Journal of Machine Learning Research 24 (2023) 1-69 Submitted 10/22; Revised 3/23; Published 4/23

An Annotated Graph Model with Differential Degree Heterogeneity for Directed Networks

Stefan Stein S.STEIN@WARWICK.AC.UK

Chenlei Leng C.LENG@WARWICK.AC.UK

University of Warwick
Coventry, CV4 7AL, UK

Editor: Pradeep Ravikumar

Abstract

Directed networks are conveniently represented as graphs in which ordered edges encode interactions between vertices. Despite their wide availability, there is a shortage of statistical models amenable for inference, specially when contextual information and degree heterogeneity are present. This paper presents an annotated graph model with parameters explicitly accounting for these features. To overcome the curse of dimensionality due to modelling degree heterogeneity, we introduce a sparsity assumption and propose a penalized likelihood approach with ℓ_1-regularization for parameter estimation. We study the estimation and selection consistency of this approach under a sparse network assumption, and show that inference on the covariate parameter is straightforward, thus bypassing the need for the kind of debiasing commonly employed in ℓ_1-penalized likelihood estimation. Simulation and data analysis corroborate our theoretical findings.

Keywords: β-model; Asymptotical normality; Degree heterogeneity; Homophily; Sparse networks.

1. Introduction

The need to examine inter-relationship of multiple entities in data is rapidly growing due to the increasing availability of datasets that can be conveniently represented as networks or graphs. This paper concerns a new random graph model for describing networks with directed edges. As a motivating example, Figure 1 depicts the lawyer friendship data in Lazega (2001) in which 71 lawyers were asked to name their friends: An edge from node i to node j exists if and only if lawyer i indicated in a survey that they socialized with lawyer j outside work. Mathematically, we denote the adjacency matrix of a network on n nodes as a binary matrix $A \in \mathbb{R}^{n \times n}$, where n is the number of nodes and $A_{ij} = 1$ if i points to j and 0 otherwise. As is common for this type of data, the lawyer data is annotated with contextual information in the form of nodal covariates. These covariates include a lawyer's status (partner or associate), their gender (man or woman), which of three offices they worked in, the number of years they had spent with the firm, their age, their practice (litigation or corporate) and the law school they had visited (Harvard and Yale, UConn or other). The main interest here is to understand how links were formed, especially the effect of the covariates between pairs of nodes for forming ties.

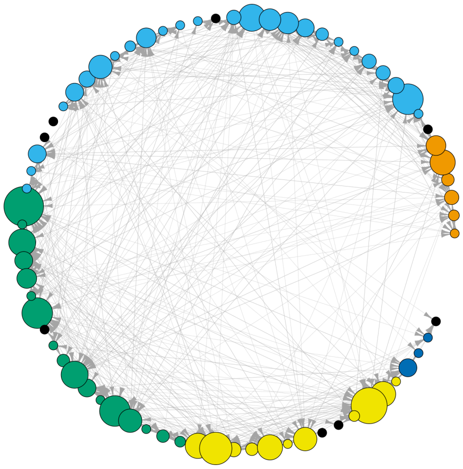


Figure 1: Lazega's lawyers friendship network. The size of the nodes corresponds to their in-degrees. For better visibility all nodes with an in-degree of five or less are plotted with the same size. The 71 lawyers are colour-coded by their age group: The lawyers aged 20-29 are represented in orange, those aged 30-39 in light-blue, those aged 40-49 in green, followed by the lawyers aged 50-59 in yellow and finally those lawyers aged 60 or older in dark-blue. The eight nodes in black correspond to lawyers with either zero in- or out-degree or both.

The lawyer network features several stylized facts of a typical real-life network. First, it exhibits degree heterogeneity, the different tendency that the nodes in a network have in participating in network activities as can be seen from Figure 1. Second, the overall network is sparse, in that the observed number of ties does not scale proportionally to the total number of possible links. In Figure 1, the average in- (and out-) degree is 8.1, whereas the maximum possible value is 70. Third, the contextual information in terms of the covariates has a role to play in determining how nodes are connected, as can be seen from the data analysis later. Here the covariates will be denoted as $Z_{ij} \in \mathbb{R}^p$ for an edge linking node i to j. In case where we only have nodal covariates denoted as $X_i \in \mathbb{R}^p$ for the ith node, a common approach is to define Z_{ij} as a function of X_i and X_j that measures their (dis)similarities.

This paper proposes a new annotated graph model that can effectively deal with the above features in directed networks. This model postulates that links are independently made with the linking probability between node i and j as

$$p_{ij} = P(A_{ij} = 1|Z_{ij}) = \frac{\exp(\alpha_i + \beta_j + \mu + \gamma^T Z_{ij})}{1 + \exp(\alpha_i + \beta_j + \mu + \gamma^T Z_{ij})}, \tag{1}$$

where $\mu \in \mathbb{R}$, $\alpha_i \in \mathbb{R}$, $\beta_j \in \mathbb{R}$ and $\gamma \in \mathbb{R}^p$ are parameters. For identifiability, we assume $\min_i\{\alpha_i\} = \min_j\{\beta_j\} = 0$, because otherwise this model becomes trivially the model in Yan et al. (2019) by absorbing μ into α_i and β_j. Clearly, to fit a model with these many parameters, we will require the number of edges of a network to be relatively large which renders (1) less useful or not even applicable for the kind of networks encountered in practice. To reduce the dimensionality of the model, we assume that both $\alpha = (\alpha_1, ..., \alpha_n)^T$ and $\beta = (\beta_1, ..., \beta_n)^T$

are sparse. While it may seem appealing to impose $\min_{1\leq i\leq n}|\alpha_i| = 0, \min_{1\leq j\leq n}|\beta_j| = 0$ instead of $\min_i\{\alpha_i\} = \min_j\{\beta_j\} = 0$, restricting the degree heterogeneity parameters only in absolute value would result in an unidentifiable parameter.

In (1), μ can be seen as a global density parameter that is allowed to diverge to $-\infty$ aiming to model sparse networks. The two node-specific parameters α_i and β_i are used to explicitly capture out- and in-degree heterogeneity respectively. The sparsity assumption on α and β introduces a notion of differential degree heterogeneity, in the sense that we only include them for nodes that are important. The effect of covariates is captured by γ. When a covariate encodes the similarity of a node attribute, a positive γ implies homophily, the tendency of nodes similar in attributes to connect, which is a widely observed phenomenon in real-life networks.

Statistical analysis of random networks has attracted enormous research attention in recent years thanks to a deluge of network data (Kolaczyk, 2009; Fienberg, 2012; Kolaczyk, 2017). Among others, there are several major classes of statistical models that have been successfully applied to model degree heterogeneity, including the stochastic block model aiming to identify groups of nodes as communities (Holland et al., 1983; Bickel and Chen, 2009; Karrer and Newman, 2011; Rohe et al., 2011; Zhao et al., 2012; Lei and Rinaldo, 2015; Gao et al., 2017; Amini and Levina, 2018; Abbe, 2018; Zhang et al., 2021), the β-model assigning individual parameters to different nodes (Chatterjee et al., 2011; Yan and Xu, 2013; Yan et al., 2016; Chen et al., 2020), and the exponential random graph models using network motifs as sufficient statistics (Holland and Leinhardt, 1981; Frank and Strauss, 1986; Robins et al., 2007).

The increasing prevalence of covariates in network data calls for models that can effectively account for their effects. For the stochastic block model, we point to Binkiewicz et al. (2017); Zhang et al. (2016); Huang and Feng (2018); Yan and Sarkar (2020). For the β-model, see Graham (2017); Yan et al. (2019), and Stein and Leng (2022). Additional references include Ma et al. (2020) where a latent space model approach is investigated and Zhang et al. (2021) that proposes a model-free approach to study the dependence of links on covariates.

Yan et al. (2019) proposed a model similar to (1) with the critical difference that α and β are dense. As a result, their model only handles relatively dense networks, and the inference on γ, the covariate parameter, warrants a bias correction step. Remarkable, the inference for this parameter in our model does not require debiasing, even when the network has vanishing link probabilities and hence are sparse. The techniques involved in deriving this result extends substantially many similar ones developed, for example in M-estimation (van der Vaart, 1998) and LASSO theory (Ravikumar et al., 2010; van de Geer and Bühlmann, 2011; van de Geer et al., 2014), where the probabilities in similar models are typically assumed to be bounded away from zero.

Another model similar to (1) has been developed for undirected networks by Stein and Leng (2022). The model in this paper is more complex due to the presence of two sets of heterogeneity parameters and thus more delicate analysis is needed (Yan et al., 2016, 2019). More importantly, this paper focuses more on variable selection while Stein and Leng (2022) focused exclusively on estimation consistency. In particular, we place special focus on the study of the interplay of the rates of convergence for the network sparsity, the parameter sparsity and the penalty we use. Each of our main results requires these three quantities to be balanced in a suitable manner, made explicit in our *balancing assumptions* (Assumptions B1, B2, B3). These assumptions highlight the effect of different sparsity regimes much more clearly than Stein and Leng (2022).

It has been claimed by Barabási (2016) that the degrees of real world networks often follow a power-law distribution. We show in the Appendix D that the degree sequence of the *sparse β-model* in Chen et al. (2020) (as a representative of the family of models introduced in Chen et al. (2020); Stein and Leng (2022) and this paper), follows a power-law distribution under the right assumptions.

1.1 Notation

Denote $N = n(n-1)$ and $[n] \coloneqq \{1, \dots, n\}$. Denote $\mathbb{R}_+ = [0, +\infty)$ as the non-negative real line. For a vector $v \in \mathbb{R}^n$, we use $S(v) = \{i : v_i \neq 0\}$ to denote its support and $\operatorname{diag}(v) \in \mathbb{R}^{n \times n}$ the n-by-n diagonal matrix with v on the diagonal. Let $\| \, . \, \|_1, \| \, . \, \|_2 . \| \, . \, \|_\infty$ denote the vector ℓ_1-, ℓ_2- and ℓ_∞-norm respectively and define $\| \, . \, \|_0$ denotes the ℓ_0-"norm". That is, $\|v\|_0 = |S(v)|$. For a vector $v \in \mathbb{R}^N$, we number its elements as $v = (v_{ij})_{i \neq j}$.

For brevity, we write $\vartheta = (\alpha^T, \beta^T)^T$ and $\xi = (\mu, \gamma^T)^T \in \mathbb{R}^{p+1}$. Thus, $\theta = (\vartheta^T, \xi^T)^T$ with its true value denoted as $\theta_0 = (\vartheta_0^T, \xi_0^T)^T$. We write $S_0 = S(\vartheta_0)$ and denote its cardinality by $s_0 = |S_0|$. We write $S_{0,+} \coloneqq S_0 \cup \{2n+1, 2n+2, \dots 2n+1+p\}$ with cardinality $s_{0,+} = |S_{0,+}| = s_0 + p + 1$ to refer to all active indices including those of μ and γ. Thus, s_0 and $s_{0,+}$ can be understood as the parameter sparsity of our model. Let $S_\alpha = \{i : \alpha_{0,i} > 0\}, S_\beta = \{j : \beta_{0,j} > 0\}$ and $s_\alpha = |S_\alpha|, s_\beta = |S_\beta|$. When we want to make the dependence of the link probabilities given Z_{ij} on different values of θ explicit, we write $p_{ij}(\theta) = \frac{\exp(\alpha_i + \beta_j + \mu + \gamma^T Z_{ij})}{1 + \exp(\beta_i + \beta_j + \mu + \gamma^T Z_{ij})}$. Finally, we use C for some generic, strictly positive constant that may change between displays.

For the convenience of the reader, we provide a (non-exhaustive) table of the most important quantities encountered in this paper in Table 1.

2. Estimation

A directed network on n nodes is represented as a directed graph $G_n = (V, E)$, consisting of a node set V with cardinality n and an edge set $E \subseteq V \times V$. Without loss of generality, we assume $V = [n]$ and that G_n is simple, having no self-loops nor multiple edges between any pair of nodes. Such a graph G_n is represented as a binary adjacency matrix $A \in \mathbb{R}^{n \times n}$, where $A_{i,j} = 1$, if $(i,j) \in E$ and $A_{i,j} = 0$ otherwise. By assumption $A_{ii} = 0$ for all i.

Given A and the covariates $\{Z_{ij}\}_{i \neq j}$, the negative log-likelihood of the model (1) is

$$\begin{aligned}\mathcal{L}(\alpha, \beta, \mu, \gamma) = &- \sum_{i=1}^n \alpha_i b_i - \sum_{i=1}^n \beta_i d_i - d_+ \mu - \sum_{\substack{i,j=1 \\ i \neq j}}^n (\gamma^T Z_{ij}) A_{ij} \\ &+ \sum_{\substack{i,j=1 \\ i \neq j}}^n \log(1 + \exp(\alpha_i + \beta_j + \mu + \gamma^T Z_{ij})),\end{aligned}$$

where $b_i = \sum_{j=1, j \neq i}^n A_{ij}$ is the out-degree of vertex i and $d_i = \sum_{j=1, j \neq i}^n A_{ji}$ its in-degree. Write $d = (d_1, \dots, d_n)^T$ and $b = (b_1, \dots, b_n)^T$ as the corresponding degree sequences. De-

Notation	Description
$\alpha, \beta \in \mathbb{R}^n$	Degree heterogeneity parameters for incomingness and outgoingness
$\mu \in \mathbb{R}$	Global sparsity parameter, may diverge to $-\infty$
$\gamma \in \mathbb{R}^p$	Covariate parameter, captures homophily, if covariates measure similarity
$\theta \in \mathbb{R}^{2n+1+p}$	Shorthand for $(\alpha^T, \beta^T, \mu, \gamma^T)^T$
$\theta_0, \alpha_0, \beta_0, \mu_0, \gamma_0$	The true parameter values
$\vartheta \in \mathbb{R}^{2n}$	Shorthand for $= (\alpha^T, \beta^T)^T$
$\xi \in \mathbb{R}^{p+1}$	Shorthand for $(\mu, \gamma^T)^T$
S_0	The support of ϑ_0 with cardinality $s_0 = \lvert S_0 \rvert$
$S_{0,+}$	$= S_0 \cup \{2n+1, 2n+2, \dots 2n+1+p\}$, with cardinality $s_{0,+} = s_0 + p + 1$
λ	The penalty used in (4)
$\bar{\theta} = (\bar{\vartheta}, \mu, \gamma)$	$= \left(\frac{1}{\sqrt{n}}\vartheta, \mu, \gamma\right)$, the rescaled parameter values
$\bar{\lambda}$	$= \sqrt{n}\lambda$, the rescaled penalty value
b_i, d_i	out- and in-degree of node i respectively
d_+	$= \sum_{i=1}^n d_i$
$\mathcal{E}(\theta)$	$= \frac{1}{N}\mathbb{E}[\mathcal{L}(\theta) - \mathcal{L}(\theta_0)]$, the excess risk at θ
$\lambda_{\min}, \lambda_{\max}$	Minimum and maximum eigenvalue of $\frac{1}{N}\mathbb{E}[Z^T Z]$ respectively
$c_{\min}$	Constant independent of n such that $\lambda_{\min} \geq c_{\min} > 0$ and $c_{\min} < 1/2$.

Table 1: List of most important definitions and notations.

note $d_+ := \sum_{i=1}^n d_i$ and $b_+ := \sum_{i=1}^n b_i$ for which we have $b_+ = d_+$. As is common in the literature, we call a network sparse if $\mathbb{E}[d_+] \sim n^\kappa$ for some $\kappa \in (0, 2)$, where $\mathbb{E}$ is the expectation with regard to the data generating process. A network is dense if $\mathbb{E}[d_+] \sim n^2$.

Since α and β are sparse, an estimate of $\theta = (\alpha^T, \beta^T, \mu, \gamma^T)^T$ can be obtained via the following penalized likelihood

$$\arg\min_{\theta \in \Theta} \frac{1}{N}\mathcal{L}(\alpha, \beta, \mu, \gamma) + \lambda(\|\alpha\|_1 + \|\beta\|_1), \tag{2}$$

where λ is a tuning parameter and $\Theta = \mathbb{R}_+^n \times \mathbb{R}_+^n \times \mathbb{R} \times \mathbb{R}^p$ is the parameter space. For simplicity, we have used the same amount of penalty on α and β because $b_+ = d_+$. The objective function in (2) is similar to the penalized logistic regression with an ℓ_1 penalty and thus can be easily solved. In this paper, we use the solver in the R package `glmnet` (Friedman et al., 2010). The similarity of our estimator to the LASSO estimator makes our estimation approach extremely scalable.

Since our focus is on sparse networks, we assume the existence of a non-random sequence $\rho_{n,0} \in (0, 1/2]$, allowing $\rho_{n,0} \to 0$ as $n \to \infty$, such that for all i, j,

$$1 - \rho_{n,0} \geq p_{ij} \geq \rho_{n,0},$$

where $\rho_{n,0}$ is referred to as the *network sparsity parameter*. The above constraint is equivalent to

$$|\alpha_{0,i} + \beta_{0,j} + \mu_0 + \gamma_0^T Z_{ij}| \leq -\text{logit}(\rho_{n,0}) =: r_{n,0}, \quad \forall i, j,$$

where $\text{logit}(p) = \log(p/(1-p))$ for $p \in (0, 1)$ and $r_{n,0} \geq 0$ since $\rho_n \leq 1/2$. This inequality can also be expressed in terms of the design matrix D associated with the corresponding logistic regression problem, defined in (5) below, and is equivalent to $\|D\theta_0\|_\infty \le r_{n,0}$. This motivates the following tweak to the estimation procedure in (2): Given a sufficiently large constant r_n, we define the local parameter space

$$\Theta_{\text{loc}} = \Theta_{\text{loc}}(r_n) := \left\{\theta \in \Theta : \|D\theta\|_\infty \le r_n\right\}, \tag{3}$$

which is convex, and propose to estimate the parameters as

$$\hat{\theta} = (\hat{\alpha}^T, \hat{\beta}^T, \hat{\mu}, \hat{\gamma}^T)^T = \underset{\theta=(\alpha^T,\beta^T,\mu,\gamma^T)^T \in \Theta_{\text{loc}}}{\arg\min} \frac{1}{N}\mathcal{L}(\alpha,\beta,\mu,\gamma) + \lambda(\|\alpha\|_1 + \|\beta\|_1), \tag{4}$$

which is more amenable for theoretical analysis.

We now give an explicit form of the associated design matrix D. Since we have the presence/ absence of $N = n(n-1)$ directed edges and $2n+1+p$ parameters, D has dimension $N\times(2n+1+p)$. Define the out-matrix $X^{\text{out}} \in \mathbb{R}^{N\times n}$ with rows $X_{ij}^{\text{out}} \in \mathbb{R}^{1\times n}, i \ne j$, such that for each component $k = 1, \dots, n$, $X_{ij,k}^{\text{out}} = 1$ if $k = i$ and zero otherwise. Likewise, define the in-matrix $X^{\text{in}} \in \mathbb{R}^{N\times n}$ with rows $X_{ij}^{\text{in}} \in \mathbb{R}^{1\times n}, i \ne j$, such that for each component $k = 1, \dots, n$, $X_{ij,k}^{\text{in}} = 1$ if $k = j$ and zero otherwise. Let $Z = (Z_{ij}^T)_{i\ne j} \in \mathbb{R}^{N\times p}$ be the matrix of the covariate vectors written below each other. Then, the design matrix D consists of four blocks, written next to each other:

$$D = \left[\begin{array}{c|c|c|c} X^{\text{out}} & X^{\text{in}} & \mathbf{1} & Z \end{array}\right] \in \mathbb{R}^{N\times(2n+p+1)}, \tag{5}$$

where $\mathbf{1} \in \mathbb{R}^N$ is a vector of all ones. We use the shorthand $X = [X^{\text{out}} \mid X^{\text{in}}] \in \mathbb{R}^{N\times 2n}$.

The design matrix D reveals an important property of our model (1). While the columns of the the global parameters μ and γ have non-zero entries in all $N \sim n^2$ rows of D, the local parameters α and β only have n non-zero entries in their respective columns. Thus, the effective sample size for α and β is only n, whereas it is N, that is, of order n larger, for the global parameters μ and γ. This will also be reflected in the different rates of convergence we obtain in Theorem 5 below.

A key quantity in the theory of high-dimensional statistics is the population Gram matrix Σ, closely linked to the Hessian of $\mathcal{L}$ and the precision matrix. Loosely speaking, it is given by

$$\Sigma = \frac{1}{\text{sample size}}\mathbb{E}[D^T D].$$

Were we to naively ignore the differing sample sizes between local and global parameters and choose $\Sigma = 1/N \cdot \mathbb{E}[D^T D]$, our proofs would fail, due to the top-left corner of Σ rapidly converging to the zero-matrix, making Σ singular in the limit. In particular, the *compatibility condition* (cf. Section 3.2), crucial for proofs for LASSO-type problems, would not hold. We need to account for this fact and therefore propose to use a *sample-size adjusted* Gram matrix. To that end, we introduce the matrix

$$T = \begin{bmatrix} \sqrt{n-1}I_{2n} & 0 \\ 0 & \sqrt{N}I_{p+1} \end{bmatrix},$$

where I_m is the $m \times m$ identity matrix and define the *sample size adjusted Gram matrix* Σ as

$$\Sigma = T^{-1}\mathbb{E}[D^T D]T^{-1}. \tag{6}$$

It will be convenient to cast problem (4) in terms of rescaled parameters $\bar\theta$ which adjust for the discrepancy in effective sample sizes. This new formulation is equivalent to the one in (4), but gives us a unified framework for treating convergence properties of our estimators. We will rely heavily on that rescaled version in our proofs. For now, we simply remark that for any parameter $\theta = (\vartheta^T, \mu, \gamma^T)^T \in \Theta$, we introduce the notation

$$\bar\theta = (\bar\vartheta^T, \mu, \gamma^T)^T = \left(\frac{1}{\sqrt{n}}\vartheta^T, \mu, \gamma^T\right)^T \tag{7}$$

and refer the reader to Section A.2 in the appendix for a derivation and interpretation of this formulation. Our original estimation problem (4) can then equivalently be rewritten in terms of these rescaled parameters, giving rise to a sample-size adjusted estimator

$$\hat{\bar\theta} = \left(\hat{\bar\vartheta}^T, \hat\mu, \hat\gamma^T\right)^T,$$

that solves a problem similar to (4) with penalty parameter $\bar\lambda = \sqrt{n}\lambda$ (see (13) in Section A.2). We denote the negative log-likelihood with respect to a rescaled parameter $\bar\theta$ by $\bar{\mathcal{L}}(\bar\theta)$. Then, given a solution $\hat{\bar\theta}$ for a given penalty parameter $\bar\lambda$ to this modified problem (13), we can obtain a solution to our original problem (4) with penalty parameter $\lambda = \bar\lambda/\sqrt{n}$, by setting

$$(\hat\vartheta, \hat\mu, \hat\gamma) = \left(\sqrt{n}\hat{\bar\vartheta}, \hat\mu, \hat\gamma\right). \tag{8}$$

3. Theory

We outline the main assumptions first.

Assumption 1 *The Z_{ij}'s are independent with $\mathbb{E}[Z_{ij}] = 0$ and $|Z_{ij}|$ uniformly bounded. We also assume that γ_0 lies in some compact, convex set $\Gamma \subset \mathbb{R}^p$ with a fixed p. Further assume that there are constants $C > c_{\min} > 0$ such that the minimum eigenvalue $\lambda_{\min}$ and the maximum eigenvalue $\lambda_{\max}$ of $\frac{1}{N}\mathbb{E}[Z^TZ]$ fulfil $c_{\min} \le \lambda_{\min} \le \lambda_{\max} \le C$. Without loss of generality we assume $c_{\min} < 1/2$.*

As a result of Assumption 1, there exist constants $\kappa, c > 0$ such that $|Z_{ij}^T\gamma| \le \kappa$ for all $1 \le i \ne j \le n$ and $|Z_{ij,k}| \le c$ for all $1 \le i \ne j \le n, k = 1, \dots, p$.

Assumption 2 *We assume that $\theta_0 \in \Theta_{\text{loc}}$ or equivalently $r_n \ge r_{n,0}$. Therefore, without loss of generality we assume $r_n = r_{n,0}$ and consequently $\rho_n = \rho_{n,0}$.*

Assumption 1 is standard. Note that Z_{ij}'s are not necessarily i.i.d., possibly having correlated entries and that Z_{ij} can be asymmetric in that $Z_{ij} \ne Z_{ji}$. We have chosen to focus on the random-design assumption which is somewhat more interesting than a fixed-design one. We have assumed a fixed p and leave the study of diverging p to future study. Assumption 2 ensures no model misspecification.

Assumption B1 $\sqrt{n}s_+^2\bar\lambda\rho_n^{-2} \to 0, n \to \infty.$

For all of our theorems, striking the right balance between parameter sparsity s_+, network sparsity ρ_n and penalty parameter $\bar{\lambda}$ is crucial. The restrictiveness of these balancing assumptions will depend on the complexity of the results being proven and we number them separately from the general assumptions as "Assumption Bi", $i = 1, 2, 3$, to make their special standing explicit in our notation. Our main result on model selection consistency, Theorem 1, is the most refined of our theorems and hence Assumption B1 is the strongest such balancing assumption. In particular, the weaker balancing assumptions required to establish parameter estimation consistency, Theorem 5 (Assumption B2), and asymptotic normality of the homophily parameter estimator $\hat{\gamma}$, Theorem 6 (Assumption B3), follow from Assumptions B1 above and 3 below. Thus, our estimator $\hat{\theta}$ in (4) can simultaneously recover the correct support, consistently estimate the parameter values and produce an asymptotically normal estimate of γ_0.

3.1 Model selection consistency

In this section we study under which conditions our estimator (4) identifies the correct subset of active variables S_0. Our main result, Theorem 1, is that under the appropriate conditions, our estimator $\hat{\theta}$ will correctly exclude all the truly inactive parameters and correctly include all those truly active parameters whose value exceeds a certain threshold. The latter minimal signal condition is typical for model selection (Ravikumar et al., 2010; Chen et al., 2020, e.g.).

Recall that we use S_0 to refer to the active set of indices associated with $\vartheta_0 = (\alpha_0^T, \beta_0^T)^T$, whereas $S_{0,+} = S_0 \cup \{2n+1, \dots, 2n+1+p\}$. In the following derivations it will be crucial to distinguish the two correctly. We use $S_{0,+}^c$ to denote the complement of $S_{0,+}$ in $[2n+1+p]$, that is $S_{0,+}^c = [2n+1+p] \backslash S_{0,+}$. Let S_0^c refer to the complement of S_0 in $[2n]$ *only*: $S_0^c = [2n] \backslash S_0$. While this may seem like a potential notational pitfall, this allows for much cleaner notation in our proofs.

We first state the main theorem of this section before giving more details on its derivation. Recall that $\bar{\lambda} = \sqrt{n}\lambda$ is the penalty parameter in the rescaled version of our problem (4). Also notice that $\hat{S} \coloneqq \{i : \hat{\bar{\vartheta}}_i > 0\} = \{i : \hat{\vartheta}_i > 0\}$, that is the estimators (4) and (8) will always select the same active set of parameters.

Assumption 3 $-\frac{N\bar{\lambda}^2}{18} + \log(n) \to -\infty, n \to \infty.$

Assumption 3 suggests we pick our penalty of order $\bar{\lambda} \asymp \sqrt{\frac{\log(n)}{N}}$. Thus, informally speaking, Assumption 3 requires that the rescaled penalty parameter $\bar{\lambda}$ must be at least of order $\sqrt{\log(\text{number variables})/(\text{effective sample size})}$, which is the typical rate for the penalty we would expect from classical LASSO literature (van de Geer and Bühlmann, 2011).

Theorem 1 *Under Assumptions 1, 2, B1 and 3, and for n sufficiently large, with probability approaching one, the penalized likelihood estimator $\hat{\theta}$ from (4):*

1. *excludes all the truly inactive parameters: $\hat{S} \cap S^c = \emptyset$ and,*

2. *with penalty of order* $\bar{\lambda} \asymp \sqrt{\frac{\log(n)}{N}}$*, it includes all those truly active parameters whose value is larger than* $C \cdot \rho_n^{-1} \frac{\sqrt{\log(n)}}{\sqrt{n}}$*:*

$$\left\{ i : \vartheta_{0,i} > C \cdot \rho_n^{-1} \frac{\sqrt{\log(n)}}{\sqrt{n}} \right\} \subseteq \hat{S},$$

where the form of C *and the exact probability are given in the proof.*

We have the following remarks.

Remark 2 1. *Notice that Assumption 3 requires* $\bar{\lambda} > 3\sqrt{2} \cdot \sqrt{\log(n)/N}$*. This is the same regime as specified by Theorems 5 and 6 which ensure consistent parameter estimation and asymptotic normality of* $\hat{\gamma}$*. Hence, consistent parameter estimation, inference on* γ *and support recovery are all possible simultaneously.*

2. *If we choose* $\bar{\lambda} \asymp \sqrt{\frac{\log(n)}{N}}$ *and* s_+ *is of lower order, such as growing logarithmically or constant, then, up to log-terms, Assumption B1 implies that we must have for the permissible network sparsity,* $\rho_n = o(n^{-1/4})$*.*

Our tool of choice for proving Theorem 1 is a *primal-dual witness construction*, similar to the one in Ravikumar et al. (2010). The idea is to construct a tuple $(\bar{\theta}^\dagger, \bar{z}^\dagger)$, such that $\bar{\theta}^\dagger$ solves the rescaled version of (4), while also identifying the correct support S_0 and $\bar{z}^\dagger$ is a solution to the Karush-Kuhn-Tucker conditions (9) as outlined below. In the construction of $(\bar{\theta}^\dagger, \bar{z}^\dagger)$, we make use of knowledge of the true active set S_0, which makes it infeasible to use in practice. However, by Lemma 3 below, if the construction succeeds – we make precise what we mean by that below – any solution to (4) must have the same support as $\bar{\theta}^\dagger$. In summary, if the construction succeeds, our estimator $\hat{\theta}$ must identify the correct support S_0, too. The bulk of the work in proving Theorem 1 is to show that the construction of $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ will be successful with high probability for large n.

It is important to point out that due to the mixture of deterministic and random columns in D and the differing sample sizes between ϑ and ξ, the standard assumptions in Ravikumar et al. (2010) imposed on the Hessian of $\mathcal{L}$ cannot simply be imposed in our model. Rather, a careful argument is needed to prove that analogous properties hold for sufficiently large n with high probability. See Section C.1 in the appendix for details.

Our starting point for proving Theorem 1 are the Karush-Kuhn-Tucker conditions (Bertsekas, 1995, Chapter 5): Equation (4) and its rescaled version (13) are a convex optimization problems. Hence, by subdifferential calculus, a vector $\bar{\theta}$ is a minimizer of (13) if and only if zero is contained in the subdifferential of $\frac{1}{N}\bar{\mathcal{L}}(\bar{\theta}) + \bar{\lambda}\|\bar{\vartheta}\|_1$ at $\bar{\theta}$. That is, if and only if there is a vector $\bar{z} \in \mathbb{R}^{2n+1+p}$ such that

$$0 = \frac{1}{N}\nabla\bar{\mathcal{L}}(\bar{\theta}) + \bar{\lambda}\bar{z}, \tag{9}$$

and

$$\bar{z}_i = 1, \text{ if } \bar{\vartheta}_i > 0, i = 1, \ldots, 2n, \tag{10a}$$

$$\bar{z}_i \in [-1, 1], \text{ if } \bar{\vartheta}_i = 0, i = 1, \ldots, 2n, \tag{10b}$$

$$\bar{z}_i = 0, i = 2n+1, \ldots, 2n+1+p. \tag{10c}$$

We call such a pair $(\bar{\theta}, \bar{z}) \in \mathbb{R}^{2n+1+p} \times \mathbb{R}^{2n+1+p}$ *primal-dual optimal* for the rescaled problem (13). Note that in the first $2n$ components of $\nabla \bar{\mathcal{L}}$ we are taking the derivative with respect to $\bar{\vartheta}$ instead of ϑ. This means we need to pay attention to additional $\sqrt{n}$-factors. For such a pair to identify the correct support S_0, it is sufficient for

$$\bar{\theta}_i > 0, \text{ for all } i \in S_0, \text{ and} \tag{11a}$$

$$\|\bar{z}_{S_{0,+}^c}\|_\infty < 1 \tag{11b}$$

to hold. Where (11a) ensures that all truly active indices are included and (11b) ensures that all truly inactive indices are excluded (due to (10a)). We call (11b) the *strict feasibility condition* as in Ravikumar et al. (2010).

We will proceed to construct a pair $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ that satisfies condition (9), (10a) - (10c) and (11a) - (11b) with high probability and for sufficiently large n. We say the construction *succeeds*, if $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ fulfils (9) - (11b), which in particular implies that $\bar{\theta}^\dagger$ identifies the correct support S_0 and also is a solution to (13).

By the following lemma, if the construction succeeds, any solution to (13) in the appendix must have the same support as $\bar{\theta}^\dagger$. Thus, if the construction succeeds, our estimator $\hat{\bar{\theta}}$ must identify the correct support S_0, too.

Lemma 3 *Suppose the construction* $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ *fulfils equations* (9) *and* (10a) - (10c) *and* (11b)*. Let* $S^\dagger = \{i : \bar{\vartheta}_i^\dagger > 0\}$*. Then,*

$$\hat{S} = S^\dagger.$$

In particular, if $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ *additionally fulfils* (11a)*, then* $S^\dagger = S_0$*, and thus,* $\hat{S} = S_0$*.*

We now give a detailed description of the primal-dual witness construction.
Primal-dual witness construction.

1. Solve the restricted penalized likelihood problem

$$\bar{\theta}^\dagger = (\bar{\vartheta}^{\dagger,T}, \mu^\dagger, \gamma^{\dagger,T})^T = \arg\min \frac{1}{N} \bar{\mathcal{L}}(\bar{\theta}) + \bar{\lambda} \|\bar{\vartheta}\|_1, \tag{12}$$

   where the argmin is taken over all $\bar{\theta} = (\bar{\vartheta}^T, \mu, \gamma^T)^T \in \Theta_{\text{loc}}$ with support $S_{0,+}$, i.e. $\bar{\theta}_{S_{0,+}}^\dagger = \bar{\theta}^\dagger$ or equivalently $\bar{\theta}_{S_{0,+}^c}^\dagger = 0$. Thus, by construction, $\bar{\theta}^\dagger$ correctly excludes all inactive indices.

2. Since (12) is a convex problem, zero must be contained in its subdifferential at $\bar{\theta}^\dagger$. Thus, we set $\bar{z}_i^\dagger = 1$, if $\bar{\vartheta}_i^\dagger > 0$ such that (10a) holds and $\bar{z}_i^\dagger = 0, i = 2n{+}1, \ldots, 2n{+}1{+}p$, such that (10c) holds. By subdifferential calculus we find $\bar{z}_i^\dagger \in [-1, 1]$, for those $i \in S$ with $\bar{\vartheta}_i^\dagger = 0$ (in case there are any), such that (9) holds for those components in S.

3. Plug $\bar{\theta}^\dagger$ and $\bar{z}^\dagger$ into (9) and solve for the remaining components of $\bar{z}^\dagger$, such that (9) holds for $(\bar{\theta}^\dagger, \bar{z}^\dagger)$.

The challenge will be proving that (11a) and (11b) also hold, which together ensure that (10b) holds, too.

3.2 Consistency

In this section we will show that under assumptions similar to those of Theorem 1, our estimator $\hat{\theta}$ will also be consistent in terms of excess risk (cf. Greenshtein and Ritov (2004), Koltchinskii (2011)) and ℓ_1-error. To that end, define the excess risk for a parameter θ as

$$\mathcal{E}(\theta) := \frac{1}{N}\mathbb{E}[\mathcal{L}(\theta) - \mathcal{L}(\theta_0)].$$

By construction, $\theta_0 = \arg\min_{\theta\in\Theta} \mathcal{E}(\theta) = \arg\min_{\theta\in\Theta_{\mathrm{loc}}(r_{n,0})} \mathcal{E}(\theta)$, where the second equality follows from Assumption 2.

A compatibility condition. A crucial identifiability assumption in LASSO theory is the so called *compatibility condition* (van de Geer and Bühlmann, 2011; van de Geer et al., 2014). It relates the quantities $\|(\hat{\theta} - \theta_0)_{S_{0,+}}\|_1$ and

$$(\hat{\theta} - \theta_0)\Sigma(\hat{\theta} - \theta_0),$$

in a suitable sense made precise below and is crucial for deriving consistency results. In our model, similar to the sparse β-model in Stein and Leng (2022), the classical compatibility condition as for example defined for generalized linear models in van de Geer et al. (2014) does not hold. The reason for this is that ϑ and $(\mu, \gamma^T)^T$ have different effective sample sizes. Therefore, it is crucial that we use the sample size adjusted Gram matrix (6). Using similar techniques as in Stein and Leng (2022), we can now show that the sample size adjusted Gram matrix fulfils the compatibility condition.

Proposition 4 (Compatibility condition) *Under Assumption 1, for $s_0 = o(\sqrt{n})$ and n large enough, it holds for every $\theta \in \mathbb{R}^{2n+1+p}$ with $\|\theta_{S_{0,+}^c}\|_1 \leq 3\|\theta_{S_{0,+}}\|_1$, that*

$$\|\theta_{S_{0,+}}\|_1^2 \leq \frac{2s_{0,+}}{c_{\min}}\theta^T\Sigma\theta,$$

where Σ is the sample size adjusted Gram matrix defined in (6).

Parameter estimation consistency is the most lenient of our theorems in terms of restrictions that we have to impose on the parameter sparsity s_0 and the network sparsity ρ_n. We may replace the stricter assumption B1 by the following.

Assumption B2 *$\sqrt{n}s_0\rho_n^{-1}\bar{\lambda} \to 0, n \to \infty$.*

Theorem 5 below suggests a choice of $\bar{\lambda} \asymp \sqrt{\log(n)/N}$. Under these conditions, Assumption B2 becomes $s_0\rho_n^{-1}\sqrt{\log(n)/n} \to 0$. That is, up to an additional factor ρ_n^{-1}, which is the price we have to pay for allowing our link probabilities to go to zero, the permissible sparsity for ϑ_0 is the permissible sparsity in classical LASSO theory for an effective sample size of order n. This makes sense, considering the discussion of the differing effective sample sizes in Section 3. Also, this choice of $\bar{\lambda}$ together with Assumption B2 implies $s_0 = o(\sqrt{n})$, as is required by Proposition 4 and which thus is not a restriction.

Theorem 5 *Let Assumptions 1, 2 and B2 hold. Fix a confidence level t and let*

$$a_n \coloneqq \sqrt{\frac{2\log(2(2n+p+1))}{N}}(1 \vee c).$$

Choose $\bar{\lambda} = \sqrt{n}\lambda$ such that

$$\bar{\lambda} \geq 8 \cdot \left(8a_n + 2\sqrt{\frac{t}{N}(11(1 \vee (c^2 p)) + 16(1 \vee c)\sqrt{n}a_n)} + \frac{4t(1 \vee c)\sqrt{n}}{3N} \right)$$

Then, with probability at least $1 - \exp(-t)$ we have

$$\mathcal{E}(\hat{\theta}) + \bar{\lambda}\left(\frac{1}{\sqrt{n}}\|\hat{\vartheta} - \vartheta_0\|_1 + |\hat{\mu} - \mu_0| + \|\hat{\gamma} - \gamma_0\|_1 \right) \leq C\frac{s_{0,+}\bar{\lambda}^2}{\rho_{n,0}}.$$

Theorem 5 implies a lower bound on $\bar{\lambda}$ of order $\sqrt{\log(n)/N}$, suggesting that we may choose $\bar{\lambda}$ of the same order. Thus, up to the additional factor ρ_n^{-1}, we obtain the classical LASSO rates of convergence for a parameter of effective sample size N for μ and γ and those for a parameter of effective sample size n for α and β. If s_0 is a lower order term, such as growing logarithmically or constant, then up to log-factors Assumption B2 requires that ρ_n tend to zero at rate at most as fast as $1/\sqrt{n}$, which is faster and thus allows for sparser networks than what we obtained for model selection consistency in Theorem 1. Under Assumption B2, we can see that the above average rate of convergence for estimating ϑ is slower than the rate for estimating γ. The rate of convergence for consistency of the estimator $\hat{\theta}$ is of the same order as the rate obtained for the analogous estimator for undirected networks in Stein and Leng (2022). We refer to Stein and Leng (2022), Section 2.2, for a comparison between ℓ_1-penalized estimation procedures as discussed here and ℓ_0-penalized estimators as advocated by Chen et al. (2020).

3.3 Inference

Finally, we derive the limiting distribution of our estimator of the covariates weights, $\hat{\gamma}$. We will see that the same arguments used for deriving the limiting distribution for $\hat{\gamma}$ also work for $\hat{\mu}$ and as a by-product of our proofs we also obtain an analogous limiting result for $\hat{\mu}$.

Our strategy will be inverting the Karush-Kuhn-Tucker conditions, similar to van de Geer et al. (2014). See Section B.2 in the appendix for details. Denote by $H(\hat{\theta}) \coloneqq H_{\xi\times\xi}(\theta)|_{\theta=\hat{\theta}}$, the Hessian of $\frac{1}{N}\mathcal{L}(\theta)$ with respect to $\xi = (\mu, \gamma^T)^T$ only, evaluated at $\hat{\theta}$.

Comment 1 *Consider the entries of $H(\hat{\theta})$: For all $k, l = 1, \ldots, (p+1)$,*

$$H(\hat{\theta})_{k,l} = \frac{1}{N}\partial_{\xi_k,\xi_l}\mathcal{L}(\hat{\theta}) = \frac{1}{N}\sum_{i\neq j} D_{ij,2n+k}D_{ij,2n+l} \cdot p_{ij}(\hat{\theta})(1 - p_{ij}(\hat{\theta})),$$

where D_{ij}^T is the (i,j)-th row of the design matrix D, i.e. in particular $D_{ij,2n+k} = 1$ if $k = 1$ and $D_{ij,2n+k} = Z_{ij,k-1}$ for $k = 2, \ldots, (p+1)$.

Let $D_\xi = [\mathbf{1}|Z]$ be the part of the design matrix D, defined in (5), corresponding to ξ with rows $D_{\xi,ij}^T = (1, Z_{ij}^T), i \neq j$. Also, let $\hat{W} = \operatorname{diag}\left(\sqrt{p_{ij}(\hat{\theta})(1-p_{ij}(\hat{\theta}))}, i \neq j\right)$. It is then easy to see

$$H(\hat{\theta}) = \frac{1}{N} D_\xi^T \hat{W}^2 D_\xi.$$

Let $W_0 = \operatorname{diag}(\sqrt{p_{ij}(\theta_0)(1-p_{ij}(\theta_0))}, i \neq j)$ and consider the corresponding population version:

$$\mathbb{E}[H(\theta_0)] = \frac{1}{N}\mathbb{E}[D_\xi^T W_0^2 D_\xi].$$

To be consistent with commonly used notation, call $\hat{\Sigma}_\xi = H(\hat{\theta}) = \frac{1}{N} D_\xi^T \hat{W}^2 D_\xi$ and $\Sigma_\xi = \mathbb{E}[H(\theta_0)] = \frac{1}{N}\mathbb{E}[D_\xi^T W_0^2 D_\xi]$ and $\hat{\Theta}_\xi := \hat{\Sigma}_\xi^{-1}, \Theta_\xi := \Sigma_\xi^{-1}$.

For the proof of asymptotic normality, we need to invert $\hat{\Sigma}_\xi$ and Σ_ξ and show that these inverses are close to each other in an appropriate sense. It is commonly assumed in LASSO theory (cf. van de Geer et al. (2014)) that the minimum eigenvalues of these matrices stay bounded away from zero. In our case, however, such an assumption is invalid. Since we allow for the lower bound $\rho_{n,0}$ on the link probabilities to go to zero, any lower bound on the entries in W_0 will go to zero with n and as a consequence our lower bound on the minimum eigenvalue of Σ_ξ will tend to zero as n goes to infinity as well. The best we can achieve is a strict positive definiteness of Σ_ξ for finite n, but not uniformly in n. Since these lower bounds tend to zero with increasing n, a careful argument is needed and we need to impose a slightly stricter balancing assumption than Assumption B2.

Assumption B3 $\sqrt{n} s_0 \rho_n^{-2} \bar{\lambda} \to 0, n \to \infty.$

Theorem 6 *Under Assumptions 1, 2 and B3, with $\lambda \asymp \sqrt{\log(n)/N}$ fulfilling the conditions of Theorem 5, we have for any $k = 1, \ldots, p$, as $n \to \infty$,*

$$\sqrt{N}\frac{\hat{\gamma}_k - \gamma_{0,k}}{\sqrt{\hat{\Theta}_{\vartheta,k+1,k+1}}} \xrightarrow{d} \mathcal{N}(0,1).$$

We also have for our estimator of the global sparsity parameter, $\hat{\mu}$, as $n \to \infty$,

$$\sqrt{N}\frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\Theta}_{\vartheta,1,1}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Contrary to what is commonly seen in the penalized likelihood literature (Zhang and Zhang, 2014; van de Geer et al., 2014), no debiasing of $\hat{\gamma}$ and $\hat{\mu}$ is needed. The reason for this is that columns of D pertaining to those parameters which are indeed biased, that is to ϑ, and those pertaining to $\xi = (\mu, \gamma^T)^T$ become asymptotically orthogonal, meaning that the bias in $\hat{\xi}$ vanishes fast enough for the derivation of Theorem 6 to be possible. For a lower order s_0, Assumption B3 essentially allows for the same level of network sparsity as Assumption B1, up to lower order factors.

4. Simulation

In this section we demonstrate the effectiveness of our estimator (4) in performing simultaneous parameter estimation and model selection consistently. To this end, we test its performance on networks of varying sizes. Specifically, we let n vary between 150 and 800 in steps of 50 and choose the sparsity level s_0 to be close to $\sqrt{n}/2$. We let $s_0 = 6, 6, 6, 8, 8, 10, 10, 10, 10, 12, 12, 12, 12, 14$ and chose $s_\alpha = s_\beta = s_0/2$ in each case. We selected a heterogeneous configuration for the assignment of non-zero α and β values. That is, we included dedicated 'spreader' nodes, with large α and zero β value as well as 'attractor' nodes with large β and zero α as well as some nodes with both active α and β. In detail, we let

$$\alpha = (2, 1.5, 1, 0.8, \ldots, 0.8, 0, \ldots, 0),$$
$$\beta = (0, \ldots, 0, 2, 1.5, 1, 0.8, \ldots, 0.8, 0, \ldots, 0),$$

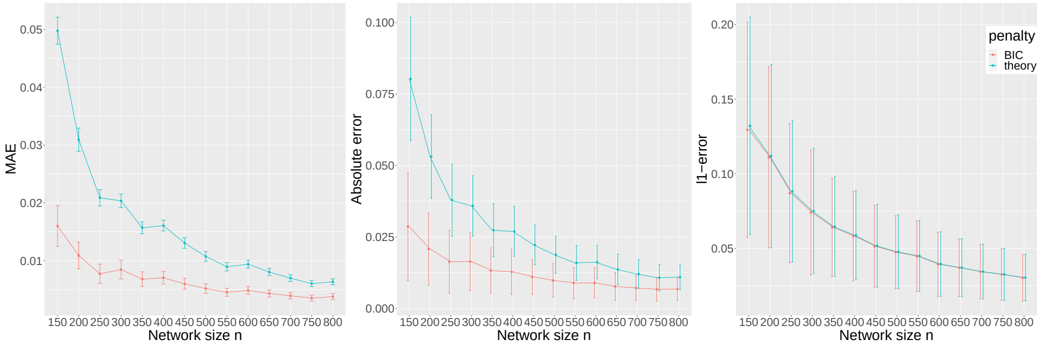
where the number of entries with value 0.8 was chosen to match the aforementioned sparsity level (zero for the first three values of n) and the number of leading zeros in β was chosen such that there were exactly two nodes with both active α and β. We let the networks get progressively sparser and set $\mu = -1.2 \cdot \log(\log(n))$. In all cases we used $p = 2$, sampled the covariate values $Z_{ij,k}, k = 1, 2, i \neq j$ from a centred Beta$(2, 2)$ distribution, and set $\gamma = (1, 0.8)^T$. Our estimator requires us to choose a tuning parameter λ and we explored the use of the Bayesian Information Criterion (BIC) as well as a heuristic based on our developed theory for model selection. While the former criterion is purely data-driven, the use of the latter is to ensure that our theoretical results are about right in terms of the rates. Specifically, our BIC is defined as

$$\text{BIC} = 2\mathcal{L}(\hat{\theta}(\lambda)) + s(\lambda)\log(N)$$

where $\hat{\theta}(\lambda)$ is the solution with λ as the tuning parameter, and $s(\lambda)$ is the cardinality of its support. On the other hand, our heuristic is motivated by the theory developed in the previous sections. We construct $\bar{\lambda}$ as in Theorem 5 based on confidence level $t = 3$, choosing to drop the leading factor eight prescribed by Theorem 5. It is known that in high-dimensional settings the penalty values prescribed by mathematical theory in practice tend to over-penalize the parameter values (Yu et al., 2019). Decreasing the penalty by removing that factor is thus in line with these empirical findings.

We drew $M = 500$ realizations for each value of n and recorded the mean absolute error for estimation of $(\alpha^T, \beta^T)^T$, the absolute error for estimation of μ and the ℓ_1-error for estimation of γ. We also constructed confidence intervals as prescribed by Theorem 6 and recorded the empirical coverage at the nominal 95% level. Finally, we studied how well BIC and our heuristic did in terms of identifying the correct model.

Consistency. We display the various error statistics for estimation of $\vartheta_0 = (\alpha_0^T, \beta_0^T)^T, \mu_0$ and γ_0 in Figures 2a, 2b and 2c respectively. We see that the error decreases with increasing network size for both model selection procedures. We see that especially for small n, BIC outperforms the heuristic for ϑ_0 and μ_0, while they both give essentially the same results for estimation of γ_0. The better performance of BIC is less prominent as n increases. BIC selects the penalty in a purely data driven manner, which allows it to adapt to differing



(a) MAE for $\vartheta_0 = (\alpha_0^T, \beta_0^T)^T$. (b) Absolute error for μ_0 (c) ℓ_1-error for γ_0.

Figure 2: Mean absolute error for $\vartheta_0 = (\alpha_0^T, \beta_0^T)$, absolute error for μ_0 and ℓ_1-error for γ_0 for varying n. The results for BIC are presented in red, the ones for our heuristic in green. The dots are the mean errors and the error bars are of length one standard deviation.

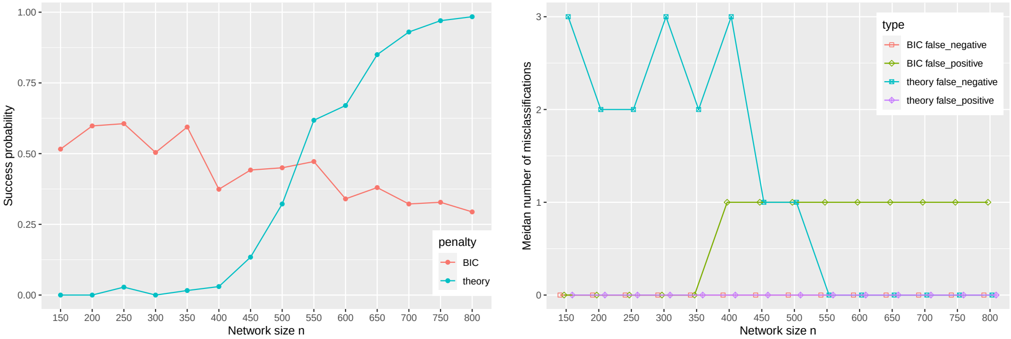
degrees of sparsity in the network, while for the heuristic the penalty value only depends on n and p. This additional flexibility is what allows BIC to achieve lower error values.

Asymptotic normality. We construct confidence intervals at the nominal 95% level for our estimators of $\gamma_{0,1}$ and $\gamma_{0,2}$ as prescribed by Theorem 6. Table 2 shows the results for $\gamma_{0,1}$ across three values of n. The results for other n and $\gamma_{0,2}$ are similar and are omitted to save space. The coverage is very close to the 95%-level across all network sizes, independent of which model selection criterion we use. This is to be expected, considering that there was hardly any difference for the estimation of γ between our two model selection criteria. This empirically illustrates the validity of the asymptotic results derived in Theorem 6. As expected, the median length of the confidence interval decreases with increasing network size.

n	Coverage	CI	Coverage	CI
	Heuristic λ		BIC	
200	0.952	0.265	0.962	0.266
400	0.950	0.141	0.964	0.141
800	0.946	0.075	0.952	0.075

Table 2: Empirical coverage under nominal 95% coverage and median lengths of confidence intervals (CIs).

Model selection. Finally we compare model selection performance between BIC and our heuristic. Figure 3a shows the empirical probability of selecting the correct model versus the various network sizes. We can see very clearly that asymptotically, as n grows, our heuristic outperforms BIC, achieving correct model selection almost all the time. Nonetheless, it is worth pointing out that even though BIC may not select the exact correct model, the number of misclassifications it does on average is not very large, as shown in Figure 3b. Figure 3b also shows that the heuristic, by virtue of selecting a larger penalty than BIC, will



(a) Probability of correct model selection. (b) Median number of misclassifications.

Figure 3: (a): The empirical probability of selecting the correct subset of active indices. (b): The median number of misclassifications for each model selection procedure, split up into false positives and false negatives.

on average incur more false negatives for small n. On the other hand, as n grows, BIC will incur false positives, resulting in the decreasing probability of selecting the correct subset.

4.1 The lawyer data

We return to our motivating example by comparing our estimates of the regression coefficients with those in Yan et al. (2019). For the seven covariates in this dataset, we followed Yan et al. (2019) in using the absolute differences of the continuous variables and the indicators whether the categorical variables are equal as our covariates. The edge density in this network is 11.6%.

To apply the model in Yan et al. (2019), one needs to remove the eight nodes in black in Figure 1 that have zero in-degree or out-degree. Otherwise the maximum likelihood estimates would be $-\infty$ for α_i if node i has no outgoing connections or for β_i if the node has no incoming links. Another interesting aspect of the model in Yan et al. (2019) lies in the inference for the fixed-dimensional parameter γ. Because the rate of convergence of its estimate is slowed down by those of the growing-dimensional heterogeneity parameters α and β, the estimator of γ requires a bias correction to be asymptotically normal. In contrast, by making a sparsity assumption on α and β in our model, we estimate the parameters via penalized likelihood and the inference of γ is straightforward as seen in Theorem 6.

When the Bayesian information criterion is used to choose the tuning parameter in the penalized likelihood estimation, our model gives 7 nonzero α_i's and 7 nonzero β_i's. Four pairs of these nonzeros come from the same nodes. In Table 3 we present the estimated γ and their standard errors when our model and the model in Yan et al. (2019) are fitted. We remark that since Yan et al. (2019) removed eight nodes, akin to biased sampling, their estimates can be biased. In terms of the parameter estimates themselves, although generally similar, we can see a few differences. First we can see that the standard errors of our estimates are smaller than those in Yan et al. (2019), reflecting that our estimates are based on a larger sample size (a network with 71 nodes compared to one with 63 nodes in the latter paper) with fewer parameters (22 versus 132). Second, the effect of age difference is not significant in our model while it is in the model in Yan et al. (2019). To explore the age effect graphically, we colour-coded the lawyers by their age group in Figure 1. We can see that plenty of connections are made between age groups and ”across the circle”, i.e. between lawyers with a large difference in age, suggesting that age may not have played an important role. Indeed, a third (33.9%) of all friendships are formed between lawyers with an age difference of ten or more years. Third, we estimate the effect of attending the same law school as positive, implying that the lawyers tend to befriend those who graduated from the same school, while Yan et al.'s model states the opposite. The former conforms better to our intuition about social networks.

	This Paper		Yan et al. (2019)	
Covariate	Estimate	SE	Estimate	SE
Same status	1.52	0.10	1.76	0.16
Same gender	0.44	0.09	0.96	0.14
Same office	2.02	0.10	3.23	0.18
Same practice	0.58	0.09	1.11	0.12
Same law school	0.29	0.10	−0.48	0.12
Difference in years with firm	−0.01	0.006	−0.064	0.014
Difference in age	0.003	0.006	−0.027	0.011

Table 3: Estimated regression coefficients and their standard errors (SE) for Lazega's lawyer friendship network.

4.2 Link prediction for the lawyer data

In this section we give a brief illustration of how our model can be used for link prediction. We remark that in general it is not possible to conduct vanilla train-test splits or cross-validation on network data, since randomly removing nodes or edges can destroy part of the network structure (Li et al., 2020). While some prior work on network cross-validation exists, notably the aforementioned paper, developing a rigorous cross-validation scheme for our model is beyond the scope of the current paper. Therefore, we present the results in this section as a guideline that the present model shows promising performance for link prediction, even when using vanilla cross-validation. We leave the rigorous mathematical treatment of this interesting problem for a future paper.

In the following, we mimic traditional K-fold cross-validation with K=10 for this dataset as follows. We randomly split the entries of the adjacency matrix A of the lawyer data into 10 folds (ignoring the diagonal). This corresponds to randomly selecting node pairs, as advocated in Li et al. (2020). For each fold, we fit our model on the other 90% of of the entries of A and use the fitted model to predict the values for the 10% holdout fold. We run 10 repetitions of this modified 10-fold cross-validation and present the results in Figure 4. As we can see, even when using this vanilla approach, we achieve a decent performance in the high 80% range usually, which suggests that with some further adjustments it might be possible to derive strong theoretical guarantees for link prediction for the present model. The key ingredient for why we believe this could be a fruitful research avenue is that link formation happens independently, conditional on the values of the covariates Z.

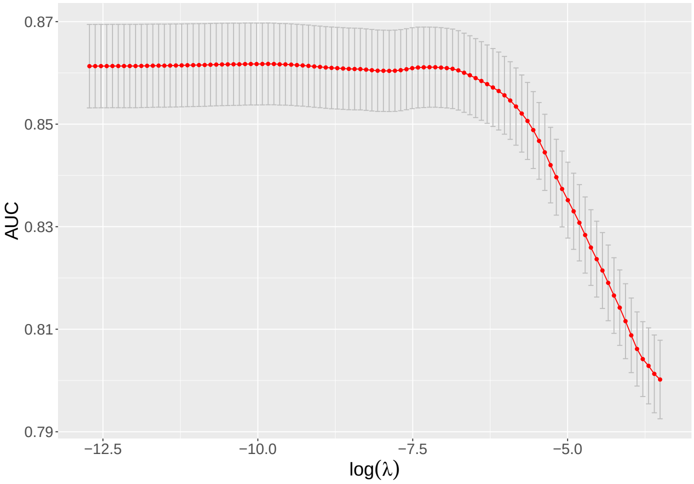


Figure 4: Area under the curve as a function of λ, averaged over 10 runs of 10-fold cross validations. Error bars are the length of one standard error, averaged over the 10 runs of 10-fold cross-validation.

4.3 Sina Weibo data

In our second case study we explore how well our estimation procedure scales to large networks. Towards this, we study the Sina Weibo data collected by Cai et al. (2018), which was also analysed in Yan et al. (2019). Sina Weibo is a Chinese social media platform similar to Twitter. In the original data set there are 4077 nodes representing MBA students and directed links represent who follows whom. Following Yan et al. (2019), we focus on the largest strongly connected component consisting of 2242 nodes, leaving us with a very sparse network in which only 0.8% of all possible edges are observed. The resulting in- and out-degree sequences have heavy tails, meaning this network also exhibits a high degree of degree heterogeneity, as illustrated in Figure 5. The in-degrees range from 1 to 253, with the first quartile, median and third quartile equal to 4, 9 and 22 respectively. For the out-degrees we observe values between 1 and 715, with first quartile, median and third quartile equal to 2, 5 and 19 respectively.

For each node we observe the number of posts they have written, their tenure (measured in months since they joined the platform) and the number of characters in personal labels which were created by the users to describe their lifestyle. We used the absolute difference between these variables as covariates and standardized their values before fitting our model to the data using BIC. In Table 4 we compare the estimates for γ and their standard errors when our model is fitted with the results obtained by Yan et al. (2019). While both papers estimate all covariates as significant with a negative sign, it is noteworthy that Yan et al. (2019) initially obtained positive covariate estimates for the difference in the number of posts and labels. Only after running their bias-correction procedure did they obtain the estimates presented here, which illustrates that indeed this ad-hoc procedure is necessary for their model to work properly. On the other hand, we obtain the the negative signs right off-the bat, while also giving us smaller standard errors.

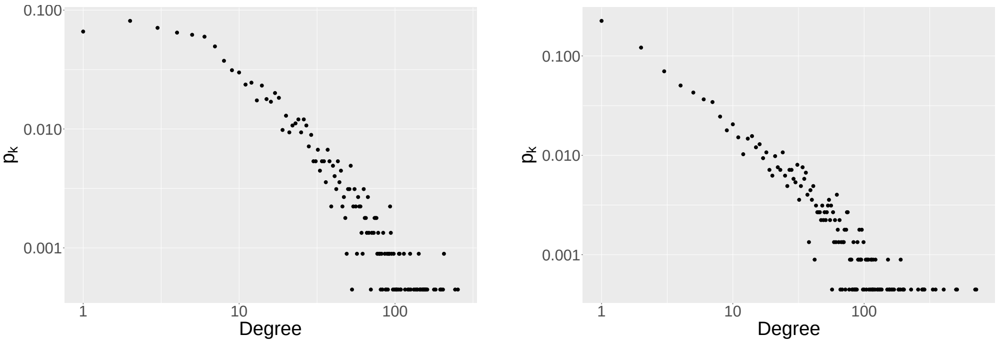


Figure 5: (a): The in-degree distribution in the Sina Weibo dataset, log-log scale (b): The out-degree distribution in the Sina Weibo dataset, log-log scale. The y-axis is the empirical frequency with which a give in- / out- degree was observed.

	This Paper		Yan et al. (2019)	
Covariate	Estimate	SE	Estimate	SE
Difference in posts	−0.127	0.006	−0.391	0.018
Difference in tenure	−0.050	0.005	−0.143	0.008
Difference in number of labels	−0.080	0.005	−0.158	0.008

Table 4: Estimated regression coefficients and their standard errors (SE) for the Sina Weibo network.

5. Discussion

We have assumed that links are formed independently between node pairs. This is a limitation because empirically reciprocity, a measure of the likelihood of vertices in a directed network to be mutually linked, may be present. In the motivating lawyers data for example, lawyer j will be more likely to call lawyer i a friend if the converse is true. To address this layer of sophistication, the next natural step is to add a reciprocity parameter to the model. In this paper, we have focused on the inference of the covariate parameter. In some applications inference on α and β may be of interest. Since their estimates are biased due to the shrinkage incurred by our ℓ_1 penalty, this will require debiasing, possibly coupled with suitable balancing assumptions. We leave the exploration of these two interesting research questions for future work.

Acknowledgments

We thank the Action Editor and two reviewers for their very helpful comments that have led to a much improved paper.

Appendix A. Appendix A

We introduce the following additional notation.

For any $a, b \in \mathbb{R}$ we use the notation $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. For any subset $S \subset [n]$, we denote by v_S the vector v with components not belonging to S set to zero. For a matrix $A \in \mathbb{R}^{d\times d}$ and a subset $S \subseteq [d]$, denote by $A_{S,S} \in R^{|S|\times|S|}$ the submatrix of A obtained by only taking the rows and columns belonging to S. Denote by $A_{-,S} \in \mathbb{R}^{d\times|S|}$ the submatrix obtained by keeping all the rows and taking only those columns belonging to S and by $A_{S,-} \in \mathbb{R}^{|S|\times d}$ the submatrix obtained by taking only those rows belonging to S and keeping all the columns. For any square matrix A, we denote by $\operatorname{maxeval}(A)$ its maximum eigenvalue and by $\operatorname{mineval}(A)$ its minimum eigenvalue.

A.1 A compatibility condition

In this section we first prove a *sample compatibility condition* before providing a proof for the population compatibility condition in Proposition 4. That is, we first want to find a suitable relation between the quantities $\|\hat\theta - \theta_0\|_1$ and $(\hat\theta - \theta_0)\hat\Sigma(\hat\theta - \theta_0)$ where $\hat\Sigma = T^{-1}D^T DT^{-1}$ is the sample version of the sample size adjusted Gram matrix Σ.

We make this mathematically precise now: For a general matrix $A \in \mathbb{R}^{(2n+1+p)\times(2n+1+p)}$ we say the *compatibility condition holds*, if A has the following property: There is a constant b independent of n such that for every $\theta \in \mathbb{R}^{2n+1+p}$ with $\|\theta_{S^c_{0,+}}\|_1 \le 3\|\theta_{S_{0,+}}\|_1$ it holds that

$$\|\theta_{S_{0,+}}\|_1^2 \le \frac{s_{0,+}}{b}\theta^T A\theta.$$

Notice that the compatibility condition is clearly equivalent to the condition that

$$\kappa^2(A, s_0) := \min_{\substack{\theta\in\mathbb{R}^{2n+1+p}\setminus\{0\}\\ \|\theta_{S^c_{0,+}}\|_1\le 3\|\theta_{S_{0,+}}\|_1}} \frac{\theta^T A\theta}{\frac{1}{s_{0,+}}\|\theta_{S_{0,+}}\|_1^2}$$

stays bounded away from zero.

We first show that the compatibility condition holds for the matrix

$$\Sigma_A := \begin{bmatrix} I_{2n} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbb{E}[Z^T Z/N] \end{bmatrix} \in \mathbb{R}^{(2n+1+p)\times(2n+1+p)},$$

where I_{2n} is the $(2n)\times(2n)$ identity matrix.

Recall that by assumption 1, the minimum eigenvalue $\lambda_{\min} = \lambda_{\min}(n)$ of $\frac{1}{N}\mathbb{E}[Z^T Z]$ stays uniformly bounded away from zero. That is, there is a finite constant $c_{\min} > 0$ independent of n, such that $\lambda_{\min} > c_{\min} > 0$ for all n. Then, clearly, for any $\theta = (\vartheta^T, \mu, \gamma^T)^T$,

$$\theta^T\Sigma_A\theta = \|\vartheta\|_2^2 + \mu^2 + \gamma^T\frac{1}{N}\mathbb{E}[Z^T Z]\gamma \ge \|\vartheta\|_2^2 + \mu^2 + c_{\min}\|\gamma\|_2^2 \ge (1 \wedge c_{\min})\|\theta\|_2^2.$$

Thus, Σ_A is strictly positive definite. Furthermore, by Cauchy-Schwarz' inequality, for any $\theta \in \mathbb{R}^{2n+1+p}$ with $\|\theta_{S^c_{0,+}}\|_1 \le 3\|\theta_{S_{0,+}}\|_1$,

$$\frac{1}{s_{0,+}}\|\theta_{S_{0,+}}\|_1^2 \le \|\theta_{S_{0,+}}\|_2^2 \le \|\theta\|_2^2.$$

Thus,

$$\kappa^2(\Sigma_A, s_0) = \min_{\substack{\theta\in\mathbb{R}^{2n+1+p}\backslash\{0\} \\ \|\theta_{S_{0,+}^c}\|_1\le 3\|\theta_{S_{0,+}}\|_1}} \frac{\theta^T\Sigma_A\theta}{\frac{1}{s_{0,+}}\|\theta_{S_{0,+}}\|_1^2} \geq \frac{(1\wedge c_{\min})\|\theta\|_2^2}{\|\theta\|_2^2} > 0.$$

We conclude that the compatibility condition holds for Σ_A. Now, we need to show that with high probability $\kappa(\hat{\Sigma}, s_0) \geq \kappa(\Sigma_A, s_0)$, which would imply that the compatibility condition holds with high probability for $\hat{\Sigma}$. To that end, we have the following auxiliary lemma found in Kock and Tang (2019). For completeness, we give the short proof of it. The notation is adapted to our setting.

Lemma 7 (Lemma 6 in Kock and Tang (2019)) *Let A and B be two positive semi-definite $(2n+1+p)\times(2n+1+p)$ matrices and $\delta = \max_{ij}|A_{ij} - B_{ij}|$. For any set $S_0 \subset \{1,\dots,2n\}$ with cardinality s_0, one has*

$$\kappa^2(B, s_0) \geq \kappa^2(A, s_0) - 16\delta(s_0 + p + 1).$$

Proof Denote by $S_{0,+} = S_0 \cup \{2n+1,\dots,2n+1+p\}$ and $s_{0,+} = s_0 + (1+p)$. Let $\theta \in \mathbb{R}^{2n+1+p}\backslash\{0\}$, with $\|\theta_{S_{0,+}^c}\|_1 \leq 3\|\theta_{S_{0,+}}\|_1$. Then,

$$\begin{aligned}|\theta^T A\theta - \theta^T B\theta| &= |\theta^T(A-B)\theta| \leq \|\theta\|_1\|(A-B)\theta\|_\infty \leq \delta\|\theta\|_1^2 \\ &= \delta(\|\theta_{S_{0,+}}\|_1 + \|\theta_{S_{0,+}^c}\|_1)^2 \leq \delta(\|\theta_{S_{0,+}}\|_1 + 3\|\theta_{S_{0,+}}\|_1)^2 \\ &\leq 16\delta\|\theta_{S_{0,+}}\|_1^2.\end{aligned}$$

Hence, $\theta^T B\theta \geq \theta^T A\theta - 16\delta\|\theta_{S_{0,+}}\|_1^2$ and thus

$$\frac{\theta^T B\theta}{\frac{1}{s_{0,+}}\|\theta_{S_{0,+}}\|_1^2} \geq \frac{\theta^T A\theta}{\frac{1}{s_{0,+}}\|\theta_{S_{0,+}}\|_1^2} - 16\delta s_{0,+} \geq \kappa^2(A, s_0) - 16\delta s_{0,+}.$$

Minimizing the left-hand side over all $\theta \neq 0$ with $\|\theta_{S_{0,+}^c}\|_1 \leq 3\|\theta_{S_{0,+}}\|_1$ proves the claim. ■

This shows that to control $\kappa^2(\hat{\Sigma}, s_0)$, we need to control the maximum element-wise distance between $\hat{\Sigma}$ and Σ_A: $\max_{ij}|\hat{\Sigma}_{ij} - \Sigma_{A,ij}|$. Introduce the set

$$\mathcal{J} = \left\{\max_{ij}|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| \leq \frac{c_{\min}}{32 s_{0,+}}\right\}.$$

On the set $\mathcal{J}$, by Lemma 7, we have $\kappa^2(\hat{\Sigma}, s_0) \geq \kappa(\Sigma_A, s_0) - \frac{c_{\min}}{2} \geq \frac{c_{\min}}{2} > 0$ and thus the compatibility condition holds for $\hat{\Sigma}$ on $\mathcal{J}$.

Lemma 8 *If $s_0 = o(\sqrt{n})$, for n large enough, with $\delta = \frac{c_{min}}{32 s_{0,+}}$ and $\tilde{c} = c^2 \vee (2c^4)$, where $c > 0$ is the universal constant such that $|Z_{k,ij}| \leq c$ for all k, i, j, we have*

$$P(\mathcal{J}) = P\left(\max_{ij}|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| \leq \frac{c_{min}}{32 s_{0,+}}\right) \geq 1 - p(p+3)\exp\left(-N\frac{c_{\min}^2}{2048 s_{0,+}^2\tilde{c}}\right).$$

Proof To make referencing of sections of $\hat{\Sigma}$ easier, we number its blocks as follows

$$\hat{\Sigma} = T^{-1} \begin{bmatrix} \underbrace{X^T X}_{\textcircled{1}} & \underbrace{X^T \mathbf{1}}_{\textcircled{2}} & \underbrace{X^T Z}_{\textcircled{3}} \\ \underbrace{\mathbf{1}^T X}_{\textcircled{4}} & \underbrace{\mathbf{1}^T \mathbf{1}}_{\textcircled{5}} & \underbrace{\mathbf{1}^T Z}_{\textcircled{6}} \\ \underbrace{Z^T X}_{\textcircled{7}} & \underbrace{Z^T \mathbf{1}}_{\textcircled{8}} & \underbrace{Z^T Z}_{\textcircled{9}} \end{bmatrix} T^{-1}.$$

For block ①, i.e. $i, j = 1, \ldots, 2n$, notice that $(X^{\text{out}})^T X^{\text{out}} = (X^{\text{in}})^T X^{\text{in}} = (n-1) I_n$ and $(X^{\text{out}})^T X^{\text{in}}$ is a matrix with zero on the diagonal and ones everywhere else. Therefore, we have either $\hat{\Sigma}_{ij} = \Sigma_{A,ij}$ or

$$|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| = \frac{1}{n-1} < \frac{c_{\min}}{32 s_{0,+}},$$

for n large enough, since $s_{0,+} = o(\sqrt{n})$. Blocks ② and ④ are a $2n$ dimensional column and row vector respectively in which each entry is equal to $n-1$. Thus, for i, j corresponding to these blocks,

$$|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| = \frac{n-1}{\sqrt{(n-1)N}} = \frac{1}{\sqrt{n}} \leq \frac{c_{\min}}{32 s_{0,+}},$$

for n large enough, since $s_{0,+} = o(\sqrt{n})$. For i, j corresponding to blocks ③ and ⑦, we have

$$|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| = \frac{c}{\sqrt{n}} < \frac{c_{\min}}{32 s_{0,+}},$$

for n large enough. Block ⑤ is a single real number and equal for $\hat{\Sigma}$ and Σ_A.

The only cases left to consider are those entries corresponding to blocks ⑥, ⑧ and ⑨. For the blocks ⑥ and ⑧, that is for $i = 2n+1, j = 2n+2, \ldots, 2n+1+p$ and $i = 2n+2, \ldots, 2n+1+p, j = 2n+1$, $\hat{\Sigma}_{ij} - \Sigma_{A,ij} = \hat{\Sigma}_{ij}$ is the scaled sum of all the entries of some column Z_k of the matrix Z for an appropriate k. That is, there is a $1 \leq k \leq p$ such that

$$\hat{\Sigma}_{ij} - \Sigma_{A,ij} = \frac{1}{N} Z_k^T \mathbf{1} = \frac{1}{N} \sum_{s \neq t} Z_{k,st}.$$

Note, that thus by model assumption $\mathbb{E}[\hat{\Sigma}_{ij} - \Sigma_{A,ij}] = 0$. We know that for each k, s, t : $Z_{k,st} \in [-c, c]$. Hence, by Hoeffding's inequality, for all $\delta > 0$,

$$P\left(|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| \geq \delta\right) = P\left(\left|\sum_{s \neq t} Z_{k,st}\right| \geq N\delta\right) \leq 2\exp\left(-\frac{2N^2\delta^2}{\sum_{i \neq j}(2c)^2}\right) = 2\exp\left(-N\frac{\delta^2}{2c^2}\right).$$

For block ⑨, that is for $i, j = 2n+2, \ldots, 2n+1+p$, a typical element has the form

$$\hat{\Sigma}_{ij} - \Sigma_{A,ij} = \frac{1}{N} \sum_{s \neq t} \left\{ Z_{k,st} Z_{l,st} - \mathbb{E}[Z_{k,st} Z_{l,st}] \right\},$$

for appropriate k, l. In other words, $\hat{\Sigma}_{ij} - \Sigma_{A,ij}$ is the inner product of two columns of Z, minus their expectation, scaled by $1/N$. Since $Z_{k,st} Z_{l,st} \in [-c^2, c^2]$ for all k, l, s, t, we have that for all k, l, s, t: $Z_{k,st} Z_{l,st} - \mathbb{E}[Z_{k,st} Z_{l,st}] \in [-2c^2, 2c^2]$. Thus, by Hoeffding's inequality, for all $\delta > 0$,

$$P\left(|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| \geq \delta\right) = P\left(\left|\sum_{s \neq t} \{Z_{k,st} Z_{l,st} - \mathbb{E}[Z_{k,st} Z_{l,st}]\}\right| \geq N\delta\right) \leq 2\exp\left(-N\frac{\delta^2}{8c^4}\right).$$

Thus, with $\tilde{c} = c^2 \vee (2c^4)$, we have for any entry in blocks ⑥, ⑧, ⑨, that for any $\delta > 0$,

$$P\left(|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| \geq \delta\right) \leq 2\exp\left(-N\frac{\delta^2}{2\tilde{c}}\right).$$

Choosing $\delta = \frac{c_{\min}}{32 s_{0,+}}$, by the exposition above we know that all entries in blocks ① - ⑤ and ⑦ are bounded by δ for $n \gg 0$. Also, because block ⑥ is the transpose of block ⑧, it is sufficient to control one of them. By symmetry of block ⑨ it suffices to control the upper triangular half, including the diagonal, of block ⑨. Thus, we only need to control the entries $\hat{\Sigma}_{ij} - \Sigma_{A,ij}$ for i, j in the following index set

$$\begin{aligned}\mathcal{A} &= \{i, j : i, j \text{ belong to block ⑧ or the upper triangular half or the diagonal of block ⑨}\} \\ &= \{(i, j) \in \{n+2, \ldots, n+1+p\} \times \{n+1\}\} \cup \{i \leq j : i, j = n+2, \ldots, n+1+p\}.\end{aligned}$$

Keep in mind that block ⑧ has p elements, while the upper triangular part of block ⑨ plus its diagonal has $\binom{p}{2} + p = \binom{p+1}{2}$ elements. Thus, for $n \gg 0$,

$$\begin{aligned} P(\mathcal{J}^c) &= P\left(\max_{ij} |\hat{\Sigma}_{ij} - \Sigma_{A,ij}| \geq \frac{c_{\min}}{32 s_{0,+}}\right) \\ &\leq \sum_{i,j \in \mathcal{A}} P\left(|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| \geq \frac{c_{\min}}{32 s_{0,+}}\right) \\ &\leq 2p \exp\left(-N\frac{\delta^2}{2c^2}\right) + 2\binom{p+1}{2} \exp\left(-N\frac{\delta^2}{8c^4}\right) \\ &\leq 2\left(p + \binom{p+1}{2}\right) \exp\left(-N\frac{\delta^2}{2\tilde{c}}\right) \\ &= p(p+3) \exp\left(-N\frac{\delta^2}{2\tilde{c}}\right). \end{aligned}$$

This proves the claim. ■

We summarize these results in the following proposition

Proposition 9 *Under Assumption 1, for $s_0 = o(\sqrt{n})$ and n large enough, with $\tilde{c} = c^2 \vee (2c^4)$, where $c > 0$ is the universal constant such that $|Z_{k,ij}| \leq c$ for all k, i, j: With probability at least*

$$1 - p(p+3)\exp\left(-N\frac{c_{\min}^2}{2048 s_{0,+}^2 \tilde{c}}\right)$$

it holds that for every $\theta \in \mathbb{R}^{2n+1+p}$ *with* $\|\theta_{S_{0,+}^c}\|_1 \leq 3\|\theta_{S_{0,+}}\|_1$,

$$\|\theta_{S_{0,+}}\|_1^2 \leq \frac{2s_{0,+}}{c_{\min}} \theta^T \hat{\Sigma} \theta.$$

Proof This follows from Lemma 8. ■

Proof [Proof of Proposition 4] To prove that the compatibility condition holds for the population sample-size adjusted Gram matrix Σ we may follow the same steps as in the proof of Proposition 9: Number the blocks of Σ as ① - ⑨ as we did for $\hat{\Sigma}$. Σ and Σ_A are equal on blocks ③, ⑤, ⑥, ⑦, ⑧ and ⑨. For blocks ①, ② and ④ we use the exact same arguments as in the proof of Proposition 9 to find that for n sufficiently large, almost surely,

$$\max_{ij} |\Sigma_{ij} - \Sigma_{A,ij}| \leq \frac{c_{\min}}{32 s_{0,+}}.$$

The claim follows from Lemma 7. ■

A.2 A rescaled estimation problem

We now formally introduce the notion of sample-size adjusted parameters $\bar{\theta}$. Precisely, define the *sample size adjusted design matrix* $\bar{D}$ as

$$\bar{D} = \left[\begin{array}{c|c|c} \bar{X} & \mathbf{1} & Z \end{array} \right] \in \mathbb{R}^{N \times (2n+p+1)},$$

where

$$\bar{X} = \left[\begin{array}{c|c} \bar{X}^{\text{out}} & \bar{X}^{\text{in}} \end{array} \right] = \left[\begin{array}{c|c} \sqrt{n} X^{\text{out}} & \sqrt{n} X^{\text{in}} \end{array} \right],$$

is blowing up the entries in D belonging to ϑ. Recall that for any parameter $\theta = (\vartheta^T, \mu, \gamma^T)^T \in \Theta$, we use

$$\bar{\theta} = (\bar{\vartheta}, \mu, \gamma) = \left(\frac{1}{\sqrt{n}} \vartheta, \mu, \gamma \right)$$

to refer to its sample-size adjusted version. In particular we use the notation $\bar{\theta}_0 = (\bar{\vartheta}_0^T, \mu_0, \gamma_0^T)^T$, to denote the re-parametrized true parameter value. The blow-up factor $\sqrt{n}$ was chosen precisely such that we can now reformulate our problem as a problem in which each parameter effectively has sample size N in the sense that

$$\Sigma = \frac{1}{N} \mathbb{E}[\bar{D}^T \bar{D}].$$

Our original penalized likelihood problem can be rewritten as

$$\begin{aligned} \hat{\bar{\theta}} = (\hat{\bar{\vartheta}}, \hat{\mu}, \hat{\gamma}) = \underset{\substack{\bar{\vartheta} = (\bar{\alpha}^T, \bar{\beta}^T)^T, \\ \mu, \gamma}}{\arg\min} \frac{1}{N} \Bigg(& - \sum_{i=1}^n \sqrt{n} \bar{\alpha}_i b_i - \sum_{i=1}^n \sqrt{n} \bar{\beta}_i d_i - d_+ \mu - \sum_{i \neq j} (Z_{ij}^T \gamma) A_{ij} \\ & + \sum_{i \neq j} \log\left(1 + \exp\left(\sqrt{n} \bar{\alpha}_i + \sqrt{n} \bar{\beta}_j + \mu + Z_{ij}^T \gamma\right)\right) \Bigg) \\ & + \bar{\lambda} \|\bar{\vartheta}\|_1, \end{aligned} \tag{13}$$

where $\bar{\lambda} = \sqrt{n}\lambda$ and the argmin is taken over $\bar{\Theta}_{\text{loc}} = \{\bar{\theta} : \theta \in \Theta, \|\bar{D}\bar{\theta}\|_\infty \leq r_n\}$. Note that by the same arguments as before, $\bar{\Theta}_{\text{loc}}$ is convex. Then, given a solution $\hat{\bar{\theta}}$ for a given penalty parameter $\bar{\lambda}$ to this modified problem (13), we can obtain a solution to our original problem (4) with penalty parameter $\lambda = \bar{\lambda}/\sqrt{n}$, by setting

$$(\hat{\vartheta}, \hat{\mu}, \hat{\gamma}) = \left(\sqrt{n}\hat{\bar{\vartheta}}, \hat{\mu}, \hat{\gamma}\right).$$

Note that for any $\theta \in \Theta$, $D\theta = \bar{D}\bar{\theta}$, and hence the bound r_n is the same in the definitions of Θ_{loc} and $\bar{\Theta}_{\text{loc}}$. Note also that $\theta \in \Theta_{\text{loc}}$ if and only if $\bar{\theta} \in \bar{\Theta}_{\text{loc}}$. For any $\bar{\theta} = (\bar{\vartheta}^T, \mu, \gamma)^T$, denote the negative log-likelihood function corresponding to the rescaled problem (13) as $\bar{\mathcal{L}}(\bar{\theta})$. Then, clearly $\bar{\mathcal{L}}(\bar{\theta}) = \mathcal{L}(\theta)$ and $\mathbb{E}[\bar{\mathcal{L}}(\bar{\theta})] = \mathbb{E}[\mathcal{L}(\theta)]$. Thus, $\bar{\theta}_0$ satisfies that $\bar{\theta}_0 = \arg\min_{\theta \in \bar{\Theta}} \mathbb{E}[\bar{\mathcal{L}}(\bar{\theta})]$. We define the excess risk for a sample-size adjusted parameter $\bar{\theta}$ as

$$\bar{\mathcal{E}}(\bar{\theta}) = \frac{1}{N}\mathbb{E}[\bar{\mathcal{L}}(\bar{\theta}) - \bar{\mathcal{L}}(\bar{\theta}_0)]$$

By construction, $\bar{\mathcal{E}}(\bar{\theta}) = \mathcal{E}(\theta)$.

A.3 A basic Inequality

A key result in the consistency proofs in classical LASSO settings is the so called *basic inequality* (cf. van de Geer and Bühlmann (2011), Chapter 6). Let P_n denote the empirical measure with respect to our observations (A_{ij}, Z_{ij}), that is, for any suitable function g,

$$P_n g \coloneqq \frac{1}{N}\sum_{i \neq j} g(A_{ij}, Z_{ij}).$$

In particular, if we let for each $\theta \in \Theta$, $l_\theta(A_{ij}, Z_{ij}) = -A_{ij}(\alpha_i + \beta_j + \mu + \gamma^T Z_{ij}) + \log(1 + \exp(\alpha_i + \beta_j + \mu + \gamma^T Z_{ij}))$, then $P_n l_\theta = \mathcal{L}(\theta)/N$. Similarly, we define the theoretical risk as $P = \mathbb{E}P_n$. In particular,

$$P l_\theta = \mathbb{E}P_n l_\theta = \frac{1}{N}\mathbb{E}[\mathcal{L}(\theta)],$$

where we suppress the dependence of the theoretical risk on n in our notation. We may write the excess risk as

$$\mathcal{E}(\theta) \coloneqq P(l_\theta - l_{\theta_0}).$$

We define the *empirical process* as

$$\{v_n(\theta) = (P_n - P)l_\theta : \theta \in \Theta\}.$$

Lemma 10 (Basic Inequality) *For any $\theta = (\beta^T, \mu, \gamma^T)^T \in \Theta_{\text{loc}}$ it holds*

$$\mathcal{E}(\hat{\theta}) + \lambda\|\hat{\beta}\|_1 \leq -[v_n(\hat{\theta}) - v_n(\theta)] + \mathcal{E}(\theta) + \lambda\|\beta\|_1.$$

Proof By plugging in the definitions and rearranging, we see that the above equation is equivalent to

$$\frac{1}{N}\mathcal{L}(\hat{\theta}) + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{N}\mathcal{L}(\theta) + \lambda\|\beta\|_1,$$

which is true by definition of $\hat{\theta}$. ■

Notice that since the basic inequality in Lemma 10 only relies on the argmin property of the estimator $\hat{\theta}$, an analogous result follows line by line for the rescaled parameter $\hat{\bar{\theta}}$. Writing

$$\bar{v}_n(\bar{\theta}) := \frac{1}{N}(\bar{\mathcal{L}}(\bar{\theta}) - \mathbb{E}[\bar{\mathcal{L}}(\bar{\theta})]) = v_n(\theta).$$

for the rescaled empirical process, we have the following.

Lemma 11 *For any $\bar{\theta} \in \bar{\Theta}_{\mathrm{loc}}$ it holds*

$$\bar{\mathcal{E}}(\hat{\bar{\theta}}) + \bar{\lambda}\|\hat{\bar{\vartheta}}\|_1 \leq -[\bar{v}_n(\hat{\bar{\theta}}) - \bar{v}_n(\bar{\theta})] + \bar{\mathcal{E}}(\bar{\theta}) + \bar{\lambda}\|\bar{\vartheta}\|_1.$$

Remark 12 *For any $0 < t < 1$ and $\theta \in \Theta_{loc}$, let $\tilde{\theta} = t\hat{\theta} + (1-t)\theta$. Since Γ is convex, $\tilde{\theta} \in \Theta_{loc}$ and since $\theta \to l_\theta$ and $\| \,.\, \|_1$ are convex functions, we can replace $\hat{\theta}$ by $\tilde{\theta}$ in the basic inequality and still obtain the same result. Plugging in the definitions, we see that the basic inequality is equivalent to the following:*

$$\mathcal{E}(\tilde{\theta}) + \lambda\|\tilde{\beta}\|_1 \leq -[v_n(\tilde{\theta}) - v_n(\theta)] + \lambda\|\beta\|_1 + \mathcal{E}(\theta)$$
$$\iff \frac{1}{N}\mathcal{L}(\tilde{\theta}) + \lambda\|\tilde{\beta}\|_1 \leq \frac{1}{N}\mathcal{L}(\theta) + \lambda\|\beta\|_1$$

and by convexity

$$\frac{1}{N}\mathcal{L}(\tilde{\theta}) + \lambda\|\tilde{\beta}\|_1 \leq \frac{1}{N}t\mathcal{L}(\hat{\theta}) + \frac{1}{N}(1-t)\mathcal{L}(\theta) + t\lambda\|\hat{\beta}\|_1 + (1-t)\lambda\|\beta\|_1 \leq \frac{1}{N}\mathcal{L}(\theta) + \lambda\|\beta\|_1,$$

where the last inequality follows by definition of $\hat{\theta}$. In particular, for any $M > 0$, choosing

$$t = \frac{M}{M + \|\hat{\theta} - \theta\|_1},$$

gives $\|\tilde{\theta} - \theta\|_1 \leq M$. The completely analogous result holds for $\bar{\theta}$.

A.4 Two norms and one function space

To give us a more compact way of writing, for any $\bar{\theta} \in \Theta$ we introduce functions $f_{\bar{\theta}} : \mathbb{R}^{2n+1+p} \to \mathbb{R}, f_{\bar{\theta}}(v) = v^T\bar{\theta}$ and denote the function space of all such $f_{\bar{\theta}}$ by $\bar{\mathbb{F}} := \{f_{\bar{\theta}} : \bar{\theta} \in \Theta\}$. We endow $\bar{\mathbb{F}}$ with two norms as follows:
Denote the law of the rows of $\bar{D}$ on $\mathbb{R}^{2n+1+p}$, i.e. the probability measure induced by $(\bar{X}_{ij}^T, 1, Z_{ij}^T)^T, i \neq j$, by $\bar{Q}$. That is, for a measurable set $A = A_1 \times A_2 \subset \mathbb{R}^{2n+1} \times \mathbb{R}^p$,

$$\bar{Q}(A) = \frac{1}{N}\sum_{i \neq j} P(\bar{D}_{ij} \in A) = \frac{1}{N}\sum_{i \neq j} \delta_{ij}(A_1) \cdot P(Z_{ij} \in A_2),$$

where $\delta_{ij}(A_1) = 1$ if $(\bar{X}_{ij}^T, 1)^T \in A_1$ and zero otherwise, is the Dirac-measure. We are interested in the L_2 and L_∞ norm on $\bar{\mathbb{F}}$ with respect to the measure $\bar{Q}$ on $\mathbb{R}^{2n+1} \times \mathbb{R}^p$.

Denote the $L_2(\bar{Q})$-norm of $f \in \bar{\mathbb{F}}$ simply by $\| \, . \, \|_{\bar{Q}}$ and let $\mathbb{E}_Z$ be the expectation with respect to Z:

$$\|f\|_{\bar{Q}}^2 := \|f\|_{L_2(\bar{Q})}^2 = \int_{\mathbb{R}^{2n+1} \times \mathbb{R}^p} f(v)^2 \bar{Q}(dv) = \frac{1}{N} \sum_{i \neq j} \mathbb{E}_Z[f((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T)^2]$$

and define the $L_\infty(\bar{Q})$-norm as usual as the $\bar{Q}$-a.s. smallest upper bound of f:

$$\|f\|_{\bar{Q},\infty} = \inf\{C \geq 0 : |f(v)| \leq C \text{ for } \bar{Q}\text{-almost every } v \in \mathbb{R}^{2n+1+p}\}.$$

Notice in particular, that for any $f_{\bar{\theta}} \in \bar{\mathbb{F}}, \bar{\theta} \in \bar{\Theta}_{\text{loc}}$: $\|f_{\bar{\theta}}\|_\infty \leq \sup_{Z_{ij}} \|\bar{D}\bar{\theta}\|_\infty \leq r_n$.

We make the analogous definitions for the unscaled design matrix. Let Q denote the probability measure induced by the rows of D. Since $\bar{D}\bar{\theta} = D\theta$, for any θ with rescaled version $\bar{\theta}$, we have

$$\|f_{\bar{\theta}}\|_{L_2(\bar{Q})} = \|f_\theta\|_{L_2(Q)}, \quad \|f_{\bar{\theta}}\|_{\bar{Q},\infty} = \|f_\theta\|_{Q,\infty}.$$

We want to apply the compatibility condition to vectors of the form $\theta = \theta_1 - \theta_2, \theta_1, \theta_2 \in \Theta_{\text{loc}}$.

Notice, that we have the following relation between the $L_2(Q)$-norm and the sample size adjusted Gram matrix Σ: For any θ we have

$$\|f_\theta\|_Q^2 = \mathbb{E}_Z \left[\frac{1}{N} \sum_{i \neq j} (D_{ij}^T \theta)^2 \right] = \bar{\theta}^T \Sigma \bar{\theta}. \tag{14}$$

We have the following corollary which follows immediately from Proposition 4 (see e.g. van de Geer and Bühlmann (2011), section 6.12 for a general treatment).

Corollary 13 *Under assumption 1, for $s_0 = o(\sqrt{n})$ and n large enough and with $\tilde{c} = c^2 \vee (2c^4)$, where $c > 0$ is the universal constant such that $|Z_{k,ij}| \leq c$ for all k, i, j, it holds that for every $\bar{\theta} = \bar{\theta}_1 - \bar{\theta}_2, \bar{\theta}_1, \bar{\theta}_2 \in \bar{\Theta}_{\text{loc}}$ with $\|\bar{\theta}_{S_{0,+}^c}\|_1 \leq 3\|\bar{\theta}_{S_{0,+}}\|_1$,*

$$\|\bar{\theta}_{S_{0,+}}\|_1^2 \leq \frac{s_{0,+}}{C} \|f_{\theta_1} - f_{\theta_2}\|_Q^2,$$

where $C = c_{\min}/2$.

Proof By Proposition 4,

$$\|\bar{\theta}_{S_{0,+}}\|_1^2 \leq \frac{2s_{0,+}}{c_{\min}} \bar{\theta} \Sigma \bar{\theta}.$$

The claim follows from (14) and the fact that $\theta \mapsto f_\theta$ is linear. ■

A.5 Lower quadratic margin for $\mathcal{E}$

In this section we will derive a lower quadratic bound on the excess risk $\mathcal{E}(\theta)$ if the parameter θ is close to the truth θ_0. This is a necessary property for the proof to come and is referred to as the *margin condition* in classical LASSO theory (cf. van de Geer and Bühlmann (2011)).

The proof mainly relies on a second order Taylor expansion of the function l_θ of introduced in section 3. Given a fixed θ, we treat l_θ as a function in $\theta^T x$ and define new functions $l_{ij} : \mathbb{R} \to \mathbb{R}, i \neq j$,

$$l_{ij}(a) = \mathbb{E}[l_\theta(A_{ij}, a)|Z_{ij}] = -p_{ij}a + \log(1 + \exp(a)),$$

where $p_{ij} = P(A_{ij} = 1|Z_{ij})$ and by slight abuse of notation we use $l_\theta(A_{ij}, a) \coloneqq -A_{ij}a + \log(1 + \exp(a))$. Taking derivations, it is easy to see that

$$f_{\theta_0}((X_{ij}^T, 1, Z_{ij}^T)^T) \in \arg\min_a l_{ij}(a).$$

All l_{ij} are clearly twice continuously differentiable with derivative

$$\frac{\partial^2}{\partial a^2} l_{ij}(a) = \frac{\exp(a)}{(1 + \exp(a))^2} > 0, \forall a \in \mathbb{R}.$$

Using a second order Taylor expansion around $a_0 = f_0((X_{ij}^T, 1, Z_{ij}^T)^T)$ we get

$$l_{ij}(a) = l_{ij}(a_0) + l'(a_0)(a - a_0) + \frac{l''(\bar{a})}{2}(a - a_0)^2 = l_{ij}(a_0) + \frac{l''(\bar{a})}{2}(a - a_0)^2,$$

with an $\bar{a}$ between a and a_0. Note that $|a_0| \leq r_n$. Then, for any a with $|a| \leq r_n$, we must have that for any intermediate point $\bar{a}$ between a_0 and a it also holds that $|\bar{a}| \leq r_n$. Also note that $\frac{\exp(a)}{(1+\exp(a))^2}$ is symmetric and monotone decreasing for $a \geq 0$. Thus, for any a with $|a| \leq r_n$,

$$\begin{aligned} l_{ij}(a) - l_{ij}(a_0) &= \frac{\exp(\bar{a})}{(1 + \exp(\bar{a}))^2} \frac{(a - a_0)^2}{2} \\ &= \frac{\exp(|\bar{a}|)}{(1 + \exp(|\bar{a}|))^2} \frac{(a - a_0)^2}{2}, \quad \text{by symmetry} \\ &\geq \frac{\exp(r_n)}{(1 + \exp(r_n))^2} \frac{(a - a_0)^2}{2}. \end{aligned} \tag{15}$$

In particular, if we pick any θ and let $a = f_\theta((X_{ij}^T, 1, Z_{ij}^T)^T)$, we have

$$\begin{aligned} &l_{ij}(f_\theta((X_{ij}^T, 1, Z_{ij}^T)^T)) - l_{ij}(f_0((X_{ij}^T, 1, Z_{ij}^T)^T)) \\ &\qquad \geq \frac{\exp(r_n)}{(1 + \exp(r_n))^2} \frac{(f_\theta((X_{ij}^T, 1, Z_{ij}^T)^T) - f_0((X_{ij}^T, 1, Z_{ij}^T)^T))^2}{2}. \end{aligned}$$

Let

$$K_n = \frac{2(1 + \exp(r_n))^2}{\exp(r_n)}. \tag{16}$$

Define a subset $\mathbb{F}_{\text{local}} \subset \mathbb{F}$ as $\mathbb{F}_{\text{local}} = \{f_\theta : \theta \in \Theta_{\text{loc}}\}$. Now, for all $f_\theta \in \mathbb{F}_{\text{local}}$:

$$\begin{aligned}\mathcal{E}(\theta) &= \frac{1}{N}\sum_{i\neq j}\mathbb{E}[l_\theta(A_{ij}, D_{ij}) - l_{\theta_0}(A_{ij}, D_{ij})] \\ &= \frac{1}{N}\sum_{i\neq j}\mathbb{E}[(l_{ij}(f_\theta(D_{ij}) - l_{ij}(f_0(D_{ij})))] \\ &\geq \frac{1}{K_n}\cdot\frac{1}{N}(\theta-\theta_0)^T\mathbb{E}_Z[D^TD](\theta-\theta_0) \\ &= \frac{1}{K_n}\cdot\|f_\theta - f_0\|_Q^2.\end{aligned}$$

Thus, we have obtained a lower bound for the excess risk given by the quadratic function $G_n(\|f_\theta - f_0\|)$ where $G_n(u) = 1/K_n \cdot u^2$. Recall that the convex conjugate of a strictly convex function G on $[0,\infty)$ with $G(0) = 0$ is defined as the function

$$H(v) = \sup_u\{uv - G(u)\}, \quad v > 0,$$

and in particular, if $G(u) = cu^2$ for a positive constant c, we have $H(v) = v^2/(4c)$. Hence, the convex conjugate of G_n is

$$H_n(v) = \frac{v^2K_n}{4}.$$

Keep in mind that by definition for any u, v

$$uv \leq G(u) + H(v).$$

A.6 Consistency on a special set

In this section we will show that the penalized likelihood estimator is consistent. We will first define a set $\mathcal{I}$ and show that consistency holds on $\mathcal{I}$. It will then suffice to show that the probability of $\mathcal{I}$ tends to one as well. The proof follows in spirit van de Geer and Bühlmann (2011), Theorem 6.4.

We define some objects that we will need for the proof of consistency. We want to use the quadratic margin condition derived in section A.5. Recall that the quadratic margin condition holds for any $\theta \in \Theta_{\text{loc}}$. Define

$$\epsilon^* = H_n\left(\frac{4\sqrt{2}\sqrt{s_{0,+}}\bar{\lambda}}{\sqrt{c_{\min}}}\right).$$

Recall the definition of $\bar{\theta}$ in equation (7) and let for any $M > 0$

$$Z_M := \sup_{\substack{\theta\in\Theta_{\text{loc}},\\ \|\bar{\theta}-\bar{\theta}_0\|_1\leq M}} |v_n(\theta) - v_n(\theta_0)|,$$

where v_n denotes the empirical process. The set over which we are maximizing in the definition of Z_M can be expressed in terms of parameters θ on the original scale as

$$\left\{\theta = (\vartheta^T, \mu, \gamma^T)^T \in \Theta_{\text{loc}} : \frac{1}{\sqrt{n}}\|\vartheta - \vartheta_0\|_1 + |\mu - \mu_0| + \|\gamma - \gamma_0\|_1 \leq M\right\}.$$

Set

$$M^* := \epsilon^* / \lambda_0,$$

where λ_0 is a lower bound on $\bar\lambda$ that will be made precise in the proof showing that $\mathcal{I}$ has large probability. Define

$$\mathcal{I} := \{Z_{M^*} \leq \lambda_0 M^*\} = \{Z_{M^*} \leq \epsilon^*\}. \tag{17}$$

Theorem 14 *Assume that assumptions 1 and B2 hold and that $\bar\lambda \geq 8\lambda_0$. Then, on the set $\mathcal{I}$, we have*

$$\mathcal{E}(\hat\theta) + \bar\lambda \left(\frac{1}{\sqrt{n}} \|\hat\vartheta - \vartheta_0\|_1 + |\hat\mu - \mu_0| + \|\hat\gamma - \gamma_0\|_1 \right) \leq 4\epsilon^* = 4H_n \left(\frac{4\sqrt{2}\sqrt{s_{0,+}}\bar\lambda}{\sqrt{c_{\min}}} \right).$$

Proof [Proof of Theorem 14] We assume that we are on the set $\mathcal{I}$ throughout. Set

$$t = \frac{M^*}{M^* + \|\hat{\bar\theta} - \bar\theta_0\|_1}$$

and $\tilde\theta = (\tilde\vartheta^T, \tilde\mu, \tilde\gamma^T)^T = t\hat{\bar\theta} + (1-t)\bar\theta_0$. Then,

$$\|\tilde\theta - \bar\theta_0\|_1 = t\|\hat{\bar\theta} - \bar\theta_0\| \leq M^*.$$

Since $\hat{\bar\theta}, \bar\theta_0 \in \bar\Theta_{\text{loc}}$ and by the convexity of $\bar\Theta_{\text{loc}}$, $\tilde\theta \in \bar\Theta_{\text{loc}}$, and by the remark after Lemma 11, the basic inequality holds for $\tilde\theta$. Also, recall that $\bar{\mathcal{E}}(\bar\theta_0) = 0$:

$$\begin{aligned}
\bar{\mathcal{E}}(\tilde\theta) + \bar\lambda\|\tilde\vartheta\|_1 &\leq -(\bar v_n(\tilde\theta) - \bar v_n(\bar\theta_0)) + \bar{\mathcal{E}}(\bar\theta_0) + \bar\lambda\|\bar\vartheta_0\|_1 \\
&\leq Z_{M^*} + \bar\lambda\|\bar\vartheta_0\|_1 \\
&\leq \epsilon^* + \bar\lambda\|\bar\vartheta_0\|_1.
\end{aligned}$$

From now on write $\tilde{\mathcal{E}} = \bar{\mathcal{E}}(\tilde\theta)$. Note, that $\|\tilde\vartheta\|_1 = \|\tilde\vartheta_{S_0^c}\|_1 + \|\tilde\vartheta_{S_0}\|_1$ and thus, by the triangle inequality,

$$\begin{aligned}
\tilde{\mathcal{E}} + \bar\lambda\|\tilde\vartheta_{S_0^c}\|_1 &\leq \epsilon^* + \bar\lambda(\|\bar\vartheta_0\|_1 - \|\tilde\vartheta_{S_0}\|_1) \\
&\leq \epsilon^* + \bar\lambda(\|\bar\vartheta_0 - \tilde\vartheta_{S_0}\|_1) \\
&\leq \epsilon^* + \bar\lambda(\|\bar\vartheta_0 - \tilde\vartheta_{S_0}\|_1 + \|(\mu_0, \gamma_0^T)^T - (\tilde\mu, \tilde\gamma^T)^T\|_1) \\
&= \epsilon^* + \bar\lambda\|(\tilde\theta - \bar\theta_0)_{S_{0,+}}\|_1.
\end{aligned} \tag{18}$$

Case i) If $\bar\lambda\|(\tilde\theta - \bar\theta_0)_{S_{0,+}}\|_1 \geq \epsilon^*$, then

$$\bar\lambda\|\tilde\vartheta_{S_0^c}\|_1 \leq \tilde{\mathcal{E}} + \bar\lambda\|\tilde\vartheta_{S_0^c}\|_1 \leq 2\bar\lambda\|(\tilde\theta - \bar\theta_0)_{S_{0,+}}\|_1. \tag{19}$$

Since $\|(\tilde\theta - \bar\theta_0)_{S_{0,+}^c}\|_1 = \|\tilde\vartheta_{S_0^c}\|_1$, we may thus apply the compatibility condition corollary 13 (note that $\bar\vartheta_0 = \bar\vartheta_{0,S_0}$) to obtain

$$\|(\tilde\theta - \bar\theta_0)_{S_{0,+}}\|_1 \leq \sqrt{2} \cdot \frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}} \|f_{\tilde\theta} - f_{\bar\theta_0}\|_{\bar Q},$$

where we have used that $\theta \mapsto f_\theta$ is linear and hence $f_{\tilde{\theta}-\bar{\theta}_0} = f_{\tilde{\theta}} - f_{\bar{\theta}_0}$. Observe that

$$\|\tilde{\theta} - \theta_0\|_1 = \|\tilde{\vartheta}_{S_0^c}\|_1 + \|(\tilde{\theta} - \theta_0)_{S_{0,+}}\|_1. \tag{20}$$

Hence,

$$\begin{aligned}
\tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\theta} - \bar{\theta}_0\|_1 &= \tilde{\mathcal{E}} + \bar{\lambda}(\|\tilde{\vartheta}_{S_0^c}\|_1 + \|(\tilde{\theta} - \bar{\theta}_0)_{S_{0,+}}\|_1) \\
&\le \epsilon^* + 2\bar{\lambda}\|(\tilde{\theta} - \bar{\theta}_0)_{S_{0,+}}\|_1 \\
&\le \epsilon^* + 2\sqrt{2}\bar{\lambda}\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\|f_{\tilde{\theta}} - f_{\bar{\theta}_0}\|_{\bar{Q}}.
\end{aligned}$$

Recall that for a convex function G and its convex conjugate H we have $uv \le G(u) + H(v)$. Thus, we obtain

$$\begin{aligned}
2\sqrt{2}\bar{\lambda}\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\|f_{\tilde{\theta}} - f_{\bar{\theta}_0}\|_{\bar{Q}} &= 4\sqrt{2}\bar{\lambda}\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\frac{\|f_{\tilde{\theta}} - f_{\bar{\theta}_0}\|_{\bar{Q}}}{2} \\
&\le H_n\left(4\sqrt{2}\bar{\lambda}\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\right) + G_n\left(\frac{\|f_{\tilde{\theta}} - f_{\bar{\theta}_0}\|_{\bar{Q}}}{2}\right) \\
&\overset{G_n \text{ convex}}{\le} H_n\left(4\sqrt{2}\bar{\lambda}\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\right) + \frac{G_n(\|f_{\tilde{\theta}} - f_{\bar{\theta}_0}\|_{\bar{Q}})}{2} \\
&\overset{\text{margin condition}}{\le} H_n\left(4\sqrt{2}\bar{\lambda}\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\right) + \frac{\tilde{\mathcal{E}}}{2}.
\end{aligned}$$

It follows

$$\tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\theta} - \bar{\theta}_0\|_1 \le \epsilon^* + H_n\left(4\sqrt{2}\bar{\lambda}\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\right) + \frac{\tilde{\mathcal{E}}}{2} = 2\epsilon^* + \frac{\tilde{\mathcal{E}}}{2}$$

and therefore

$$\frac{\tilde{\mathcal{E}}}{2} + \bar{\lambda}\|\tilde{\theta} - \bar{\theta}_0\|_1 \le 2\epsilon^*. \tag{21}$$

Finally, this gives

$$\|\tilde{\theta} - \bar{\theta}_0\|_1 \le \frac{2\epsilon^*}{\bar{\lambda}} = \frac{2\lambda_0 M^*}{\bar{\lambda}} \underbrace{\le}_{\bar{\lambda} \ge 4\lambda_0} \frac{M^*}{2}.$$

From this, by using the definition of $\tilde{\theta}$, we obtain

$$\|\tilde{\theta} - \bar{\theta}_0\|_1 = t\|\hat{\bar{\theta}} - \bar{\theta}_0\|_1 = \frac{M^*}{M^* + \|\hat{\bar{\theta}} - \bar{\theta}_0\|_1}\|\hat{\bar{\theta}} - \bar{\theta}_0\|_1 \le \frac{M^*}{2}.$$

Rearranging gives

$$\|\hat{\bar{\theta}} - \bar{\theta}_0\|_1 \le M^*.$$

Case ii) If $\bar{\lambda}\|(\bar{\theta}_0 - \tilde{\theta})_{S_{0,+}}\|_1 \le \epsilon^*$, then from (18)

$$\tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\vartheta}_{S_0^c}\|_1 \le 2\epsilon^*.$$

Using once more (20), we get

$$\tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\theta} - \bar{\theta}_0\|_1 = \tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\vartheta}_{S_0^c}\|_1 + \bar{\lambda}\|(\tilde{\theta} - \bar{\theta}_0)_{S_{0,+}}\|_1 \leq 3\epsilon^*. \tag{22}$$

Thus,

$$\|\tilde{\theta} - \bar{\theta}_0\|_1 \leq 3\frac{\epsilon^*}{\bar{\lambda}} = 3\frac{\lambda_0}{\bar{\lambda}}M^* \leq \frac{M^*}{2}$$

by choice of $\lambda \geq 6\lambda_0$. Again, plugging in the definition of $\tilde{\theta}$, we obtain

$$\|\hat{\bar{\theta}} - \bar{\theta}_0\|_1 \leq M^*.$$

Hence, in either case we have $\|\hat{\bar{\theta}} - \bar{\theta}_0\|_1 \leq M^*$. That means, we can repeat the above steps with $\hat{\bar{\theta}}$ instead of $\tilde{\theta}$: Writing $\hat{\mathcal{E}} := \bar{\mathcal{E}}(\hat{\bar{\theta}})$, following the same reasoning as above we arrive once more at (18):

$$\hat{\mathcal{E}} + \bar{\lambda}\|\hat{\bar{\vartheta}}_{S_0^c}\|_1 \leq \epsilon^* + \bar{\lambda}\|\bar{\vartheta}^* - \hat{\bar{\vartheta}}_{S_0}\|_1 \leq 2\epsilon^* + \bar{\lambda}\|(\hat{\bar{\theta}} - \bar{\theta}_0)_{S_{0,+}}\|_1.$$

From this, in **case i)** we obtain (19) which allows us to use the compatibility assumption to arrive at (21):

$$\frac{\hat{\mathcal{E}}}{2} + \bar{\lambda}\|\hat{\bar{\theta}} - \bar{\theta}_0\|_1 \leq 2\epsilon^*,$$

resulting in

$$\hat{\mathcal{E}} + \bar{\lambda}\|\hat{\bar{\theta}} - \bar{\theta}_0\|_1 \leq 4\epsilon^*.$$

In **case ii)** on the other hand, we arrive directly at (22), and hence

$$\hat{\mathcal{E}} + \bar{\lambda}\|\hat{\bar{\theta}} - \bar{\theta}_0\|_1 \leq 3\epsilon^*.$$

Plugging in the definitions of $\hat{\bar{\theta}}$ and $\bar{\theta}_0$ and using the fact that $\hat{\mathcal{E}} = \bar{\mathcal{E}}(\hat{\bar{\theta}}) = \mathcal{E}(\hat{\theta})$ proves the claim. ■

A.7 Controlling the special set $\mathcal{I}$

We now show that $\mathcal{I}$ has probability tending to one. Recall some results on concentration inequalities.

Concentration inequalities

We first recall some probability inequalities that we will need. This is based on chapter 14 in van de Geer and Bühlmann (2011). Throughout let $Z_1, \dots, Z_n$ be a sequence of independent random variables in some space $\mathcal{Z}$ and $\mathcal{G}$ be a class of real valued functions on $\mathcal{Z}$.

Definition 15 *A Rademacher sequence is a sequence $\epsilon_1, \dots, \epsilon_n$ of i.i.d. random variables with $P(\epsilon_i = 1) = P(\epsilon_i = -1) = 1/2$ for all i.*

Theorem 16 (Symmetrization Theorem as in van der Vaart and Wellner (1996), abridged) *Let $\epsilon_1, \dots, \epsilon_n$ be a Rademacher sequence independent of $Z_1, \dots, Z_n$. Then*

$$\mathbb{E}\left(\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}\{g(Z_i)-\mathbb{E}[g(Z_i)]\}\right|\right) \leq 2\mathbb{E}\left(\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}\epsilon_i g(Z_i)\right|\right).$$

Theorem 17 (Contraction Theorem as in Ledoux and Talagrand (1991)) *Let $z_1, \dots, z_n$ be non-random elements of $\mathcal{Z}$ and let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{Z}$. Consider Lipschitz functions $g_i : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant $L = 1$, i.e. for all i*

$$|g_i(s) - g_i(s')| \leq |s - s'|, \forall s, s' \in \mathbb{R}.$$

Let $\epsilon_1, \dots, \epsilon_n$ be a Rademacher sequence. Then for any function $f^* : \mathcal{Z} \to \mathbb{R}$ we have

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i\{g_i(f(z_i)) - g_i(f^*(z_i))\}\right|\right) \leq 2\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i\{f(z_i) - f^*(z_i)\}\right|\right).$$

The last theorem we need is a concentration inequality due to Bousquet (2002). We give a version as presented in van de Geer (2008).

Theorem 18 (Bousequet's concentration theorem) *Suppose $Z_1, \dots, Z_n$ and all $g \in \mathcal{G}$ satisfy the following conditions for some real valued constants η_n and τ_n*

$$\|g\|_\infty \leq \eta_n, \ \forall g \in \mathcal{G}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(g(Z_i)) \leq \tau_n^2, \ \forall g \in \mathcal{G}.$$

Define

$$\boldsymbol{Z} \coloneqq \sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}g(Z_i) - \mathbb{E}[g(Z_i)]\right|.$$

Then for any $z > 0$

$$P\left(\boldsymbol{Z} \geq \mathbb{E}[\boldsymbol{Z}] + z\sqrt{2(\tau_n^2 + 2\eta_n\mathbb{E}[\boldsymbol{Z}])} + \frac{2z^2\eta_n}{3}\right) \leq \exp(-nz^2).$$

Remark 19 *Looking at the original paper of Bousquet (2002), their result looks quite different at first. To see that the above falls into their framework, set the variables in Bousquet (2002) as follows*

$$f(Z_i) = (g(Z_i) - \mathbb{E}[g(Z_i)])/(2\eta_n), \qquad \tilde{Z}_k = \sup_f |\sum_{i\neq k} f(Z_i)|,$$

$$f_k = \arg\sup_f |\sum_{i\neq k} f(Z_i)|, \qquad \tilde{Z}'_k = |\sum_{i=1}^{n} f_k(Z_i)| - \tilde{Z}_k$$

$$\tilde{Z} = \frac{2\eta_n}{n}\boldsymbol{Z}.$$

Now apply Theorem 2.1 in Bousquet (2002), choosing for their $(Z, Z_1, \ldots, Z_n)$ the above defined $(\tilde{Z}, \tilde{Z}_1, \ldots, \tilde{Z}_n)$, for their $(Z_1', \ldots, Z_n')$ the above defined $(\tilde{Z}_1', \ldots, \tilde{Z}_n')$ and setting $u = 1$ and $\sigma^2 = \frac{\tau_n^2}{4\eta_n^2}$ in their theorem: The result is exactly Theorem 18 above.

Finally we have a Lemma derived from Hoeffding's inequality. The proof can be found in van de Geer and Bühlmann (2011), Lemma 14.14 (here we use the special case of their Lemma for $m = 1$).

Lemma 20 *Let $\mathcal{G} = \{g_1, \ldots, g_p\}$ be a set of real valued functions on $\mathcal{Z}$ satisfying for all $i = 1, \ldots, n$ and all $j = 1, \ldots, p$*

$$\mathbb{E}[g_j(Z_i)] = 0, \ |g_j(Z_i)| \leq c_{ij}$$

for some positive constants c_{ij}. Then

$$\mathbb{E}\left[\max_{1\leq j\leq p}\left|\sum_{i=1}^{n} g_j(Z_i)\right|\right] \leq [2\log(2p)]^{1/2} \max_{1\leq j\leq p}\left[\sum_{i=1}^{n} c_{ij}^2\right]^{1/2}.$$

The expectation of Z_M

Recall the definition of Z_M

$$Z_M := \sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\text{loc}}, \\ \|\bar{\theta}-\bar{\theta}_0\|_1 \leq M}} |\bar{v}_n(\bar{\theta}) - \bar{v}_n(\bar{\theta}_0)|,$$

where $\bar{v}_n$ denotes the re-parametrized empirical process. Recall, that there is a constant $c \in \mathbb{R}$ such that uniformly $|Z_{ij,k}| \leq c, 1 \leq i \neq j \leq n, k = 1, \ldots, p$.

Lemma 21 *For any $M > 0$ we have*

$$\mathbb{E}[Z_M] \leq 8M(1 \vee c)\sqrt{\frac{2\log(2(2n+p+1))}{N}}.$$

Proof Let $\epsilon_{ij}, i \neq j$, be a Rademacher sequence independent of $A_{ij}, Z_{ij}, i \neq j$. We first want to use the symmetrization theorem 16: For the random variables $Z_1, \ldots, Z_n$ we choose $T_{ij} = (A_{ij}, \bar{X}_{ij}^T, 1, Z_{ij}^T)^T \in \{0,1\} \times \mathbb{R}^{2n+1+p}$. For any $\bar{\theta} \in \bar{\Theta}_{\text{loc}}$ we consider the functions

$$g_{\bar{\theta}}(T_{ij}) = \frac{1}{N}\left\{-A_{ij}\bar{D}_{ij}^T(\bar{\theta} - \bar{\theta}_0) + \log(1 + \exp(\bar{D}_{ij}^T\bar{\theta})) - \log(1 + \exp(\bar{D}_{ij}^T\bar{\theta}_0))\right\}$$

and the function set $\mathcal{G} = \mathcal{G}(M) := \{g_{\bar{\theta}} : \bar{\theta} \in \bar{\Theta}_{\text{loc}}, \|\bar{\theta} - \bar{\theta}_0\|_1 \leq M\}$. Note, that

$$\bar{v}_n(\bar{\theta}) - \bar{v}_n(\bar{\theta}_0) = \sum_{i\neq j}\{g_{\bar{\theta}}(T_{ij}) - \mathbb{E}[g_{\bar{\theta}}(T_{ij})]\}.$$

Then, the symmetrization theorem gives us

$$\begin{aligned}\mathbb{E}[Z_M] &= \mathbb{E}\left[\sup_{g_{\bar{\theta}}\in\mathcal{G}}\left|\sum_{i\neq j} g_{\bar{\theta}}(T_{ij}) - \mathbb{E}[g_{\bar{\theta}}(T_{ij})]\right|\right] \\ &\leq 2\mathbb{E}\left[\sup_{g_{\bar{\theta}}\in\mathcal{G}}\left|\sum_{i\neq j} \epsilon_{ij} g_{\bar{\theta}}(T_{ij})\right|\right].\end{aligned}$$

Next, we want to apply the contraction Theorem 17. Denote $T = (T_{ij})_{i \neq j}$ and let $\mathbb{E}_T$ be the conditional expectation given T. We need the conditional expectation at this point, because Theorem 17 requires non-random arguments in the functions. This does not hinder us, as later we will simply take iterated expectations, cancelling out the conditional expectation, see below. For the functions g_i in Theorem 17 we choose

$$g_{ij}(x) = \frac{1}{2}\{-A_{ij}x + \log(1 + \exp(x))\}$$

Note, that $\log(1 + \exp(x))$ has derivative bounded by one and thus is Lipschitz continuous with constant one by the Mean Value Theorem. Thus, all g_{ij} are also Lipschitz continuous with constant 1:

$$|g_{ij}(x) - g_{ij}(x')| \leq \frac{1}{2}\{|A_{ij}(x - x')| + |\log(1 + \exp(x)) - \log(1 + \exp(x'))|\} \leq |x - x'|.$$

For the function class $\mathcal{F}$ in Theorem 17 we choose $\mathcal{F} = \mathcal{F}_M := \{f_{\bar{\theta}} : \bar{\theta} \in \bar{\Theta}_{\text{loc}}, \|\bar{\theta} - \bar{\theta}_0\|_1 \leq M\}$ and pick $f^* = f_{\bar{\theta}_0}$. Then, by Theorem 17

$$\begin{aligned}
&\mathbb{E}_T\left[\sup_{\substack{\bar{\theta} \in \bar{\Theta}_{\text{loc}}, \\ \|\bar{\theta} - \bar{\theta}_0\|_1 \leq M}} \left| \frac{1}{N} \sum_{i \neq j} \epsilon_{ij}(g_{ij}(f_{\bar{\theta}}((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T)) - g_{ij}(f_{\bar{\theta}_0}((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T))) \right| \right] \\
&\quad \leq 2\mathbb{E}_T\left[\sup_{\substack{\bar{\theta} \in \bar{\Theta}_{\text{loc}}, \\ \|\bar{\theta} - \bar{\theta}_0\|_1 \leq M}} \left| \frac{1}{N} \sum_{i \neq j} \epsilon_{ij}(f_{\bar{\theta}}((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T) - f_{\bar{\theta}_0}((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T)) \right| \right].
\end{aligned}$$

Recall that we can express the functions $f_{\bar{\theta}} = f_{\bar{\alpha}, \bar{\beta}, \mu, \gamma}$ as

$$f_{\bar{\alpha}, \bar{\beta}, \mu, \gamma}(\,.\,) = \sum_{i=1}^{n} \bar{\alpha}_i e_i(\,.\,) + \sum_{i=n+1}^{2n} \bar{\beta}_{i-n} e_i(\,.\,) + \mu e_{2n+1}(\,.\,) + \sum_{i=1}^{p} \gamma_i e_{2n+1+i}(\,.\,),$$

where $e_i(\,.\,)$ is the projection on the i-th coordinate. Consider any $\bar{\theta} \in \bar{\Theta}_{\text{loc}}$ with $\|\bar{\theta} - \bar{\theta}_0\|_1 \leq M$. For the sake of a compact representation we use our shorthand notation $\bar{\theta} = (\bar{\theta}_i)_{i=1}^{2n+1+p}$ where the components θ_i are defined in the canonical way and we also simply write $e_k(\bar{X}_{ij}, 1, Z_{ij})$ for the projection of the the vector $(\bar{X}_{ij}^T, 1, Z_{ij}^T)^T \in \mathbb{R}^{2n+p+1}$ to its k-th component, i.e. instead of $e_k((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T)$. Then,

$$\begin{aligned}
&\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}(f_{\bar{\theta}}((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T) - f_{\bar{\theta}_0}((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T))\right| \\
&= \left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}\left(\sum_{k=1}^{2n+p+1}(\bar{\theta}_k - \bar{\theta}_{0,k})e_k(\bar{X}_{ij}, 1, Z_{ij})\right)\right| \\
&\leq \frac{1}{N}\sum_{k=1}^{2n+p+1}\left\{|\bar{\theta}_k - \bar{\theta}_{0,k}| \max_{1\leq l\leq 2n+p+1}\left|\sum_{i\neq j}\epsilon_{ij}e_l(\bar{X}_{ij}, 1, Z_{ij})\right|\right\} \\
&\leq M \max_{1\leq l\leq 2n+p+1}\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}e_l(\bar{X}_{ij}, 1, Z_{ij})\right|.
\end{aligned}$$

Note, that the last expression no longer depends on $\bar{\theta}$. To bind the right hand side in the last expression we use Lemma 20: In the language of the Lemma, choose $Z_1, \dots, Z_n$ as $T_{ij} = (\epsilon_{ij}, \bar{X}_{ij}^T, 1, Z_{ij}^T)^T$. We choose for the p in the formulation of the Lemma $2n+p+1$ and pick for our functions

$$g_k(T_{ij}) = \frac{1}{N}\epsilon_{ij}e_k(\bar{X}_{ij}, 1, Z_{ij}), k = 1, \dots, 2n+p+1.$$

Note, that then $\mathbb{E}[g_k(T_{ij})] = 0$. We want to employ Lemma 20 which requires us to bound $|g_k(T_{ij})| \leq c_{ij,k}$ for all $i \neq j$ and $k = 1, \dots, n+1+p$.

For any fixed $1 \leq k \leq n$ we have

$$|g_k(T_{ij})| \leq \begin{cases} \frac{\sqrt{n}}{N} = \frac{1}{(n-1)\sqrt{n}}, & i \text{ or } j = k \\ 0, & \text{otherwise.} \end{cases}$$

Note that the first case occurs exactly $(n-1)$ times for each k. Thus, for any $k \leq 2n$,

$$\sum_{i\neq j} c_{ij,k}^2 = \left(\frac{1}{(n-1)\sqrt{n}}\right)^2 (n-1) = \frac{1}{N}.$$

If $k = 2n+1$, $|g_k(T_{ij})| = 1/N$ and hence

$$\sum_{i\neq j} c_{ij,2n+1}^2 = \frac{1}{N}.$$

Finally, if $k > 2n+1$, $|g_k(T_{ij})| \leq c/N$ and therefore,

$$\sum_{i\neq j} c_{ij,k}^2 \leq \frac{c^2}{N}.$$

In total, this means

$$\max_{1\leq k\leq 2n+1+p}\sum_{i\neq j} c_{ij,k}^2 \leq \frac{1 \vee c^2}{N}.$$

Therefore, an application of Lemma 20 results in

$$\begin{aligned}\mathbb{E}\left[\max_{1\leq l\leq 2n+p+1}\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}e_l(\bar{X}_{ij},Z_{ij})\right|\right] &\leq \sqrt{2\log(2(2n+1+p))}\max_{1\leq k\leq 2n+1+p}\left[\sum_{i\neq j}c_{ij,k}^2\right]^{1/2}\\ &\leq \sqrt{2\log(2(2n+1+p))}\sqrt{\frac{1\vee c^2}{N}}\\ &= \sqrt{\frac{2\log(2(2n+1+p))}{N}}(1\vee c).\end{aligned}$$

Putting everything together, we obtain

$$\begin{aligned}\mathbb{E}[Z_M] &\leq 2\mathbb{E}\left[\sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\mathrm{loc}},\\ \|\bar{\theta}-\bar{\theta}_0\|_1\leq M}}\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}(-A_{ij}(f_{\bar{\theta}}(\bar{X}_{ij},1,Z_{ij})-f_{\bar{\theta}_0}(\bar{X}_{ij},1,Z_{ij})))\right|\right]\\ &= 2\mathbb{E}\left[\mathbb{E}_T\left[\sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\mathrm{loc}},\\ \|\bar{\theta}-\bar{\theta}_0\|_1\leq M}}\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}(-A_{ij}(f_{\bar{\theta}}(\bar{X}_{ij},1,Z_{ij})-f_{\bar{\theta}_0}(\bar{X}_{ij},1,Z_{ij})))\right|\right]\right]\\ &\leq 8\mathbb{E}\left[\mathbb{E}_T\left[\sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\mathrm{loc}},\\ \|\bar{\theta}-\bar{\theta}_0\|_1\leq M}}\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}(f_{\bar{\theta}}(\bar{X}_{ij},1,Z_{ij})-f_{\bar{\theta}_0}(\bar{X}_{ij},1,Z_{ij}))\right|\right]\right]\\ &\leq 8M\mathbb{E}\left[\mathbb{E}_T\left[\max_{1\leq l\leq 2n+p+1}\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}e_l(\bar{X}_{ij},1,Z_{ij})\right|\right]\right]\\ &\leq 8M\sqrt{\frac{2\log(2(2n+1+p))}{N}}(1\vee c).\end{aligned}$$

This concludes the proof. ■

We now want to show that Z_M does not deviate too far from its expectation. The proof relies on the concentration theorem due to Bousquet, Theorem 18.

Corollary 22 *Pick any confidence level $t>0$. Let*

$$a_n := \sqrt{\frac{2\log(2(2n+p+1))}{N}}(1\vee c)$$

and choose $\lambda_0=\lambda_0(t,n)$ as

$$\lambda_0 = 8a_n + 2\sqrt{\frac{t}{N}(11(1\vee(c^2p))+16(1\vee c)\sqrt{n}a_n)}+\frac{4t(1\vee c)\sqrt{n}}{3N}$$

Then, we have the inequality

$$P\left(Z_M\geq M\lambda_0\right)\leq\exp(-t).$$

Proof We want to apply Bousquet's concentration Theorem 18. For the random variables Z_i in the formulation of the theorem we choose once more $T_{ij} = (A_{ij}, \bar{X}_{ij}, 1, Z_{ij}), i \neq j$, and as functions we consider

$$\begin{aligned} g_{\bar{\theta}}(T_{ij}) &= -A_{ij}\bar{D}_{ij}^T(\bar{\theta} - \bar{\theta}_0) + \log(1 + \exp(\bar{D}_{ij}^T\bar{\theta})) - \log(1 + \exp(\bar{D}_{ij}^T\bar{\theta}_0)), \\ \mathcal{G} = \mathcal{G}_M &:= \{g_{\bar{\theta}} : \bar{\theta} \in \bar{\Theta}_{\text{loc}}, \|\bar{\theta} - \bar{\theta}_0\|_1 \leq M\}. \end{aligned}$$

Then, we have

$$Z_M = \sup_{g_{\bar{\theta}} \in \mathcal{G}} \frac{1}{N} \left| \sum_{i \neq j} \{g_{\bar{\theta}}(T_{ij}) - \mathbb{E}[g_{\bar{\theta}}(T_{ij})]\} \right|.$$

To apply Theorem 18, we need to bound the infinity norm of $g_{\bar{\theta}}$. Recall that we denote the distribution of $[\bar{X}|1|Z]$ by $\bar{Q}$ and the infinity norm is defined as the $\bar{Q}$-almost sure smallest upper bound on the value of $g_{\bar{\theta}}$. We have for any $g_{\bar{\theta}} \in \mathcal{G}$, using the Lipschitz continuity of $\log(1 + \exp(x))$:

$$\begin{aligned} |g_{\bar{\theta}}(T_{ij})| &\leq |\bar{D}_{ij}^T(\bar{\theta} - \bar{\theta}_0)| + |\log(1 + \exp(\bar{D}_{ij}^T\bar{\theta})) - \log(1 + \exp(\bar{D}_{ij}^T\bar{\theta}_0))| \\ &\leq 2|\bar{D}_{ij}^T(\bar{\theta} - \bar{\theta}_0)| \\ &\leq 2\|\vartheta - \vartheta_0\|_1 + |\mu - \mu_0| + c\|\gamma - \gamma_0\|_1. \end{aligned}$$

Thus,

$$\begin{aligned} \|g_{\bar{\theta}}\|_\infty &\leq 2\|\vartheta - \vartheta_0\|_1 + |\mu - \mu_0| + c\|\gamma - \gamma_0\|_1 \\ &\leq 2(1 \vee c)\|\theta - \theta_0\|_1 \\ &\leq 2(1 \vee c)\sqrt{n}M =: \eta_n. \end{aligned}$$

For the last inequality we used that for any θ with $\|\bar{\theta} - \bar{\theta}_0\|_1 \leq M$ it follows that $\|\theta - \theta_0\|_1 \leq \sqrt{n}M$, which is possibly a very generous upper bound. This does not matter, however, as the term associated with the above bound will be negligible, as we shall see.

The second requirement of Theorem 18 is that the average variance of $g_{\bar{\theta}}(T_{ij})$ has to be uniformly bounded. To that end we calculate

$$\begin{aligned} \frac{1}{N}\sum_{i \neq j} \text{Var}(g_{\bar{\theta}}(T_{ij})) &= \frac{1}{N}\sum_{i \neq j} \text{Var}(-A_{ij}D_{ij}^T(\theta - \theta_0)) \\ &+ \frac{1}{N}\sum_{i \neq j} \text{Var}(\log(1 + \exp(\bar{D}_{ij}^T\bar{\theta})) - \log(1 + \exp(\bar{D}_{ij}^T\bar{\theta}_0))) \\ &+ \frac{2}{N}\sum_{i \neq j} \text{Cov}(-A_{ij}D_{ij}^T(\theta - \theta_0), \log(1 + \exp(\bar{D}_{ij}^T\bar{\theta})) - \log(1 + \exp(\bar{D}_{ij}^T\bar{\theta}_0))). \end{aligned}$$

Let us look at these terms in term. For the first term, we obtain

$$\frac{1}{N}\sum_{i \neq j} \text{Var}(-A_{ij}D_{ij}^T(\theta - \theta_0)) \leq \frac{1}{N}\sum_{i \neq j} \mathbb{E}[(-A_{ij}D_{ij}^T(\theta - \theta_0))^2] \leq \mathbb{E}\left[\frac{1}{N}\sum_{i \neq j}(D_{ij}^T(\theta - \theta_0))^2\right].$$

For the second term we get

$$\begin{aligned}
&\frac{1}{N}\sum_{i\neq j}\mathrm{Var}(\,\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0)))\\
&\qquad\qquad \leq \frac{1}{N}\sum_{i\neq j}\mathbb{E}[(\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0)))^2]\\
&\qquad\qquad \leq \mathbb{E}\left[\frac{1}{N}\sum_{i\neq j}(D_{ij}^T(\theta-\theta_0))^2\right].
\end{aligned}$$

The last term decomposes as

$$\begin{aligned}
&\frac{2}{N}\sum_{i\neq j}\mathrm{Cov}(-A_{ij}D_{ij}^T(\theta-\theta_0),\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0)))\\
&\qquad = \frac{2}{N}\sum_{i\neq j}\mathbb{E}[-A_{ij}D_{ij}^T(\theta-\theta_0)\cdot(\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0)))]\\
&\qquad\quad - \frac{2}{N}\sum_{i\neq j}\mathbb{E}[-A_{ij}D_{ij}^T(\theta-\theta_0)]\cdot\mathbb{E}[\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0))]
\end{aligned}$$

For the first term in that decomposition we have

$$\begin{aligned}
&\frac{2}{N}\sum_{i\neq j}\left|\mathbb{E}[-A_{ij}D_{ij}^T(\theta-\theta_0)\cdot(\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0)))]\right|\\
&\qquad \leq \frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|\cdot|\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0))|]\\
&\qquad \leq \frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|^2]
\end{aligned}$$

and for the second term using the same arguments, we get

$$\frac{2}{N}\sum_{i\neq j}\mathbb{E}[-A_{ij}D_{ij}^T(\theta-\theta_0)]\cdot\mathbb{E}[\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0))] \leq \frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|]^2.$$

Meaning that in total

$$\begin{aligned}
&\frac{2}{N}\sum_{i\neq j}\left|\mathrm{Cov}(-A_{ij}D_{ij}^T(\theta-\theta_0),\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0)))\right|\\
&\qquad\quad \leq \frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|^2]+\frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|]^2.
\end{aligned}$$

In total, we thus get

$$\frac{1}{N}\sum_{i\neq j}\mathrm{Var}(g_{\bar{\theta}}(T_{ij}))\leq 4\cdot\mathbb{E}\left[\frac{1}{N}\sum_{i\neq j}(D_{ij}^T(\theta-\theta_0))^2\right]+\frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|]^2. \tag{23}$$

Furthermore,

$$\begin{aligned}\frac{1}{N}\sum_{i\neq j}(D_{ij}^T(\theta-\theta_0))^2 &= \frac{1}{N}\sum_{i\neq j}(\alpha_i+\beta_j+\mu-\alpha_{0,i}-\beta_{0,j}-\mu_0+(\gamma-\gamma_0)^T Z_{ij})^2\\ &\overset{\text{Cauchy-Schwarz}}{\leq} \frac{4}{N}\sum_{i\neq j}\left\{(\alpha_i-\alpha_{0,i})^2+(\beta_j-\beta_{0,j})^2+(\mu-\mu_0)^2+((\gamma-\gamma_0)^T Z_{ij})^2\right\}.\end{aligned}$$

Recall that for any $x\in\mathbb{R}^p$, $\|x\|_2\leq\|x\|_1\leq\sqrt{p}\|x\|_2$ and note that

$$|(\gamma-\gamma_0)^T Z_{ij}|\leq c\|\gamma-\gamma_0\|_1\leq c\sqrt{p}\|\gamma-\gamma_0\|_2.$$

Then, from the above

$$\begin{aligned}\frac{1}{N}\sum_{i\neq j}(D_{ij}^T(\theta-\theta_0))^2 &\leq \frac{4}{N}\sum_{i\neq j}\left\{(\alpha_i-\alpha_{0,i})^2+(\beta_j-\beta_{0,j})^2+(\mu-\mu_0)^2+c^2p\|\gamma-\gamma_0\|_2^2\right\}\\ &= 4\left((\mu-\mu_0)^2+c^2p\|\gamma-\gamma_0\|_2^2+\frac{1}{N}\sum_{i\neq j}\left\{(\alpha_i-\alpha_{0,i})^2+(\beta_j-\beta_{0,j})^2\right\}\right)\\ &= 4\left((\mu-\mu_0)^2+c^2p\|\gamma-\gamma_0\|_2^2+\frac{1}{N}(n-1)\|\vartheta-\vartheta_0\|_2^2\right)\\ &= 4\left((\mu-\mu_0)^2+c^2p\|\gamma-\gamma_0\|_2^2+\left\|\frac{1}{\sqrt{n}}(\vartheta-\vartheta_0)\right\|_2^2\right)\\ &= 4\left((\mu-\mu_0)^2+c^2p\|\gamma-\gamma_0\|_2^2+\|\bar{\vartheta}-\bar{\vartheta}_0\|_2^2\right)\\ &\leq 4(1\vee(c^2p))\|\bar{\theta}-\bar{\theta}_0\|_2^2\\ &\leq 4(1\vee(c^2p))\|\bar{\theta}-\bar{\theta}_0\|_1^2\\ &\leq 4(1\vee(c^2p))M^2.\end{aligned} \tag{24}$$

Notice that for the second summand on the right-hand side in (23), we have

$$\begin{aligned}\frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|]^2 &= \frac{2}{N}\sum_{i\neq j}(\alpha_i+\beta_j+\mu-\alpha_{0,i}-\beta_{0,j}-\mu_0+(\gamma-\gamma_0)^T\mathbb{E}[Z_{ij}])^2\\ &= \frac{2}{N}\sum_{i\neq j}(\alpha_i+\beta_j+\mu-\alpha_{0,i}-\beta_{0,j}-\mu_0)^2.\end{aligned}$$

So that we may use the same steps as in (24) to conclude that

$$\frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|]^2\leq 6(1\vee(c^2p))M^2.$$

Such that in total,

$$\frac{1}{N}\sum_{i\neq j}\mathrm{Var}(g_{\bar{\theta}}(T_{ij}))\leq 22(1\vee(c^2p))M^2:=\tau_n^2.$$

Applying Bousquet's concentration Theorem 18 with η_n, τ_n defined above, we obtain for all $z > 0$

$$\begin{aligned}\exp\left(-Nz^2\right) &\geq P\left(Z_M \geq \mathbb{E}[Z_M] + z\sqrt{2(\tau_n^2 + 2\eta_n \mathbb{E}[Z_M])} + \frac{2z^2\eta_n}{3}\right)\\ &= P\left(Z_M \geq \mathbb{E}[Z_M] + z\sqrt{2(22(1 \vee (c^2p))M^2 + 4(1\vee c)\sqrt{n}M\mathbb{E}[Z_M])} + \frac{4z^2(1\vee c)\sqrt{n}M}{3}\right). \end{aligned} \tag{25}$$

From Lemma 21, we know

$$\mathbb{E}[Z_M] \leq 8M\sqrt{\frac{2\log(2(2n+p+1))}{N}}(1\vee c) = 8Ma_n.$$

Using this, we obtain from (25)

$$\begin{aligned}\exp\left(-Nz^2\right) &\geq P\left(Z_M \geq 8Ma_n + z\sqrt{2(22(1 \vee (c^2p))M^2 + 32(1\vee c)\sqrt{n}M^2a_n)} + \frac{4z^2(1\vee c)\sqrt{n}M}{3}\right)\\ &= P\left(Z_M \geq M\left(8a_n + 2z\sqrt{11(1\vee(c^2p)) + 16(1\vee c)\sqrt{n}a_n} + \frac{4z^2(1\vee c)\sqrt{n}}{3}\right)\right).\end{aligned}$$

Now, pick $z = \sqrt{t/N}$ to get

$$P\left(Z_M \geq M\left(8a_n + 2\sqrt{\frac{t}{N}(11(1\vee(c^2p)) + 16(1\vee c)\sqrt{n}a_n)} + \frac{4t(1\vee c)\sqrt{n}}{3N}\right)\right) \leq \exp(-t),$$

which is the claim. ■

A.8 Putting it all together

Proof [Proof of Theorem 5] Theorem 5 now follows from Theorem 14 and corollary 22. Recall the definition of K_n in (16), which simplifies to

$$K_n = 2\frac{(1+\exp(r_{n,0}))^2}{\exp(r_{n,0})} = 2\frac{\left(1+\exp\left(-\text{logit}(\rho_{n,0})\right)\right)^2}{\exp\left(-\text{logit}(\rho_{n,0})\right)} = \frac{2}{\rho_{n,0}}.$$

Thus, under the conditions of Theorem 5, we have with high probability by Theorem 14 and Corollary 22,

$$\mathcal{E}(\hat{\theta}) + \bar{\lambda}\left(\frac{1}{\sqrt{n}}\|\hat{\vartheta} - \vartheta_0\|_1 + |\hat{\mu} - \mu_0| + \|\hat{\gamma} - \gamma_0\|_1\right) \leq C\frac{s_{0,+}\bar{\lambda}^2}{\rho_{n,0}}.$$

with constant $C = 64/c_{\min}$. ■

Appendix B. Proof of Theorem 6

B.1 Inverting population and sample Gram matrices

Note that the function $f(x) = x(1-x)$ is monotonically increasing in x for $x \leq 1/2$ and monotonically decreasing in x for $x \geq 1/2$. Thus, by considering the cases $p_{ij} \leq 1/2$ and $p_{ij} \geq 1/2$ separately and using that $\rho_n \leq 1/2$, we may employ the following lower bound for all $i \neq j$: $p_{ij}(\theta_0)(1-p_{ij}(\theta_0)) \geq 1/2\rho_n$. Also, recall that by assumption 1, the minimum eigenvalue $\lambda_{\min}$ of $\mathbb{E}[Z^TZ/N]$ stays uniformly bounded away from zero for all n. Then, for any n and $v \in \mathbb{R}^{p+1}\backslash\{0\}$ with components $v = (v_1, v_R^T)^T, v_R \in \mathbb{R}^p$, we have

$$\begin{aligned} v^T\Sigma_\xi v &\geq \frac{1}{2}\rho_n v^T \frac{1}{N}\mathbb{E}[D_\xi^T D_\xi]v = \frac{1}{2}\rho_n v^T \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \frac{1}{N}\mathbb{E}[Z^TZ] \end{pmatrix} v \\ &= \frac{1}{2}\rho_n \left(v_1^2 + v_R^T \frac{1}{N}\mathbb{E}[Z^TZ]v_R \right) \\ &\geq \frac{1}{2}\rho_n (v_1^2 + \lambda_{\min}\|v_R\|_2^2) \geq \frac{1}{2}\rho_n(1 \wedge c_{\min})\|v\|_2^2 > 0. \end{aligned}$$

Hence, for finite n all eigenvalues of Σ_ξ are strictly positive and consequently this matrix is invertible. We now want to show that the same hold with high probability for the sample matrix $\hat{\Sigma}_\xi$. Using the tools deployed in the proofs of Lemma 7 and 8 we can now show that with high probability the minimum eigenvalue of $D_\xi^T D_\xi/N$ is also strictly larger than zero, which means that $D_\xi^T D_\xi/N$ is invertible with high probability, from which the desired properties of $\hat{\Sigma}_\xi$ follow. More precisely, recall the definition of $\kappa(A, m)$ for square matrices A and dimensions m. We want to consider the expression $\kappa^2\left(\frac{1}{N}\mathbb{E}[D_\xi^T D_\xi], p+1\right)$ which simplifies to

$$\kappa^2\left(\frac{1}{N}\mathbb{E}[D_\xi^T D_\xi], p+1\right) := \min_{v \in \mathbb{R}^{p+1}\backslash\{0\}} \frac{v^T \frac{1}{N}\mathbb{E}[D_\xi^T D_\xi]v}{\frac{1}{p+1}\|v\|_1^2}$$

and compare it to $\kappa^2\left(\frac{1}{N}D_\xi^T D_\xi, p+1\right)$. By assumption 1 and the argument above, we have

$$\kappa^2\left(\frac{1}{N}\mathbb{E}[D_\xi^T D_\xi], p+1\right) \geq C > 0$$

for a universal constant C independent of n. With $\delta = \max_{kl}\left|\left(\frac{1}{N}D_\xi^T D_\xi\right)_{kl} - \left(\frac{1}{N}\mathbb{E}[D_\xi^T D_\xi]\right)_{kl}\right|$, by Lemma 7, we have

$$\kappa^2\left(\frac{1}{N}D_\xi^T D_\xi, p+1\right) \geq \kappa^2\left(\frac{1}{N}\mathbb{E}[D_\xi^T D_\xi], p+1\right) - 16\delta(p+1).$$

By looking at the proof of Lemma 7, we see that in this particular case we do not even need the factor $16(p+1)$ on the right hand side above, but this does not matter anyways, so we keep it. By the exact same arguments we have used in the proof of Lemma 8 for the blocks ⑤, ⑥, ⑧ and ⑨, we now get

$$\delta = O_P\left(N^{-1/2}\right).$$

Thus, for n large enough, we have with high probability $\delta \leq \frac{\lambda_{\min}}{32}$. Then, by Lemma 7, with high probability and uniformly in n,

$$\kappa^2\left(\frac{1}{N}D_\xi^T D_\xi, p+1\right) \geq \kappa^2\left(\frac{1}{N}\mathbb{E}[D_\xi^T D_\xi], p+1\right) - 16\delta(p+1) \geq \frac{\lambda_{\min}(p+1)}{2} \geq C > 0.$$

Yet, if $\kappa^2\left(\frac{1}{N}D_\xi^T D_\xi, p+1\right) \geq C > 0$ uniformly in n, then for any $v \neq 0, v^T\frac{1}{N}D_\xi^T D_\xi v \geq C\|v\|_2^2$. But we also know that the minimum eigenvalue of $\frac{1}{N}D_\xi^T D_\xi$ is the largest possible C such that this bound holds (it is actually tight with equality for the eigenvectors corresponding to the minimum eigenvalue). Therefore, with high probability, the minimum eigenvalue of $\frac{1}{N}D_\xi^T D_\xi$ stays uniformly bounded away from zero. Thus, for any $v \in \mathbb{R}^{p+1}\backslash\{0\}$ and any finite n:

$$\frac{1}{N}v^T D_\xi^T \hat{W}^2 D_\xi v \geq \min_{i\neq j}\{p_{ij}(\hat{\theta})(1-p_{ij}(\hat{\theta}))\}\left(v^T\frac{1}{N}D_\xi^T D_\xi v\right) \geq C\rho_n\|v\|_2^2 > 0.$$

Thus, $\text{mineval}\left(\frac{1}{N}D_\xi^T \hat{W}^2 D_\xi\right) \geq C\rho_n \text{mineval}\left(\frac{1}{N}D_\xi^T D_\xi\right) > 0$. That is, for every finite n, $\frac{1}{N}D_\xi^T \hat{W}^2 D_\xi$ is invertible with high probability.

B.2 Goal and approach

Our strategy will be inverting the KKT conditions, similar to van de Geer et al. (2014). Recall our discussion of the KKT conditions in Section 3.1. By the same arguments, we find that 0 has to be contained in the subdifferential of $\frac{1}{N}\mathcal{L}(\theta) + \lambda\|\beta\|_1$ at $\hat{\theta}$, where this time we consider the KKT conditions with respect to the original parameters θ. That is, there exists a $\hat{z} \in \mathbb{R}^{2n+1+p}$ such that

$$0 = \frac{1}{N}\nabla\ \mathcal{L}(\theta)|_{\theta=\hat{\theta}} + \lambda\hat{z},$$

where $\nabla\ \mathcal{L}(\theta)|_{\theta=\hat{\theta}}$ is the gradient of $\mathcal{L}(\theta)$ evaluated at $\hat{\theta}$ and for $i = 1, \ldots, 2n, \hat{z}_i = 1$ if $\hat{\vartheta}_i > 0$ and $\hat{z}_i \in [-1, 1]$ if $\hat{\vartheta}_i = 0$, and for $i = 2n+1, \ldots, 2n+1+p, \hat{z}_i = 0$.

Denoting $\nabla_\xi\ \mathcal{L}(\theta)|_{\theta=\hat{\theta}} \in \mathbb{R}^{p+1}$ the gradient of $\mathcal{L}$ with respect to the unpenalized parameters $\xi = (\mu, \gamma^T)^T$ only, evaluated at $\hat{\theta}$, we have

$$0 = \nabla_\xi\ \mathcal{L}(\theta)|_{\theta=\hat{\theta}}. \tag{26}$$

Goal: We want to show that for $k = 1, \ldots, p+1$,

$$\sqrt{N}\frac{\hat{\xi}_k - \xi_{0,k}}{\sqrt{\hat{\Theta}_{\xi,k,k}}} \to \mathcal{N}(0,1).$$

Approach: Recall the definition of the "one-sample-version" of $\mathcal{L}$, i.e. $l_\theta : \{0,1\} \times \mathbb{R}^{2n+1+p} \to \mathbb{R}$, for $\theta = (\alpha^T, \beta^T, \mu, \gamma^T)^T \in \Theta$,

$$l_\theta(y, x) := -y\theta^T x + \log(1 + \exp(\theta^T x)).$$

Then, the negative log-likelihood is given by

$$\mathcal{L}(\theta) = \sum_{i \neq j} l_\theta(A_{ij}, D_{ij}^T)$$

and

$$\nabla \mathcal{L}(\theta) = \sum_{i \neq j} \nabla l_\theta(A_{ij}, D_{ij}^T), \quad H\mathcal{L}(\theta) = \sum_{i \neq j} Hl_\theta(A_{ij}, D_{ij}^T),$$

where H denotes the Hessian with respect to θ. Consider l_θ as a function in $\theta^T x$ and introduce:

$$l(y, a) := -ya + \log(1 + \exp(a)), \tag{27}$$

with second derivative: $\ddot{l}(y, a) = \partial_{a^2} l(y, a) = \frac{\exp(a)}{(1+\exp(a))^2}$. Note, that $\partial_{a^2} l(y, a)$ is Lipschitz continuous (it has bounded derivative $|\partial_{a^3} l(y, a)| \leq 1/(6\sqrt{3})$; Lipschitz continuity then follows by the Mean Value Theorem). Doing a first order Taylor expansion in a of $\dot{l}(y, a) = \partial_a l(y, a)$ in the point $(A_{ij}, D_{ij}^T \theta_0)$ evaluated at $(A_{ij}, D_{ij}^T \hat{\theta})$, we get

$$\partial_a l(A_{ij}, D_{ij} \hat{\theta}) = \partial_a l(A_{ij}, D_{ij}^T \theta_0) + \partial_{a^2} l(A_{ij}, \alpha) D_{ij}^T (\hat{\theta} - \theta_0), \tag{28}$$

for an α between $D_{ij}^T \hat{\theta}$ and $D_{ij}^T \theta_0$. By Lipschitz continuity of $\partial_{a^2} l$, we also find

$$\begin{aligned} |\partial_{a^2} l(A_{ij}, \alpha) D_{ij}^T (\hat{\theta} - \theta_0) - \partial_{a^2} l(A_{ij}, D_{ij}^T \hat{\theta}) D_{ij}^T (\hat{\theta} - \theta_0)| &\leq |\alpha - D_{ij}^T \hat{\theta}| |D_{ij}^T (\hat{\theta} - \theta_0)| \\ &\leq |D_{ij}^T (\hat{\theta} - \theta_0)|^2, \end{aligned} \tag{29}$$

where the last inequality follows, because α is between $D_{ij}^T \hat{\theta}$ and $D_{ij}^T \theta_0$.

Consider the vector $P_n \nabla l_{\hat{\theta}}$: By equation (28), with α_{ij} between $D_{ij}^T \hat{\theta}$ and $D_{ij}^T \theta_0$,

$$\begin{aligned} P_n \nabla l_{\hat{\theta}} &= \frac{1}{N} \sum_{i \neq j} \left(\partial_{\theta_k} l(A_{ij}, D_{ij}^T \hat{\theta}) \right)_{k=1,\ldots,2n+1+p}, \quad \text{as a } (2n+1+p) \times 1\text{-vector} \\ &= \frac{1}{N} \sum_{i \neq j} \dot{l}(A_{ij}, D_{ij}^T \hat{\theta}) D_{ij} \\ &= \frac{1}{N} \sum_{i \neq j} (\dot{l}(A_{ij}, D_{ij}^T \theta_0) + \ddot{l}(A_{ij}, \alpha_{ij}) D_{ij}^T (\hat{\theta} - \theta_0)) D_{ij} \end{aligned}$$

which by (29) gives

$$= P_n \nabla l_{\theta_0} + \frac{1}{N} \sum_{i \neq j} D_{ij} \left\{ \ddot{l}(A_{ij}, D_{ij}^T \hat{\theta}) D_{ij}^T (\hat{\theta} - \theta_0) + O(|D_{ij}^T (\hat{\theta} - \theta_0)|^2) \right\}.$$

Noticing that $\ddot{l}(A_{ij}, D_{ij}^T \hat{\theta}) = p_{ij}(\hat{\theta})(1 - p_{ij}(\hat{\theta}))$ and thus $\sum_{i \neq j} \ddot{l}(A_{ij}, D_{ij}^T \hat{\theta}) D_{ij} D_{ij}^T (\hat{\theta} - \theta_0) = D^T \hat{W}^2 D(\hat{\theta} - \theta_0)$:

$$\begin{aligned} &= P_n \nabla l_{\theta_0} + P_n H l_{\hat{\theta}} (\hat{\theta} - \theta_0) + O\left(\frac{1}{N} \sum_{i \neq j} D_{ij} |D_{ij}^T (\hat{\theta} - \theta_0)|^2 \right) \\ &= P_n \nabla l_{\theta_0} + \frac{1}{N} D^T \hat{W}^2 D(\hat{\theta} - \theta_0) + O\left(\frac{1}{N} \sum_{i \neq j} D_{ij} |D_{ij}^T (\hat{\theta} - \theta_0)|^2 \right), \end{aligned}$$

where the O notation is to be understood componentwise. Above, we have equality of two $((2n+1+p)\times 1)$-vectors. We are only interested in the portion relating to $\xi = (\mu, \gamma^T)^T$, that is, in the last $p+1$ entries. Introduce the $((2n+1+p)\times(2n+1+p))$-matrix

$$A = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \hat{\Theta}_\xi \end{pmatrix},$$

where $\mathbf{0}$ are zero-matrices of appropriate dimensions. Multiplying the above with A on both sides gives:

$$AP_n \nabla l_{\hat{\theta}} = AP_n \nabla l_{\theta_0} + A\frac{1}{N} D^T \hat{W}^2 D(\hat{\theta}-\theta_0) + AO\left(\frac{1}{N}\sum_{i\neq j} D_{ij} |D_{ij}^T(\hat{\theta}-\theta_0)|^2\right). \tag{30}$$

Let us consider these terms in turn: Multiplication by A means that the first n entries of any of the vectors above are zero. Hence we only need to consider the last $p+1$ entries. The left-hand side of (30) is equal to zero by (26). The last $p+1$ entries of the first term on the right-hand side are $\hat{\Theta}_\xi P_n \nabla_\xi l_{\theta_0}$. For the second term on the right hand side, notice that

$$\frac{1}{N} D^T \hat{W}^2 D = \frac{1}{N}\begin{bmatrix} X^T\hat{W}^2X & X^T\hat{W}^2\mathbf{1} & X^T\hat{W}^2Z \\ \mathbf{1}^T\hat{W}^2X & \mathbf{1}^T\hat{W}^2\mathbf{1} & \mathbf{1}^T\hat{W}^2Z \\ Z^T\hat{W}^2X & Z^T\hat{W}^2\mathbf{1} & Z^T\hat{W}^2Z \end{bmatrix}.$$

$\hat{\Theta}_\xi$ is the exact inverse of $\hat{\Sigma}_\xi$ which is the lower-right $(p+1)\times(p+1)$ block of above matrix. Thus,

$$A\frac{1}{N} D^T \hat{W}^2 D = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \hat{\Theta}_\xi \frac{1}{N} D_\xi^T \hat{W}^2 X & I_{(p+1)\times(p+1)} \end{bmatrix}.$$

Then, for the last $p+1$ entries of $A\frac{1}{N}D^T\hat{W}^2D(\hat{\theta}-\theta_0)$

$$\left(A\frac{1}{N}D^T\hat{W}^2D(\hat{\theta}-\theta_0)\right)_{\text{last } p+1 \text{ entries}} = \hat{\Theta}_\xi \frac{1}{N} D_\xi^T \hat{W}^2 X(\hat{\vartheta}-\vartheta_0) + \begin{pmatrix} \hat{\mu}-\mu_0 \\ \hat{\gamma}-\gamma_0 \end{pmatrix}.$$

Thus, (30) implies

$$0 = \hat{\Theta}_\xi P_n \nabla_\gamma l_{\theta_0} + \hat{\Theta}_\xi \frac{1}{N} D_\xi^T \hat{W}^2 X(\hat{\vartheta}-\vartheta_0) + \begin{pmatrix} \hat{\mu}-\mu_0 \\ \hat{\gamma}-\gamma_0 \end{pmatrix} + O\left(\hat{\Theta}_\xi \frac{1}{N}\sum_{i\neq j}\begin{pmatrix} 1 \\ Z_{ij}\end{pmatrix} |D_{ij}^T(\hat{\theta}-\theta_0)|^2\right),$$

which is equivalent to

$$\begin{pmatrix} \hat{\mu}-\mu_0 \\ \hat{\gamma}-\gamma_0 \end{pmatrix} = -\hat{\Theta}_\xi P_n \nabla_\xi l_{\theta_0} - \hat{\Theta}_\xi \frac{1}{N} D_\xi^T \hat{W}^2 X(\hat{\vartheta}-\vartheta_0) + O\left(\hat{\Theta}_\xi \frac{1}{N}\sum_{i\neq j}\begin{pmatrix} 1 \\ Z_{ij}\end{pmatrix} |D_{ij}^T(\hat{\theta}-\theta_0)|^2\right). \tag{31}$$

Our goal is now to show that for each component $k = 1, \dots, p+1$,

$$\sqrt{N}\frac{\hat{\xi}_k - \xi_{0,k}}{\sqrt{\hat{\Theta}_{\xi,k,k}}} \xrightarrow{d} \mathcal{N}(0,1).$$

as described in the **Goal** section. To that end, by equation (31), we now need to solve the following three problems: Writing $\hat{\Theta}_{\xi,k}$ for the k-th row of $\hat{\Theta}_\xi$,

1. $\sqrt{N}\frac{\hat{\Theta}_{\xi,k}P_n\nabla_\xi l_{\theta_0}}{\sqrt{\hat{\Theta}_{\xi,k,k}}} \xrightarrow{d} \mathcal{N}(0,1)$,

2. $\frac{1}{\sqrt{\hat{\Theta}_{\xi,k,k}}}\hat{\Theta}_{\xi,k}\frac{1}{N}D_\xi^T\hat{W}^2X(\hat{\vartheta}-\vartheta_0) = o_P\left(N^{-1/2}\right)$,

3. $O\left(\frac{1}{\sqrt{\hat{\Theta}_{\xi,k,k}}}\hat{\Theta}_{\xi,k}\frac{1}{N}\sum_{i\neq j}\begin{pmatrix}1\\Z_{ij}\end{pmatrix}|D_{ij}^T(\hat{\theta}-\theta_0)|^2\right) = o_P\left(N^{-1/2}\right)$.

B.3 Bounding inverses

The problems (1) - (3) above suggest that it will be essential to bound the norm and the distance of $\hat{\Theta}_\xi$ and Θ_ξ in an appropriate manner. Notice that for any invertible matrices $A, B \in \mathbb{R}^{m\times m}$ we have

$$A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}.$$

Thus, for any sub-multiplicative matrix norm $\| \,.\, \|$, we get

$$\|A^{-1} - B^{-1}\| \leq \|A^{-1}\|\|B^{-1}\|\|B-A\|. \tag{32}$$

We are particularly interested in the matrix ∞-norm, defined as

$$\|A\|_\infty := \sup\left\{\frac{\|Ax\|_\infty}{\|x\|_\infty}, x\neq 0\right\} = \sup\left\{\|Ax\|_\infty, \|x\|_\infty = 1\right\} = \max_{1\leq i\leq m}\sum_{j=1}^m |A_{i,j}|,$$

i.e. $\|A\|_\infty$ is the maximal row ℓ_1-norm of A. It is well-known, that any such matrix norm induced by a vector norm is sub-multiplicative ($\|AB\|_\infty \leq \|A\|_\infty\|B\|_\infty$) and consistent with the inducing vector norm ($\|Ax\|_\infty \leq \|A\|_\infty\|x\|_\infty$ for any vector x of appropriate dimension). We first want to bound the matrix ∞-norm in terms of the largest eigenvalue.

Lemma 23 *For any symmetric, positive semi-definite $(m\times m)$-matrix A with maximal eigenvalue $\lambda > 0$, we have $\|A\|_\infty \leq \sqrt{m}\lambda$.*

Proof

$$\begin{aligned}\|A\|_\infty &= \sup\left\{\|Ax\|_\infty, \|x\|_\infty = 1\right\}\\ &\leq \sup\left\{\|Ax\|_2, \|x\|_\infty = 1\right\}, \quad \|Ax\|_\infty \leq \|Ax\|_2\\ &= \sup\left\{\frac{\|Ax\|_2}{\|x\|_2}\|x\|_2, \|x\|_\infty = 1\right\}\\ &\leq \sqrt{m}\sup\left\{\frac{\|Ax\|_2}{\|x\|_2}, \|x\|_\infty = 1\right\}, \qquad \text{if } \|x\|_\infty = 1, \text{ then } \|x\|_2 \leq \sqrt{m},\\ &\leq \sqrt{m}\sup\left\{\frac{\|Ax\|_2}{\|x\|_2}, x\neq 0\right\}\\ &= \sqrt{m}\|A\|_2 = \sqrt{m}\lambda,\end{aligned}$$

where $\|A\|_2$ is the spectral norm of the matrix A and we have used that for symmetric matrices, the spectral norm is equal to the modulus of the largest eigenvalue of A. ■

Also, recall that the inverse of a symmetric matrix A is itself symmetric:

$$I = AA^{-1} = A^T A^{-1} \overset{\text{transpose}}{\Longrightarrow} I = (A^{-1})^T A^T \overset{\text{symmetry}}{=} (A^{-1})^T A \overset{\text{uniqueness of inverse}}{\Longrightarrow} (A^{-1})^T = A^{-1}.$$

Hence, $\hat{\Theta}_\xi$ and Θ_ξ are symmetric and we may apply Lemma 23. Using that $\lambda_{\max}(\Sigma_\xi^{-1}) = \frac{1}{\lambda_{\min}(\Sigma_\xi)}$, we get

$$\|\Theta_\xi\|_\infty \leq \sqrt{p}\lambda_{\max}(\Sigma_\xi^{-1}) \leq C\frac{1}{\rho_n},$$

and with high probability

$$\|\hat{\Theta}_\xi\|_\infty \leq \sqrt{p}\lambda_{\max}(\hat{\Sigma}_\xi^{-1}) \leq C\frac{1}{\rho_n},$$

with some absolute constant C. Finally, by (32),

$$\|\hat{\Theta}_\xi - \Theta_\xi\|_\infty \leq \|\hat{\Theta}_\xi\|_\infty \|\Theta_\xi\|_\infty \|\hat{\Sigma}_\xi - \Sigma_\xi\|_\infty \leq \frac{C}{\rho_n^2}\|\hat{\Sigma}_\xi - \Sigma_\xi\|_\infty.$$

It remains to control $\|\hat{\Sigma}_\xi - \Sigma_\xi\|_\infty$. We have

$$\begin{aligned}\hat{\Sigma}_\xi - \Sigma_\xi &= \frac{1}{N}\left(D_\xi^T \hat{W}^2 D_\xi - \mathbb{E}[D_\xi^T W_0^2 D_\xi]\right)\\ &= \underbrace{\frac{1}{N}\left(D_\xi^T(\hat{W}^2 - W_0^2)D_\xi\right)}_{(I)} + \underbrace{\frac{1}{N}\left(D_\xi^T W_0^2 D_\xi - \mathbb{E}[D_\xi^T W_0^2 D_\xi]\right)}_{(II)}.\end{aligned}$$

Recall that $\hat{w}_{ij}^2 = p_{ij}(\hat{\theta})(1 - p_{ij}(\hat{\theta})) = \frac{\exp(D_{ij}^T\hat{\theta})}{(1+\exp(D_{ij}^T\hat{\theta}))^2} = \partial_{a^2} l(A_{ij}, D_{ij}^T\hat{\theta})$, with the function l defined in (27). Also recall that $\partial_{a^2} l$ is Lipschitz with constant one, by the Mean Value Theorem and the fact that it has derivative $\partial_{a^3} l$ bounded by one. Thus, considering the

(k, l)-th element of (I) above, we get:

$$\begin{aligned}
\left|\frac{1}{N}\left(D_\xi^T(\hat{W}^2 - W_0^2)D_\xi\right)_{kl}\right| &= \left|\frac{1}{N}\sum_{i\neq j} D_{ij,n+k}D_{ij,n+l}(\hat{w}_{ij}^2 - w_{0,ij}^2)\right| \\
&\leq C\frac{1}{N}\sum_{i\neq j}|\hat{w}_{ij}^2 - w_{0,ij}^2|, \quad \text{by unifrom boundedness of } Z_{ij} \\
&\leq C\frac{1}{N}\sum_{i\neq j}|D_{ij}^T(\hat{\theta} - \theta_0)|, \quad \text{by Lipschitz continuity} \\
&\leq \frac{C}{N}\sum_{i\neq j}\left\{|\hat{\alpha}_i - \alpha_{0,i}| + |\hat{\beta}_j - \beta_{0,j}| + |\hat{\mu} - \mu_0| + |Z_{ij}^T(\hat{\gamma} - \gamma_0)|\right\} \\
&\leq \underbrace{\frac{C}{N}\left\{\sum_{i\neq j}|\hat{\alpha}_i - \alpha_{0,i}| + |\hat{\beta}_j - \beta_{0,j}|\right\}}_{=(n-1)\|\hat{\vartheta}-\vartheta_0\|_1} + C|\hat{\mu} - \mu_0| + C\|\hat{\gamma} - \gamma_0\|_1 \\
&\leq C\left\{\frac{1}{n}\|\hat{\vartheta} - \vartheta_0\|_1 + |\hat{\mu} - \mu_0| + \|\hat{\gamma} - \gamma_0\|_1\right\} \\
&= O_P\left(s_+^*\sqrt{\frac{\log(n)}{N}}\rho_n^{-1}\right), \text{ under the conditions of theorem 5.}
\end{aligned}$$

Since the dimension of (I) is $(p+1)\times(p+1)$ and thus remains fixed, any row of (I) has ℓ_1 norm of order $O_P\left(s_+^*\sqrt{\frac{\log(n)}{N}}\rho_n^{-1}\right)$ and thus

$$\|(I)\|_\infty = O_P\left(s_+^*\sqrt{\frac{\log(n)}{N}}\rho_n^{-1}\right).$$

Taking a look at the (k, l)-th element in (II):

$$\left|\frac{1}{N}\left(D_\xi^T W_0^2 D_\xi - \mathbb{E}[D_\xi^T W_0^2 D_\xi]\right)_{kl}\right| = \left|\frac{1}{N}\sum_{i\neq j}\left\{D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2 - \mathbb{E}[D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2]\right\}\right|.$$

Note that the random variables $D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2$ are bounded uniformly in i, j, k, l. Thus, by Hoeffding's inequality, for any $t \geq 0$,

$$P\left(\left|\frac{1}{N}\sum_{i\neq j}\left\{D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2 - \mathbb{E}[D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2]\right\}\right| \geq t\right) \leq 2\exp\left(-CNt^2\right).$$

This means, $\left|\frac{1}{N}\left(D_\xi^T W_0^2 D_\xi - \mathbb{E}[D_\xi^T W_0^2 D_\xi]\right)_{kl}\right| = O_P\left(N^{-1/2}\right)$. Again, since the dimension $p+1$ is fixed, we get by a simple union bound

$$\|(II)\|_\infty = O_P\left(N^{-1/2}\right).$$

In total, we thus get

$$\|\hat{\Sigma}_\xi - \Sigma_\xi\|_\infty = O_P\left(s_+^* \sqrt{\frac{\log(n)}{N}}\rho_n^{-1} + \frac{1}{\sqrt{N}}\right) = O_P\left(s_+^* \sqrt{\frac{\log(n)}{N}}\rho_n^{-1}\right).$$

We can now obtain a rate for $\|\hat{\Theta}_\xi - \Theta_\xi\|_\infty$.

$$\|\hat{\Theta}_\xi - \Theta_\xi\|_\infty \leq \frac{C}{\rho_n^2}\|\hat{\Sigma}_\xi - \Sigma_\xi\|_\infty = O_P\left(s_+^* \sqrt{\frac{\log(n)}{N}}\rho_n^{-3}\right).$$

By assumption B3, we have $s_+^* \frac{\sqrt{\log(n)}}{\sqrt{n}\rho_n^2} \to 0, n \to \infty$, which in particular also implies that the above is $o_P(1)$. Notice in particular, that we have now managed to get for $k = 1, \dots, p+1$,

- $\|\hat{\Theta}_{\xi,k} - \Theta_{\xi,k}\|_1 = o_P(1)$,
- $\hat{\Theta}_{\xi,k,k} = \Theta_{\xi,k,k} + o_p(1)$.

B.4 Problem 1

We can now take a look at the problems (1) - (3) outlined above. For problem (1), we want to show:

$$\sqrt{N}\frac{\hat{\Theta}_{\xi,k} P_n \nabla_\xi l_{\theta_0}}{\sqrt{\hat{\Theta}_{\xi,k,k}}} \to \mathcal{N}(0,1).$$

Step 1: Show that

$$\hat{\Theta}_{\xi,k} P_n \nabla_\xi l_{\theta_0} = \Theta_{\xi,k} P_n \nabla_\xi l_{\theta_0} + o_P\left(N^{-1/2}\right). \tag{33}$$

We have

$$\begin{aligned}|(\hat{\Theta}_{\xi,k} - \Theta_{\xi,k}) P_n \nabla_\xi l_{\theta_0}| &\leq \|\hat{\Theta}_{\xi,k} - \Theta_{\xi,k}\|_1 \left\| \frac{1}{N}\sum_{i \neq j} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} (p_{ij}(\theta_0) - A_{ij}) \right\|_\infty \\ &\leq \|\hat{\Theta}_\xi - \Theta_\xi\|_\infty \left\| \frac{1}{N}\sum_{i \neq j} D_{\xi,ij}(p_{ij}(\theta_0) - A_{ij}) \right\|_\infty .\end{aligned}$$

Consider the vector $\sum_{i \neq j} D_{\xi,ij}(p_{ij}(\theta_0) - A_{ij}) \in \mathbb{R}^{p+1}$. The k-th component of it has the form $\sum_{i \neq j}(p_{ij}(\theta_0) - A_{ij})$ for $k = 1$ and $\sum_{i \neq j} Z_{ij,k-1}(p_{ij}(\theta_0) - A_{ij}), k = 2, \dots, p+1$. Notice that for these components are all centred:

$$\mathbb{E}[D_{\xi,ij,k}(p_{ij}(\theta_0) - A_{ij})] = \mathbb{E}[D_{\xi,ij,k}\mathbb{E}[(p_{ij}(\theta_0) - A_{ij})|Z_{ij}]] = \mathbb{E}[D_{\xi,ij,k} \cdot 0] = 0,$$

as well as $|D_{\xi,ij,k}(p_{ij}(\theta_0) - A_{ij})| \leq c$, where $c > 1$ is a universal constant bounding $|Z_{ij,k}|$ for all i, j, k. Thus, by Hoeffding's inequality, for any $t > 0$,

$$P\left(\left|\frac{1}{N}\sum_{i \neq j} D_{\xi,ij,k}(p_{ij}(\theta_0) - A_{ij})\right| \geq t\right) \leq 2\exp\left(-2\frac{Nt^2}{c^2}\right)$$

and thus,

$$\frac{1}{N}\sum_{i\neq j} D_{\xi,ij}(p_{ij}(\theta_0) - A_{ij}) = O_P\left(N^{-1/2}\right).$$

Since we have $\|\hat{\Theta}_Z - \Theta_Z\|_\infty = o_P(1)$, by section B.3, step 1 is now concluded.
Step 2: Show that

$$\hat{\Theta}_{\xi,k,k} = \Theta_{\xi,k,k} + o_P(1).$$

Since $\|\hat{\Theta}_\xi - \Theta_\xi\|_\infty = o_P(1)$, by section B.3, for all k

$$|\hat{\Theta}_{\xi,k,k} - \Theta_{\xi,k,k}| \leq \|\hat{\Theta}_\xi - \Theta_\xi\|_\infty = o_P(1)$$

and step 2 is concluded.
Step 3: Show that

$$\left|\frac{1}{\Theta_{\xi,k,k}}\right| \leq C < \infty,$$

for some universal constant $C > 0$. Then, we may conclude from step 1 and step 2 that

$$\sqrt{N}\frac{\hat{\Theta}_{\xi,k}P_n\nabla_\xi l_{\theta_0}}{\sqrt{\hat{\Theta}_{\xi,k,k}}} = \sqrt{N}\frac{\Theta_{\xi,k}P_n\nabla_\xi l_{\theta_0}}{\sqrt{\Theta_{\xi,k,k}}} + o_P(1).$$

To prove step 3, notice that Θ_ξ is symmetric and hence has only real eigenvalues. Therefore it is unitarily diagonalizable and for any $x \in \mathbb{R}^{p+1}$, we have $x^T\Theta_\xi x \geq \lambda_{\min}(\Theta_\xi)\|x\|_2^2$. We also know that

$$\lambda_{\min}(\Theta_\xi) = \frac{1}{\lambda_{\max}(\Sigma_\xi)}.$$

Under assumption 1 we can now deduce an upper bound on the maximum eigenvalue of Σ_ξ: For any $x \in \mathbb{R}^p$,

$$x^T\Sigma_\xi x = x^T\frac{1}{N}\mathbb{E}[D_\xi^T W_0^2 D_\xi]x \leq x^T\frac{1}{N}\mathbb{E}[D_\xi^T D_\xi]x \leq (1 \vee \lambda_{\max})\|x\|_2^2,$$

where we have used that any entry in W_0^2 is bounded above by one. Since $x^T\Sigma_\xi x \leq \lambda_{\max}(\Sigma_\xi)\|x\|_2^2$ and since this bound is tight (we have equality if x is an eigenvector corresponding to $\lambda_{\max}$), we can conclude by assumption 1 that $\lambda_{\max}(\Sigma_\xi) \leq (1 \vee \lambda_{\max}) \leq C < \infty$ for some universal constant $C > 0$.

In particular, since $\Theta_{\xi,k,k} = e_k^T\Theta_\xi e_k$, we get

$$\Theta_{\xi,k,k} \geq \lambda_{\min}(\Theta_\xi)\|e_k\|_2^2 = \frac{1}{\lambda_{\max}(\Sigma_\xi)} \geq C > 0,$$

uniformly for all n. Consequently,

$$0 < \frac{1}{\Theta_{\xi,k,k}} \leq C < \infty.$$

Step 3 is thus concluded.

Step 4: Finally, show that

$$\sqrt{N}\frac{\Theta_{\xi,k}P_n\nabla_\xi l_{\theta_0}}{\sqrt{\Theta_{\xi,k,k}}}\xrightarrow{d}\mathcal{N}(0,1),$$

Such that by all the above

$$\sqrt{N}\frac{\hat{\Theta}_{\xi,k}P_n\nabla_\xi l_{\theta_0}}{\sqrt{\hat{\Theta}_{\xi,k,k}}}\xrightarrow{d}\mathcal{N}(0,1).$$

For brevity, we write p_{ij} for the true link probabilities $p_{ij}(\theta_0)$. Also keep in mind that $\Theta_{\xi,k}$ denotes the k-th *row* of Θ_ξ, while $D_{\xi,ij}$ denote $((p+1)\times 1)$-*column* vectors. We want to apply the Lindeberg-Feller Central Limit Theorem. The random variables we study are the summands in

$$\sqrt{N}\Theta_{\xi,k}P_n\nabla_\xi l_{\theta_0}=\sum_{i\neq j}\left\{\frac{1}{\sqrt{N}}\Theta_{\xi,k}D_{\xi,ij}(p_{ij}-A_{ij})\right\}.$$

First, notice that these random variables are centred:

$$\mathbb{E}\left[\frac{1}{\sqrt{N}}\Theta_{\xi,k}D_{\xi,ij}(p_{ij}-A_{ij})\right]=\mathbb{E}\left[\frac{1}{\sqrt{N}}\Theta_{\xi,k}D_{\xi,ij}\mathbb{E}[p_{ij}-A_{ij}|Z_{ij}]\right]=\mathbb{E}\left[\frac{1}{\sqrt{N}}\Theta_{\xi,k}D_{\xi,ij}\cdot 0\right]=0.$$

For the Lindeberg-Feller CLT we need to sum up the variances of these random variables. We claim that

$$\sum_{i\neq j}\mathrm{Var}\left(\frac{1}{\sqrt{N}}\Theta_{\xi,k}D_{\xi,ij}(p_{ij}-A_{ij})\right)=\Theta_{\xi,k,k}.$$

Indeed, consider the vector-valued random variable $\sum_{i\neq j}\left\{\frac{1}{\sqrt{N}}D_{\xi,ij}(p_{ij}-A_{ij})\right\}\in\mathbb{R}^{p+1}$. It has covariance matrix

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i\neq j}\left\{\frac{1}{\sqrt{N}}D_{\xi,ij}(p_{ij}-A_{ij})\right\}\sum_{i\neq j}\left\{\frac{1}{\sqrt{N}}D_{\xi,ij}(p_{ij}-A_{ij})\right\}^T\right]\\
&=\mathbb{E}\left[\sum_{i\neq j}\frac{1}{\sqrt{N}}D_{\xi,ij}(p_{ij}-A_{ij})\frac{1}{\sqrt{N}}D_{\xi,ij}^T(p_{ij}-A_{ij})\right],\quad\text{by independence accross } i,j\\
&=\frac{1}{N}\sum_{i\neq j}\left[\mathbb{E}[D_{\xi,ij,k}D_{\xi,ij,l}(p_{ij}-A_{ij})^2]\right]_{k,l=1,\dots,p+1},\quad\text{as a } ((p+1)\times(p+1))\text{-matrix}\\
&=\frac{1}{N}\mathbb{E}[D_\xi^TW_0^2D_\xi]\\
&=\Sigma_\xi.
\end{aligned}$$

Thus, by independence across i,j,

$$\sum_{i\neq j}\mathrm{Var}\left(\frac{1}{\sqrt{N}}\Theta_{\xi,k}D_{\xi,ij}(p_{ij}-A_{ij})\right)=\mathrm{Var}\left(\Theta_{\xi,k}\sum_{i\neq j}\frac{1}{\sqrt{N}}D_{\xi,ij}(p_{ij}-A_{ij})\right)=\Theta_{\xi,k}\Sigma_\xi\Theta_{\xi,k}^T=\Theta_{\xi,k,k},$$

where for the last equality we have used that Θ_ξ is the inverse of Σ_ξ and thus, $\Sigma_\xi \Theta_{\xi,k}^T = e_k$. Now, we need to show that the Lindeberg condition holds. That is, we want that for any $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{\Theta_{\xi,k,k}} \sum_{i\neq j} \mathbb{E}\left[\left\{ \frac{1}{\sqrt{N}} \Theta_{\xi,k} D_{\xi,ij}(p_{ij} - A_{ij}) \right\}^2 \mathbb{1}\left(|\Theta_{\xi,k} D_{\xi,ij}(p_{ij} - A_{ij})| > \epsilon \sqrt{N\Theta_{\xi,k,k}} \right) \right] = 0. \tag{34}$$

We have

$$|\Theta_{\xi,k} D_{\xi,ij}(p_{ij} - A_{ij})| \le p \cdot c \cdot \|\Theta_{\xi,k}\|_1 \le C\|\Theta_\xi\|_\infty \le C\rho_n^{-1}.$$

At the same time, we know from step 3 that $\Theta_{Z,k,k} \geq C > 0$ for some universal C. Then, as long as ρ_n^{-1} goes to infinity at a rate slower than n, which is enforced by assumption B3, we must have for n large enough

$$|\Theta_{\xi,k} D_{\xi,ij}(p_{ij} - A_{ij})| < \epsilon\sqrt{N\Theta_{\xi,k,k}}$$

uniformly in i, j. Thus, the indicator function and therefore each summand in (34) is equal to zero for n large enough. Hence, (34) holds. Then, by the Lindeberg-Feller CLT,

$$\sqrt{N} \frac{\Theta_{\xi,k} P_n \nabla_\xi l_{\theta_0}}{\sqrt{\Theta_{\xi,k,k}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Now, by steps 1-4,

$$\sqrt{N} \frac{\hat{\Theta}_{\xi,k} P_n \nabla_\xi l_{\theta_0}}{\sqrt{\hat{\Theta}_{\xi,k,k}}} \xrightarrow{d} \mathcal{N}(0,1).$$

This concludes solving problem 1.

B.5 Problem 2

For problem 2 we must show

$$\frac{1}{\sqrt{\hat{\Theta}_{\xi,k,k}}} \hat{\Theta}_{\xi,k} \frac{1}{N} D_\xi^T \hat{W}^2 X(\hat{\vartheta} - \vartheta_0) = o_P\left(N^{-1/2}\right).$$

Since we have $\|\hat{\Theta}_\xi - \Theta_\xi\|_\infty = o_P(1)$, we do not need to worry about $\frac{1}{\sqrt{\hat{\Theta}_{\xi,k,k}}}$, because $\hat{\Theta}_{\xi,k,k} = \Theta_{\xi,k,k} + o_P(1)$ and $\frac{1}{\sqrt{\Theta_{\xi,k,k}}} \le C < \infty$, i.e. $\frac{1}{\sqrt{\hat{\Theta}_{\xi,k,k}}} = O_P(1)$. By Theorem 5 we also have a high-probability error bound on $\|\hat{\vartheta} - \vartheta_0\|_1$. The problem will be bounding the corresponding matrix norms.

$$\left| \hat{\Theta}_{\xi,k} \frac{1}{N} D_\xi^T \hat{W}^2 X(\hat{\vartheta} - \vartheta_0) \right| \le \left\| \frac{1}{N} X^T \hat{W}^2 D_\xi \hat{\Theta}_{\xi,k}^T \right\|_\infty \|\hat{\vartheta} - \vartheta_0\|_1.$$

Notice that in the display above we have the vector ℓ_∞-norm. Also,

$$\left\| \frac{1}{N} X^T \hat{W}^2 D_\xi \hat{\Theta}_{\xi,k}^T \right\|_\infty \le \|\hat{\Theta}_{\xi,k}^T\|_\infty \left\| \frac{1}{N} X^T \hat{W}^2 D_\xi \right\|_\infty.$$

Here we used the compatibility of the matrix ℓ_∞-norm with the vector ℓ_∞-norm. The first term is the vector norm, the second the matrix norm. We know,

$$\|\hat{\Theta}_{\xi,k}^T\|_\infty \leq \|\hat{\Theta}_\xi\|_\infty \leq C\rho_n^{-1},$$

where on the left hand side we have the vector norm and in the middle display the matrix norm. Finally, $1/N \cdot X^T\hat{W}^2D_\xi$ is a $(n \times (p+1))$-matrix. The (k,l)-th element looks like $1/N \cdot S_{k,l}$, where $S_{k,l}$ is the sum of $n-1$ terms of the form $D_{\xi,il,k}\hat{w}_{il}^2$, summed over the appropriate indices i,j, all of which are uniformly bounded. Thus,

$$\left|\left(\frac{1}{N}X^T\hat{W}^2D_\xi\right)_{k,l}\right| \leq \frac{1}{N}\cdot(n-1)\cdot c = \frac{C}{n}.$$

Thus, the ℓ_1-norm of any row of $\frac{1}{N}X^T\hat{W}^2D_\xi$ is bounded by pC/n and thus

$$\left\|\frac{1}{N}X^T\hat{W}^2D_\xi\right\|_\infty \leq \frac{C}{n}.$$

Recall that $\|\hat{\vartheta}-\vartheta_0\|_1 = O_P\left(s_+^*\frac{\sqrt{\log(n)}}{\sqrt{n}}\rho_n^{-1}\right)$ by Theorem 5. Then,

$$\begin{aligned}\left|\hat{\Theta}_{\xi,k}\frac{1}{N}X^T\hat{W}^2D_\xi(\hat{\vartheta}-\vartheta_0)\right| &\leq \|\hat{\Theta}_{\xi,k}^T\|_\infty\left\|\frac{1}{N}D_\xi^T\hat{W}^2X\right\|_\infty\|\hat{\vartheta}-\vartheta_0\|_1\\ &= O_P\left(\frac{s_+^*}{\rho_n^2\cdot n}\cdot\frac{\sqrt{\log(n)}}{\sqrt{n}}\right).\end{aligned}$$

Multiplying by $\sqrt{N}=O(n)$, gives

$$\sqrt{N}\left|\hat{\Theta}_{\vartheta,k}\frac{1}{N}D_\vartheta^T\hat{W}^2X(\hat{\vartheta}-\vartheta_0)\right| = O_P\left(\frac{s_+^*}{\rho_n^2}\cdot\frac{\sqrt{\log(n)}}{\sqrt{n}}\right),$$

which is $o_P(1)$ under Assumption B3.

B.6 Problem 3

Finally, we must show

$$O\left(\frac{1}{\sqrt{\hat{\Theta}_{\xi,k,k}}}\hat{\Theta}_{\xi,k}\frac{1}{N}\sum_{i\neq j}\left(\frac{1}{Z_{ij}}\right)|D_{ij}^T(\hat{\theta}-\theta_0)|^2\right) = o_P\left(N^{-1/2}\right).$$

Again, since $\hat{\Theta}_{\xi,k,k} = \Theta_{\xi,k,k}+o_P(1)$ and $\Theta_{\xi,k,k}\geq C>0$ uniformly in n, we do not need to worry about the factor $\frac{1}{\sqrt{\hat{\Theta}_{\xi,k,k}}}$ and it remains to show

$$O\left(\hat{\Theta}_{\xi,k}\frac{1}{N}\sum_{i\neq j}D_{\xi,ij}|D_{ij}^\top(\hat{\theta}-\theta_0)|^2\right) = o_P\left(N^{-1/2}\right).$$

We have

$$\begin{aligned}\left|\hat{\Theta}_{\xi,k}\frac{1}{N}\sum_{i\neq j}D_{\xi,ij}|D_{ij}^{\top}(\hat{\theta}-\theta_0)|^2\right| &\leq \frac{1}{N}\sum_{i\neq j}|\hat{\Theta}_{\xi,k}D_{\xi,ij}||D_{ij}^{T}(\hat{\theta}-\theta_0)|^2\\ &\leq c\|\hat{\Theta}_{\xi,k}\|_1\frac{1}{N}\sum_{i\neq j}|D_{ij}^{T}(\hat{\theta}-\theta_0)|^2\\ &\leq C\frac{1}{\rho_n}\frac{1}{N}\sum_{i\neq j}|D_{ij}^{T}(\hat{\theta}-\theta_0)|^2,\end{aligned}$$

where for the last inequality we have used that $\|\hat{\Theta}_{\xi,k}\|_1 \leq \|\hat{\Theta}_{\xi}\|_\infty \leq C\frac{1}{\rho_n}$. Now remember from (24) that

$$\frac{1}{N}\sum_{i\neq j}|D_{ij}^{T}(\hat{\theta}-\theta_0)|^2 \leq C\|\hat{\bar{\theta}}-\bar{\theta}_0\|_1^2,$$

where we make use of the fact that $\bar{D}\bar{\theta} = D\theta$. From Theorem 5 we know that under the assumptions of Theorem 6, $\|\hat{\bar{\theta}}-\bar{\theta}_0\|_1 = O_P\left(s_{0,+}\sqrt{\frac{\log(n)}{N}}\rho_n^{-1}\right)$. Thus,

$$\sqrt{N}\left|\hat{\Theta}_{\xi,k}\frac{1}{N}\sum_{i\neq j}D_{\xi,ij}|D_{ij}^{T}(\hat{\theta}-\theta_0)|^2\right| = O_P\left((s_{0,+})^2\frac{\log(n)}{\sqrt{N}}\rho_n^{-3}\right).$$

We see that this is $o_P(1)$ by applying assumption B3 twice. Problem 3 is solved.

Proof [of Theorem 6] Theorem 6 now follows from the solved problems (1) - (3). ■

Appendix C. Proof of Theorem 1

C.1 Proof of Lemmas

To make the representation cleaner, for the remainder of Section C we will simply write S for S_0 and S_+ for $S_{0,+}$. Recall that we use S_+^c to denote the complement of S_+ in $[2n+1+p]$, that is $S_+^c = [2n+1+p]\backslash S_+$. We also use S^c to refer to the complement of S in $[2n]$ *only*: $S^c = [2n]\backslash S$.

We begin by providing proofs of the lemmas in section 3.1.

Proof [of Lemma 3] Since $\bar{\theta}^\dagger$ and $\hat{\bar{\theta}}$ both solve (13), we must have

$$\frac{1}{N}\bar{\mathcal{L}}(\bar{\theta}^\dagger) + \bar{\lambda}\|\bar{\vartheta}^\dagger\|_1 = \frac{1}{N}\bar{\mathcal{L}}(\hat{\bar{\theta}}) + \bar{\lambda}\|\hat{\bar{\vartheta}}\|_1.$$

Denote by $\bar{z}_\vartheta^\dagger$ the first $2n$ components of $\bar{z}^\dagger$. Then, by (10a) and (10b), $\langle\bar{z}_\vartheta^\dagger, \bar{\vartheta}^\dagger\rangle = \|\bar{\vartheta}^\dagger\|_1$. Thus,

$$\frac{1}{N}\bar{\mathcal{L}}(\bar{\theta}^\dagger) + \bar{\lambda}\langle\bar{z}_\vartheta^\dagger, \bar{\vartheta}^\dagger\rangle = \frac{1}{N}\bar{\mathcal{L}}(\hat{\bar{\theta}}) + \bar{\lambda}\|\hat{\bar{\vartheta}}\|_1.$$

Hence, using that the last $p+1$ components of $\bar{z}^\dagger$ are zero,

$$\frac{1}{N}\bar{\mathcal{L}}(\bar{\theta}^\dagger) + \bar{\lambda}\langle \bar{z}^\dagger, \bar{\theta}^\dagger - \hat{\bar{\theta}}\rangle = \frac{1}{N}\bar{\mathcal{L}}(\hat{\bar{\theta}}) + \bar{\lambda}\left(\|\hat{\bar{\vartheta}}\|_1 - \langle \bar{z}^\dagger, \hat{\bar{\theta}}\rangle\right).$$

But by (9), $\bar{\lambda}\bar{z}^\dagger = -1/N \cdot \nabla\bar{\mathcal{L}}(\bar{\theta}^\dagger)$ and therefore

$$\frac{1}{N}\bar{\mathcal{L}}(\bar{\theta}^\dagger) - \langle 1/N \cdot \nabla\bar{\mathcal{L}}(\bar{\theta}^\dagger), \bar{\theta}^\dagger - \hat{\bar{\theta}}\rangle - \frac{1}{N}\bar{\mathcal{L}}(\hat{\bar{\theta}}) = \bar{\lambda}\left(\|\hat{\bar{\vartheta}}\|_1 - \langle \bar{z}^\dagger, \hat{\bar{\theta}}\rangle\right).$$

By the convexity of $\bar{\mathcal{L}}$, the left-hand side in the above display is negative. Therefore,

$$\|\hat{\bar{\vartheta}}\|_1 \le \langle \bar{z}^\dagger, \hat{\bar{\theta}}\rangle = \langle \bar{z}^\dagger_{\vartheta}, \hat{\bar{\vartheta}}\rangle \le \|\bar{z}^\dagger_{\vartheta}\|_\infty \|\hat{\bar{\vartheta}}\|_1 \le \|\hat{\bar{\vartheta}}\|_1.$$

Hence, $\langle \bar{z}^\dagger_{\vartheta}, \hat{\bar{\vartheta}}\rangle = \|\hat{\bar{\vartheta}}\|_1$. But since $\|\bar{z}^\dagger_{S^{\dagger c}}\|_\infty < 1$ by (11b), this can only hold if $\hat{\bar{\vartheta}}_{S^{\dagger c}} = 0$. The claim follows. ■

For the proof of Theorem 1 we need conditions similar to the ones in Ravikumar et al. (2010). The first condition is the so-called *dependency condition* which demands that the population Hessian of $\bar{\mathcal{L}}$ with respect to the variables contained in the active set S is invertible. For our specific case, we let

$$Q := \frac{1}{n-1}X^T W_0^2 X = H_{\bar{\vartheta}\times\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}) \in \mathbb{R}^{2n\times 2n}, \tag{35}$$

where $W_0 = \text{diag}(\sqrt{p_{ij}(\theta_0)(1-p_{ij}(\theta_0))}, i \ne j)$, be the Hessian of $\bar{\mathcal{L}}$ with respect to $\bar{\vartheta}$ only.

Lemma 24 (Dependency condition) *For any n, the minimum eigenvalue of $Q_{S,S}$ satisfies*

$$\lambda_{min}\left(Q_{S,S}\right) \ge \frac{1}{2}\rho_n \cdot \left(1 - \frac{\max\{s_\alpha, s_\beta\}}{n-1}\right) > 0.$$

Proof [of Lemma 24] Notice that

$$\frac{1}{n-1}X^T X = \begin{bmatrix} I_n & B \\ B & I_n \end{bmatrix} \in \mathbb{R}^{2n\times 2n},$$

where I_n is the $(n\times n)$ identity matrix and B is a matrix with zeros on the diagonal and $1/(n-1)$ everywhere else. Now consider the submatrix with only those rows and columns belonging to S

$$P := \frac{1}{n-1}(X^T X)_{S\times S} = \begin{bmatrix} I_{s_\alpha} & B_{S_\alpha, S_\beta} \\ B_{S_\beta, S_\alpha} & I_{s_\beta} \end{bmatrix} \in \mathbb{R}^{s\times s}.$$

This matrix P is strictly diagonally dominant. Indeed,

$$\sum_{j\in S, j\ne i} P_{ij} = \frac{s_\beta}{n-1} < 1 = P_{ii}, \quad i \in S_\alpha$$
$$\sum_{j\in S, j\ne i} P_{ij} = \frac{s_\alpha}{n-1} < 1 = P_{ii}, \quad i \in S_\beta,$$

where the strict inequalities hold because $\min_i\{\alpha_{0,i}\} = \min_j\{\beta_{0,j}\} = 0$. Thus, P is strictly positive definite. More, by the Gershgorin Circle Theorem, all the eigenvalues of P must lie in one of the discs $D(P_{ii}, R_i)$, where $R_i = \sum_{j \in S, j \neq i} P_{ij}$ and $D(P_{ii}, R_i)$ is the disc with radius R_i centred at P_{ii}. In particular,

$$\text{mineval}(P) \geq 1 - \frac{\max\{s_\alpha, s_\beta\}}{n-1}.$$

But now, for any $v \in \mathbb{R}^s$,

$$v^T Q_{S,S} v \geq \frac{1}{2}\rho_n \cdot v^T P v \geq \frac{1}{2}\rho_n \left(1 - \frac{\max\{s_\alpha, s_\beta\}}{n-1}\right) \|v\|_2^2$$

and the claim follows. ■

Lemma 25 (Incoherence condition) *For any n,*

$$\|Q_{S^c,S} Q_{S,S}^{-1}\|_\infty \leq \frac{1}{2}\rho_n^{-1} \cdot \frac{\max\{s_\alpha, s_\beta\}}{n - \max\{s_\alpha, s_\beta\}}.$$

By Lemma 24 the left-hand side of Lemma 25 is well-defined. Furthermore, under Assumption B1, the right-hand side in Lemma 25 tends to zero as n tends to infinity.
Proof [of Lemma 25] We make use of the following bound of a the infinity norm of the inverse of a diagonally dominant matrix (see for example Varah (1975))

$$\|Q_{S,S}^{-1}\|_\infty \leq \max_{i \in S} \left\{ \frac{1}{|q_{ii}| - R_i} \right\},$$

where q_{ii} is the ith diagonal entry of $Q_{S,S}$ and R_i is the sum of the off-diagonal elements of the ith row of $Q_{S,S}$. That is, for $i \in S_\alpha$,

$$q_{ii} - R_i = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} p_{ij}(1-p_{ij}) - \frac{1}{n-1} \sum_{j \in S_\beta} p_{ij}(1 - p_i j) \geq \frac{1}{2(n-1)}\rho_n(n - s_\beta),$$

and analogously for $i \in S_\beta$,

$$q_{ii} - R_i \geq \frac{1}{2(n-1)}\rho_n(n - s_\alpha).$$

Thus,

$$q_{ii} - R_i \geq \frac{1}{2(n-1)}\rho_n(n - \max\{s_\alpha, s_\beta\})$$

and therefore,

$$\|Q_{S,S}^{-1}\|_\infty \leq 2\rho_n^{-1} \cdot \frac{n-1}{n - \max\{s_\alpha, s_\beta\}}. \tag{36}$$

Furthermore, notice that any row of $Q_{S^c,S}$ has either s_α or s_β non-zero entries, each of the form $1/(n-1) \cdot p_{ij}(1-p_{ij}) \leq 1/(4(n-1))$. Hence,

$$\|Q_{S^c,S}\|_\infty \leq \frac{\max\{s_\alpha, s_\beta\}}{4(n-1)}.$$

The claim follows by the submultiplicativity of the matrix infinity norm. ■

C.2 General strategy

The proof of Theorem 1 hinges on the construction of $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ succeeding with high probability and the challenge in proving this is proving that $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ fulfils conditions (11a) and (11b). Our proof relies on the following derivations. From (9) we obtain

$$0 = \frac{1}{N}\nabla\bar{\mathcal{L}}(\bar{\theta}^\dagger) + \bar{\lambda}\bar{z}^\dagger - \frac{1}{N}\nabla\bar{\mathcal{L}}(\bar{\theta}_0) + \frac{1}{N}\nabla\bar{\mathcal{L}}(\bar{\theta}_0).$$

Doing a Taylor expansion along the same lines as (28) and (29), we obtain

$$\frac{1}{N}\nabla\bar{\mathcal{L}}(\bar{\theta}^\dagger) - \frac{1}{N}\nabla\bar{\mathcal{L}}(\bar{\theta}_0) = \frac{1}{N}\bar{D}^T W_0^2 \bar{D}(\bar{\theta}^\dagger - \bar{\theta}_0) + O\left(\frac{1}{N}\sum_{i\neq j}\bar{D}_{ij}|\bar{D}_{ij}^T(\bar{\theta}^\dagger - \bar{\theta}_0)|^2\right),$$

where we have used the fact that we are taking derivatives with respect to $\bar{\theta}$ and used $\bar{D}_{ij}\bar{\theta}_0$ in (29), to obtain W_0^2 instead of $\hat{W}^2$ above. Combining the last two equations, we obtain

$$\frac{1}{N}\bar{D}^T W_0^2 \bar{D}(\bar{\theta}^\dagger - \bar{\theta}_0) = -\bar{\lambda}\bar{z}^\dagger - \frac{1}{N}\nabla\bar{\mathcal{L}}(\bar{\theta}_0) + O\left(\frac{1}{N}\sum_{i\neq j}\bar{D}_{ij}|\bar{D}_{ij}^T(\bar{\theta}^\dagger - \bar{\theta}_0)|^2\right).$$

Taking only the first $2n$ entries of that equation we obtain

$$\frac{1}{N}\bar{X}^T W_0^2 \bar{X}(\bar{\vartheta}^\dagger - \bar{\vartheta}_0) = -\frac{1}{N}\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0) + \frac{1}{N}\bar{X}^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right](\xi^\dagger - \xi_0) - \bar{\lambda}\bar{z}^\dagger_{1:2n} + \bar{R} \tag{37}$$

where we use $\bar{z}^\dagger_{1:2n}$ to refer to the first $2n$ components of $\bar{z}^\dagger_{1:2n}$, use our shorthand notation $\xi = (\mu, \gamma^T)^T$ and let

$$\bar{R} = O\left(\frac{1}{N}\sum_{i\neq j}\bar{X}_{ij}|\bar{D}_{ij}^T(\bar{\theta}^\dagger - \bar{\theta}_0)|^2\right).$$

Notice that we the left-hand side in (37) is equal to

$$Q(\bar{\vartheta}^\dagger - \bar{\vartheta}_0) = Q_{-,S}(\bar{\vartheta}^\dagger - \bar{\vartheta}_0)_S + Q_{-,S^c}\underbrace{(\bar{\vartheta}^\dagger - \bar{\vartheta}_0)_{S^c}}_{=0}.$$

Plugging this into (37) and splitting up by rows, we get

$$Q_{S,S}(\bar{\vartheta}^\dagger - \bar{\vartheta}_0)_S = -\frac{1}{N}\left(\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right)_S + \frac{1}{N}\bar{X}_S^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right](\xi^\dagger - \xi_0) - \bar{\lambda}\bar{z}^\dagger_{1:2n,S} + \bar{R}_S \tag{38a}$$

$$Q_{S^c,S}(\bar{\vartheta}^\dagger - \bar{\vartheta}_0)_S = -\frac{1}{N}\left(\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right)_{S^c} + \frac{1}{N}\bar{X}_{S^c}^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right](\xi^\dagger - \xi_0) - \bar{\lambda}\bar{z}^\dagger_{1:2n,S^c} + \bar{R}_{S^c}, \tag{38b}$$

where it is important to remember that $S^c = [2n]\backslash S$ refers to the complement of S in $[2n]$. We solve (38a) for $(\bar{\vartheta}^\dagger - \bar{\vartheta}_0)_S$ and plug the result into (38b). Finally we rearrange for $-\bar{\lambda}\bar{z}^\dagger_{1:2n,S^c}$,

$$\begin{aligned}-\bar{\lambda}\bar{z}^\dagger_{1:2n,S^c} =& Q_{S^c,S}Q_{S,S}^{-1}\left\{-\frac{1}{N}\left(\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right)_S + \frac{1}{N}\bar{X}_S^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right](\xi^\dagger - \xi_0) - \bar{\lambda}\bar{z}^\dagger_{1:2n,S} + \bar{R}_S\right\} \\ &+ \frac{1}{N}\left(\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right)_{S^c} - \frac{1}{N}\bar{X}_{S^c}^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right](\xi^\dagger - \xi_0) - \bar{R}_{S^c}\end{aligned}$$

Now, divide by $\bar{\lambda}$ and take the ∞-norm on both sides. Rearrange corresponding terms.

$$\begin{aligned}\|\bar{z}^{\dagger}_{1:2n,S^c}\|_\infty \le{}& \frac{1}{\bar{\lambda}}\left\{\|Q_{S^c,S}Q^{-1}_{S,S}\|_\infty + 1\right\}\left\|\frac{1}{N}\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right\|_\infty && (I)\\ &+\frac{1}{\bar{\lambda}}\left\{\|Q_{S^c,S}Q^{-1}_{S,S}\|_\infty + 1\right\}\|\bar{R}\|_\infty && (II)\\ &+\frac{1}{\bar{\lambda}}\left\{\|Q_{S^c,S}Q^{-1}_{S,S}\|_\infty + 1\right\}\left\|\frac{1}{N}\bar{X}^TW_0^2\left[\ \mathbf{1}\mid Z\ \right](\xi^\dagger-\xi_0)\right\|_\infty && (III)\\ &+\left\|Q_{S^c,S}Q^{-1}_{S,S}\right\|_\infty && (IV).\end{aligned}$$

By appropriately bounding the terms $(I)-(IV)$ on the right-hand side, we will proceed to show that for sufficiently large n, with high probability, $\|\bar{z}^{\dagger}_{1:2n,S^c}\|_\infty < 1$, which is clearly equivalent to (11b). Notice that we already may control term (IV) as well as the terms $\|Q_{S^c,S}Q^{-1}_{S,S}\|_\infty + 1$ by the incoherence condition, Lemma 25.

C.3 Controlling term (I)

Notice that the ith component of $\frac{1}{N}\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)$ is of the form

$$\frac{1}{N}\sqrt{n}\sum_{j=1,j\neq i}(A_{ij}-p_{ij}) = \frac{1}{\sqrt{n}}\cdot\frac{1}{n-1}\sum_{j=1,j\neq i}(A_{ij}-p_{ij}).$$

In particular, each summand is a centred, bounded random variable. By Hoeffding's inequality, we have for every $t>0$,

$$P\left(\left|\frac{1}{n-1}\sum_{j=1,j\neq i}(A_{ij}-p_{ij})\right|\ge t\right)\le 2\exp\left(-\frac{n-1}{2}t^2\right).$$

Thus, for any $\epsilon>0$, picking $t=\epsilon\sqrt{n}\bar{\lambda}$, gives

$$P\left(\frac{1}{\bar{\lambda}}\left|\frac{1}{N}\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)_i\right|\ge\epsilon\right)\le 2\exp\left(-\frac{N\bar{\lambda}^2}{2}\epsilon^2\right).$$

Taking a union bound over all $2n$ components of $\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)$, leads to

$$P\left(\frac{1}{\bar{\lambda}}\left\|\frac{1}{N}\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right\|_\infty\ge\epsilon\right)\le 4n\cdot\exp\left(-\frac{N\bar{\lambda}^2}{2}\epsilon^2\right) = 4\cdot\exp\left(-\frac{N\bar{\lambda}^2}{2}\epsilon^2+\log(n)\right). \quad (39)$$

In the next section, when controlling term (II), we will also need a similar bound on the components of $\frac{1}{N}\nabla\bar{\mathcal{L}}(\bar{\theta}_0)$ corresponding to $\xi=(\mu,\gamma^T)^T$, which is why we derive the respective bounds now. Using analogous arguments to the above, we obtain

$$P\left(\frac{1}{\bar{\lambda}}\left\|\frac{1}{N}\nabla_{\xi}\bar{\mathcal{L}}(\bar{\theta}_0)\right\|_\infty\ge\epsilon\right)\le 2(p+1)\cdot\exp\left(-\frac{N\bar{\lambda}^2}{2(1\vee c^2)}\epsilon^2\right). \quad (40)$$

Combining (39) and (40), we obtain a bound on the infinity norm of the full gradient,

$$P\left(\frac{1}{\bar{\lambda}}\left\|\frac{1}{N}\nabla\bar{\mathcal{L}}(\bar{\theta}_0)\right\|_\infty\ge\epsilon\right)\le 4\cdot\exp\left(-\frac{N\bar{\lambda}^2}{2}\epsilon^2+\log(n)\right)+2(p+1)\cdot\exp\left(-\frac{N\bar{\lambda}^2}{2(1\vee c^2)}\epsilon^2\right), \quad (41)$$

which tends to zero, as long as $-\frac{N\bar{\lambda}^2}{2}\epsilon^2+\log(n)\to\infty$, as n tends to infinity.

C.4 Controlling term (II)

Controlling term (II) is by far the most involved step in controlling $\|\bar{z}^\dagger_{1:2n,S^c}\|_\infty$. We start by controlling the ℓ_2-error between our construction $\bar{\theta}^\dagger$ and the truth $\bar{\theta}_0$.

Lemma 26 *Under assumptions 1, 2, B1, 3, for n large enough, for any $\epsilon > 0$, with probability at least*

$$1 - 4 \cdot \exp\left(-\frac{N\bar{\lambda}^2}{2}\epsilon^2 + \log(n)\right) - 2(p+1) \cdot \exp\left(-\frac{N\bar{\lambda}^2}{2(1 \vee c^2)}\epsilon^2\right)$$
$$- p(p+3)\exp\left(-N\frac{c^2_{\min}}{2048 s_+^2 \tilde{c}}\right),$$

which tends to one as long as $-\frac{N\bar{\lambda}^2}{2}\epsilon^2 + \log(n) \to -\infty$, as n tends to infinity, we have

$$\|\bar{\theta}^\dagger - \bar{\theta}_0\|_1 \leq (1+\epsilon)\frac{9}{c_{\min}}\rho_n^{-1} s_+ \bar{\lambda}.$$

Proof Keep in mind that $\bar{\theta}^\dagger - \bar{\theta}_0 = \bar{\theta}^\dagger_{S_+} - \bar{\theta}_{0,S_+}$. Define a function $G : \mathbb{R}^{s+1+p} \to \mathbb{R}$,

$$G(u) = \frac{1}{N}\left\{\bar{\mathcal{L}}(\bar{\theta}_{0,S_+} + u) - \bar{\mathcal{L}}(\bar{\theta}_{0,S_+})\right\} + \bar{\lambda}\left(\|\bar{\theta}_{0,S} + u_S\|_1 - \|\bar{\theta}_{0,S}\|_1\right),$$

where for the addition $\bar{\theta}_{0,S_+} + u$ to be well-defined, we use the canonical embedding of $\mathbb{R}^{s+1+p} \hookrightarrow \mathbb{R}^{2n+1+p}$, by setting the components not contained in S to zero. In the following we will make use of that embedding without explicitly mentioning it if there is no chance of confusion. Also pay close attention to the distinction between S_+ and S in above display. Clearly, $G(0) = 0$ and G is minimized at $\bar{u}^\dagger = \bar{\theta}^\dagger_{S_+} - \bar{\theta}_{0,S_+}$, which implies that $G(\bar{u}^\dagger) \leq 0$. Also, G is convex.

Now suppose we manage to find some $B \in \mathbb{R}, B > 0$, such that for all $u \in \mathbb{R}^{s+1+p}$ with $\|u\|_1 = B$ it holds $G(u) > 0$. We claim that in that case it must hold $\|\bar{u}^\dagger\|_1 \leq B$. Indeed, if $\|\bar{u}^\dagger\|_1 > B$, then there exists a $t \in (0,1)$ such that for $\tilde{u} = t\bar{u}^\dagger$ we have $\|\tilde{u}\|_1 = B$. But then, by convexity of G, $G(\tilde{u}) \leq tG(\bar{u}^\dagger) + (1-t)G(0) = tG(\bar{u}^\dagger) \leq 0$. A contradiction.

Thus, we need to find an appropriate B. Let $B > 0$, the correct form to be determined later. Now, pick any $u \in \mathbb{R}^{s+1+p}$ with $\|u\|_1 = B$. We do a first order Taylor expansion of $\bar{\mathcal{L}}$ in the point $\bar{\theta}_{0,S_+}$, evaluated at $\bar{\theta}_{0,S_+} + u$. This yields

$$\begin{aligned} G(u) =& \frac{1}{N}\left\{\nabla_{S_+}\bar{\mathcal{L}}(\bar{\theta}_{0,S_+})^T(\bar{\theta}_{0,S_+} + u - \bar{\theta}_{0,S_+}) + \frac{1}{2}\cdot u^T H_{S_+,S_+}\bar{\mathcal{L}}(\bar{\theta}_{0,S_+} + u\alpha)u\right\} \\ & + \bar{\lambda}\left(\|\bar{\theta}_{0,S} + u_S\|_1 - \|\bar{\theta}_{0,S}\|_1\right), \end{aligned}$$

for some $\alpha \in [0,1]$. Now, using (41), we know that with high-probability,

$$\left|\frac{1}{N}\nabla_{S_+}\bar{\mathcal{L}}(\bar{\theta}_{0,S_+})^T u\right| \leq \left\|\frac{1}{N}\nabla_{S_+}\bar{\mathcal{L}}(\bar{\theta}_{0,S_+})\right\|_\infty \|u\|_1 \leq \epsilon\bar{\lambda}B \qquad (42)$$

with the ϵ from (41). Furthermore, by using the triangle inequality, we obtain

$$\bar{\lambda}\left(\|\bar{\theta}_{0,S} + u_S\|_1 - \|\bar{\theta}_{0,S}\|_1\right) \geq -\bar{\lambda}\|u\|_1 = -\bar{\lambda}B \qquad (43)$$

Clearly, the canonical embedding of u into $\mathbb{R}^{2n+1+p}$ fulfils the condition of the empirical compatibility condition, Proposition 9. Also, keep in mind that assumptions B1 and 3 together imply $n^{-1/2}\rho_n^{-1}s_+ \to 0$, which in particular implies $s_+ = o(\sqrt{n})$. Thus, Proposition 9 is applicable and with high probability as prescribed in Proposition 9, we have

$$\begin{aligned} \frac{1}{2} \cdot u^T H_{S_+,S_+}\bar{\mathcal{L}}(\bar{\theta}_{0,S_+} + u\alpha)u &\ge \frac{1}{4}\rho_n u^T \left\{ \frac{1}{N}\bar{D}^T\bar{D} \right\}_{S_+,S_+} u \\ &= \frac{1}{4}\rho_n u^T \Sigma u \\ &\ge \frac{1}{8}\rho_n \frac{c_{\min}}{s_+}\|u\|_1^2 \\ &= \frac{1}{8}\rho_n \frac{c_{\min}}{s_+} B^2 \end{aligned} \tag{44}$$

Combining (42), (43), (44), we find

$$G(u) \ge -\epsilon\bar{\lambda}B - \bar{\lambda}B + \frac{1}{8}\rho_n \frac{c_{\min}}{s_+}B^2.$$

The right-hand side of this equation is strictly larger zero, whenever

$$B > (1+\epsilon)\frac{8}{c_{\min}}\rho_n^{-1}s_+\bar{\lambda}.$$

Thus, the claim follows from picking

$$B = (1+\epsilon)\frac{9}{c_{\min}}\rho_n^{-1}s_+\bar{\lambda}.$$

■

Lemma 27 *Under assumptions 1, 2, B1, 3, for n large enough, for any $\epsilon > 0$, with probability at least*

$$\begin{aligned} 1 - 4 \cdot \exp\left(-\frac{N\bar{\lambda}^2}{2}\epsilon^2 + \log(n)\right) - 2(p+1) \cdot \exp\left(-\frac{N\bar{\lambda}^2}{2(1 \vee c^2)}\epsilon^2\right) \\ - p(p+3)\exp\left(-N\frac{c_{\min}^2}{2048 s_+^2 \tilde{c}}\right), \end{aligned}$$

which tends to one as long as $-\frac{N\bar{\lambda}^2}{2}\epsilon^2 + \log(n) \to -\infty$, as n tends to infinity, we have

$$\frac{1}{\bar{\lambda}}\|\bar{R}\|_\infty \le \frac{324(1 \vee (c^2 p))(1+\epsilon)^2}{c_{\min}^2} \cdot \sqrt{n}\rho_n^{-2}s_+^2\bar{\lambda}.$$

Proof Consider the ith component of $\bar{R}$, for $i \in S_\alpha$. Similar to (24) we obtain,

$$\begin{aligned}
\bar{R}_i &= \frac{1}{N} \sum_{j=1, j\neq i}^{n} \bar{X}_{ij} |\bar{D}_{ij}^T(\bar{\theta}^\dagger - \bar{\theta}_0)|^2 \\
&= \frac{1}{\sqrt{n}} \cdot \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} |\bar{D}_{ij}^T(\bar{\theta}^\dagger - \bar{\theta}_0)|^2 \\
&= \frac{1}{\sqrt{n}} \cdot \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} \Big(\alpha_i^\dagger - \alpha_{0,i} + \beta_j^\dagger - \beta_{0,j}^\dagger + \mu^\dagger - \mu_0 + Z_{ij}^T(\gamma^\dagger - \gamma_0)\Big)^2 \\
&\leq \frac{4}{\sqrt{n}} \cdot \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} \Big((\alpha_i^\dagger - \alpha_{0,i})^2 + (\beta_j^\dagger - \beta_{0,j})^2 + (\mu^\dagger - \mu_0)^2 + c^2 p \|\gamma^\dagger - \gamma_0\|_2^2\Big) \\
&= \frac{4}{\sqrt{n}} \Big\{(\alpha_i^\dagger - \alpha_{0,i})^2 + (\mu^\dagger - \mu_0)^2 + c^2 p \|\gamma^\dagger - \gamma_0\|_2^2\Big\} + \frac{4}{\sqrt{n}} \cdot \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} (\beta_j^\dagger - \beta_{0,j})^2 \\
&\leq \frac{4}{\sqrt{n}} (1 \vee (c^2 p)) \Big\{(\alpha_i^\dagger - \alpha_{0,i})^2 + (\mu^\dagger - \mu_0)^2 + \|\gamma^\dagger - \gamma_0\|_2^2\Big\} + \frac{\sqrt{n}}{n-1} \|\bar{\beta}^\dagger - \bar{\beta}_0\|_2^2.
\end{aligned}$$

We have

$$(\alpha_i^\dagger - \alpha_{0,i})^2 = n(\bar{\alpha}_i^\dagger - \bar{\alpha}_{0,i})^2 \leq n\|\bar{\alpha}^\dagger - \bar{\alpha}_0\|_2^2.$$

Thus, by Lemma 26, with at least the prescribed probability and for all $i \in S_\alpha$,

$$\begin{aligned}
\frac{R_i}{\bar{\lambda}} \leq & 4(1 \vee (c^2 p))\sqrt{n}\|\bar{\theta}^\dagger - \bar{\theta}_0\|_2^2 \leq 4(1 \vee (c^2 p))\sqrt{n}\|\bar{\theta}^\dagger - \bar{\theta}_0\|_1^2 \\
\leq & \frac{324(1 \vee (c^2 p))(1+\epsilon)^2}{c_{\min}^2} \cdot \sqrt{n}\rho_n^{-2} s_+^2 \bar{\lambda}.
\end{aligned}$$

The same bound is found for all $i \in S_\beta$ using the exact same steps. Since the right-hand side above does not depend on i the claim follows. ■

C.5 Controlling term (III)

Lemma 28 *Under assumptions 1, 2, B1, 3 for n large enough, for any $\epsilon > 0$, with probability at least*

$$\begin{aligned}
1 - 4 \cdot \exp\left(-\frac{N\bar{\lambda}^2}{2}\epsilon^2 + \log(n)\right) - 2(p+1) \cdot \exp\left(-\frac{N\bar{\lambda}^2}{2(1 \vee c^2)}\epsilon^2\right) \\
- p(p+3)\exp\left(-N\frac{c_{\min}^2}{2048 s_+^2 \tilde{c}}\right),
\end{aligned}$$

which tends to one as long as $-\frac{N\bar{\lambda}^2}{2}\epsilon^2 + \log(n) \to -\infty$, as n tends to infinity, we have

$$\frac{1}{\bar{\lambda}} \left\| \frac{1}{N} \bar{X}^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right] (\xi^\dagger - \xi_0) \right\|_\infty \leq \frac{9(1 \vee c)(1+\epsilon)(p+1)}{4c_{\min}} \cdot \frac{1}{\sqrt{n}} \rho_n^{-1} s_+.$$

Proof We have

$$\left\| \frac{1}{N} \bar{X}^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right] (\xi^\dagger - \xi_0) \right\|_\infty \leq \left\| \frac{1}{N} \bar{X}^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right] \right\|_\infty \|(\xi^\dagger - \xi_0)\|_\infty$$
$$\leq \left\| \frac{1}{N} \bar{X}^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right] \right\|_\infty \|\bar{\theta}^\dagger - \bar{\theta}_0\|_1.$$

Consider the ith row of the matrix $\frac{1}{N}\bar{X}^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right]$,

$$\left\| \left(\frac{1}{N} \bar{X}_{-,i}^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right] \right)^T \right\|_1 \leq \frac{1}{N} \sqrt{n}(n-1) \cdot \frac{1}{4}(1 \vee c)(p+1) = \frac{p+1}{4}(1 \vee c)\frac{1}{\sqrt{n}},$$

where we have used that the ith column of $\bar{X}$ has exactly $(n-1)$ non-zero entries, each with value $\sqrt{n}$, each entry of W_0^2 is upper bounded by 1/4 and any row of $\left[\ \mathbf{1} \mid Z\ \right]$ has $p+1$ entries, each of which is upper bounded by $1 \vee c$. Thus, by Lemma 26, with the prescribed probability,

$$\frac{1}{\bar{\lambda}} \left\| \frac{1}{N} \bar{X}^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right] (\xi^\dagger - \xi_0) \right\|_\infty \leq \frac{9(1 \vee c)(1+\epsilon)(p+1)}{4c_{\min}} \cdot \frac{1}{\sqrt{n}} \rho_n^{-1} s_+.$$

■

C.6 Condition 11b

Lemma 29 *Under assumptions 1, 2, B1, 3, for n large enough, with probability at least*

$$1 - 4 \cdot \exp\left(-\frac{N\bar{\lambda}^2}{18} + \log(n)\right) - 2(p+1) \cdot \exp\left(-\frac{N\bar{\lambda}^2}{18(1 \vee c^2)}\right) - p(p+3)\exp\left(-N\frac{c_{\min}^2}{2048 s_+^2 \tilde{c}}\right),$$

which tends to one as long as $-\frac{N\bar{\lambda}^2}{18} + \log(n) \to -\infty$, as n tends to infinity, we have

$$\|\bar{z}^\dagger_{1:2n,S^c}\|_\infty < 1.$$

Proof By equation (39), Lemmas 25, 27, 28, with the probability given in those Lemmas, for any $\epsilon > 0$,

$$\begin{aligned} \|\bar{z}^\dagger_{1:2n,S^c}\|_\infty \leq & \left\{ \|Q_{S^c,S}Q_{S,S}^{-1}\|_\infty + 1 \right\} \epsilon \\ & + \left\{ \|Q_{S^c,S}Q_{S,S}^{-1}\|_\infty + 1 \right\} \frac{324(1 \vee (c^2 p))(1+\epsilon)^2}{c_{\min}^2} \cdot \sqrt{n}\rho_n^{-2} s_+^2 \bar{\lambda} \\ & + \left\{ \|Q_{S^c,S}Q_{S,S}^{-1}\|_\infty + 1 \right\} \frac{9(1 \vee c)(1+\epsilon)(p+1)}{4c_{\min}} \cdot \frac{1}{\sqrt{n}}\rho_n^{-1} s_+ \\ & + \frac{1}{2}\|Q_{S^c,S}Q_{S,S}^{-1}\|_\infty. \end{aligned}$$

By Lemma 25, for n sufficiently large, we have $\|Q_{S^c,S}Q^{-1}_{S,S}\|_\infty < 1/2$. Thus, by equation (39), Lemmas 27 and 28, for n sufficiently large, with the prescribed probability,

$$\begin{aligned}\|\bar{z}^{\dagger}_{1:2n,S^c}\|_\infty \leq & \frac{3}{2}\epsilon + \frac{1}{4} \\ & + \frac{486(1 \vee (c^2 p))(1+\epsilon)^2}{c^2_{\min}} \cdot \sqrt{n}\rho_n^{-2} s_+^2 \bar{\lambda} \\ & + \frac{27(1 \vee c)(1+\epsilon)(p+1)}{8c_{\min}} \cdot \frac{1}{\sqrt{n}}\rho_n^{-1} s_+.\end{aligned}$$

Pick $\epsilon = 1/3$, to obtain

$$\|\bar{z}^{\dagger}_{1:2n,S^c}\|_\infty \leq \frac{3}{4} + \frac{486(1 \vee (c^2 p))(4/3)^2}{c^2_{\min}} \cdot \sqrt{n}\rho_n^{-2} s_+^2 \bar{\lambda} + \frac{27(1 \vee c)(4/3)(p+1)}{8c_{\min}} \cdot \frac{1}{\sqrt{n}}\rho_n^{-1} s_+.$$

The second and third term go to zero as n tends to infinity by assumption B1. Indeed, the second term is assumption B1 exactly. For the third term note that by assumption B1, $\sqrt{n}s_+^2\bar{\lambda}\rho_n^{-2} = n^{-1/2}\rho_n^{-1}s_+ \cdot n\rho_n s_+\bar{\lambda} \to 0$ as, $n \to \infty$. On the other hand, by assumption 3, $n\rho_n s_+ \bar{\lambda} \geq C\rho_n^{-1}s_+\log(n) \to \infty$. Therefore it must hold that $n^{-1/2}\rho_n^{-1}s_+ \to 0$. The claim follows. ■

C.7 Proof of Theorem 1

Proof [of Theorem 1] By Lemma 29, we know that with probability at least as large as

$$1 - 4 \cdot \exp\left(-\frac{N\bar{\lambda}^2}{18} + \log(n)\right) - 2(p+1) \cdot \exp\left(-\frac{N\bar{\lambda}^2}{18(1 \vee c^2)}\right) - p(p+3)\exp\left(-N\frac{c^2_{\min}}{2048 s_+^2 \tilde{c}}\right),$$

property (11b) holds for the construction $(\bar{\theta}^\dagger, \bar{z}^\dagger)$. Thus, by Lemma 3, $\hat{S} = S^\dagger$ and in particular $\hat{S} \cap S^c = \emptyset$.

For the second part of Theorem 1, recall that by equation (38a),

$$\bar{\vartheta}^\dagger_S = \bar{\vartheta}_{0,S} + Q^{-1}_{S,S}\left\{-\frac{1}{N}\left(\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right)_S + \frac{1}{N}\bar{X}_S^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right](\xi^\dagger - \xi_0) - \bar{\lambda}\bar{z}^\dagger_{1:2n,S} + \bar{R}_S\right\}. \tag{45}$$

Thus, $S^\dagger$ contains all those indices i with

$$\left\|Q^{-1}_{S,S}\left\{-\frac{1}{N}\left(\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right)_S + \frac{1}{N}\bar{X}_S^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right](\xi^\dagger - \xi_0) - \bar{\lambda}\bar{z}^\dagger_{1:2n,S} + \bar{R}_S\right\}\right\|_\infty < \bar{\vartheta}_{0,i}.$$

Hence, consider

$$\begin{aligned}
&\left\| Q^{-1}_{S,S}\left\{-\frac{1}{N}\left(\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right)_S + \frac{1}{N}\bar{X}_S^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right](\xi^\dagger - \xi_0) - \bar{\lambda}\bar{z}^\dagger_{1:2n,S} + \bar{R}_S\right\}\right\|_\infty \\
&\quad \leq \|Q^{-1}_{S,S}\|_\infty \left\| -\frac{1}{N}\left(\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right)_S + \frac{1}{N}\bar{X}_S^T W_0^2 \left[\ \mathbf{1} \mid Z\ \right](\xi^\dagger - \xi_0) - \bar{\lambda}\bar{z}^\dagger_{1:2n,S} + \bar{R}_S\right\|_\infty \\
&\quad \leq 2\rho_n^{-1} \cdot \frac{n-1}{n - \max\{s_\alpha, s_\beta\}} \\
&\qquad \left\{\epsilon\bar{\lambda} + \bar{\lambda}\right. \\
&\qquad + \frac{9(1 \vee c)(1+\epsilon)(p+1)}{4c_{\min}} \cdot \frac{1}{\sqrt{n}}\rho_n^{-1}s_+\bar{\lambda} \\
&\qquad \left. + \frac{324(1 \vee (c^2 p))(1+\epsilon)^2}{c_{\min}^2} \cdot \sqrt{n}\rho_n^{-2}s_+^2\bar{\lambda}^2\right\}
\end{aligned}$$

where we used (36), (41) and Lemmas 27 and 28.

By assumption $\bar{\lambda} \leq C \cdot \sqrt{\log(n)/N}$ for some $C > 0$, thus the first two terms in the bracket may be upper bound by $C \cdot \sqrt{\log(n)/N}$, for a possibly different C. The third term is $o(1) \cdot 1/n$ by assumption B1 and the last term is $o(1) \cdot \sqrt{\log(n)}/n$ by assumption B1. Since $(n-1)/(n - \max\{s_\alpha, s_\beta\}) = O(1)$, the entire right-hand side is less or equal $C\rho_n^{-1}\frac{\sqrt{\log(n)}}{n}$. Multiply (45) by $\sqrt{n}$ to transition to the unscaled parameters $\vartheta_S^\dagger$ and the claim follows. In particular, $C\rho_n^{-1}\frac{\sqrt{\log(n)}}{n}$ goes to zero as n tends to infinity, which implies that for n large enough, with at least the prescribed probability the construction fulfils (11a) and thus $\hat{S} = S^\dagger = S$. ■

Appendix D. Sparse β-models and the power law

In this appendix we show that the degrees in sparse β-models can exhibit power law distributions. For that we leverage the results in Britton et al. (2006) that show that the β-model can generate node degrees asymptotically following a power law and the empirical degree distribution converging in probability to the same power law if β_i are randomly generated in a suitable way. Recall that in the β-model, each node is associated with a degree heterogeneity parameter β_i and links are formed independently with probabilities

$$P(A_{ij} = 1) = p_{ij} = \frac{e^{\beta_i + \beta_j}}{1 + e^{\beta_i + \beta_j}}. \tag{46}$$

We show the result for the model introduced in Chen et al. (2020), which is an undirected version of out model without covariates. Recall that in this model, links are made independently with probabilities

$$P(A_{ij} = 1) = p_{ij} = \frac{e^{\mu + \beta_i + \beta_j}}{1 + e^{\mu + \beta_i + \beta_j}}. \tag{47}$$

Proposition 30 (SβM and power law) *Let $\{W_i\}_{i=1}^{\infty}$ be i.i.d. random variables supported in $[1,\infty)$ with $P(W_1 > w) \sim cw^{-\rho}$ as $w \to \infty$ for some $c > 0$ and $\rho \in (0,1)$. For the SβM in (47), suppose that $\mu = -\rho^{-1}\log n$ and β_i's are generated as $\beta_i = \log W_i$. Then the limiting distribution of each node degree d_i as $n \to \infty$ is a power law with exponent $\tau = 2$, that is,*

$$p_k := \lim_{n\to\infty} P(d_i = k) \sim k^{-2}, \ k \to \infty.$$

In addition, for $N_k := |\{i \in \{1,\dots,n\} : d_i = k\}|$, we have $N_k/n \xrightarrow{P} p_k$ as $n \to \infty$.

In the above proposition, we do not impose sparsity on β, but it is possible to do so by assuming a mass at 1 to the distribution of W_1 since the assumption only requires the tail of the distribution of W_1 to behave like $cw^{-\rho}$. The proposition follows from the results of Britton et al. (2006), which are restated in the following. We first recall the definition of a mixed Poisson distribution.

Definition 31 (Mixed Poisson distribution) *Let F be a distribution function supported in $\mathbb{R}_+$. A random variable X taking values in the nonnegative integers follows the mixed Poisson distribution with mixing distribution F if*

$$P(X = k) = \int_{[0,\infty)} e^{-w}\frac{w^k}{k!}dF(w), \quad k = 0,1,2,\dots.$$

If $W \sim F$, then we also say that X follows the mixed Poisson distribution with parameter W.

The next lemma shows that the tail behavior of a mixed Poisson distribution is determined solely by that of the mixing distribution.

Lemma 32 *Let F be a distribution function supported in $\mathbb{R}_+$ such that $c_1 x^{1-\tau} \le 1-F(x) \le c_2 x^{1-\tau}$ for large x for some $0 < c_1 < c_2 < \infty$. Then there exist $0 < c_1' < c_2' < \infty$ such that the distribution function G of a mixed Poisson distribution with mixing distribution F satisfies $c_1' x^{1-\tau} \le 1 - G(x) \le c_2' x^{1-\tau}$ for large x.*

Proof See van der Hofstad (2016) Exercise 6.12. ■

The following results are taken from Theorem 3.2 and Proposition 3.1 in Britton et al. (2006). In the following, the variables $d_1,\dots,d_n$ are indeed a triangular sequence and hence should be indexed by n, but this is suppressed for the notational convenience.

Theorem 33 (β-model and mixed Poisson distribution) *Let $\{W_i\}_{i=1}^{\infty}$ be i.i.d. positive random variables with $P(W_1 > w) \sim cw^{-\rho}$ as $w \to \infty$ for some $c > 0$ and $\rho \in (0,1)$. For the β-model in (46), suppose that $\beta_1,\dots,\beta_n$ are generated as $\beta_i = \log W_i - (\log n)/(2\rho)$ for $i = 1,\dots,n$ for each $n = 1,2,\dots$. Then:*

(i) The limiting distribution of each node degree d_i as $n \to \infty$ is the mixed Poisson distribution with parameter ϱW_1^{ρ}, where $\varrho = c\int_0^{\infty}(1+x)^{-2}x^{-\rho}dx$.

(ii) For $N_k := |\{i \in \{1, \ldots, n\} : d_i = k\}|$, we have $N_k/n \overset{P}{\to} P(\varrho W_1^\rho = k)$ as $n \to \infty$, where ϱ appears in (i).

Combined with Lemma 32, we know that

$$\begin{aligned} P(d_i \geq y) &\approx P(\varrho W_1^\rho \geq y) \\ &= P(W_1 \geq (y/\varrho)^{1/\rho}) \\ &\sim c\varrho y^{-1}, \end{aligned}$$

so that the limiting distribution of d_i is a power law with exponent $\tau = 2$ in the sense that

$$\lim_{n\to\infty} P(d_i = k) \sim k^{-2},\ k \to \infty.$$

Likewise, the empirical degree distribution converges in probability to the same power law, which proves Proposition 30.

References

E. Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18:1–86, 2018.

Arash A Amini and Elizaveta Levina. On semidefinite relaxations for the block model. *The Annals of Statistics*, 46(1):149–179, 2018.

A. Barabási. *Network Science.* Cambridge University Press, 2016.

Dimitri Bertsekas. *Nonlinear Programming.* Athena Scientific, 1995.

P. J. Bickel and J. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Science*, 106:21068–21073, 2009.

Norbert Binkiewicz, Joshua T Vogelstein, and Karl Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377, 2017.

Olivier Bousquet. A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495 – 500, 2002.

T. Britton, M. Deijfen, and A. Martin-Löf. Generating simple random graphs with prescribed degree distribution. *Journal of Statistical Physics*, 124:1377–1397, 2006.

Wei Cai, Guoyu Guan, Rui Pan, Xuening Zhu, and Hansheng Wang. Network linear discriminant analysis. *Computational Statistics & Data Analysis*, 117:32–44, 2018.

S. Chatterjee, P. Diaconis, and A. Sly. Random graphs with a given degree sequence. *Annals of Applied Probability*, 21(4):1400–1435, 2011.

Mingli Chen, Kengo Kato, and Chenlei Leng. Analysis of networks via the sparse beta model. *Journal of the Royal Statistical Society, Series B*, 2020. arXiv:1908.03152.

S. E. Fienberg. A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, 21:825–839, 2012.

Ove Frank and David Strauss. Markov graphs. *Journal of the american Statistical association*, 81(395):832–842, 1986.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block models. *The Journal of Machine Learning Research*, 18(1):1980–2024, 2017.

B. S. Graham. An econometric model of network formation with degree heterogeneity. *Econometrica*, 85:1033–1063, 2017.

E. Greenshtein and Y. Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10:971–988, 2004.

P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76:33–50, 1981.

P. W. Holland, K. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5:109–137, 1983.

Sihan Huang and Yang Feng. Pairwise covariates-adjusted block model for community detection, 2018. arXiv:1807.03469.

Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

Anders Bredahl Kock and Haihan Tang. Uniform inference in high-dimensional dynamic panel data models with approximately sparse fixed effects. *Econometric Theory*, 35(2): 295–359, 2019.

E. D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, 2009.

E. D. Kolaczyk. *Topics at the Frontier of Statistics and Network Analysis: (Re)Visiting the Foundations*. Cambridge University Press, 2017.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. École d'été de probabilités de Saint-Flour XXXVIII-2008*. Springer, 2011.

Emmanuel Lazega. *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*. Oxford University Press, 2001.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer-Verlag, 1991.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Annals of Statistics*, 43(1):215–237, 2015.

Tianxi Li, Elizaveta Levina, and Ji Zhu. Network cross-validation by edge sampling. *Biometrika*, 107(2):257–276, 2020.

Zhuang Ma, Zongming Ma, and Hongsong Yuan. Universal latent space model fitting for large networks with edge covariates. *Journal of Machine Learning Research*, 21(4):1–67, 2020.

Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional ising model selection using l1 -regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319, 06 2010. doi: 10.1214/09-AOS691.

Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2):173–191, 2007.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Annals of Statistics*, 39(4):1878–1915, 2011.

Stefan Stein and Chenlei Leng. A sparse β-model with covariates for networks. arXiv 2010.13604, 2022.

Sara van de Geer and Peter Bühlmann. *Statistics for High-Dimensional Data*. Springer Series in Statistics. Springer-Verlag, 2011.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

Sara A. van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645, 2008.

R. van der Hofstad. *Random Graphs and Complex Networks*, volume 1. Cambridge University Press, 2016.

A. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, 1996.

J.M. Varah. A lower bound for the smallest singular value of a matrix. *Linear Algebra and its Applications*, 11(1):3 – 5, 1975. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(75)90112-3.

Bowei Yan and Purnamrita Sarkar. Covariate regularized community detection in sparse graphs. *Journal of the American Statistical Association*, 0(0):1–12, 2020.

T. Yan and J. Xu. A central limit theorem in the β-model for undirected random graphs with a diverging number of vertices. *Biometrika*, 100:519–524, 2013.

T. Yan, C. Leng, and J. Zhu. Asymptotics in directed exponential random graph models with an increasing bi-degree sequence. *The Annals of Statistics*, 44:31–57, 2016.

T. Yan, B. Jiang, S. E. Fienberg, and C. Leng. Statistical inference in a directed network model with covariates. *Journal of the American Statistical Association*, 114(526):857–868, 2019.

Yi Yu, Jelena Bradic, and Richard J. Samworth. Confidence intervals for high-dimensional cox models. *Statistics Sinica (to appear)*, 2019. arXiv:1803.01150.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

Jingnan Zhang, Xin He, and Junhui Wang. Directed community detection with network embedding. *Journal of the American Statistical Association*, pages 1–11, 2021.

Yuan Zhang, Elizaveta Levina, and Ji Zhu. Community detection in networks with node features. *Electronic Journal of Statistics*, 10(2):3153–3178, 2016.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40 (4):2266–2292, 2012.