}{4\alpha_1 - 3\mu} \right\} \leq R \leq \min\left\{ \frac{\alpha_2}{8\lambda}, \alpha_2 \sqrt{\frac{n}{\log p}} \right\}.$$

The following lemma is a restatement of Theorem 1 of Reference 26.

Lemma 3. *Suppose that a regularizer satisfies Definition 1, an empirical convex loss $\mathcal{L}_n$ satisfies Definition 2, and a regularization parameter λ satisfies Assumption 1. If the sample size $n \geq \frac{16R^2 \max\{\gamma_1^2, 1\}}{\alpha^2} \log p$, for any vector $\tilde{\theta}$ satisfying the following*

first-order condition

$$\langle \nabla \mathcal{L}_n(\tilde{\theta}) + \nabla \rho_\lambda(\tilde{\theta}), \theta - \tilde{\theta} \rangle \geq 0, \quad \text{for all feasible } \theta \in \mathbb{R}^p,$$

then the following error bound holds:

$$\|\tilde{\theta} - \theta^*\|_2 \leq \frac{6\lambda\sqrt{k}}{4\alpha_1 - 3\mu},$$

where $\|\theta^*\|_0 = k$.

Assumption 2. There exists a constant $\rho_{\max}$ such that the largest eigenvalue of the sample covariance matrix satisfies

$$\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}^{(i)})(\boldsymbol{X}^{(i)})^T\right) < \rho_{\max} < \infty, \tag{A7}$$

where $\boldsymbol{X}^{(i)}$ denotes the ith row of the design matrix $\boldsymbol{X}$, and $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of a matrix.

The following lemma is known as the eigenvalue interlacing theorem.[44]

Lemma 4. *Suppose* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *is a real symmetric matrix. Let* $\boldsymbol{B} \in \mathbb{R}^{m\times m}$, $m < n$, *be a principal submatrix of symmetric matrix* $\boldsymbol{A}$. *Suppose the matrix* $\boldsymbol{A}$ *has the eigenvalues* $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$, *and the matrix* $\boldsymbol{B}$ *has the eigenvalues* $\gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_m$. *Then*

$$\lambda_k \leq \gamma_k \leq \lambda_{k+n-m}, \quad k = 1, 2, \ldots, m.$$

If $m = n - 1$, *then*

$$\lambda_1 \leq \gamma_1 \leq \lambda_2 \leq \gamma_2 \leq \cdots \leq \gamma_{n-1} \leq \lambda_n.$$

By Lemma 4, Assumption 2 implies that for any $r \in \boldsymbol{V}$,

$$\lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_{\boldsymbol{V}_r}^{(i)}\boldsymbol{X}_{\boldsymbol{V}_r}^{(i)\,T}\right) < \rho_{\max} < \infty. \tag{A8}$$

For notational simplicity, we let $\mathbf{Q}^* = \nabla^2\mathcal{L}_n(\boldsymbol{\theta}_r^*)$, and let $\boldsymbol{Q}_{A,B}^*$ denote a submatrix of $\boldsymbol{Q}^*$ formed with the rows in the set A and the columns in the set B.

Assumption 3. For each node r, there exists a constant $\rho_{\min}$ such that $\lambda_{\min}(\mathbf{Q}_{S_r,S_r}^*) > \rho_{\min} > 0$, where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of a matrix.

Assumption 4 (Incoherence condition). For each node r, there exists a constant $\eta \in (0, 1)$ such that

$$|||\mathbf{Q}_{S_r^c,S_r}^*(\mathbf{Q}_{S_r,S_r}^*)^{-1}|||_\infty \leq 1 - \eta,$$

where S_r^c denotes the complementary set of S_r.

Assumption 5. For each node r, there exists a constant $C > 0$ such that $\theta_{r,\min}^* \geq 2C\sqrt{\frac{k\log p}{n}}$, where $\theta_{r,\min}^* = \min_{i=1}^{p-1}|\theta_{r,i}^*|$.

Assumptions 6 and 7 were taken from Reference 45. For mixed graphical models, we cannot bound each variable directly, but we can bound their first and second moments.

Assumption 6. For each node r, the first and second moments of X_r are bounded, that is, there exist finite constants κ_m and κ_v such that

$$\mathbb{E}|X_r| < \kappa_m \quad \text{and} \quad \mathbb{E}(X_r^2) < \kappa_v.$$

Further, the log-partition function $A(\cdot)$ of the joint distribution satisfies

$$\max_{u:|u|\le 1} \frac{\partial^2}{\partial \theta_r^2} A(\boldsymbol{\theta}^* + ue_r) \le \kappa_h,$$

for some constant $\kappa_h < \infty$, where $\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*, \ldots, \theta_p^*)$, and $e_r \in \mathbb{R}^{p^2}$ is an indicator vector equaling to one at the index corresponding to θ_r and zero elsewhere. In addition,

$$\max_{\eta:|\eta|\le 1} \frac{\partial^2}{\partial \eta^2} \bar{A}_r(\eta, \boldsymbol{\theta}^*) \le \kappa_h,$$

where $\bar{A}_r(\eta, \boldsymbol{\theta})$ is given by

$$\bar{A}_r(\eta; \boldsymbol{\theta}) = \log \int_{\mathcal{X}} \exp\left\{ \eta x_r^2 + \sum_{u\in V} \theta_u x_u + \sum_{(u,v)\in E} \theta_{uv} x_u x_v + \sum_{u\in V} C(x_u) \right\} d\boldsymbol{x}$$

for some scalar η.

With Assumption 6, the following two lemmas can be established, whose proofs can be found in Reference 45.

Lemma 5. *Suppose Assumption 6 holds. For any node X_r of a mixed graphical model, we have*

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} (X_r^{(i)})^2 \ge \delta \right) \le \exp(-cn\delta^2),$$

where $0 < \delta \le \min\{2\kappa_v/3, \kappa_h + \kappa_v\}$, and $c > 0$ is the constant.

Lemma 6. *Suppose Assumption 6 holds. For any node X_r of a mixed graphical model and any $i \in \{1, 2, \ldots, n\}$,*

$$\mathbb{P}(|X_r^{(i)}| \ge \delta \log \eta) \le c\eta^{-\delta},$$

where $\delta > 0$ is any positive real number, and $c > 0$ is a constant.

Assumption 7. There exist constants $\kappa_2 > 0$ and $\kappa_3 > 0$ such that $\|\psi'(t)\|_\infty \le \kappa_2$ and $\|\psi''(t)\|_\infty \le \kappa_3$, for any $t \in \mathbb{R}$, where $\psi(\cdot)$ is the cumulant function for the GLM (A3).

Lemma 7 is a restatement of Theorem 1 of Reference 43, which is the key to the proof of Theorem 2.

Lemma 7 (Theorem 1 of Reference 43). *Consider the GLM (9) and the regularized M-estimator*

$$\widehat{\boldsymbol{\beta}} \in \arg\min_{\|\boldsymbol{\beta}\|_1 \le R} \{\mathcal{L}_n(\boldsymbol{\beta}) + \rho_\lambda(\boldsymbol{\beta})\}. \tag{A9}$$

Suppose $\mathcal{L}_n$ is a twice-differentiable, (α, τ)-RSC function and ρ_λ is μ-amenable for some $\mu < \frac{3}{4}\alpha_1$. Further suppose that:

(a) The parameters (λ, R) satisfy the bounds

$$4\max\left\{ \|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty, \alpha_2\sqrt{\frac{\log k}{n}} \right\} \le \lambda \le \sqrt{\frac{(4\alpha_1 - 3\mu)\alpha_2}{384k}},$$
$$\max\left\{ 2\|\boldsymbol{\beta}^*\|_1, \frac{48k\lambda}{4\alpha_1 - 3\mu} \right\} \le R \le \min\left\{ \frac{\alpha_2}{8\lambda}, \alpha_2\sqrt{\frac{n}{\log p}} \right\}, \tag{A10}$$

where $\boldsymbol{\beta}^*$ denotes the true parameter vector of the GLM (9).

(b) For some $\delta \in [4R\tau_1 \log p/(n\lambda), 1]$, the dual vector $\widehat{z}$ from the primal-dual witness (PWD) construction satisfies the strict dual feasibility condition:

$$\|\widehat{z}_{S^c}\| \le 1 - \delta, \tag{A11}$$

where S denotes the support set of $\boldsymbol{\beta}^*$ and S^c denotes the complementary set of S.

Then if $n \succ k \log p$ and $\boldsymbol{\beta}^*$ is k-sparse, (A9) has a unique stationary point given by the primal output $\widehat{\boldsymbol{\beta}}$ given by the PDW construction.

Theorem 2. *Consider a set of variables $\{X_1, X_2, \ldots, X_p\}$ and the associated graphical model $\boldsymbol{G} = (\boldsymbol{V}, \boldsymbol{E})$, for which the node-wise GLM is given by (A3). Suppose that for each node X_r, a regularization method is used to estimate $\boldsymbol{\theta}_r$ by minimizing the empirical loss (A5). If Assumptions 1 to 7 hold and $n \succ k \log p$, where k has been assumed to increase with n and p, then there exist constants c_1, c_2, and c_3 such that*

$$P(\hat{S}_r = S_r) \geq 1 - 2\exp\{-c_1 \log p\} - \exp\{-c_2 \log p\} - c_3 \min\{n, p\}^{-2}.$$

Proof. The proof is based on the PDW strategy,[43] which is to first show that the primal output $\hat{\boldsymbol{\beta}}_r$ given by the PDW construction is the unique stationary point of (A5), and then show that the support of $\boldsymbol{\beta}_r^*$ can be recovered by $\hat{\boldsymbol{\beta}}_r$ as the sample size $n \to \infty$. For the first step, by Lemma 7, it suffices to verify the strict dual feasibility condition (A11) for the conditional distribution of each node X_r as other conditions of Lemma 7 have been satisfied by our assumptions. The second step can be accomplished based on the $\beta_{\min}$-condition, that is, Assumption 5.

Part (i), which is to verify the strict dual feasibility condition. Let $\hat{\boldsymbol{\theta}}_r := ((\hat{\boldsymbol{\theta}}_r)_{S_r}, \boldsymbol{0}_{S_r^c})$ be the primal output constructed with the PDW technique. Let $q_\lambda(t) = \lambda|t| - \rho_\lambda(t)$. Differentiating the loss (A5) with respect to $\boldsymbol{\theta}_r$ leads to the equation

$$\nabla \mathcal{L}_n(\hat{\boldsymbol{\theta}}_r) - \nabla q_\lambda(\hat{\boldsymbol{\theta}}_r) + \lambda \hat{\boldsymbol{z}} = 0, \tag{A12}$$

through which $\hat{\boldsymbol{z}}_{S_r^c}$ is defined; that is, $\hat{\boldsymbol{z}} = (\hat{\boldsymbol{z}}_{S_r}, \hat{\boldsymbol{z}}_{S_r^c})$, where $\hat{\boldsymbol{z}}_{S_r} \in \partial \|(\hat{\boldsymbol{\theta}}_r)_{S_r}\|_1$, and $\hat{\boldsymbol{z}}_{S_r^c}$ is chosen to satisfy the above zero sub-gradient condition.

Equation (A12) implies that

$$\nabla \mathcal{L}_n(\hat{\boldsymbol{\theta}}_r) - \nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*) + \nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*) - \nabla q_\lambda(\hat{\boldsymbol{\theta}}_r) + \lambda \hat{\mathbf{z}} = 0.$$

Define $\mathbf{Q} = \int_0^1 \nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_r^* + t(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*))dt$. Then

$$\mathbf{Q}(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*) + \nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*) - \nabla q_\lambda(\hat{\boldsymbol{\theta}}_r) + \lambda \hat{\mathbf{z}} = 0$$

and

$$\mathbf{Q}^*(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*) + (\mathbf{Q} - \mathbf{Q}^*)(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*) + \nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*) - \nabla q_\lambda(\hat{\boldsymbol{\theta}}_r) + \lambda \hat{\mathbf{z}} = 0.$$

To simplify the equation, we define $\mathbf{Re} := (\mathbf{Q} - \mathbf{Q}^*)(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*)$. Then the equation can be expressed in the matrix form as

$$\begin{bmatrix} \mathbf{Q}^*_{S,S} & \mathbf{Q}^*_{S,S^c} \\ \mathbf{Q}^*_{S^c,S} & \mathbf{Q}^*_{S^c,S^c} \end{bmatrix} \begin{bmatrix} (\hat{\boldsymbol{\theta}}_r)_S - (\boldsymbol{\theta}_r)^*_S \\ 0 \end{bmatrix} + \begin{bmatrix} \mathbf{Re}_S \\ \mathbf{Re}_{S^c} \end{bmatrix} + \begin{bmatrix} \nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)_S - \nabla q_\lambda(\hat{\boldsymbol{\theta}}_r)_S \\ \nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)_{S^c} - \nabla q_\lambda(\hat{\boldsymbol{\theta}}_r)_{S^c} \end{bmatrix} + \lambda \begin{bmatrix} \hat{\mathbf{z}}_S \\ \hat{\mathbf{z}}_{S^c} \end{bmatrix} = 0,$$

where $S = S_r = \{i : \theta^*_{r,i} \neq 0\}$ by the PDW construction. Further, by some algebra,

$$\begin{aligned} \hat{\boldsymbol{z}}_{S^c} = \frac{1}{\lambda} \Big\{ & (\nabla q_\lambda(\hat{\boldsymbol{\theta}}_r)_{S^c} - \nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)_{S^c} - \mathbf{Re}_{S^c}) \\ & + \mathbf{Q}^*_{S^cS}(\mathbf{Q}^*)^{-1}_{S,S}(\mathbf{Re}_S + \nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)_S - \nabla q_\lambda(\hat{\boldsymbol{\theta}}_r)_S + \lambda \hat{\boldsymbol{z}}_S) \Big\}. \end{aligned}$$

By the selection property (vi) of the nonconvex regularizer, we have

$$\nabla q_\lambda(\hat{\boldsymbol{\theta}}_r)_{S^c} = \nabla q_\lambda(0_{S^c}) = 0_{S^c},$$

which leads to

$$\hat{\boldsymbol{z}}_{S^c} = -\frac{1}{\lambda}(\nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)_{S^c} + \mathbf{Re}_{S^c}) + \frac{1}{\lambda}\mathbf{Q}^*_{S^cS}(\mathbf{Q}^*)^{-1}_{S,S}(\mathbf{Re}_S + \nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)_S - \nabla q_\lambda(\hat{\boldsymbol{\theta}}_r)_S + \lambda \hat{\boldsymbol{z}}_S). \tag{A13}$$

By Lemma 1, we have

$$\|\nabla q_\lambda(\hat{\theta}_r)_S - \lambda \hat{\boldsymbol{z}}_S)\|_\infty \leq \lambda.$$

Taking the supreme norm on both sides of (A13) via applying Assumption 4, we have

$$\|\hat{\boldsymbol{z}}_{S^c}\|_\infty \leq \frac{2-\eta}{\lambda}\|\nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)\|_\infty + \frac{2-\eta}{\lambda}\|\mathbf{Re}\|_\infty + 1 - \eta. \tag{A14}$$

To have the strictly dual feasibility $\|\boldsymbol{z}_{S^c}\|_\infty < 1$, it suffices to show the following two inequalities hold:

$$\|\nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)\|_\infty < \frac{\eta}{2(2-\eta)}\lambda, \tag{A15}$$

$$\|(\mathbf{Q} - \mathbf{Q}^*)(\hat{\theta}_r - \boldsymbol{\theta}_r^*)\|_\infty < \frac{\eta}{2(2-\eta)}\lambda. \tag{A16}$$

Consider the inequality (A15). Given a constant $c > 0$, we need to show the upper bound for the probability

$$\mathbb{P}\left(\|\nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)\|_\infty \geq c\sqrt{\frac{1}{k}}\right).$$

For each observation $i \in \{1, 2, \ldots, n\}$ and each variable $j \in \boldsymbol{V}_r = \boldsymbol{V} \setminus \{r\}$, we define the random variable $W_j^{(i)} = (\psi'(\boldsymbol{x}_{V_r}^{(i)}\boldsymbol{\theta}_r^*) - x_r^{(i)})x_j^{(i)}$. It is easy to figure out that the jth component of $\nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)$ is equal to $\frac{1}{n}\sum_{i=1}^n W_j^{(i)}$. Our goal is to bound the above probability via bounding $\max_{j\in\boldsymbol{V}_r}|\frac{1}{n}\sum_{i=1}^n W_j^{(i)}|$. Toward this goal, we define the event

$$\mathcal{A} := \left\{\max_{j\in\boldsymbol{V}_r}\left\{\frac{1}{n}\sum_{i=1}^n\left(X_j^{(i)}\right)^2\right\} \leq \Delta\right\},$$

for some constant $\Delta \leq \min\{2\kappa_v/3, \kappa_h + \kappa_v\}$. Then

$$\mathbb{P}\left(\|\nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)\|_\infty \geq c\sqrt{\frac{1}{k}}\right) \leq \mathbb{P}(\mathcal{A}^c) + \mathbb{P}\left(\left\{\|\nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)\|_\infty \geq c\sqrt{\frac{1}{k}}\right\} \cap \mathcal{A}\right) \triangleq I_1 + I_2. \tag{A17}$$

To bound the term I_2, we consider the expectation of the moment function conditioned on $\boldsymbol{X}_{\boldsymbol{V}_r}$. For any $t \in \mathbb{R}$,

$$\begin{aligned}\log \mathbb{E}[\exp(tW_j^{(i)})|\boldsymbol{X}_{\boldsymbol{V}_r}] &= \log[\exp(tX_j^{(i)}\psi'(\boldsymbol{X}_{V_r}^{(i)}\boldsymbol{\theta}_r^*))] \cdot \mathbb{E}[\exp(-tX_j^{(i)}X_r^{(i)})] \\ &= tX_j^{(i)}\psi'(\boldsymbol{X}_{V_r}^{(i)}\boldsymbol{\theta}_r^*) \cdot \mathbb{E}[\exp(-tX_j^{(i)}X_r^{(i)})] \\ &= tX_j^{(i)}\psi'(\boldsymbol{X}_{V_r}^{(i)}\boldsymbol{\theta}_r^*) + (\psi(-tX_j^{(i)} + \boldsymbol{X}_{V_r}^{(i)}\boldsymbol{\theta}_r^*) - \psi(\boldsymbol{X}_{V_r}^{(i)}\boldsymbol{\theta}_r^*)),\end{aligned}$$

where the last equality follows from the fact that ψ is the cumulant generating function for the underlying exponential family. Then, by using the Taylor expansion for the function $\psi(-tX_j^{(i)} + \boldsymbol{X}_{V_r}^{(i)}\boldsymbol{\theta}_r^*)$ at $\boldsymbol{X}_{V_r}^{(i)}\boldsymbol{\theta}_r^*$,

$$\log \mathbb{E}[\exp(tW_j^{(i)})|\boldsymbol{X}_{\boldsymbol{V}_r}] = \frac{1}{2}(X_j^{(i)}t)^2\psi''(\boldsymbol{X}_{V_r}^{(i)}\boldsymbol{\theta}_r^*) \leq \frac{\kappa_2}{2}(X_j^{(i)})^2t^2,$$

where the bound κ_2 is from Assumption 7. Therefore,

$$\mathbb{E}[\exp(tW_j^{(i)})|\boldsymbol{X}_{\boldsymbol{V}_r}] \leq \exp\left\{\frac{\kappa_2}{2}(X_j^{(i)})^2t^2\right\}.$$

Since $W_j^{(1)}, W_j^{(2)}, \ldots, W_j^{(n)}$ are mutually independent,

$$\mathbb{E}\left[\exp\left(t\sum_{i=1}^n W_j^{(i)}\right)|\boldsymbol{X}_{\boldsymbol{V}_r}\right] \leq \exp\left\{\frac{\kappa_2}{2}\sum_{i=1}^n(X_j^{(i)})^2t^2\right\}.$$

Consequently, conditioned on the $\boldsymbol{X}_{V_r}$, intersected with the event $\mathcal{A}$, $\sum_{i=1}^{n} W_j^{(i)}$ is sub-Gaussian. By the tail probability inequality of sub-Gaussian (Chernoff bound),

$$\mathbb{P}\left(\left\{\|\nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)\|_\infty \geq c\sqrt{\frac{1}{k}}\right\} \cap \mathcal{A} | \boldsymbol{X}_{V_r}\right) \leq 2\exp\left\{-\frac{c^2 n^2}{2n\kappa_2 \Delta}\frac{1}{k} + \log p\right\} \leq 2\exp\{-c_1 \log p\},$$

for some constant $c_1 > 0$, where the last inequality holds as $n \succ k \log p$. Integrating over the values of $\boldsymbol{X}_{V_r}$, we get a bound for I_2:

$$I_2 = \mathbb{P}\left(\left\{\|\nabla \mathcal{L}_n(\boldsymbol{\theta}_r^*)\|_\infty \geq c\sqrt{\frac{1}{k}}\right\} \cap \mathcal{A}\right) \leq 2\exp\{-c_1 \log p\}.$$

Next, we consider to bound the term I_1 in (A17). By using Lemma 5, if $n \succ k \log p$, then there exists a constant $c_2 > 0$ such that

$$\mathbb{P}(\mathcal{A}^c) = \mathbb{P}\left(\max_{j \in V_r}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_j^{(i)})^2\right\} > \Delta\right) = 1 - \prod_{j \in V_r}\left(1 - \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}(X_j^{(i)})^2 > \Delta\right)\right) \leq e^{-cn\Delta^2 + \log(p-1)} \leq e^{-c_2 \log(p)}.$$

Combining the bounds for I_1 and I_2, we get

$$\mathbb{P}\left(\|\mathcal{L}_n(\boldsymbol{\theta}_r^*)\|_\infty \leq c\sqrt{\frac{1}{k}}\right) \leq 1 - 2\exp(-c_1 \log p) - \exp(-c_2 \log p).$$

Now we start to consider the inequality (A16), where

$$\begin{aligned}
\mathbf{Q} - \mathbf{Q}^* &= \mathbf{Q} - \nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_r^*) \\
&= \int_0^1 \left\{\nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_r^* + t(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*)) - \nabla^2 \mathcal{L}_n(\boldsymbol{\theta}_r^*)\right\} dt \\
&= \int_0^1 \frac{1}{n}\sum_{i=1}^{n}\left\{\psi''(\boldsymbol{X}_{V_r}^{(i)}(\boldsymbol{\theta}_r^* + t(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*)) - \psi''(\boldsymbol{X}_{V_r}^{(i)}\boldsymbol{\theta}_r^*)\right\}\boldsymbol{X}_{V_r}^{(i)}(\boldsymbol{X}_{V_r}^{(i)})^T dt \\
&= \int_0^1 \left\{t \cdot \frac{1}{n}\sum_{i=1}^{n}\psi'''(\boldsymbol{X}_{V_r}^{(i)}\boldsymbol{\theta}_r^m)\boldsymbol{X}_{V_r}^{(i)}(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*)^T\boldsymbol{X}_{V_r}^{(i)}(\boldsymbol{X}_{V_r}^{(i)})^T\right\} dt,
\end{aligned}$$

where $\boldsymbol{\theta}_r^m$ denotes a point between $\boldsymbol{\theta}_r^*$ and $\hat{\boldsymbol{\theta}}_r$. Given an indicator vector $\mathbf{e}_j$, $j \in V_r$, whose jth element is 1 and all other elements are 0. Therefore,

$$\begin{aligned}
\mathbf{e}_j^T(\mathbf{Q} - \mathbf{Q}^*) &= \int_0^1 \left\{t \cdot \frac{1}{n}\sum_{i=1}^{n}\psi'''(\boldsymbol{X}_{V_r}^{(i)}\boldsymbol{\theta}_r^m)\mathbf{e}_j^T\boldsymbol{X}_{V_r}^{(i)}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T\boldsymbol{X}_{V_r}^{(i)}(\boldsymbol{X}_{V_r}^{(i)})^T\right\} dt \\
&\leq \frac{\kappa_3}{2} \cdot \frac{1}{n}\sum_{i=1}^{n}\mathbf{e}_j^T\boldsymbol{X}_{V_r}^{(i)}(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*)^T\boldsymbol{X}_{V_r}^{(i)}(\boldsymbol{X}_{V_r}^{(i)})^T.
\end{aligned}$$

Define the event $\mathcal{B} := \left\{\max_{r,i}|X_r^{(i)}| \leq 4\log(\min\{n,p\})\right\}$. By using Lemma 6, we have that for any $j \in V_r$, with probability $1 - c_3 \min\{n,p\}^{-2}$, the following inequality holds:

$$\mathbf{e}_j^T(\mathbf{Q} - \mathbf{Q}^*)(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*) \leq \frac{\kappa_3}{2} \cdot \frac{1}{n}\sum_{i=1}^{n}\mathbf{e}_j^T\boldsymbol{X}_{V_r}^{(i)}(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*)^T\boldsymbol{X}_{V_r}^{(i)}(\boldsymbol{X}_{V_r}^{(i)})^T(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_r^*)$$

$$\leq 2\kappa_3 \log(\min\{n,p\}) \cdot \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_r - \theta_r^*)^T \boldsymbol{X}_{V_r}^{(i)}(\boldsymbol{X}_{V_r}^{(i)})^T(\hat{\theta}_r - \theta_r^*)$$
$$\leq 2\kappa_3 \log(\min\{n,p\})\rho_{\max}\|\hat{\theta}_r - \theta_r^*\|_2^2.$$

Further, it implies

$$\|(\mathbf{Q} - \mathbf{Q}^*)(\hat{\theta}_r - \theta_r^*)\|_\infty \leq 2\kappa_3 \log(\min\{n,p\})\rho_{\max}\|\hat{\theta}_r - \theta_r^*\|_2^2$$
$$\leq 2\kappa_3\rho_{\max}\log p \cdot C^2\lambda^2 k,$$

where the last inequality follows from Lemma 3 that $\|\hat{\theta}_r - \theta_r^*\|_2 \leq C\lambda\sqrt{k}$ with $C = \frac{6}{4\alpha_1 - 3\mu} > 0$.

Therefore, if $\lambda k \leq \frac{1}{2\kappa_3\rho_{\max}\log p \cdot C^2}\frac{\eta}{2(2-\eta)}$ holds, then, with probability $1 - c_3\min\{n,p\}^{-2}$, we have

$$\|(\mathbf{Q} - \mathbf{Q}^*)(\hat{\theta}_r - \theta_r^*)\|_\infty < \frac{\eta}{2(2-\eta)}\lambda.$$

As the consequence, with probability $1 - 2\exp\{-c_1\log p\} - \exp\{-c_2\log p\} - c_3\min\{n,p\}^{-2}$, the strictly dual feasibility holds.

Part (ii), which is to show the support of θ_r^* can be recovered as the sample size $n \to \infty$. By Lemma 3, we have

$$\|\hat{\theta}_r - \theta_r^*\|_\infty \leq \|\hat{\theta}_r - \theta_r^*\|_2 \leq C \cdot \lambda\sqrt{k}.$$

Thus, with probability $1 - 2\exp\{-c_1\log p\} - \exp\{-c_2\log p\} - c_3\min\{n,p\}^{-2}$, we have

$$\|\hat{\theta}_r - \theta_r^*\|_\infty \leq \|\hat{\theta}_r - \theta_r^*\|_2 \leq C\sqrt{\frac{k\log p}{n}}.$$

By using Assumption 5 that $\theta_{r,\min}^* > 2C\sqrt{\frac{k\log p}{n}}$, we can prove the consistency of Markov blanket estimation, that is, $\mathbb{P}(\hat{\mathbf{S}}_r = \mathbf{S}_r) > 1 - 2\exp\{-c_1\log p\} - \exp\{-c_2\log p\} - c_3\min\{n,p\}^{-2}$. ∎

A.2.3 Consistency of variable selection for the GLM

Let $\mathbf{Q} = \nabla^2\mathcal{L}_n(\boldsymbol{\beta}^*)$, where $\boldsymbol{\beta}^*$ denote the true regression coefficient vector of the GLM. Let $\boldsymbol{Q}_{A,B}$ denote a submatrix of $\boldsymbol{Q}$ formed with the rows in the set A and the columns in the set B. Let S_* denote the support of $\boldsymbol{\beta}^*$, and let $S_*^c = \boldsymbol{V} \setminus S_*$ denote the complementary set of S_*.

Assumption 8. There exist a constant $\rho_{\min}$ such that $\lambda_{\min}(\mathbf{Q}_{S_*^c,S_*}) > \rho_{\min} > 0$, where $\lambda_{\min}(\cdot)$ denotes the minimum eigenvalue of a matrix.

Assumption 9 (Incoherence condition). There exists a constant $\eta \in (0,1)$ such that

$$|||\mathbf{Q}_{S_*^c,S_*}(\mathbf{Q}_{S_*,S_*})^{-1}|||_\infty \leq 1 - \eta.$$

Assumption 10. There exists a constant C such that $\beta_{\min}^* \geq 2C\sqrt{\frac{k\log p}{n}}$, where $k = |S_*|$ denotes the size of the true model, and $\beta_{\min}^*$ is the smallest nonzero entry of $\boldsymbol{\beta}^*$.

Assumption 11. There exist constants $\kappa_2 > 0$ and $\kappa_3 > 0$ such that $\|\varphi''(t)\|_\infty \leq \kappa_2$ and $\|\varphi'''(t)\|_\infty \leq \kappa_3$, for any $t \in \mathbb{R}$, where $\varphi(\cdot)$ is the cumulant function of the GLM (9).

Corollary 1. *Consider the GLM (9) and the regularized M-estimator (A9). Suppose $\mathcal{L}_n$ is a twice-differentiable, (α,τ)-RSC function and ρ_λ is μ-amenable for some $\mu < \frac{3}{4}\alpha_1$. Further suppose that Assumptions 2, 6, and 8 to 11 hold and the parameters (λ, R) satisfy the bounds (A10). Then if $n \succ k\log p$ and $\boldsymbol{\beta}^*$ is k-sparse, the support set of $\boldsymbol{\beta}^*$ can be recovered with a probability not less than $1 - 2\exp\{-c_1\log p\} - \exp\{-c_2\log p\} - c_3\min\{n,p\}^{-2}$, where c_1, c_2, and c_3 denote some constants.*

The proof of the corollary follows from that of Theorem 2 closely and is thus omitted.

A.2.4 Proof of Theorem 1

Proof. The validity of the algorithm can be proved by verifying the conditions (6) to (8) given in the main text. The condition (7) follows from the consistency of Markov blanket estimation, which directly follows from Theorem 2. The condition (6) follows from the consistency of variable selection, which directly follows from Corollary 1. Following from (i) the sparsity of the GLM and the Markov blanket, and (ii) the consistency of variable selection and Markov blanket estimation, the condition (8) is satisfied. ■