High-dimensional Ising model selection with Bayesian information criteria

Rina Foygel*

Department of Statistics
The University of Chicago
Chicago, IL 60637, U.S.A.
e-mail: rina@uchicago.edu

Mathias Drton

Department of Statistics
University of Washington
Seattle, WA 98195, U.S.A.
e-mail: md5@uw.edu

Abstract: We consider the use of Bayesian information criteria for selection of the graph underlying an Ising model. In an Ising model, the full conditional distributions of each variable form logistic regression models, and variable selection techniques for regression allow one to identify the neighborhood of each node and, thus, the entire graph. We prove high-dimensional consistency results for this pseudo-likelihood approach to graph selection when using Bayesian information criteria for the variable selection problems in the logistic regressions. The results pertain to scenarios of sparsity and following related prior work the information criteria we consider incorporate an explicit prior that encourages sparsity.

Keywords and phrases: Bayesian information criterion, graphical model, logistic regression, model selection, neighborhood selection, variable selection.

1. Introduction

Let $Z_1, \ldots, Z_p$ be binary random variables with values in $\{-1, 1\}$, and let $G = (V,E)$ be an undirected graph with vertex set $V = [p] := \{1, \ldots, p\}$. The (symmetric) Ising model postulates that

$$\text{Prob}(Z_1 = z_1, \ldots, Z_p = z_p) \propto \exp \left\{ \sum_{\{v,w\} \in E} \theta_{vw} z_v z_w \right\},$$

(1.1)

for values $z_1, \ldots, z_p \in \{-1, 1\}$ and interaction parameters $\theta_{ij} \in \mathbb{R}$. The Ising model is a special case of more general graphical log-linear or Markov random field models (Lauritzen, 1996) but it is of importance in its own right; see e.g. the monograph of Kindermann and Snell (1980). In this paper we will treat the problem of selecting the graph $G$ based on a random sample drawn from a

*Thanks to somebody
distribution in such an Ising model, complementing recent work on this problem by Anandkumar et al. (2012), Ravikumar, Wainwright and Lafferty (2010), Santhanam and Wainwright (2012) and Loh and Wainwright (2014).

The model selection procedure we consider uses a pseudo-likelihood approach based on conditional distributions, as popularized by Besag (1972, 1974). Let

$$\text{ne}(v) = \{ w \in V \setminus \{ v \} : \{ v, w \} \in E \}$$

be the set of neighbors of node $v$ in the graph $G = (V, E)$. Assuming (1.1), the full conditional distributions satisfy

$$\log \left( \frac{\text{Prob}(Z_v = 1 | Z_w = z_w, w \neq v)}{1 - \text{Prob}(Z_v = 1 | Z_w = z_w, w \neq v)} \right) = \sum_{w \in \text{ne}(v)} \beta_{vw} z_w,$$  (1.2)

where $\beta_{vw} = 2\theta_{vw}$. Hence, for each variable $Z_v$ in an Ising model, the conditional distributions form a logistic regression model with $Z_v$ as response and the remaining variables $Z_w$, $w \neq v$, as covariates. Selection of the graph $G = (V, E)$ can thus be achieved by identifying each neighborhood $\text{ne}(v)$, $v \in V$, by variable selection in each of the $p = |V|$ logistic regression problems given by (1.2).

Strictly speaking, we have $\beta_{vw} = \beta_{wv}$ in the system of logistic regression models in (1.2). However, we will treat the neighborhood selection approach in the version that uncouples the parameters, that is, we allow the pair $(\beta_{vw}, \beta_{wv})$ to range freely in $\mathbb{R}^2$. This allows one to treat the $p$ regression problems separately, which brings about simplifications with regards to computation as well as theoretical analysis; compare the work on $\ell_1$-penalization methods by Ravikumar, Wainwright and Lafferty (2010) and by Meinshausen and Bühlmann (2006) who treat the Gaussian case. Höfling and Tibshirani (2009) demonstrated empirically that this decoupling of $\beta_{vw}$ and $\beta_{wv}$, when addressing inferential inconsistencies as described in Section 4 below, does not lead to any important loss in statistical efficiency for selection of the graph $G$ in an Ising model; at least in the higher-dimensional settings that these authors and also we have in mind here. Höfling and Tibshirani (2009) also showed that, for selection of the graph underlying an Ising model, pseudo-likelihood methods fare as well as computationally more involved methods based on the actual joint distribution.

In this paper, we explore the use of Bayesian information criteria in the logistic neighborhood selection approach. Consider a logistic regression model that includes a subset $J$ of a set of $p$ covariates. For sample size $n$, and defined for minimization, the classical Bayesian information criterion (BIC) of Schwarz (1978) is the model score

$$\text{BIC}_0(J) = -2 \log L(\hat{\beta}_J) + |J| \log(n),$$

where $\hat{\beta}_J$ is the maximum likelihood estimator in the model given by $J$. The BIC is well-known to yield variable selection consistency in the asymptotic scenario in which the sample size $n$ grows large while the number of covariates $p$ remains constant. It has been observed, however, that the BIC tends to overselect variables in regression problems in which $p$ is of substantial size compared to
To address this problem, a number of extensions have been proposed and analyzed (Bogdan, Ghosh and Doerge, 2004; Chen and Chen, 2008, 2012; Frommlet et al., 2012). The main idea for these extensions is to incorporate into the BIC an explicit prior on the set of considered models. The priors specified in the mentioned earlier work are equivalent for our purposes (Zak-Szatkowska and Bogdan, 2011), and we will treat the criterion

\[ \text{BIC}_\gamma(J) = -2 \log L(\hat{\beta}_J) + |J| \left( \log(n) + 2\gamma \log(p) \right), \tag{1.3} \]

which is associated with a choice of \( \gamma \geq 0 \). For a review and pointers to prior work that suggests and evaluates defaults for \( \gamma \), or a quantity corresponding to \( \gamma \), see Zak-Szatkowska and Bogdan (2011). In particular, the choice of \( \gamma = 1 \) is associated with assigning equal prior probability to each set

\[ J_k = \{ J \subset [p] : |J| = k \}, \quad k = 0, \ldots, q, \]

which is a prior that is also considered in Scott and Berger (2010). Indeed,

\[ |J_k| = \binom{p}{k} \]

which for small \( k \leq q \leq p/2 \) scales as \( p^k \). In (1.3), this contribution of the prior on models appears as the term \( |J| \log(p) \). Note that (1.3) has the maximum of the log-likelihood function multiplied by two and, hence, the additional factor of two. By analogy, the prior for Ising model selection has to be specified on the set of graphs with \( p \) nodes and there are

\[ \binom{p}{2} \sim p^{2k} \]

graphs with \( k \) edges. This suggests that for Ising model selection \( \gamma \) should be chosen roughly twice as large as for variable selection in a single logistic regression model. The cutoffs for \( \gamma \) that appear in our theoretical analysis are in agreement with this intuition (compare Corollary 2.1 and Theorem 3.1.)

The goal of this paper is to show that using BIC\(_\gamma\) for variable selection in the logistic neighborhood selection approach allows one to consistently estimate the graph underlying an Ising model. Our focus is on higher-dimensional problems under sparsity, that is, problems in which the number of variables \( p \) may be large, the sample size \( n \) may be comparatively moderate, but the neighborhood sizes are bounded by an integer \( q \) that is small compared to \( p \). Our works builds on ideas of Chen and Chen (2012) and Luo and Chen (2013) who analyze the performance of BIC\(_\gamma\) for variable selection in generalized linear models. Their work makes assumptions on a sequence of fixed/deterministic design matrices that ensure that the Hessian of the log-likelihood function is well-behaved. In contrast, the conditional distributions in (1.2) have random covariates. We thus develop suitable conditions on the joint distribution of random covariates in logistic regression that, in particular, ensure that the deterministic conditions
imposed in Luo and Chen (2013) hold with high probability. In this part of our work we aim to be as general as possible and consider possibly unbounded covariates. The conditions we give allow us to deduce consistency of BIC$_\gamma$ in Ising model selection. For growing $p$, this involves a growing number of logistic regression problems and requires us to make some of the intermediate results in Luo and Chen (2013) more explicit.

The paper is organized as follows. Section 2 provides finite-sample results for logistic regression. The main technical result is Theorem 2.1, which considers the setting with random covariates and gives conditions that provide control of the Hessian of the log-likelihood function. Theorem 2.2 shows how a well-behaved Hessian leads to bounds on likelihood ratios and is closely related to the prior work of Chen and Chen (2012) and Luo and Chen (2013). The proofs for both these theorems are deferred to parts B and C of the Appendix, where part D contains technical lemmas. As a consequence of Theorems 2.1 and 2.2 we can clarify in Section 2.4 the consistency of BIC$_\gamma$ in logistic regression with random covariates. In Section 3, we extend the consistency result to Ising model selection. Some of the conditions imposed in our work involve third moments, and we show in part A of the Appendix that those cannot be weakened to conditions on second moments. We conclude with numerical experiments, see Sections 4 and 5, and a discussion in Section 6.

2. Logistic regression with random covariates

2.1. Setup

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be $n$ observations that each pair a binary response $Y_i \in \{0, 1\}$ and a covariate vector $X_i \in \mathbb{R}^p$. Suppose that the pairs $(X_i, Y_i)$ are independent and identically distributed, and that the responses follow a logistic regression model conditional on the covariates. Let $\pi_i(x)$ be the conditional probability that $Y_i = 1$ given $X_i = x$. The logistic regression model states that

$$\log \left( \frac{\pi_i(x)}{1 - \pi_i(x)} \right) = x^T \beta_0$$

for some unknown parameter vector $\beta_0 \in \mathbb{R}^p$. Define the cumulant function $b(\theta) = \log(1 + e^\theta)$. Conditional on the $X_i$, the logistic regression model for the responses $Y_i$ has log-likelihood, score, and negative Hessian functions

$$\log L(\beta) = \sum_{i=1}^{n} Y_i \cdot X_i^T \beta - b(X_i^T \beta) \in \mathbb{R},$$

$$s(\beta) = \sum_{i=1}^{n} X_i \left( Y_i - b'(X_i^T \beta) \right) \in \mathbb{R}^p,$$

$$H(\beta) = \sum_{i=1}^{n} X_i X_i^T \cdot b''(X_i^T \beta) \in \mathbb{R}^{p \times p}.$$
with the derivatives of the cumulant function being
\[ b'(\theta) = \frac{e^\theta}{1 + e^\theta}, \quad b''(\theta) = \frac{e^\theta}{(1 + e^\theta)^2}. \] (2.1)

We will be interested in scenarios in which \( \beta_0 \) is sparse, and we wish to recover the support of \( \beta_0 \), that is, the set
\[ J_0 = \text{supp}(\beta_0) := \{ j \in [p] : \beta_{0j} \neq 0 \}. \]

Here and throughout, \([p] := \{1, \ldots, p\}\). We assume that an upper bound on the size of support is given, that is, \(|J_0| \leq q\). However, we will later allow the bound \(q\) to grow in an asymptotic scenario in which the number of covariates \(p\) is allowed to grow with the sample size \(n\). To avoid triviality, we assume \(n, p \geq 2\) throughout. Similarly, we assume \(q \geq 1\) without further mention.

The conditions we impose below are formulated in terms of the marginal distribution of the covariate vectors \(X_1, \ldots, X_n\) and pertain to the tail behavior of the entries of \(X_i\) as well as the possible dependences among them. We will show that our conditions entail that, with large probability, the covariates satisfy deterministic Hessian conditions that Luo and Chen (2013) used to establish consistency properties of BIC for generalized linear models with fixed design.

These conditions concern sparse submodels of our logistic regression model given by support sets \(J \subseteq [p]\). When treating such submodels, we write \(s_J(\beta)\) and \(H_J(\beta)\) for the subvector and submatrix of \(s(\beta)\) and \(H(\beta)\), respectively, obtained by extracting entries indexed by \(J\).

### 2.2. Hessian conditions when covariates are random

Luo and Chen (2013) invoke conditions on a sequence of deterministic designs to control the curvature and change of the Hessian of the log-likelihood function. Specifically, the eigenvalues of \( \frac{1}{n} H_J(\beta_0) \) for all sparse \( J \supseteq J_0 \) are assumed to be bounded above and below, and furthermore for any \( \epsilon \), there is a \( \delta > 0 \) such that
\[ (1 - \epsilon) H_J(\beta_0) \preceq H_J(\beta) \preceq (1 + \epsilon) H_J(\beta_0), \] (2.2)
for all \( J \supseteq J_0 \) and \( \beta \in \mathbb{R}^J \) with \( \|\beta - \beta_0\|_2 \leq \delta \). These conditions are assumed to hold uniformly for all large enough sample sizes \(n\) and associated values of \(p\), \(q\), and \(\beta_0\), which may change with \(n\).

In this work, we begin instead with random and i.i.d. covariates and derive stronger versions of these Hessian conditions from the below conditions on the distribution of the covariates \(X_1, \ldots, X_n\). We call a vector \(u \in \mathbb{R}^p\) is \(q\)-sparse if \(|\text{supp}(u)| \leq q\), and we recall that a random variable \(Z\) is \(\sigma\)-subgaussian if, for all \( t \in \mathbb{R} \),
\[ \mathbb{E}[e^{tZ}] \leq e^{t^2\sigma^2/2}. \]

Let \(a_1, a_2, a_3 > 0\) be constants that are fixed throughout the remainder of this section. Using \(X_1 = (X_{11}, \ldots, X_{1p})^\top\) as a representative, we will say that the i.i.d. covariates satisfy assumptions (A1)-(A3) with respect to an integer \(q \geq 1\) if the following holds:
(A1) For any $q$-sparse unit vector $u$, $E[(X_1^T u)^2] \geq a_1$.
(A2) For any $q$-sparse unit vector $u$, $E[|X_1^T u|^3] \leq a_2$.
(A3) For each $j \in [p]$, the product $X_{1j} W$ is $a_3$-subgaussian, where $W \in \{ \pm 1 \}$ is a Rademacher random variable independent of $X_{1j}$.

Rephrased, (A1) states that for any subset $J \subset [p]$ of cardinality $|J| \leq q$ the smallest eigenvalue of the $J \times J$ matrix $E[X_{1j}X_{1j}^T]$ is at least $a_1$. Assumption (A2) guarantees the existence of third moments of linear combinations of $q$ or less covariates. The subgaussianity in (A3) holds in particular if $X_{1j} - E[X_{1j}]$ is symmetric and $a_3'$-subgaussian, in which case we can take $a_3$ to be the square root of $(a_3')^2 + E[X_{1j}]^2$. It also holds if $|X_{1j}| \leq a_3$ with probability 1.

According to the following theorem, which is proved in part C of the Appendix, our assumptions entail well-behaved Hessians with large probability. To clarify notation, in this theorem and throughout the rest of the paper the norm $\|H\|$ of a matrix $H$ is the spectral norm.

**Theorem 2.1.** There exist constants $c_{\text{sample}}$, $c_{\text{change}}$, $c_{\text{prob}} > 0$, a decreasing function $c_{\text{lower}} : [0, \infty) \to (0, \infty)$ and an increasing function $c_{\text{upper}} : [0, \infty) \to (0, \infty)$, all depending only on $(a_1, a_2, a_3)$, such that the following statement is true. If the covariates satisfy (A1)-(A3) with respect to $q$ and $n \geq c_{\text{sample}} \cdot q^3 \log^3(np)$, then the event that, simultaneously for all $|J| \leq q$ and all $\beta, \beta' \in \mathbb{R}^J$,

$$c_{\text{lower}}(\|\beta\|_2) I_J \leq \frac{1}{n} H_J(\beta) \leq c_{\text{upper}}(\|\beta\|_2) I_J$$

(2.3)

and

$$\frac{1}{n} \|H_J(\beta) - H_J(\beta')\| \leq c_{\text{change}} \cdot \|\beta - \beta'\|_2$$

(2.4)

has probability at least

$$1 - \exp \left\{ - c_{\text{prob}} \cdot \frac{n}{q^3 \log^3(np)} \right\} - \frac{1}{np}.
$$

If the inequalities (2.3) and (2.4) hold and $\beta \in \mathbb{R}^J$ for a set $J \supseteq J_0$, then

$$\frac{1}{n} H_J(\beta) \leq c_{\text{change}} \cdot \|\beta - \beta_0\|_2 \cdot I_J + \frac{1}{n} H_J(\beta_0)$$

$$\leq \left( 1 + \frac{c_{\text{change}}}{c_{\text{lower}}(\|\beta_0\|_2)} \cdot \|\beta - \beta_0\|_2 \right) \frac{1}{n} H_J(\beta_0).$$

With the analogous lower bound,

$$\frac{1}{n} H_J(\beta) \geq \left( 1 - \frac{c_{\text{change}}}{c_{\text{lower}}(\|\beta_0\|_2)} \cdot \|\beta - \beta_0\|_2 \right) \frac{1}{n} H_J(\beta_0),$$

we obtain the following version of the assumption from (2.2).
Proposition 2.1. If the inequalities (2.3) and (2.4) hold, then

\[(1 - \epsilon)H_J(\beta_0) \leq H_J(\beta) \leq (1 + \epsilon)H_J(\beta_0)\]

for all \( J \supseteq J_0 \) and \( \beta \in \mathbb{R}^J \) with

\[\|\beta - \beta_0\|_2 \leq \delta := \epsilon \cdot \frac{c_{\text{lower}}(\|\beta_0\|_2)}{c_{\text{change}}}. \tag{2.5}\]

2.3. Bounds on likelihood ratios from Hessian conditions

The following theorem provides bounds on log-likelihood ratios for sparse models indexed by \( J \) versus the smallest true model indexed by \( J_0 \). The result concerns fixed values for the covariates \( X_1, \ldots, X_n \) that satisfy the Hessian conditions (2.3) and (2.4) from Theorem 2.1. The statement of the result makes reference to constants from Theorem 2.1. We also invoke an upper bound \( a_0 \) on the signal; some control of the norm of \( \beta_0 \) is needed to avoid degeneracy of the conditional distribution of the binary response variable. The proof of Theorem 2.2 is deferred to Appendix B.

Theorem 2.2. Let \( \beta_0 \) be the true parameter with \( J_0 = \text{supp}(\beta_0) \) and \( \|\beta_0\|_2 \leq a_0 \) for a constant \( a_0 > 0 \). Fix \( \epsilon, \nu > 0 \), and condition on the covariates \( X_1, \ldots, X_n \) taking values that satisfy the Hessian conditions (2.3) and (2.4) for all \( J \supseteq J_0 \) with \( |J| \leq 2q \). Then there exist constants \( C_{\text{false}}, C_{\text{dim}}, C_{\text{sample},1}, C_{\text{sample},2} > 0 \), depending only on \( (c_{\text{change}}, c_{\text{lower}}(a_0), c_{\text{upper}}(a_0)) \) and the chosen pair \( (\epsilon, \nu) \), such that if

\[p \geq C_{\text{dim}}, \quad n \geq C_{\text{sample},1} \cdot q^3 \log^3(p) \quad \text{and} \quad n \geq C_{\text{sample},2} \cdot \frac{q \log(p)}{\min_{j \in J_0} |(\beta_0)_j|^2},\]

the following two statements hold simultaneously with conditional probability at least \( 1 - p^{-\nu} \):

(a) For all \( |J| \leq q \) with \( J \supseteq J_0 \),

\[\log L(\hat{\beta}_J) - \log L(\hat{\beta}_{J_0}) \leq (1 + \epsilon)(|J\setminus J_0| + \nu) \log(p).\]

(b) For all \( |J| \leq q \) with \( J \nsubseteq J_0 \),

\[\log L(\hat{\beta}_{J_0}) - \log L(\hat{\beta}_J) \geq C_{\text{false}} n \min_{j \in J_0} |(\beta_0)_j|^2.\]

We remark that the proof of claim (a) invokes the Hessian conditions only for \( J \supseteq J_0 \) with \( |J| \leq q \). The conditions for cardinality up to \( 2q \) are used for claim (b), which is proved by considering the union of \( J_0 \) and a given set \( J \not\supseteq J_0 \).

2.4. Consistency of extended BIC in logistic regression

Having established bounds on Hessian and likelihood ratios via Theorem 2.1 and Theorem 2.2, respectively, we are able to give conditions that entail that \( \text{BIC}_\gamma \) selects the most parsimonious true model with high probability.
Theorem 2.3. Let $\beta_0$ be the true parameter with $J_0 = \text{supp}(\beta_0)$ and $\|\beta_0\|_2 \leq a_0$ for a constant $a_0 > 0$. Fix $\gamma \geq 0$ and $\epsilon, \nu > 0$. Then there exist constants $C_0, C_1, C_2, C_3 > 0$, depending only on $(a_0, a_1, a_2, a_3)$ and $(\epsilon, \nu)$, such that if the covariates satisfy (A1)-(A3) with respect to $2q$ for $q \geq |J_0|$, if

$$p \geq C_0, \quad n \geq \max \left\{ C_1 \cdot q^3 \log^3(np), C_2 \cdot \frac{q \log(np^2 \gamma)}{\min_{j \in J_0} \|\hat{\beta}_0\|_j^2} \right\},$$

and if

$$\sqrt{n} > p^{(1+\epsilon)(1+\nu)-\gamma},$$

then the event that

$$J_0 = \arg\min\{\text{BIC}_\gamma(J) : J \subseteq [p], |J| \leq q\}$$

has probability at least

$$(1 - \exp \left\{ - C_3 \cdot \frac{n}{q^3 \log^3(np)} \right\} - \frac{1}{np}) \left(1 - \frac{1}{p^\nu}\right).$$

Proof. Choosing the constants $C_0, C_1, C_2, C_3$ suitably, Theorem 2.1 and Theorem 2.2 become applicable and imply that, with the claimed probability, the following statement is true simultaneously for all $|J| \leq q$:

$$\log L(\hat{\beta}_J) - \log L(\hat{\beta}_{J_0}) \leq \begin{cases} (1 + \epsilon)(|J| \setminus J_0| + \nu) \log(p) & \text{if } J \supseteq J_0, \\ -C_{\text{false}} n \min_{j \in J_0} \|\hat{\beta}_0\|_j^2 & \text{if } J \not\supseteq J_0, \end{cases}$$

(2.7)

where $C_{\text{false}} > 0$ is a constant from Theorem 2.2. Condition on (2.7) being true for all $|J| \leq q$. We claim that under our assumptions

$$\text{BIC}_\gamma(J) - \text{BIC}_\gamma(J_0) = -2 \left( \log L(\hat{\beta}_J) - \log L(\hat{\beta}_{J_0}) \right) + (|J| - |J_0|)(\log(n) + 2\gamma \log(p))$$

is positive for any model given by a set $J \not= J_0$ of cardinality $|J| \leq q$.

If $J \not\supseteq J_0$, that is, if the model is false, then (2.7) yields the bound

$$\text{BIC}_\gamma(J) - \text{BIC}_\gamma(J_0) \geq 2C_{\text{false}} n \min_{j \in J_0} \|\hat{\beta}_0\|_j^2 - q \log(np^2 \gamma).$$

For $C_2$ large enough, the lower bound is positive.

For $J \supseteq J_0$ with $|J| \leq q$, we have

$$\text{BIC}_\gamma(J) - \text{BIC}_\gamma(J_0) \geq -(1 + \epsilon)(|J| \setminus J_0| + \nu) \log(p) + |J \setminus J_0|(\log(n) + 2\gamma \log(p)),$$

which can be lower-bounded further as

$$\text{BIC}_\gamma(J) - \text{BIC}_\gamma(J_0) \geq |J \setminus J_0| \cdot (\log(n) + 2[\gamma - (1 + \epsilon)(1 + \nu)] \log(p)) > 0.$$

This is positive by the assumed inequality from (2.6). \hfill \Box
Based on Theorem 2.3, we can identify asymptotic scenarios under which BIC\(\gamma\) yields consistent variable selection. To this end, consider a sequence of variable selection problems indexed by the sample size \(n\), where the \(n\)-th problem has \(p_n\) covariates and true parameter \(\beta_0(n)\) with support \(J_0(n)\). Let \(q_n\) be the bound on the size of the considered models, and let

\[
\beta_{\text{min}}(n) = \min_{j \in J_0(n)} |\beta_0(n)_j|
\]

be the smallest absolute value of any non-zero coefficient in \(\beta_0(n)\).

**Corollary 2.1.** Suppose that \(p_n = n^\kappa\) for \(\kappa > 0\), that \(q_n = n^\psi\) for \(0 \leq \psi < 1/3\), and that \(\beta_{\text{min}}(n) = n^{-\phi/2}\) for \(0 \leq \phi < 1 - \psi\). Assume that the covariates satisfy (A1)-(A3) with respect to \(2q_n\) for some constants \(a_1, a_2, a_3 > 0\), and that \(|J_0(n)| \leq q_n\) and \(\|\beta_0(n)\|_2 \leq a_0\) for a constant \(a_0 > 0\). Then for any \(\gamma > 1 - \frac{1}{2\kappa}\), variable selection with BIC\(\gamma\) is consistent in the sense that the event

\[
J_0(n) = \arg \min \{\text{BIC}_{\gamma}(J) : J \subset [p_n], |J| \leq q_n\}
\]

has probability tending to one as \(n \to \infty\).

**Proof.** Since \(p_n = n^\kappa\), condition (2.6) in Theorem 2.3 holds for all \(n\) if

\[
\frac{1}{2\kappa} > (1 + \epsilon)(1 + \nu) - \gamma.
\]

Having assumed \(\gamma > 1 - \frac{1}{2\kappa}\) here, the condition is satisfied for \(\epsilon\) and \(\nu\) sufficiently small. Fix a suitable choice of \((\epsilon, \nu)\) for the rest of the argument.

Our scaling assumptions for \(p_n, q_n\) and \(\beta_{\text{min}}(n)\) are such that the conditions involving the constants \(C_0, C_1\) and \(C_2\) in Theorem 2.3 are met for \(n\) large enough. Hence, Theorem 2.3 applies for all large \(n\). And, as \(n \to \infty\), the probability in Theorem 2.3 tends to one. \(\square\)

**Remark 2.1.** Corollary 2.1 assumes \(p_n\) to grow polynomially with \(n\). An analogous consistency result could be stated for faster subexponential growth of \(p_n\) with \(n\) by choosing \(\gamma > 1\) and ensuring that \(q_n^3 \log^3(np_n) = o(n)\).

### 3. Consistency of extended BIC for Ising models

We now turn to neighborhood selection for Ising models. So consider an i.i.d. sample \(Z_1, \ldots, Z_n\), where each \(Z_i = (Z_i_1, \ldots, Z_{ip})\) is a vector of binary random variables with values in \([-1, 1]\) that follows an Ising model as in (1.1), with graph \(G = (V, E)\) on the vertex set \(V = [p]\), and interaction parameters \(\theta_{vw} \in \mathbb{R}\) for \((v, w) \in E\). We will consider selection of \(G\) by means of variable selection in the logistic regression models, where the \(v\)-th regression problem has response variable \(Z_v\) and \(p - 1\) covariates \(Z_w, w \neq v\). We write BIC\(\gamma(J, v)\) for the BIC score from (1.3) evaluated for the logistic regression model with response \(Z_v\) and covariates \(Z_w, w \in J, w \neq J\). Correct inference of \(G\) is achieved if, for each \(v \in [p]\), the neighborhood

\[
\text{ne}(v) = \{w \in [p] \setminus \{v\} : \theta_{vw} \neq 0\},
\]
(uniquely) minimizes $\text{BIC}_\gamma(\cdot, v)$.

Using $Z_1 = (Z_{11}, \ldots, Z_{1p})^T$ as a representative, we will say that $Z_1, \ldots, Z_n$ satisfy assumptions (B1)-(B3) with respect to an integer $q \geq 1$ if the following holds for fixed constants $b_0, b_1, b_2 > 0$:

(B1) The interaction parameters $\theta_{vw}$ are bounded in absolute value as $|\theta_{vw}| \leq b_0$ for all $\{v, w\} \in E$.

(B2) For any $q$-sparse unit vector $u$, $E \left[ (Z_1^T u)^2 \right] \geq b_1$.

(B3) For any $q$-sparse unit vector $u$, $E \left[ |Z_1^T u|^3 \right] \leq b_2$.

As explained in Santhanam and Wainwright (2012), the graph selection problem is ill-posed without some upper bound on the absolute values of the interaction parameters as in (B1). Assumption (B2) constitutes a lower bound on the eigenvalues of the $q \times q$ principal submatrices of the covariance matrix $E [Z_1 Z_1^T]$ and akin to requirements in Ravikumar, Wainwright and Lafferty (2010); Loh and Wainwright (2014). As we clarify at the end of this section, condition (B2) is implied by (B1) for asymptotic scenarios in which all neighborhoods $\text{ne}(v)$ have cardinality bounded by a constant, that is, the graph $G$ has bounded degree. Assumption (B3) is the final piece needed to invoke our result on general logistic regression.

To formulate a consistency result for neighborhood selection in Ising models, we consider a sequence of neighborhood selection problems indexed by the sample size $n$. The $n$-th problem has $p_n$ variables and interaction parameters $\theta_{vw}(n)$, with associated neighborhoods $\text{ne}_n(v)$ and edge set $E(n)$. Let $d_n$ be the maximum cardinality of any neighborhood $\text{ne}_n(v)$, $v \in [p_n]$, and let

$$\theta_{\text{min}}(n) = \min_{\{v, w\} \in E(n)} |\theta_{vw}(n)|$$

be the non-zero interaction of smallest magnitude.

**Theorem 3.1.** Suppose that $p_n = n^\kappa$ for $\kappa > 0$, that $q_n = n^\psi$ for $0 \leq \psi < 1/3$, and that $\theta_{\text{min}}(n) = n^{-\phi/2}$ for $0 \leq \phi < 1 - \psi$. Assume that the sample $Z_1, \ldots, Z_n$ satisfies (B1)-(B3) with respect to $2q_n$ and that $d_n \leq q_n$. Then for any $\gamma > 2 - \frac{1}{2\kappa}$, Ising neighborhood selection with $\text{BIC}_\gamma$ is consistent in the sense that the event that, simultaneously for all $v \in [p_n]$,

$$\text{ne}_n(v) = \arg \min \{ \text{BIC}_\gamma(J, v) : J \subset [p_n] \setminus \{v\}, |J| \leq q_n \}$$

has probability tending to one as $n \to \infty$.

**Proof.** We will show that the result follows from Theorem 2.3 together with a union bound over the $p_n$ logistic regression problems.

First, we observe that with $p_n = n^\kappa$, condition (2.6) in Theorem 2.3 holds for all $n$ if

$$\frac{1}{2\kappa} > (1 + \epsilon)(1 + \nu) - \gamma.$$
Having assumed $\gamma > 2 - \frac{1}{2\nu}$ here, the condition can be satisfied with a choice of $\epsilon > 0$ and $\nu > 1$. We fix such a choice of $(\epsilon, \nu)$ for the rest of the argument.

Next, note that Theorem 2.3 is applicable to each one of the $p_n$ logistic regression problems in neighborhood selection. Indeed, since $Z_1, \ldots, Z_n$ are bounded assumption (A3) holds. Conditions (A1) and (A2) are ensured by (B2) and (B3), respectively, and (B1) yields the bounded signal assumed in Theorem 2.3. Moreover, the scaling assumptions on $p_n$, $q_n$ and $\theta_{\min}(n)$ are such that the assumptions on the corresponding quantities in Theorem 2.3 are met.

Applying Theorem 2.3 a total of $p_n$ times, we obtain that, separately for each $v \in [p_n]$, the event that

$$\text{ne}_n(v) = \arg \min \{ \text{BIC}_\gamma(J,v) : J \subset [p_n] \setminus \{v\}, |J| \leq q_n \}$$

occurs with at least the probability from Theorem 2.3. Ignoring smaller terms of higher order in $1/p_n$, this probability is

$$1 - \frac{1}{np_n} - \frac{1}{p_n^\nu}.$$ 

Since $\nu > 1$, we have that

$$p_n \cdot \left( \frac{1}{np_n} + \frac{1}{p_n^\nu} \right) \to 0$$

as $n$, and thus also $p_n$, tends to infinity. Hence, a union bound yields the desired claim that all events hold simultaneously with probability tending to one.

Finally, we observe that conditions (B2) and (B3) do not present a restriction when considering problems in which there is a fixed bound on the degree of the graph underlying the Ising model and a bound on the interaction parameters as in (B1). Indeed, (B3) holds trivially in this case since the coordinate of the random vectors are bounded by one in absolute value. The sparse eigenvalue condition (B2) is addressed in the next lemma.

**Lemma 3.1.** Suppose the random vector $Z = (Z_1, \ldots, Z_p)$ follows an Ising model with $|\text{ne}(v)| \leq q$ for all $v \in [p]$. If the interaction parameters $\theta_{vw}$ for $Z$ satisfy (B1) then it holds for any $q$-sparse unit vector $u$ that

$$\mathbb{E} [(Z^\top u)^2] \geq \frac{4}{q} \cdot \frac{e^{2q\theta_0}}{(1 + e^{2q\theta_0})^2}.$$ 

**Proof.** Without loss of generality, we consider a $q$-sparse unit vector $u$ that has $\text{supp}(u) = \{1, \ldots, q\}$ and $|u_1| \geq |u_2| \geq \cdots \geq |u_q|$. Then $u_1^2 \geq 1/q$. Let $Z_{-1} = (Z_2, \ldots, Z_p)^\top$. For a random variable $X$ with finite variance,

$$\text{Var}[X] = \min_{a \in \mathbb{R}} \mathbb{E} [(X - a)^2].$$
Therefore,

\[ E \left[ (Z^\top u)^2 \mid Z_{-1} \right] \geq \text{Var} \left[ Z_1 u_1 \mid Z_{-1} \right] \geq \frac{1}{q} \text{Var} \left[ Z_1 \mid Z_{-1} \right]. \]

Since \( Z_1 \) takes values in \( \{-1, 1\} \), we rescale to \((Z_1 + 1)/2\) for values in \( \{0, 1\} \). Then the conditional distribution of \((Z_1 + 1)/2\) given \( Z_{-1} \) is a Bernoulli distribution with success probability

\[ \frac{\exp \left\{ 2 \sum_{w \in \text{ne}(1)} \theta_{1w} Z_w \right\}}{1 + \exp \left\{ 2 \sum_{w \in \text{ne}(1)} \theta_{1w} Z_w \right\}}; \]

recall (1.2). We obtain that

\[ \text{Var} \left[ Z_1 \mid Z_{-1} \right] = 4 \text{Var} \left[ (Z_1 + 1)/2 \mid Z_{-1} \right] = \frac{4 \exp \left\{ 2 \sum_{w \in \text{ne}(1)} \theta_{1w} Z_w \right\}}{(1 + \exp \left\{ 2 \sum_{w \in \text{ne}(1)} \theta_{1w} Z_w \right\})^2}. \]

By assumption (B1),

\[ -qb_0 \leq \sum_{w \in \text{ne}(1)} \theta_{1w} Z_w \leq qb_0. \]

It follows that

\[ E \left[ (Z^\top u)^2 \right] = E \left[ E \left[ (Z^\top u)^2 \mid Z_{-1} \right] \right] \geq \frac{4}{q} \cdot \frac{e^{2qb_0}}{(1 + e^{2qb_0})^2}. \]

4. Practical considerations when applying information criteria

Theorem 3.1 shows that, with sufficient data, application of BIC\(_\gamma\) allows one to identify the correct set of edges, simultaneously at each node, with high probability. Application of the information criterion in practice, however, faces two issues:

(i) At an individual node, in order to find the sparse model that minimizes BIC\(_\gamma\), we must fit a large number of models. With sparsity bounded by \( q \), there are on the order of \( p^q \) models, preventing an exhaustive search when the number of variables \( p \) is larger.

(ii) After performing neighborhood selection for each node, our results may be asymmetrical, that is, we might find that our estimates of the coefficients in (1.2) satisfy \( \hat{\beta}_{vw} \neq 0 \) but \( \hat{\beta}_{wv} = 0 \) for some pair of nodes \( v, w \).

To resolve the issue of the large number of possible models at each node, it is common to use a computationally efficient procedure to first produce a short list of candidate models, and then apply BIC\(_\gamma\) to select from this list. We use an
ℓ₁-penalized likelihood (the ‘logistic Lasso’) with varying levels of penalization ρ to produce the candidate models:

$$\hat{\beta}_v(\rho) = \arg\min_{\beta \in \mathbb{R}^{V \setminus \{v\}}} \left\{ -\sum_{i=1}^n \log P\{Z_{iw} \mid \{Z_{iw} : w \neq v\}\} + \rho \|\beta\|_1 \right\}. \quad (4.1)$$

To account for potential asymmetries when we compile information across nodes, we follow the work of Meinshausen and Bühlmann (2006) and draw an edge connecting nodes v and w based on either an AND rule (requiring both $\hat{\beta}_{vw} \neq 0$ and $\hat{\beta}_{wv} \neq 0$) or an OR rule (requiring $\hat{\beta}_{vw} \neq 0$ or $\hat{\beta}_{wv} \neq 0$); recall the discussion from the introduction and, in particular, the empirical study of Höfling and Tibshirani (2009).

5. Experiments: Regional weather patterns

5.1. Data and methods for model selection

We apply BICγ, and other competing methods, to the task of inferring dependencies among binary indicators of precipitation at $p = 92$ weather stations across four states in the Midwest region of the U.S. The four states are Illinois, Indiana, Iowa, and Missouri. We fit models without taking the geographical locations of the 92 stations into account, but then assess the performance of different methods by referring to the distance between weather stations. Our rationale is that plausible graphs should primarily link neighboring stations. (One could argue that longer links in East-West direction might be more reasonable than longer North-South links but it seems difficult to quantify this and we did not attempt to make such refined distinctions.)

The binary variables we consider indicate the existence of precipitation at each station on a given day. We model their joint distribution with an Ising model as in (1.1) such that the precipitation indicator at each node (weather station), conditional on the observations from the other nodes, follows the logistic regression model from (1.2). We compute a set of (sparse) candidate models for each node using the logistic Lasso and then select a model from the set using either the ordinary BIC = BIC₀, the modification to BICγ with $\gamma = 0.25$ or $\gamma = 0.5$, as well as 10-fold cross-validation and stability selection as in Meinshausen and Bühlmann (2010). For each method, the node-wise edge selections are compiled across all nodes to form a graph. Performance is measured relative to the true geographical layout of the weather stations, which as mentioned above is “unknown” to the procedures we compare.

To give more specifics, we used data from the United States Historical Climatology Network (Menne, Williams Jr. and Vose, 2011).¹ The data consists of weather-related variables that were recorded on a daily basis. We specifically gathered the precipitation data, which gives the total amount of precipitation for each day. Seasonality effects on precipitation are not as pronounced in the

¹Available at http://cdiac.ornl.gov/ftp/ushcn_daily/
Fig 1. Delaunay triangulation for 92 weather stations in Illinois, Indiana, Iowa, and Missouri.

Midwest as in other parts of the U.S., and we thus simply consider data from the entire year. However, to limit the effects of temporal dependencies between successive observations, we took data from only the 1st and 16th day of each month. The resulting multivariate observations are then treated as independent. We removed weather stations where data availability was low and discarded observations with missing values for any of the remaining weather stations. A total of \( n = 370 \) days and \( p = 92 \) stations remained in the final data set. Figure 1 shows a map of the 92 stations, along with an undirected graph representing the Delaunay triangulation of the 92 locations.

For each weather station \( j \), we define binary variables \( Z_{ij} \) taking values 1 or 0 depending on whether or not there was a positive amount of rainfall at weather station \( j \) on day \( i \). For each one of the stations \( j \), we then used each of the above mentioned five methods to perform a sparse logistic regression that has response vector \( Z_{\cdot j} \) and covariates \( \{Z_{\cdot k} : k \neq j\} \). We applied the logistic Lasso (4.1) with a range of 100 penalty-parameter values to the data, using the \texttt{glmnet} package for \texttt{R} (Friedman, Hastie and Tibshirani, 2010; R Core Team, 2013). This produced a list of 100 (not necessarily distinct) support sets, \( J_1, \ldots, J_{100} \). When applying the information criteria, we refitted each candidate model without \( \ell_1 \)-penalization using the function \texttt{glm} in \texttt{R}, and applied BIC\( _{\gamma} \) with \( \gamma = 0, 0.25, 0.5 \) to each candidate model, optimizing the resulting scores to select a single model for each choice of \( \gamma \). We also applied 10-fold cross-validation, selecting the model that minimizes average error on the test sets over the 10 folds. Finally, for stability selection, we used the \texttt{stabsel} function in the \texttt{mboost} package for \texttt{R} (Hothorn et al., 2013), setting the expected support size to 10. As noted by Meinshausen and Bühlmann (2010), changing this setting within a reasonable range did not have a large effect on the output.
Table 1

| Method           | AND rule | OR rule |
|------------------|----------|---------|
|                  | PSR      | FDR     | PSR      | FDR     |
| BIC_0            | 41.98    | 32.93   | 55.73    | 46.72   |
| BIC_0.25         | 37.4     | 27.94   | 52.67    | 42.02   |
| BIC_0.5          | 34.73    | 26.61   | 50.38    | 38.89   |
| Cross-validation | 69.08    | 75.83   | 79.39    | 85.05   |
| Stability selection | 30.53  | 32.77   | 45.42    | 64.26   |

5.2. Results

To evaluate the model selection methods, we first compare the inferred graphs to
the geographic layout of the 92 stations by treating the Delaunay triangulation
as a “true” underlying graph for the considered Ising model. Table 1 shows
the results we obtain for each method, stated in terms of positive selection
rate (PSR) and false discovery rate (FDR), relative to the “true” Delaunay
triangulation graph. The results are also displayed in Figure 2, and Figure 3
shows the recovered graphs under the AND and OR combination rules.

We see that cross-validation leads to a somewhat higher PSR than the other
methods, under either an AND or an OR rule. However, this comes at the cost
of a drastically higher FDR. For BIC_γ, as we increase γ, we reduce the FDR at
a cost of a lower PSR, as expected. Stability selection does not perform as well
as the information criteria under this evaluation, with higher FDR and slightly
lower PSR than BIC_0, but was substantially more computationally expensive.

While it does not seem unreasonable to assume that the edges of the Delau-
nay triangulation capture most of the strongest dependencies, there might be
additional dependencies that are not captured by the edges in the triangulation.
For a different comparison of the methods that more directly uses the geographic
distances between the weather stations, we apply Gaussian smoothing (scale:
standard deviation = 10 miles) to estimate, as a function of d, the probability
that a method will infer an edge between two nodes that are d miles apart.
The resulting functions are plotted in Figure 4, which also includes the same
smoothed function calculation for the graph from the Delaunay triangulation.

We observe that the smoothed function for the cross-validation methods (un-
der either the OR or the AND rule) does not decay to zero as distance increases.
That is, in this experiment, cross-validation selects a nonnegligible proportion
of edges between nodes that are arbitrarily far apart, which is undesirable. To a
lesser extent, the same problem occurs for stability selection combined with the
OR rule for mid-range distances. The other methods, in contrast, yield functions
that do decay to zero relatively quickly as distance increases. Comparing the
methods that show the decay to zero, we see also that for two nearby weather
stations, the BIC_γ methods combined with the OR rule are more likely to select
an edge than any of the remaining methods. Overall, we find that the informa-
tion criteria perform well while requiring the least amount of computation, and
increasing γ provides a useful trade-off between PSR and FDR.
6. Discussion

As suggested by our numerical experiments and supported by our theoretical analysis, Bayesian information criteria extended to include a penalization term involving the number of covariates are useful tools for variable selection in logistic regression as well as neighborhood selection for Ising models. The additional penalty term can be motivated via a particular class of prior distributions on the set of considered models. We aim to discuss the formal connection between fully Bayesian approaches and $\text{BIC}_\gamma$ in a subsequent paper; preliminary results under bounded sparsity are described in the Ph.D. thesis of the first author (Foygel, 2012) and in a preprint (Foygel and Drton, 2011).

At the heart of this paper is an analysis of logistic regression with random covariates. While logistic regression has special properties, our technical results can be extended to other generalized linear models. The main challenge for such generalizations is control of the third derivative of the cumulant function which might no longer be bounded. Preliminary results under bounded sparsity can again be found in Foygel (2012) and Foygel and Drton (2011).
Fig 3. Graphs recovered under each method. (Black edges indicate true positives, red edges indicate false positives, and light gray edges indicate false negatives, i.e. true edges that were not recovered by the method, where the true graph is defined via the Delaunay triangulation.)

Acknowledgments

This work was supported by the U.S. National Science Foundation under Grant No. DMS-1305154. We would like to thank Kean Ming Tan for very helpful comments on draft versions of this paper.

Appendix A: Why are second moments not sufficient?

Returning to the setup of Section 2, we recall that our results on general logistic regression rely on assumption (A2), which places an upper bound on third moments. In contrast, the lower bound in assumption (A1) concerns second moments (or, put differently, eigenvalues of small submatrices of the covariance matrix). It is tempting to try and weaken our condition (A2) to a sparse eigenvalue upper bound:

(A2') For any q-sparse unit vector \( u \), \( \mathbb{E} \left[ (X_1^T u)^2 \right] \leq a_2' \).

However, we now show that this assumption will not be sufficient to get the desired results in any asymptotic scenario where \( q \) grows with \( n \), no matter how slow the rate of growth is assumed to be. In particular, we construct an example where, even though sparse eigenvalues are bounded above and below, the Hessian conditions assumed by Luo and Chen (2013) do not hold at \( \beta_0 = 0 \) (i.e. \( J_0 = \emptyset \)), recall (2.2).

For simplicity, let \( p = q \), and let \( Z \) be a random vector that follows a uniform distribution on \( \{ \pm 1 \}^q \). Let \( 1_q = (1, \ldots, 1)^T \). Then define a random vector \( X \) by
Fig 4. Smoothed probability of selecting edges as a function of distance, for each method under the OR and AND rule.

setting

\[ X = \begin{cases} 
  1_q \quad \text{with prob. } \frac{1}{2q}, \\
  -1_q \quad \text{with prob. } \frac{1}{2q}, \\
  Z \quad \text{with prob. } 1 - \frac{1}{q}.
\end{cases} \]

Clearly, \( E[Z] = E[X] = 0 \), and \( E[ZZ^\top] = I_q \). Therefore,

\[ E[XX^\top] = \frac{1}{q} \cdot 1_q 1_q^\top + \left(1 - \frac{1}{q}\right) \cdot I_q, \]

has minimal and maximal eigenvalue equal to

\[ \lambda_{\text{min}}(E[XX^\top]) = 1 - \frac{1}{q}, \quad \lambda_{\text{max}}(E[XX^\top]) = 2 - \frac{1}{q}, \]

respectively. We observe that the eigenvalues are bounded above and below as required by (A1) and (A2’).

Now take the unit vector \( u = \frac{1}{\sqrt{q}} 1_q \). We see that \( |X^\top u| = \sqrt{q} \) with probability at least \( 1/q \). For independent random vectors \( X_1, \ldots, X_n \) that all have the same distribution as \( X \), it follows that

\[ \sum_{i=1}^{n} (X_i^\top u)^2 \cdot 1_{\{|X_i^\top u| \geq \sqrt{q}\}} \geq q \cdot \sum_{i=1}^{n} 1_{\{|X_i^\top u| = \sqrt{q}\}} \geq q \cdot \text{Binomial}\left(n, \frac{1}{q}\right). \]

Assume for simplicity that \( n/q \) is an integer. Then \( n/q \) is the median of the Binomial\((n, 1/q)\) distribution, and so with probability at least \( 1/2 \),

\[ \sum_{i=1}^{n} (X_i^\top u)^2 \cdot 1_{\{|X_i^\top u| \geq \sqrt{q}\}} \geq n. \]
In the remainder of this section, we prove that this property contradicts the inequalities in (2.2) when assuming an upper bound on the eigenvalues of the Hessian $\frac{1}{n}H(0)$. (Note that since we have simplified the problem by setting $p = q$, we do not need to make reference to submatrices of $H(0)$.)

Let $\delta \leq 1/\sqrt{q}$. Since $b''(0) \geq b''(z)$ for all $z \in \mathbb{R}$, we have

$$u^\top (H(0) - H(\beta)) u = \sum_{i=1}^{n} (X_i^\top u)^2 (b''(0) - b''(X_i^\top \beta)) \geq \sum_{i: |X_i^\top u| \geq \delta^{-1}} (X_i^\top u)^2 (b''(0) - b''(X_i^\top \beta)).$$

Since $b''(z) \leq b''(1)$ for all $|z| \geq 1$, we may further bound this to get

$$u^\top (H(0) - H(\beta)) u \geq \sum_{i: |X_i^\top u| \geq \delta^{-1}} (X_i^\top u)^2 (b''(0) - b''(1)).$$

Since $\delta^{-1} \geq \sqrt{q}$, we deduce that it holds with probability at least $1/2$ that

$$u^\top (H(0) - H(\beta)) u \geq n(b''(0) - b''(1)).$$

Now, in accordance with the conditions used by Luo and Chen (2013), suppose that (2.2) holds and that the Hessian is bounded from above as $H(0) \preceq n \cdot c^2_l I_q$. Then for any choice of $\epsilon > 0$, we require that there exists some $\delta > 0$ such that

$$H(\beta) \succeq (1 - \epsilon)H(0) \succeq H(0) - n \cdot \epsilon c^2_l I_q,$$

with high probability, for all $\beta \in \mathbb{R}^q$ with $\|\beta\|_2 \leq \delta$. In particular, this implies that for the vector $u$ chosen above, with high probability, for all $\beta \in \mathbb{R}^J$ with $\|\beta\|_2 \leq \delta$,

$$n \cdot \epsilon c^2_l \geq u^\top (H(0) - H(\beta)) u.$$

In particular, this must be true for all $\|\beta\|_2 \leq \min\{\delta, \frac{1}{\sqrt{q}}\}$. But from the work above, the bound can only be true if

$$\epsilon c^2_l \geq (b''(0) - b''(1)),$$

which is false for $\epsilon$ small enough.

**Appendix B: Proofs for likelihood and score results**

This appendix is devoted to the proof of Theorem 2.2, which gives bounds on likelihood ratios for models postulating sparsity in the coefficient vector $\beta$. The bounds are for fixed values of the covariates $X_1, \ldots, X_n$ that satisfy the Hessian conditions from Theorem 2.1. All probability statements in this section are tacitly understood to be conditional on $X_1, \ldots, X_n$. 

B.1. Bounding the score function

In this section, we prove bounds on the score function at the true parameter \( \beta_0 \) that hold with high probability. These bounds concern the score function of true sparse models given by sets \( J \supseteq J_0 \) with \( |J| \leq q \).

Let \( \epsilon' < \epsilon \) be a positive value that will be specified later. For integer \( r \geq 1 \), let \( \tau, \tilde{\tau} > 0 \) be the square roots of

\[
\tau_r^2 := \frac{2}{(1-\epsilon')^3} \left( (|J_0| + r) \log \left( \frac{3}{\epsilon'} \right) + \log(4p^r) + r \log(2p) \right)
\]

and

\[
\tilde{\tau}_r^2 := \frac{2}{(1-\epsilon')^3} \left( r \log \left( \frac{3}{\epsilon'} \right) + \log(4p^r) + r \log(2p) \right),
\]

respectively. Assume that

\[
\tau_r \leq \sqrt{\frac{n \cchange(\|\beta_0\|_2)^3}{(1-\epsilon')^3}}
\]

for \( r \leq q - |J_0| \). This assumption can be guaranteed to hold by choosing \( C_{\text{sample},1} \) in the statement of Theorem 2.2 appropriately.

**Lemma B.1.** Fix values for the observations \( X_1, \ldots, X_n \) that satisfy the Hessian conditions (2.3) and (2.4) from Theorem 2.1. Assume further that the inequality in (B.1) holds. Then with conditional probability at least \( 1 - p - \nu \), we have for all \( J \supseteq J_0 \) with \( |J| \leq q \) that both

\[
\left\| H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) \right\|_2 \leq \tau_{|J \setminus J_0|} \tag{B.2}
\]

and

\[
\left\| \Proj_{S_J^J} \left( H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) \right) \right\|_2 \leq \tilde{\tau}_{|J \setminus J_0|}, \tag{B.3}
\]

where the projection is on the orthogonal complement of the subspace

\[
S_J = \left\{ H_J(\beta_0)^{\frac{1}{2}} z : z \in \mathbb{R}^{J_0} \right\} \subset \mathbb{R}^J.
\]

To be clear, in the definition of \( S_J \), we use \( \mathbb{R}^{J_0} \) to denote the coordinate subspace of vectors \( z \in \mathbb{R}^J \) with \( z_j = 0 \) for all \( j \in J \setminus J_0 \).

**Proof.** We will establish the bounds in (B.2) and (B.3) by using an \( \epsilon \)-net argument based on the fact that for any vector \( z \in \mathbb{R}^p \),

\[
\|z\|_2 = \sup \left\{ u^T z : u \in \mathbb{R}^p, \|u\|_2 = 1 \right\}.
\]

To prepare for the argument, fix a superset \( J \supseteq J_0 \), a vector \( u \in \mathbb{R}^J \), and a scalar \( \tau > 0 \). Observe that

\[
\Prob \left\{ u^T H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) > \tau | X \right\} 
\]

\[
\leq \mathbb{E} \left[ \exp \left\{ \tau \cdot u^T H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) - \tau^2 \right\} | X \right]. \tag{B.4}
\]
By definition,
\[ s_J(\beta_0) = \sum_{i=1}^n X_i J(Y_i - b'(X_i^T \beta_0)), \quad (B.5) \]
and since the conditional distribution of \( Y_i \) given \( X_i \) belongs to an exponential family, we have
\[ \mathbb{E} \{ \exp \{ s Y_i \} | X_i \} = \exp \{ b(X_i^T \beta_0 + s) - b(X_i^T \beta_0) \}. \quad (B.6) \]
Plugging (B.5) into (B.4) and using (B.6), we obtain that
\[
\log \text{Prob} \left\{ \left| u^T H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) > \tau \right| X \right\} \\
\leq \sum_{i=1}^n \left[ b \left( X_i^T (\beta_0 + \tau H_J(\beta_0)^{-\frac{1}{2}} u) \right) - b(X_i^T \beta_0) \right] \\
- \sum_{i=1}^n \left[ b'(X_i^T \beta_0) \cdot \tau X_i J H_J(\beta_0)^{-\frac{1}{2}} u \right] - \tau^2 \\
= \frac{1}{2} \sum_{i=1}^n \left[ b'' \left( X_i^T (\beta_0 + \xi \cdot \tau H_J(\beta_0)^{-\frac{1}{2}} u) \right) \cdot \left( \tau X_i J H_J(\beta_0)^{-\frac{1}{2}} u \right)^2 \right] - \tau^2,
\]
where the last equation is a 2nd-order Taylor expansion with \( \xi \in [0, 1] \). We may rewrite the inequality just obtained as
\[
\log \text{Prob} \left\{ \left| u^T H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) > \tau \right| X \right\} \leq \\
\frac{\tau^2}{2} u^T H_J(\beta_0)^{-\frac{1}{2}} H_J \left( \beta_0 + \xi \cdot \tau H_J(\beta_0)^{-\frac{1}{2}} u \right) H_J(\beta_0)^{-\frac{1}{2}} u - \tau^2.
\]
Now, for \( \tau = \tau'_r := \tau_r (1 - \epsilon') \) with \( r = |J \setminus J_0| \) and a vector \( u \in \mathbb{R}^J \) with \( \| u \|_2 \leq 1 \), it holds that
\[
\| \xi \cdot \tau'_r H_J(\beta_0)^{-\frac{1}{2}} u \|_2 \leq \tau_r (1 - \epsilon') \cdot \sqrt{\frac{1}{n c_{\text{lower}}(\| \beta_0 \|_2)}} \leq \epsilon' \cdot \frac{c_{\text{lower}}(\| \beta_0 \|_2)}{c_{\text{change}}};
\]
recall (B.1). Via (2.2) and (2.5), the assumed Hessian conditions imply that
\[
H_J(\beta_0)^{-\frac{1}{2}} H_J \left( \beta_0 + \xi \cdot \tau'_r H_J(\beta_0)^{-\frac{1}{2}} u \right) H_J(\beta_0)^{-\frac{1}{2}} \\
\leq H_J(\beta_0)^{-\frac{1}{2}} \left[ (1 + \epsilon') H_J(\beta_0) \right] H_J(\beta_0)^{-\frac{1}{2}} = (1 + \epsilon') \cdot I_J,
\]
and thus
\[
\text{Prob} \left\{ \left| u^T H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) > \tau'_r \right| X \right\} \\
\leq \exp \left\{ \frac{\tau'^2}{2} (1 + \epsilon') - \tau'^2 \right\} = \exp \left\{ -\frac{\tau'^2}{2} (1 - \epsilon') \right\}. \quad (B.7)
\]
Next, let $U_J$ be an $\epsilon'$-net for the unit sphere in $\mathbb{R}^J$ with respect to the Euclidean norm, that is, $U_J$ is a subset of the sphere such that for any unit vector $v$ there exists a (unit) vector $u \in U_J$ such that $\|u - v\|_2 < \epsilon'$. In particular, for the unit vector

$$v = \frac{H_J(\beta_0)^{-\frac{1}{2}}s_J(\beta_0)}{\|H_J(\beta_0)^{-\frac{1}{2}}s_J(\beta_0)\|_2}$$

and corresponding $u \in U_J$ with $\|u - v\|_2 \leq \epsilon'$, we see that

$$u^\top v = v^\top v + (u - v)^\top v \geq \|v\|_2^2 - \|u - v\|_2 \cdot \|v\|_2 \geq 1 - \epsilon',$$

and so

$$u^\top H_J(\beta_0)^{-\frac{1}{2}}s_J(\beta_0) \geq (1 - \epsilon') \left\|H_J(\beta_0)^{-\frac{1}{2}}s_J(\beta_0)\right\|_2. \quad (B.8)$$

We can take the $\epsilon$-net such that $|U_J| \leq \left(1 + \frac{2}{\epsilon'}\right)^{|J|} \leq \left(\frac{3}{\epsilon'}\right)^{|J|}; \quad (B.9)$

see Proposition 1.3 in Chapter 15 of Lorentz, Golitschek and Makovoz (1996) or Lemma 14.27 in Bühlmann and van de Geer (2011). Inequality (B.8) and a union bound yield that

$$\text{Prob}\left\{\left\|H_J(\beta_0)^{-\frac{1}{2}}s_J(\beta_0)\right\|_2 > \tau_r\right\} \leq |U_J| \cdot \text{Prob}\left\{u^\top H_J(\beta_0)^{-\frac{1}{2}}s_J(\beta_0) \geq \tau'_r\right\},$$

Applying inequalities (B.7) and (B.9), and plugging in the definition of $\tau'_r$, we obtain that

$$\text{Prob}\left\{\left\|H_J(\beta_0)^{-\frac{1}{2}}s_J(\beta_0)\right\|_2 > \tau_r\right\} \leq \left(\frac{3}{\epsilon'}\right)^{|J|} \cdot \exp\left\{-\frac{\tau'^2_r}{2}(1 - \epsilon')\right\} = \exp\left\{-\log(4p') - r\log(2p)\right\} = \frac{1}{4(2p)^r} \cdot \frac{1}{p'^r}. \quad (B.10)$$

Finally, to consider all sets $J \supseteq J_0$ with $|J| \leq q$ simultaneously, we apply the union bound

$$\text{Prob}\left\{\left\|H_J(\beta_0)^{-\frac{1}{2}}s_J(\beta_0)\right\|_2 > \tau_{|J\setminus J_0|}\right\} \leq \sum_{r=0}^{q-|J_0|} \text{Prob}\left\{\left\|H_J(\beta_0)^{-\frac{1}{2}}s_J(\beta_0)\right\|_2 \geq \tau_r\right\} \text{ for some } J \supseteq J_0, |J| \leq q.$$
Using the fact that there are at most \( \binom{p}{r} \leq \frac{p^r}{r!} \) sets \( J \supseteq J_0 \) with \( |J \setminus J_0| = r \), inequality (B.10) and another union bound imply that

\[
\text{Prob} \left\{ \left\| H_J(\beta_0)^{-\frac{1}{2}} S_J(\beta_0) \right\|_2 > \tau |J \setminus J_0| \text{ for some } J \supseteq J_0, |J| \leq q \right\} \\
\leq \sum_{r=0}^{q-|J_0|} \frac{p^r}{4(2p^r)} \cdot \frac{1}{p^{\nu}} \leq \frac{1}{4p^{\nu}} \sum_{r=0}^{\infty} \frac{1}{2^r} = \frac{1}{2p^\nu}.
\]

To prove the analogous statement about the projection operator, we instead take \( U_J \) to be an \( \epsilon' \)-net of the unit sphere in the orthogonal complement \( S_J^\perp \subset \mathbb{R}^J \), which has dimension \( |J \setminus J_0| \). Consequently, we have \( |U_J| \leq (3/\epsilon')^{|J \setminus J_0|} \). The rest of the argument proceeds identically with a bound of \( 1/(2p^\nu) \) for the probability of both inequalities holding.

**B.2. Bounding the likelihood function**

In this subsection we analyze the log-likelihood ratios of sparse models given by sets \( |J| \leq q \), proving Theorem 2.2. It suffices to show that the two statements (a) and (b) in Theorem 2.2 are implied by the bounds (B.2) and (B.3) from Lemma B.1. The probability of the latter bounds holding was shown to be large in the previous subsection. In our proof we consider a fixed vector \( \beta_0 \). The statement being true uniformly for vectors with \( \|\beta_0\|_2 \) bounded by \( a_0 \) follows from the monotonicity of the functions \( c_{\text{lower}} \) and \( c_{\text{upper}} \).

Fix any \( J \supseteq J_0 \) with \( |J| \leq q \). Consider any \( \beta \in \mathbb{R}^J \) and let \( \gamma = \beta - \beta_0 \). Let

\[
\tilde{\gamma} = H_J(\beta_0)^{-\frac{1}{2}} \cdot \text{Proj}_{S_J} \left( H_J(\beta_0)^{\frac{1}{2}} \gamma \right) \in \mathbb{R}^J,
\]

where \( S_J \subset \mathbb{R}^J \) is the \( |J_0| \)-dimensional subspace defined in Lemma B.1. By definition \( H_J(\beta_0)^{\frac{1}{2}} \tilde{\gamma} = \text{Proj}_{S_J} \left( H_J(\beta_0)^{\frac{1}{2}} \gamma \right) \), and thus

\[
\left\| H_J(\beta_0)^{\frac{1}{2}} \gamma \right\|_2^2 = \left\| H_J(\beta_0)^{\frac{1}{2}} \tilde{\gamma} \right\|_2^2 + \left\| H_J(\beta_0)^{\frac{1}{2}} (\gamma - \tilde{\gamma}) \right\|_2^2.
\]

Using (2.3), we obtain that

\[
\left\| \tilde{\gamma} \right\|_2 \leq \frac{1}{\sqrt{n}c_{\text{lower}}(\|\beta_0\|_2)} \left\| \text{Proj}_{S_J} \left( H_J(\beta_0)^{\frac{1}{2}} \gamma \right) \right\|_2 \\
\leq \frac{1}{\sqrt{n}c_{\text{lower}}(\|\beta_0\|_2)} \left\| H_J(\beta_0)^{\frac{1}{2}} \gamma \right\|_2 \\
\leq \sqrt{\frac{c_{\text{upper}}(\|\beta_0\|_2)}{c_{\text{lower}}(\|\beta_0\|_2)}} \left\| \gamma \right\|_2.
\]
We now compare the values of the log-likelihood function at \( \hat{\beta}_0, \hat{\beta} = \beta_0 + \gamma, \) and \( \hat{\beta}_0 + \tilde{\gamma}, \) using Taylor-expansions. Using Proposition 2.1, we calculate

\[
\log L(\hat{\beta}_0 + \gamma) - \log L(\beta_0) = s_J(\beta_0)^\top \gamma - \frac{1}{2} \gamma^\top H_J(\beta_0 + \xi) \gamma \\
\leq s_J(\beta_0)^\top \gamma - \frac{1}{2} \left( 1 - \frac{c_{\text{change}}}{c_{\text{lower}}(\|\beta_0\|_2)} \cdot \|\gamma\|_2 \right) \gamma^\top H_J(\beta_0) \gamma
\]  
(B.13)

and

\[
\log L(\hat{\beta}_0 + \tilde{\gamma}) - \log L(\beta_0) = s_J(\beta_0)^\top \tilde{\gamma} - \frac{1}{2} \tilde{\gamma}^\top H_J(\hat{\beta}_0 + \xi) \tilde{\gamma} \\
\geq s_J(\beta_0)^\top \tilde{\gamma} - \frac{1}{2} \left( 1 + \frac{c_{\text{change}}}{c_{\text{lower}}(\|\beta_0\|_2)} \cdot \|\tilde{\gamma}\|_2 \right) \tilde{\gamma}^\top H_J(\beta_0) \tilde{\gamma},
\]  
(B.14)

where \( \xi, \tilde{\xi} \in [0, 1]. \) Subtracting (B.14) from (B.13) and using (B.11), we find

\[
\log L(\hat{\beta}_0 + \gamma) - \log L(\beta_0) \leq s_J(\beta_0)^\top (\gamma - \tilde{\gamma}) \\
- \frac{1}{2} \left\| H_J(\beta_0)^{\frac{1}{2}} (\gamma - \tilde{\gamma}) \right\|_2^2 + \frac{c_{\text{change}}}{c_{\text{lower}}(\|\beta_0\|_2)} \frac{2}{\|\beta_0\|_2} \|\gamma\|_2 \|\tilde{\gamma}\|_2.
\]  
(B.15)

Inequalities (2.3) and (B.12) yield that

\[
\log L(\hat{\beta}_0 + \gamma) - \log L(\beta_0) \leq s_J(\beta_0)^\top (\gamma - \tilde{\gamma}) \\
- \frac{1}{2} \left\| H_J(\beta_0)^{\frac{1}{2}} (\gamma - \tilde{\gamma}) \right\|_2^2 + n \cdot c_{\text{change}} \left( \frac{c_{\text{upper}}(\|\beta_0\|_2)}{c_{\text{lower}}(\|\beta_0\|_2)} \right) \|\gamma\|_2^2.
\]  
(B.16)

Writing

\[
s_J(\beta_0)^\top (\gamma - \tilde{\gamma}) = \left( H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) \right)^\top \left( H_J(\beta_0)^{\frac{1}{2}} (\gamma - \tilde{\gamma}) \right)
\]

and noting that \( H_J(\beta_0)^{\frac{1}{2}} (\gamma - \tilde{\gamma}) \in S_J^\perp, \) we see that the first two terms of the bound in (B.15) can be bounded as

\[
s_J(\beta_0)^\top (\gamma - \tilde{\gamma}) - \frac{1}{2} \left\| H_J(\beta_0)^{\frac{1}{2}} (\gamma - \tilde{\gamma}) \right\|_2^2 \leq \sup_{z \in S_J^\perp} \left( H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) \right)^\top z - \frac{1}{2} \|z\|_2^2 \leq \frac{1}{2} \left\| \text{Proj}_{S_J^\perp} \left( H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) \right) \right\|_2^2,
\]

which is at most \( \frac{\hat{\gamma}^2_{\beta_J} J}{2} / 2 \) by the assumed inequality (B.3).

Consider now the MLE \( \hat{\beta} = \hat{\beta}_J = \beta_0 + \gamma, \) and define \( \hat{\gamma} \in \mathbb{R}^{J} \) as before. Then

\[
\log L(\hat{\beta}_J) - \log L(\hat{\beta}_0) \leq \log L(\hat{\beta}_J) - \log L(\hat{\beta}) \\
\leq \frac{1}{2} \hat{\gamma}_{\beta_J}^2 z_J J / 2 + n \cdot c_{\text{change}} \left( \frac{c_{\text{upper}}(\|\beta_0\|_2)}{c_{\text{lower}}(\|\beta_0\|_2)} \right) \|\hat{\beta}_J - \beta_0\|_2^2.
\]  
(B.16)
We can thus bound the difference between the maxima of the log-likelihood functions if we can bound the distance \( \| \hat{\beta}_J - \beta_0 \|_2 \).

To bound \( \| \hat{\beta}_J - \beta_0 \|_2 \), we return to (B.13). The assumed inequality (B.2) implies that

\[
\begin{align*}
    s_J(\beta_0) \gamma &= \left( H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) \right) \left( H_J(\beta_0)^{\frac{1}{2}} \right) \\
    &\leq \sqrt{n c_{\text{upper}}(\| \beta_0 \|_2) \cdot \tau_{J \setminus J_0}} |\gamma|_2.
\end{align*}
\]

Therefore, for \( \| \gamma \|_2 \leq \frac{c_{\text{lower}}(\| \beta_0 \|_2)}{c_{\text{change}}} \), the inequality (B.13) with another application of (2.3) gives

\[
\begin{align*}
    \log L(\beta_0 + \gamma) - \log L(\beta_0) &\leq \sqrt{n c_{\text{upper}}(\| \beta_0 \|_2) \cdot \tau_{J \setminus J_0}} |\gamma|_2 \\
    &\quad - \frac{n c_{\text{lower}}(\| \beta_0 \|_2)}{2} (1 - c_{\text{lower}}(\| \beta_0 \|_2)^{-1} c_{\text{change}} \cdot |\gamma|_2^2) |\gamma|_2^2.
\end{align*}
\]

In particular, for \( \| \gamma \|_2 \leq \frac{c_{\text{lower}}(\| \beta_0 \|_2)}{2 c_{\text{change}}} \), we have

\[
\begin{align*}
    \log L(\beta_0 + \gamma) - \log L(\beta_0) &\leq \| \gamma \|_2 \left( \sqrt{n c_{\text{upper}}(\| \beta_0 \|_2) \cdot \tau_{J \setminus J_0}} - \frac{n c_{\text{lower}}(\| \beta_0 \|_2)}{4} \right),
\end{align*}
\]

and so by concavity of the log-likelihood function, for all \( \gamma \in \mathbb{R}^d \),

\[
\begin{align*}
    \log L(\beta_0 + \gamma) - \log L(\beta_0) &\leq \| \gamma \|_2 \left( \sqrt{n c_{\text{upper}}(\| \beta_0 \|_2) \cdot \tau_{J \setminus J_0}} - \frac{n c_{\text{lower}}(\| \beta_0 \|_2)}{4} \min \left\{ \| \gamma \|_2, \frac{c_{\text{lower}}(\| \beta_0 \|_2)}{2 c_{\text{change}}} \right\} \right). \tag{B.17}
\end{align*}
\]

Since \( \log L(\hat{\beta}_J) - \log L(\hat{\beta}_J) \geq 0 \), this shows that

\[
\| \hat{\beta}_J - \beta_0 \|_2 \leq \frac{4 \sqrt{c_{\text{upper}}(\| \beta_0 \|_2) \cdot \tau_{J \setminus J_0}}}{\sqrt{n c_{\text{lower}}(\| \beta_0 \|_2)}},
\]

as long as we assume that

\[
\frac{4 \sqrt{c_{\text{upper}}(\| \beta_0 \|_2) \cdot \tau_{J \setminus J_0}}}{\sqrt{n c_{\text{lower}}(\| \beta_0 \|_2)}} \leq \frac{c_{\text{lower}}(\| \beta_0 \|_2)}{2 c_{\text{change}}}. \tag{B.16}
\]

Taking up (B.16), we get

\[
\begin{align*}
    \log L(\hat{\beta}_J) - \log L(\hat{\beta}_J) &\leq \frac{1}{2} \tau_{J \setminus J_0} \\
    &\quad + n \cdot c_{\text{change}} \left( \frac{c_{\text{upper}}(\| \beta_0 \|_2)}{c_{\text{lower}}(\| \beta_0 \|_2)} \right)^\frac{3}{2} \left( \frac{4 \sqrt{c_{\text{upper}}(\| \beta_0 \|_2) \cdot \tau_{J \setminus J_0}}}{\sqrt{n c_{\text{lower}}(\| \beta_0 \|_2)}} \right)^3.
\end{align*}
\]
For $\sqrt{n}$ sufficiently large relative to $\tau_{|J \setminus J_0|^2}$, we get

$$\log L(\hat{\beta}_J) - \log L(\hat{\beta}_{J_0}) \leq \frac{1}{2} \tau_{|J \setminus J_0|^2} \cdot (1 + \epsilon')$$

$$= \frac{(1 + \epsilon')^2}{(1 - \epsilon')^3} \cdot \left( |J \setminus J_0| \log \left( \frac{6p}{\epsilon'} \right) + \log(4p') \right). \quad (B.18)$$

Hence, this inequality holds by choosing the constant $C_{\text{sample}}$ appropriately. Now fix $\epsilon' \in (0, \epsilon)$ such that

$$\frac{(1 + \epsilon')^2}{(1 - \epsilon')^3} < 1 + \epsilon.$$  

(B.19)

Choosing the constant $C_{\dim}$ to ensure that $p$ is large enough, we have that

$$\frac{|J \setminus J_0| \log \left( \frac{6p}{\epsilon'} \right) + \log(4p')}{(|J \setminus J_0| + \nu) \log(p)} = 1 + \frac{|J \setminus J_0| \log \left( \frac{6p}{\epsilon'} \right) + \log(4)}{(|J \setminus J_0| + \nu) \log(p)} \leq 1 + \epsilon',$$

which implies, by (B.18) and (B.19), that

$$\log L(\hat{\beta}_J) - \log L(\hat{\beta}_{J_0}) \leq (1 + \epsilon)(|J \setminus J_0| + \nu) \log(p).$$

This proves statement (a) of Theorem 2.2.

To show the remaining claim (b) of Theorem 2.2, we first note that for any $J \not\supset J_0$, it holds that

$$\|\hat{\beta}_J - \beta_0\|_2 \geq \min_{j \in J_0} |(\beta_0)_j|.$$

Having assumed that the Hessian conditions hold for true models with up to $2q$ covariates, we may apply (B.17) to the model given by $(J \cup J_0) \supseteq J$. We deduce that

$$\log L(\hat{\beta}_J) - \log L(\hat{\beta}_0) \leq \min_{j \in J_0} |(\beta_0)_j| \left( \sqrt{n_{\text{upper}}(\|\beta_0\|_2 \cdot \tau_{|J \setminus J_0|})} \cdot n_{\text{upper}}(\|\beta_0\|_2) \cdot \frac{\tau_{|J \setminus J_0|}}{4} \cdot \min_{j \in J_0} \left\{ \frac{|(\beta_0)_j|}{2 \cdot \text{change}} \right\} \right),$$

as long as the term in the parentheses is non-positive. However, this can be guaranteed to be the case, by appropriate choice of the constant $C_{\text{sample}}$.

Appendix C: Proof of Hessian conditions (Theorem 2.1)

This part of the appendix provides the proof of Theorem 2.1, according to which the assumptions (A1)-(A3) from Section 2 yield a well-behaved Hessian matrix for the log-likelihood function of all sparse submodels of a logistic regression model. The proof is split into three parts. First, we address the inequality (2.4), next the upper bound in (2.3) and then the lower bound in (2.3). In each case we provide an explicit probability for an event that ensures the desired conclusion. A union bound over the three cases implies that all inequalities hold simultaneously with a probability large enough to conform with the assertion of Theorem 2.1.
C.1. Upper bound on change in Hessian

Define the constant

\[ c_{\text{change}} = 1 + a_2 + 48\sqrt{a_3}. \]  

(C.1)

We claim that if \( n \geq \max\{2, q^3 \log^3(np)\} \), then with probability at least

\[
1 - \exp\left\{ -\frac{n}{128a_3^3q^3 \log^3(np)} \right\} \leq \frac{1}{2np},
\]

(C.2)

we have

\[
\sup_{J \supseteq J_0, |J| \leq q} \sup_{\beta \neq \beta' \in \mathbb{R}^d} \frac{\|H_J(\beta) - H_J(\beta')\|}{\|\beta - \beta'\|_2} \leq c_{\text{change}} \cdot n.
\]

To show this claim, take any superset \( J \supseteq J_0 \) with \( |J| \leq q \), any unit vector \( u \in \mathbb{R}^d \) and any pair of distinct vectors \( \beta \neq \beta' \in \mathbb{R}^d \). Define the unit vector \( v = \frac{\beta - \beta'}{\|\beta - \beta'\|_2} \in \mathbb{R}^d \). Then we have

\[
|u^\top (H(\beta) - H(\beta')) u| \leq \sum_{i=1}^n (X_i^\top u)^2 \cdot |b''(X_i^\top \beta) - b''(X_i^\top \beta')|
\]

\[
\leq \sum_{i=1}^n (X_i^\top u)^2 \cdot |X_i^\top \beta - X_i^\top \beta'| \cdot \max_{t \in [0,1]} |b''(X_i^\top (t\beta + (1-t)\beta'))|.
\]

In the logistic regression model, \( |b'''(\theta)| \leq 1 \) for all \( \theta \). Hence,

\[
|u^\top (H(\beta) - H(\beta')) u| \leq \|\beta - \beta'\|_2 \cdot n \left( \frac{1}{n} \sum_{i=1}^n (X_i^\top u)^2 \cdot |X_i^\top v| \right)
\]

\[
\leq \|\beta - \beta'\|_2 \cdot n \left( \frac{1}{n} \sum_{i=1}^n |X_i^\top u|^3 \right)^{\frac{2}{3}} \left( \frac{1}{n} \sum_{i=1}^n |X_i^\top v|^3 \right)^{\frac{1}{3}}
\]

\[
\leq \|\beta - \beta'\|_2 \cdot n \cdot \left( \sup_{\text{q-sparse unit } w} \frac{1}{n} \sum_{i=1}^n |X_i^\top w|^3 \right).
\]

Applying Corollary D.1 for exponent \( k = 3 \), we find that with at least the claimed probability from (C.2),

\[
\sup_{\text{q-sparse unit } w} \frac{1}{n} \sum_{i=1}^n |X_i^\top w|^3 \leq 1 + a_2 + 48\sqrt{a_3} = c_{\text{change}},
\]

as long as \( n \geq \max\{2, q^3 \log^3(np)\} \). Since \( H(\beta) - H(\beta') \) is symmetric, this implies that

\[
\frac{\|H_J(\beta) - H_J(\beta')\|}{\|\beta - \beta'\|_2} \leq \sup_{\text{q-sparse unit } u} \frac{|u^\top (H(\beta) - H(\beta')) u|}{\|\beta - \beta'\|_2} \leq c_{\text{change}} \cdot n
\]

for all sets \( J \supseteq J_0 \) of cardinality \( |J| \leq q \) and all \( \beta \neq \beta' \in \mathbb{R}^d \), as claimed.

---

\(^2\)When treating another exponential family, one could instead bound the \( b'''(\cdot) \) term by taking \( q \) to be constant, taking the \( X_{ij} \)'s to be bounded, and only considering \( \beta \) and \( \beta' \) of bounded norm.
C.2. Upper bound on Hessian

In this subsection, we prove that if inequality (2.4) holds, then with probability at least
\[ 1 - \exp \left\{ -\frac{n}{32a_4^4q^2\log^2(np)} \right\} - \frac{1}{2np}, \]  
(C.3)
it also holds that
\[ H_J(\beta) \preceq n \cdot c_{\text{upper}}(\|\beta\|_2) \cdot I_J \]
for all \( J \supseteq J_0 \) with \( |J| \leq q \) and all \( \beta \in \mathbb{R}^J \). Here, we define
\[
c_{\text{upper}}(r) := b''(0) \cdot \left( 1 + a_2 + 16\sqrt{2}a_3^2 \right) + c_{\text{change}} \cdot r
\]
where \( c_{\text{change}} \) is the constant from (C.1) and \( b''(0) = 1/4 \) for logistic regression.

The idea for our proof is to show that, on a suitable event, \( \sup_{|J| \leq q} \| H_J(0) \| = O(n) \). Then, combined with the bounded change condition (2.4), we will be able to bound \( \| H_J(\beta) \| \) for any \( \beta \in \mathbb{R}^J \).

First, for any \( q \)-sparse unit \( u \), we have
\[
E \left[ (X^\top u)^2 \right] \leq E \left[ |X^\top u|^3 \right]^\frac{2}{3} \leq a_2^2.
\]
Then we have, with at least the probability in (C.3),
\[
\sup_{|J| \leq q} \| H_J(0) \| = \sup_{|J| \leq q} \left\| \sum_{i=1}^{n} X_{iJ}X_{iJ}^\top b''(X_{iJ}^\top 0) \right\|
\]
\[
= b''(0) \cdot \sup_{|J| \leq q} \left\| \sum_{i=1}^{n} X_{iJ}X_{iJ}^\top \right\|
\]
\[
= b''(0) \cdot \sup_{|J| \leq q, \text{ unit } u \in \mathbb{R}^J} \sum_{i=1}^{n} (X_{iJ}^\top u)^2
\]
\[
\leq b''(0) \cdot n \left( 1 + a_2 + 16\sqrt{2}a_3^2 \right),
\]
where for the last step we apply Corollary D.1 with \( k = 2 \), using the assumption that \( n \geq \max\{2, q^a\log^2(np)\} \). The bounded change condition from (2.4) now implies the desired conclusion, namely, that for all \( J \supseteq J_0 \) with \( |J| \leq q \) and all \( \beta \in \mathbb{R}^J \),
\[
\| H_J(\beta) \| \leq \| H_J(0) \| + \| H_J(0) - H_J(\beta) \| \leq n \cdot c_{\text{upper}}(\|\beta\|_2).
\]

C.3. Lower bound on Hessian

Finally, we prove that with probability at least
\[ 1 - 2 \exp \left\{ -\frac{n}{2} \cdot \left( \frac{a_3^3}{512a_2^2} \right)^2 \right\}, \]  
(C.4)
it holds for all $|J| \leq q$, for all $|J| \leq q$ and for all $\beta \in \mathbb{R}^J$ that
\[
H_J(\beta) \geq n \cdot c_{\text{lower}}(\|\beta\|_2) \cdot I_J,
\]
where
\[
c_{\text{lower}}(r) := \frac{a_1^4}{2048\alpha_2^2} \cdot \min_{|\theta| \leq r \cdot 2 \sqrt{256\alpha_2}/a_1} b''(\theta) \\
= \frac{a_1^4}{2048\alpha_2^2} \frac{\exp\{r \cdot 2 \sqrt{256\alpha_2}/a_1\}}{\left(1 + \exp\{r \cdot 2 \sqrt{256\alpha_2}/a_1\}\right)^2}
\]
for the case of logistic regression; recall (2.1). We show this for triples $(n, p, q)$ that have $n$ larger than the product of $q \log(2p)$ and a constant that is determined through (C.8).

For a proof, since
\[
H_J(\beta) = \sum_{i=1}^{n} X_i^T X_i J b''(X_i^T \beta),
\]
we consider the quantity
\[
\sum_{i=1}^{n} (X_i^T u)^2 b''(X_i^T \beta)
\]
where $u \in \mathbb{R}^J$ is a unit vector. For any choice of $w_1, w_2 \geq 0$, we have
\[
\sum_{i=1}^{n} (X_i^T u)^2 b''(X_i^T \beta) \geq \sum_{i=1}^{n} w_1^2 \min_{|\theta| \leq \|\beta\|_2 w_2} b''(\theta) \cdot \mathbb{1}\{|X_i^T u| \geq w_1, |X_i^T \beta| \leq \|\beta\|_2 w_2\}.
\]
Using the symmetry and monotonicity of $b''$ for logistic regression we find
\[
\sum_{i=1}^{n} (X_i^T u)^2 b''(X_i^T \beta) \geq n \cdot w_1^2 b''(\|\beta\|_2 w_2) \\
\times \left(1 - \frac{\# \{ i : |X_i^T u| < w_1 \}}{n} - \frac{\# \{ i : |X_i^T \beta| / \|\beta\|_2 > w_2 \}}{n}\right). \quad (C.5)
\]
We now show how to choose $w_1$ and $w_2$ such that the two relative frequencies are sufficiently small, with high probability, for any choice of $u$ and $\beta$.

By Lemma D.3, for any $t > 0$, with probability at least $1 - 2e^{-nt^2/2}$, for all $q$-sparse unit vectors $u$,
\[
\frac{\# \{ i : |X_i^T u| < w_1 \}}{n} \leq \text{Prob}\left\{ |X_1^T u| < w_1 + \frac{1}{t} \sqrt{32a_3^2 q \log(2p)/n} \right\} + 2t \quad (C.6)
\]
and
\[
\frac{\# \{ i : |X_i^T u| > w_2 \}}{n} \leq \text{Prob}\left\{ |X_1^T u| > w_2 - \frac{1}{t} \sqrt{32a_3^2 q \log(2p)/n} \right\} + 2t. \quad (C.7)
\]
Now set 

\[ w_1 = \sqrt{\frac{a_1}{8}}, \quad w_2 = \frac{2\sqrt[3]{256a_2}}{a_1}, \quad t = \frac{a_1^3}{512a_2^2} \]

and assume

\[ \frac{1}{t} \sqrt{\frac{32a_3^2q \log(2p)}{n}} \leq \min \left\{ \sqrt{\frac{a_1}{8}}, \frac{\sqrt[3]{256a_2}}{a_1} \right\}. \] \tag{C.8}

Then, for shorter notation, define the two scalars

\[ w_1' = w_1 + \frac{1}{t} \sqrt{\frac{32a_3^2q \log(2p)}{n}}, \quad w_2' = w_2 - \frac{1}{t} \sqrt{\frac{32a_3^2q \log(2p)}{n}}. \]

We begin by bounding (C.6). By (C.8), \( w_1'^2 \leq a_1^2 \), and so applying Lemma D.5 with \( Z = (X_1^\top u)^2 \), \( h(Z) = \sqrt{Z} \) and \( a = a_1^2 \) yields that for all \( q \)-sparse unit vectors \( u \),

\[
\text{Prob}\left\{ |X_1^\top u| < w_1' \right\} \leq 1 - \frac{\mathbb{E}[(X_1^\top u)^2] - w_1'^2}{2 \sup \left\{ x : \sqrt{x} \leq \frac{2\mathbb{E}[(X_1^\top u)^2]}{\mathbb{E}[(X_1^\top u)^2] - w_1'^2} \right\}} \leq 1 - \frac{a_1 - w_1'^2}{2 \sup \left\{ x : \sqrt{x} \leq \frac{2a_2}{a_1 - w_1'^2} \right\}}.
\]

The term involving the supremum satisfies

\[
\frac{a_1 - w_1'^2}{2 \sup \left\{ x : \sqrt{x} \leq \frac{2a_2}{a_1 - w_1'^2} \right\}} = \frac{a_1 - w_1'^2}{2 \left( \frac{2a_2}{a_1 - w_1'^2} \right)^2} = \frac{(a_1 - w_1'^2)^3}{8a_2^2} \geq \frac{a_1^3}{64a_2^2},
\]

and so

\[
\sup_{q\text{-sparse unit } u} \frac{\# \{ i : |X_i^\top u| < w_1' \}}{n} \leq \left( 1 - \frac{a_1^3}{64a_2^2} \right) + 2t \leq 1 - \frac{3a_1^3}{256a_2^2}.
\]

Next, we bound (C.7). By (C.8), for any \( u \),

\[
\text{Prob}\left\{ |X_1^\top u| > w_2' \right\} \leq \text{Prob}\left\{ |X_1^\top u| > \frac{\sqrt[3]{256a_2}}{a_1} \right\} = \text{Prob}\left\{ |X_1^\top u|^3 > \frac{256a_2^3}{a_1^3} \right\}.
\]

By Markov’s inequality,

\[
\text{Prob}\left\{ |X_1^\top u| > w_2' \right\} \leq \frac{\mathbb{E}[(X_1^\top u)^3]}{256a_2^3/a_1^3} \leq \frac{a_2}{256a_2^3/a_1^3} = \frac{a_1^3}{256a_2^2}.
\]

We obtain that

\[
\sup_{q\text{-sparse unit } u} \frac{\# \{ i : |X_i^\top u| > w_2' \}}{n} \leq \frac{a_1^3}{256a_2^2} + 2t \leq \frac{a_1^3}{128a_2^2}.
\]
Returning to (C.5), we conclude that, with at least the probability from (C.4), for all $|J| \leq q$ and all unit $u \in \mathbb{R}^J$ and all $\beta \in \mathbb{R}^J$,

$$\sum_{i=1}^{n} (X_i^T u)^2 b''(X_i^T \beta) \geq n \cdot w_1^2 b''(\|\beta\|_2 w_2) \cdot \left[ 1 - \left( 1 - \frac{3a_1^3}{256a_2^2} \right) - \frac{a_1^3}{128a_2^2} \right]$$

$$\geq n \cdot \frac{a_1}{8} \cdot \frac{a_1^3}{256a_2^2} b'' \left( \|\beta\|_2 \cdot \frac{2\sqrt[3]{256a_2}}{a_1} \right)$$

$$= n \cdot c_{\text{lower}}(\|\beta\|_2).$$

Appendix D: Technical lemmas

This section of the appendix provides the lemmas that were used in previous parts of the paper to control the behavior of sparse linear combinations of the covariates.

D.1. Concentration bound and subgaussian maxima

The lemmas we establish subsequently make use of the following general concentration bound.

**Lemma D.1.** Let $X, X_1, \ldots, X_n$ be i.i.d. random variables drawn from a set $\mathcal{X}$, and let $\mathcal{F}$ be a class of functions $f: \mathcal{X} \to \mathbb{R}$. Consider an $L$-Lipschitz function $g: \mathbb{R} \to \mathbb{R}$ with $g(0) = 0$ and $|g(f(X))| \leq M$ almost surely. Then, for any $t \geq 0$, with probability at least $1 - e^{-t^2/2}$,

$$\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \left( g(f(X_i)) - E_X [g(f(X))] \right) \right| \leq 4L E_{\nu_1, \ldots, \nu_n, X_1, \ldots, X_n} \left[ \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{n} \nu_i f(X_i) \right| \right] + t \cdot M \sqrt{n},$$

where $\nu_1, \ldots, \nu_n \in \{\pm 1\}$ are independent Rademacher random variables that are also independent of $X_1, \ldots, X_n$.

_Proof._ The claim follows by combining known bounded difference, symmetrization, and contraction results. Specifically, it is a consequence of Theorems 2.5, 2.1, and 2.3 in Koltchinskii (2011).

The next lemma states a well-known property of subgaussian random variables (Koltchinskii, 2011, Prop. 3.1).

**Lemma D.2.** Suppose $Z_1, \ldots, Z_m$ are, not necessarily independent, random variables with a common distribution that is $\sigma$-subgaussian for $\sigma > 0$. Then

$$E \left[ \max_{1 \leq i \leq m} |Z_i| \right] \leq \sigma \cdot \sqrt{2 \log(2m)}.$$
D.2. Sparse unit linear combinations falling in an interval

We return to the setting where \( X, X_1, \ldots, X_n \) are i.i.d. random vectors in \( \mathbb{R}^p \) and satisfy assumptions (A1)-(A3).

**Lemma D.3.** Fix any \( a > 0 \) and \( t > 0 \). With probability at least \( 1 - e^{-t^2 n^2} / 2 \), for all \( q \)-sparse unit vectors \( u \),

\[
\frac{1}{n} \# \{ i : |X_i^\top u| < a \} \leq \text{Prob} \left\{ |X_i^\top u| < a + \frac{1}{t} \sqrt{\frac{32\alpha q \log(2p)}{n}} \right\} + 2t.
\]

Similarly, with probability at least \( 1 - e^{-t^2 n^2} / 2 \), for all \( q \)-sparse unit vectors \( u \),

\[
\frac{1}{n} \# \{ i : |X_i^\top u| > a \} \leq \text{Prob} \left\{ |X_i^\top u| > a - \frac{1}{t} \sqrt{\frac{32\alpha q \log(2p)}{n}} \right\} + 2t.
\]

**Proof.** The proofs of the two statements are essentially identical, so we prove only the first one. Let

\[
c := t^{-1} \sqrt{\frac{32\alpha q \log(2p)}{n}},
\]

and define the piece-wise linear function

\[
g(z) = \begin{cases} 
1 & \text{if } |z| \leq a, \\
0 & \text{if } |z| \geq a + c, \\
(a + c - z)/c & \text{if } a \leq |z| \leq a + c.
\end{cases}
\]

Then \( g \) is \( 1/c \)-Lipschitz, has values in \([0, 1]\), and satisfies

\[
\mathbb{1}_{\{|z|<a\}} \leq g(z) \leq \mathbb{1}_{\{|z|<a+c\}}.
\]

By the concentration bound from Lemma D.1, with probability at least \( 1 - e^{-t^2 n^2} / 2 \),

\[
\sup_{q\text{-sparse unit } u} \left| \sum_{i=1}^n \left( g(X_i^\top u) - \mathbb{E}[g(X_i^\top u)] \right) \right| \leq 4c^{-1} \mathbb{E}_{\nu, X} \left[ \sup_{q\text{-sparse unit } u} \left| \sum_{i=1}^n \nu_i X_i^\top u \right| \right] + nt. \quad (D.1)
\]

Assume from now on that this event occurs.

We may bound the expectation in \( (D.1) \) as

\[
\mathbb{E}_{\nu, X} \left[ \sup_{q\text{-sparse unit } u} \left| \sum_{i=1}^n \nu_i X_i^\top u \right| \right] \leq \mathbb{E}_{\nu, X} \left[ \sup_{|J|\leq q} \left| \sum_{i=1}^n \nu_i X_{iJ} \right| \right] \leq \sqrt{q} \mathbb{E}_{\nu, X} \left[ \sup_{1\leq j\leq p} \left| \sum_{i=1}^n \nu_i X_{ij} \right| \right]. \quad (D.2)
\]
By assumption (A3), $\nu_i X_{ij}$ is $a_3$-subgaussian, that is, for all $t \in \mathbb{R}$,
$$
\mathbb{E} \left[ e^{t \nu_i X_{ij}} \right] \leq e^{t^2 a_3^2 / 2}.
$$

Then $\sum_{i=1}^n \nu_i X_{ij}$ is $(a_3 \sqrt{n})$-subgaussian because, by independence,
$$
\mathbb{E} \left[ e^{t \sum_{i=1}^n \nu_i X_{ij}} \right] = \prod_{i=1}^n \mathbb{E} \left[ e^{t \nu_i X_{ij}} \right] \leq e^{nt^2 a_3^2 / 2}.
$$

Applying Lemma D.2 to (D.2), we obtain that
$$
\mathbb{E}_{\nu, X} \left[ \sup_{q \text{-sparse unit } u} \left| \sum_{i=1}^n \nu_i X_i^\top u \right| \right] \leq \sqrt{q} \cdot a_3 \sqrt{n} \cdot \sqrt{2 \log(2p)}.
$$

We deduce that
$$
\sup_{q \text{-sparse unit } u} \left| \sum_{i=1}^n g(X_i^\top u) - \mathbb{E} \left[ g(X_i^\top u) \right] \right| \leq 4c^{-1} \sqrt{q} \cdot a_3 \sqrt{n} \cdot \sqrt{2 \log(2p)} + nt = 2nt,
$$
by our choice of $c$. Returning to (D.1), we have shown that, as desired,
$$
\sup_{q \text{-sparse unit } u} \# \{ i : |X_i^\top u| < a \} \leq \sup_{q \text{-sparse unit } u} \sum_{i=1}^n g(X_i^\top u) \leq n \mathbb{E} \left[ g(X_i^\top u) \right] + 2nt \leq n \text{Prob}\{|X_i^\top u| < a + c\} + 2nt.
$$

\[ \square \]

**D.3. Bounding functions of sparse unit linear combinations**

**Lemma D.4.** Suppose $f : [0, \infty) \to [0, \infty)$ is a nondecreasing function with

$$
f(z) \leq M, \quad |f(z) - f(z')| \leq L|z - z'|,
$$

for all $0 \leq z, z' \leq 2a_3 \sqrt{q \log(np)}$. If $\mathbb{E} \left[ g(|X_i^\top u|) \right] \leq a_2$ for all $q$-sparse unit vectors $u$, then it holds with probability at least $1 - e^{-n/(2M^2)} - 1/(2np)$ that

$$
\sup_{q \text{-sparse unit } u} \sum_{i=1}^n f(|X_i^\top u|) \leq n \left( 1 + a_2 + 4La_3 \sqrt{\frac{2 \log(2p)}{n}} \right).
$$

**Proof.** Let $T = 2a_3 \sqrt{q \log(np)}$, and define $h(z) = \min \{ f(|z|), f(T) \}$ for $z \in \mathbb{R}$. By our assumptions on $f$, the function $h$ is $M$-bounded and $L$-Lipschitz on $\mathbb{R}$.
Applying the concentration bound from Lemma D.1 to \( h \), with \( t = \sqrt{n}/M \), we obtain that with probability at least \( 1 - e^{-n/(2M^2)} \),

\[
\sup_{|J| \leq q, \text{ unit } u \in \mathbb{R}^J} \left| \sum_{i=1}^{n} \left( h(X_i^\top u) - \mathbb{E}[h(X_i^\top u)] \right) \right| \leq n + 4L \mathbb{E}_{\nu, X} \sup_{|J| \leq q, \text{ unit } u \in \mathbb{R}^J} \left| \sum_{i=1}^{n} \nu_i X_i^\top u \right| \leq n + 4L \sqrt{q} \cdot a_3 \sqrt{n} \cdot \sqrt{2 \log(2p)},
\]

where the second inequality follows from (D.3). Since \( h(z) \leq f(|z|) \) for all \( z \in \mathbb{R} \), we have

\[
\mathbb{E}[h(X_i^\top u)] \leq \mathbb{E}[f(|X_i^\top u|)] \leq a_2.
\]

Hence, with probability at least \( 1 - e^{-n/(2M^2)} \),

\[
\sup_{|J| \leq q, \text{ unit } u \in \mathbb{R}^J} \sum_{i=1}^{n} h(X_i^\top u) \leq n \left( 1 + a_2 + 4L a_3 \sqrt{2 \log(2p) / n} \right) \tag{D.4}
\]

For \( z \in [-T, T] \), it holds by definition that \( h(z) = f(|z|) \). Consequently, if \( x \in \mathbb{R}^p \) satisfies \( \|x\|_\infty \leq T \), then \( h(x^\top u) = f(|x_i^\top u|) \) for any \( q \)-sparse unit vector \( u \). Since each \( \nu_i X_{ij} \) is \( a_3 \)-subgaussian when \( \nu \) is Rademacher, the union bound yields that

\[
\text{Prob}\{\exists 1 \leq i \leq n: \|X_i\|_\infty > T\} \leq n p \sup_{1 \leq j \leq p} \text{Prob}\{|\nu X_{ij}| > T\}
\leq n p \text{Prob}\{|N(0, 1)| > T/a_3\}
\leq n p a_3 \frac{a_3}{T} \exp \left\{ - \frac{T^2}{2a_3^2} \right\},
\]

where the last inequality follows from the usual bound on the Mills ratio for the normal distribution (Shorack, 2000, p. 237). For our choice of \( T = 2a_3 \sqrt{\log(np)} \), we obtain

\[
\text{Prob}\{\exists 1 \leq i \leq n: \|X_i\|_\infty > T\} \leq \frac{np}{2 \sqrt{\log(np)}} e^{-2 \log(np)} = \frac{1}{2np \sqrt{\log(np)}} \leq \frac{1}{2np}.
\]

So, by a union bound, with probability at least

\[
1 - e^{-n/(2M^2)} \geq \frac{1}{2np},
\]

the inequality

\[
\sup_{|J| \leq q, \text{ unit } u \in \mathbb{R}^J} \sum_{i=1}^{n} f(|X_i^\top u|) \leq \sup_{|J| \leq q, \text{ unit } u \in \mathbb{R}^J} \sum_{i=1}^{n} h(X_i^\top u)
\]

and the inequality from (D.4) hold simultaneously, which proves the claim. \( \square \)
We obtain the following corollary about moments of sparse unit linear combinations.

**Corollary D.1.** Let $k > 0$. If $E[|X^\top u|^k] \leq a_2$ for all $q$-sparse unit vectors $u$, then

$$
sup_{|J| \leq q, \text{ unit } u \in \mathbb{R}^n} \sum_{i=1}^{n} |X_i^\top u|^k \leq n \left( 1 + a_2 + k2^{k+1/2}a_3^k \sqrt{\frac{q^k \log(2p)[\log(np)]^{k-1}}{n}} \right)
$$

holds with probability at least

$$1 - \exp \left\{ - \frac{n}{2 \left( 2a_3 \sqrt{q \log(np)} \right)^{2k}} \right\} - \frac{1}{2np}.$$

**Proof.** Apply Lemma D.4 to $f(|z|) = |z|^k$, setting

$$M = f \left( 2a_3 \sqrt{q \log(np)} \right) \left( 2a_3 \sqrt{q \log(np)} \right)^k,$$

$$L = f' \left( 2a_3 \sqrt{q \log(np)} \right) = k \left( 2a_3 \sqrt{q \log(np)} \right)^{k-1},$$

and collecting terms to find the upper bound.

---

**D.4. Bounding a variable away from zero using expectations**

**Lemma D.5.** Let $h : [0, \infty) \rightarrow [0, \infty)$ be an increasing function. Let $Z \geq 0$ be a random variable with $E[Z] < \infty$ and $E[Z \cdot h(Z)] < \infty$. Then for any $a \leq E[Z],$

$$\text{Prob}\{Z \geq a\} \geq \frac{E[Z] - a}{2 \sup \{x \geq 0 : h(x) \leq \frac{2E[Z \cdot h(Z)]}{E[Z] - a} \}}.$$

**Proof.** Let $b = \frac{E[Z] - a}{2 \text{Prob}\{Z \geq a\}} \geq 0$. We claim that

$$E[Z \cdot 1_{Z \geq b}] \geq \frac{E[Z] - a}{2}.$$

Indeed, if $b < a$, then

$$E[Z \cdot 1_{Z \geq b}] \geq E[Z \cdot 1_{Z \geq a}] = E[Z] - E[Z \cdot 1_{Z < a}] \geq E[Z] - a \geq \frac{E[Z] - a}{2}.$$

And if $b \geq a$, we have

$$E[Z] = E[Z \cdot 1_{Z < a}] + E[Z \cdot 1_{a \leq Z < b}] + E[Z \cdot 1_{Z \geq b}] \leq a + \text{Prob}\{Z \geq a\} \cdot b + E[Z \cdot 1_{Z \geq b}],$$
which implies
\[ \mathbb{E} [Z \cdot 1_{Z \geq b}] \geq \mathbb{E} [Z] - a - \text{Prob}(Z \geq a) \cdot b = \frac{\mathbb{E} [Z] - a}{2} . \]

Next, since \( h \) is monotonically increasing,
\[ \mathbb{E} [Z \cdot h(Z)] \geq \mathbb{E} [Z \cdot h(Z) \cdot 1_{Z \geq b}] \geq \mathbb{E} [Z \cdot 1_{Z \geq b}] \cdot h(b) \geq \frac{\mathbb{E} [Z] - a}{2} \cdot h(b) , \]
and so
\[ b \leq \sup \left\{ x \geq 0 : h(x) \leq \frac{2 \mathbb{E} [Z \cdot h(Z)]}{\mathbb{E} [Z] - a} \right\} . \]

Therefore,
\[ \text{Prob}(Z \geq a) = \frac{\mathbb{E} [Z] - a}{2b} \geq \frac{\mathbb{E} [Z] - a}{2 \sup \left\{ x \geq 0 : h(x) \leq \frac{2 \mathbb{E} [Z \cdot h(Z)]}{\mathbb{E} [Z] - a} \right\}} . \]

\[ \square \]

References

ANANDKUMAR, A., TAN, V. Y. F., HUANG, F. and WILLSKY, A. S. (2012). High-dimensional structure estimation in Ising models: local separation criterion. \textit{Ann. Statist.} 40 1346–1375. MR3015028

BESAG, J. E. (1972). Nearest-neighbour systems and the auto-logistic model for binary data. \textit{J. Roy. Statist. Soc. Ser. B} 34 75–83. MR0323276 (48 ##1634)

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. \textit{J. Roy. Statist. Soc. Ser. B} 36 192–236. With discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley, and M. S. Bartlett and with a reply by the author. MR0373208 (51 ##9409)

BOGDAN, M., GHOSH, J. K. and DOERGE, R. W. (2004). Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci. \textit{Genetics} 167 989–999.

BROMAN, K. W. and SPEED, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. \textit{J. R. Stat. Soc. Ser. B Stat. Methodol.} 64 641–656. MR1979381

BÜHLMANN, P. and VAN DE GEER, S. (2011). \textit{Statistics for high-dimensional data. Springer Series in Statistics.} Springer, Heidelberg. Methods, theory and applications. MR2807761 (2012e:62006)

CHEN, J. and CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. \textit{Biometrika} 95 759–771. MR2443189

CHEN, J. and CHEN, Z. (2012). Extended BIC for small-\( n \)-large-\( P \) sparse GLM. \textit{Statist. Sinica} 22 555–574. MR2954352

FOYGEL, R. (2012). Prediction and model selection for high-dimensional data with sparse or low-rank structure PhD thesis, The University of Chicago.
R. Foygel, M. Drton/Ising model selection

Foygel, R. and Drton, M. (2011). Bayesian model choice and information criteria in sparse generalized linear models. ArXiv e-prints.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software 33 1–22.

Frommlet, F., Ruhaltinger, F., Twaróg, P. and Bogdan, M. (2012). Modified versions of Bayesian information criterion for genome-wide association studies. Comput. Statist. Data Anal. 56 1038–1051. MR2897552

Höfling, H. and Tibshirani, R. (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. J. Mach. Learn. Res. 10 883–906. MR2505138 (2011b:62041)

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M. and Hofner, B. (2013). mboost: Model-based boosting. URL: http://CRAN.R-project.org/package=mboost, R package version 2.2-3.

Kindermann, R. and Snell, J. L. (1980). Markov random fields and their applications. Contemporary Mathematics 1. American Mathematical Society, Providence, R.I. MR620955 (82j:60183)

Koltchinskii, V. (2011). Oracle inequalities in empirical risk minimization and sparse recovery problems. Lecture Notes in Mathematics 2033. Springer, Heidelberg. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d’Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. MR2829871 (2012i:91165)

Lauritzen, S. L. (1996). Graphical models. Oxford Statistical Science Series 17. The Clarendon Press Oxford University Press, New York. Oxford Science Publications. MR1419991 (98g:62001)

Loh, P.-L. and Wainwright, M. J. (2014). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. Ann. Statist. 42 to appear.

Lorentz, G. G., Golitschek, M. V. and Makovoz, Y. (1996). Constructive approximation. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences] 304. Springer-Verlag, Berlin. Advanced problems. MR1393437 (97k:41002)

Luo, S. and Chen, Z. (2013). Selection consistency of EBIC for GLIM with non-canonical links and diverging number of parameters. Statistics and Its Interface 6 275–284.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. Ann. Statist. 34 1436–1462. MR2278363 (2008b:62044)

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. J. R. Stat. Soc. Ser. B Stat. Methodol. 72 417–473.

Menne, M. J., Williams Jr., C. N. and Vose, R. S. (2011). United States Historical Climatology Network Daily Temperature, Precipitation, and Snow Data.

Ravikumar, P., Wainwright, M. J. and Lafferty, J. D. (2010). High-dimensional Ising model selection using $\ell_1$-regularized logistic regression. Ann.
Santhanam, N. P. and Wainwright, M. J. (2012). Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory* **58** 4117–4134.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014 (57 #7855)

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450 (2011h:62268)

Shorack, G. R. (2000). *Probability for statisticians. Springer Texts in Statistics*. Springer-Verlag, New York. MR1762415 (2001d:60002)

R Core Team (2013). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

˙Zak-Szatkowska, M. and Bogdan, M. (2011). Modified versions of the Bayesian information criterion for sparse generalized linear models. *Comput. Statist. Data Anal.* **55** 2908–2924. MR2813055 (2012g:62313)