Electronic Journal of Statistics
Vol. 9 (2015) 567–607
ISSN: 1935-7524
DOI: 10.1214/15-EJS1012

High-dimensional Ising model selection with Bayesian information criteria

Rina Foygel Barber

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

and

Mathias Drton

Department of Statistics
University of Washington
Seattle, WA 98195, USA
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.

MSC 2010 subject classifications: 62F12, 62J12.
Keywords and phrases: Bayesian information criterion, graphical model, logistic regression, log-linear model, neighborhood selection, variable selection.

Received February 2014.

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 $V = [p] := \{1, \ldots, p\}$ and edge set E whose elements are unordered pairs of distinct vertices that we denote by a set of two nodes $\{v, w\}$. The (symmetric) Ising model associated to G postulates that

$$\text{Prob}(Z_1 = z_1, \ldots, Z_p = z_p) \propto \exp\Big\{ \sum_{\{v,w\} \in E} \theta_{vw} z_v z_w \Big\}, \tag{1.1}$$

for values $z_1, \ldots, z_p \in \{-1, 1\}$ and interaction parameters $\theta_{vw} \in \mathbb{R}$. 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. Roudi, Aurell and Hertz (2009) or 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 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 (2013).

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

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

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

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

where $\beta_{vw} = 2\theta_{vw}$. Hence, for each variable Z_v, the conditional distributions form a logistic regression model with Z_v as response and the remaining variables Z_w for all $w \neq v$ as covariates. Selection of the graph $G = (V, E)$ can thus be achieved by identifying each neighborhood $\mathrm{ne}(v)$ 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 (β_{vw}, β_{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 ℓ_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 β_{vw} and β_{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. We remark that while we focus on ℓ_1-penalization techniques in our later numerical experiments, the problem of recovering the edges of G in a high-dimensional setting can also be solved by greedy search methods (Jalali, Johnson and Ravikumar, 2011).

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

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

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 n (Broman and Speed, 2002). 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, as shown in Żak-Szatkowska and Bogdan (2011). Following Żak-Szatkowska and Bogdan (2011), we will treat the criterion

$$\mathrm{BIC}_\gamma(J) = -2\log L(\hat{\beta}_J) + |J|\big(\log(n) + 2\gamma\log(p)\big), \tag{1.3}$$

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

$$\mathcal{J}_k = \{J \subset [p] \,:\, |J| = k\}, \qquad k = 0, \ldots, q,$$

where q is an a priori bound on the size of the models; therefore for each $k \leq q$, any given model of size k has probability proportional to $1/|\mathcal{J}_k|$ of being chosen. The connection to this prior, which is also considered in Scott and Berger (2010), is due to the fact that

$$|\mathcal{J}_k| = \binom{p}{k}$$

scales as p^k for small $k \leq q \leq p/2$. 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. This justifies the criterion (1.3) for model selection in regression.

Now we turn back to the graphical model setting. By analogy, the prior for Ising model selection has to be specified on the set of graphs with p nodes and there are

$$\binom{\binom{p}{2}}{k} \sim p^{2k}$$

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

In this paper we show that using BIC_γ for variable selection in the logistic neighborhood selection approach allows one to consistently estimate the graph of 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. Briefly put, under the conditions we impose, BIC_γ can successfully identify the graph if n exceeds a constant multiple of $q^3 \log(p)$, which agrees with the rates found in Ravikumar, Wainwright and Lafferty (2010) and Santhanam and Wainwright (2012).

Our work builds on ideas of Chen and Chen (2012) and Luo and Chen (2013) who analyze the performance of BIC_γ 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. The conditions we give allow us to deduce consistency of BIC_γ 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_γ in logistic regression with random covariates. In Section 3, we extend the consistency result to Ising models. 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 on simulated and real data, 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

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

for some unknown parameter vector $\beta_0 \in \mathbb{R}^p$. Define the cumulant function $\mathrm{b}(z) = \log(1 + e^z)$. Conditional on the X_i, the logistic regression model for the

responses Y_i has log-likelihood, score, and negative Hessian functions

$$\log \mathrm{L}(\beta) = \sum_{i=1}^{n} Y_i \cdot X_i^\top \beta - \mathrm{b}(X_i^\top \beta) \ \in \ \mathbb{R},$$
$$s(\beta) = \sum_{i=1}^{n} X_i \left(Y_i - \mathrm{b}'(X_i^\top \beta)\right) \ \in \ \mathbb{R}^p,$$
$$H(\beta) = \sum_{i=1}^{n} X_i X_i^\top \cdot \mathrm{b}''(X_i^\top \beta) \ \in \ \mathbb{R}^{p \times p},$$

with the derivatives of the cumulant function being

$$\mathrm{b}'(z) = \frac{e^z}{1+e^z}, \quad \mathrm{b}''(z) = \frac{e^z}{(1+e^z)^2}. \tag{2.1}$$

We will be interested in scenarios in which β_0 is sparse, and we wish to recover the support of β_0, that is, the set

$$J_0 = \mathrm{supp}(\beta_0) := \{j \in [p] \ : \ \beta_{0j} \neq 0\},$$

which gives the most parsimonious (most sparse) true model. We assume that an upper bound q on the size of the support is given, that is, $|J_0| \leq q$. Later, the bound q is allowed to grow in an asymptotic scenario in which the number of covariates p may 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_i 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_γ for generalized linear models with fixed design. These conditions concern sparse submodels of our logistic regression model given by support sets $J \subseteq [p]$.

Notation for submodels The parameters of the submodel given by a set J are regression coefficients that form a vector of length $|J|$. We index such vectors β by the elements of J, that is, $\beta = (\beta_j : j \in J)$, and similarly write $\mathbb{R}^J$ for the parameter space comprising all these coefficient vectors. This way the index of a coefficient always coincides with the index of the covariate it belongs to. In other words, the coefficient for the j-th coordinate of covariate vector X_i is denoted by β_j in any model J with $j \in J$.

Furthermore, it is at times convenient to identify a vector $\beta \in \mathbb{R}^J$ with the vector in $\mathbb{R}^p$ that is obtained from β by filling in zeros outside of the set J. As this is clear from the context, we simply write β again when referring to this sparse vector in $\mathbb{R}^p$. Finally, $s_J(\beta)$ and $H_J(\beta)$ denote 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 $\epsilon > 0$, there is a $\delta > 0$ such that

$$(1-\epsilon)H_J(\beta_0) \preceq H_J(\beta) \preceq (1+\epsilon)H_J(\beta_0), \tag{2.2}$$

for all sparse $J \supseteq J_0$ and $\beta \in \mathbb{R}^J$ with $\|\beta - \beta_0\|_2 \leq \delta$. The notation "$\preceq$" refers to the ordering in the positive semidefinite cone with $A \preceq B$ whenever $0 \preceq B - A$, i.e., $B - A$ is positive semidefinite. The above conditions are assumed to hold uniformly for all large enough sample sizes n and associated values of p, q and β_0, which may change with n.

In this work, we begin instead with random and i.i.d. covariates $X_1, \ldots, X_n$ and derive stronger versions of these Hessian conditions from the below conditions on the distribution of each covariate X_i. We refer to a vector $u \in \mathbb{R}^p$ as q-sparse if $|\operatorname{supp}(u)| \leq q$. 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, $\mathbb{E}[(X_1^\top u)^2] \geq a_1$.
(A2) For any q-sparse unit vector u, $\mathbb{E}[|X_1^\top u|^3] \leq a_2$.
(A3) For each $j \in [p]$, the variable X_{1j} is bounded as $|X_{1j}| \leq a_3$.

Rephrased, (A1) states that for any subset $J \subset [p]$ of cardinality $|J| \leq q$ the smallest eigenvalue of the matrix $\mathbb{E}[X_{1J}X_{1J}^\top]$ is at least a_1. (Here, $X_{1J} = (X_{1j} : j \in J)$ is the subvector of X_1 induced by J.) Assumption (A2) guarantees the existence of third moments of linear combinations of q or fewer covariates. In an Ising model all variables are bounded and thus (A3) always holds.[1]

According to the following theorem, our assumptions entail well-behaved Hessians with large probability. In this theorem and throughout the rest of the paper, the norm $\|H\|$ of a matrix H is the spectral norm.

Theorem 2.1. *Suppose that the covariates satisfy conditions (A1)–(A3) for some sparsity level q and some constants $a_1, a_2, a_3 > 0$. Then 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 if*

$$n \geq c_{\text{sample}} \cdot q^3 \log(p),$$

then the event that, simultaneously for all $|J| \leq q$ and all $\beta, \beta' \in \mathbb{R}^J$,

$$c_{\text{lower}}(\|\beta\|_2)\mathbf{I}_J \preceq \frac{1}{n}H_J(\beta) \preceq c_{\text{upper}}(\|\beta\|_2)\mathbf{I}_J \tag{2.3}$$

[1] A weaker condition requiring only that each X_{1j} is subgaussian was considered in a preprint version of this paper (Foygel and Drton, 2014). The same results were obtained, but at the cost of additional log factors in the sample size—specifically, with a sample size requirement of $n \gtrsim q^3 \log^3(np)$ instead of $n \gtrsim q^3 \log(p)$ as in the theorems in this paper.

and

$$\frac{1}{n}\|H_J(\beta) - H_J(\beta')\| \le c_{\text{change}} \cdot \|\beta - \beta'\|_2 \tag{2.4}$$

has probability at least

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

The proof of Theorem 2.1 is given in Appendix C.

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

$$\begin{aligned}\frac{1}{n}H_J(\beta) &\preceq c_{\text{change}} \cdot \|\beta - \beta_0\|_2 \cdot \mathbf{I}_J + \frac{1}{n}H_J(\beta_0) \\ &\preceq \left(1 + \frac{c_{\text{change}}}{c_{\text{lower}}(\|\beta_0\|_2)} \cdot \|\beta - \beta_0\|_2\right)\frac{1}{n}H_J(\beta_0).\end{aligned}$$

We also have the analogous lower bound,

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

Combining these two bounds, we have proved the following version of the assumption from (2.2):

Proposition 2.1. *If the inequalities (2.3) and (2.4) hold for all $J \supseteq J_0$ with $|J| \le q$, then*

$$(1-\epsilon)H_J(\beta_0) \preceq H_J(\beta) \preceq (1+\epsilon)H_J(\beta_0)$$

holds for all such J and for all $\beta \in \mathbb{R}^J$ with

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

Remark 2.1. Although this proposition only treats true models (i.e., models J that contain the true support J_0), it will be used also for proving that the BIC will not select a false model (i.e., a model $J \not\supset J_0$). The connection lies in observing that, for a model $J \not\supset J_0$, the proposition can be applied to analyze the model given by the union $J \cup J_0$, which is a true model.

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 β_0 is needed to avoid degeneracy of the conditional distribution of the binary response variable.

Theorem 2.2. *Let β_0 be the true parameter with $J_0 = \mathrm{supp}(\beta_0)$ and $\|\beta_0\|_2 \le a_0$ for a constant $a_0 > 0$. Fix $\epsilon, \nu > 0$, and condition on the covariates $X_1, \dots, X_n$ satisfying the Hessian conditions (2.3) and (2.4) for all $J \supseteq J_0$ with $|J| \le 2q$, where $q \ge |J_0|$. Then there exist constants $C_{\mathrm{false}}, C_{\mathrm{dim}}, C_{\mathrm{sample},1}, C_{\mathrm{sample},2} > 0$, depending only on $(c_{\mathrm{change}}, c_{\mathrm{lower}}(a_0), c_{\mathrm{upper}}(a_0))$ and on the chosen pair (ϵ, ν), such that if*

$$p \ge C_{\mathrm{dim}} \quad \text{and} \quad n \ge \max\left\{ C_{\mathrm{sample},1} \cdot q^3 \log(p), C_{\mathrm{sample},2} \cdot \frac{q \log(p)}{\min_{j \in J_0} |(\beta_0)_j|^2} \right\},$$

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

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

$$\log \mathrm{L}(\widehat{\beta}_J) - \log \mathrm{L}(\widehat{\beta}_{J_0}) \ \le \ (1+\epsilon)(|J \backslash J_0| + \nu) \log(p).$$

(b) For all $|J| \le q$ with $J \not\supseteq J_0$,

$$\log \mathrm{L}(\widehat{\beta}_{J_0}) - \log \mathrm{L}(\widehat{\beta}_J) \ \ge \ C_{\mathrm{false}}\, n \min_{j \in J_0} |(\beta_0)_j|^2.$$

The proof of Theorem 2.2 is deferred to Appendix B. We remark that the proof of claim (a) invokes the Hessian conditions only for $J \supseteq J_0$ with $|J| \le q$. The conditions for cardinality up to $2q$ are used for claim (b), which is proved by considering the union $J_0 \cup J$ for the given false model $J \not\supseteq J_0$.

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 BIC_γ selects the most parsimonious true model with high probability.

Theorem 2.3. *Let β_0 be the true parameter with $J_0 = \mathrm{supp}(\beta_0)$ and $\|\beta_0\|_2 \le a_0$ for a constant $a_0 > 0$. Fix $\gamma \ge 0$ 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 (ϵ, ν), such that if the covariates satisfy (A1)–(A3) with respect to $2q$ for $q \ge |J_0|$, if*

$$p \ge C_0, \quad n \ge \max\left\{ C_1 \cdot q^3 \log(p),\ C_2 \cdot \frac{q \log(np^{2\gamma})}{\min_{j \in J_0} |(\beta_0)_j|^2} \right\},$$

and if

$$\sqrt{n} > p^{(1+\epsilon)(1+\nu)-\gamma}, \tag{2.6}$$

then the event that

$$J_0 = \arg\min\{\mathrm{BIC}_\gamma(J) \ : \ J \subset [p],\ |J| \le q\}$$

has probability at least

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

Proof. First, examining the statements of Theorem 2.1 and Theorem 2.2, we see that we can choose the constants C_0, C_1, C_2, C_3 large enough that the conditions in Theorems 2.1 and 2.2 are satisfied. These theorems then imply that, with the claimed probability, the following statement is true simultaneously for all $|J| \leq q$:

$$\log \mathrm{L}(\widehat{\beta}_J) - \log \mathrm{L}(\widehat{\beta}_{J_0}) \leq \begin{cases} (1+\epsilon)(|J \backslash J_0| + \nu) \log(p) & \text{if } J \supseteq J_0, \\ -C_{\text{false}} n \min_{j \in J_0} |(\beta_0)_j|^2 & \text{if } J \not\supseteq J_0, \end{cases} \tag{2.7}$$

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

$$\begin{aligned} \mathrm{BIC}_\gamma(J) - \mathrm{BIC}_\gamma(J_0) = -2 \left(\log \mathrm{L}(\widehat{\beta}_J) - \log \mathrm{L}(\widehat{\beta}_{J_0}) \right) \\ + (|J| - |J_0|) \big(\log(n) + 2\gamma \log(p) \big) \end{aligned}$$

is positive for any model given by a set $J \neq J_0$ of cardinality $|J| \leq q$.

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

$$\mathrm{BIC}_\gamma(J) - \mathrm{BIC}_\gamma(J_0) \;\geq\; 2C_{\text{false}} n \min_{j \in J_0} |(\beta_0)_j|^2 - q \log(np^{2\gamma}).$$

Since we require that $n \geq C_2 \cdot \frac{q \log(np^{2\gamma})}{\min_{j \in J_0} |(\beta_0)_j|^2}$, this lower bound on $\mathrm{BIC}_\gamma(J) - \mathrm{BIC}_\gamma(J_0)$ is positive for a sufficiently large choice of the constant C_2.

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

$$\begin{aligned} \mathrm{BIC}_\gamma(J) - \mathrm{BIC}_\gamma(J_0) \;\geq\; -2(1+\epsilon)(|J \backslash J_0| + \nu) \log(p) \\ + |J \backslash J_0| \left(\log(n) + 2\gamma \log(p) \right), \end{aligned}$$

which can be lower-bounded further as

$$\mathrm{BIC}_\gamma(J) - \mathrm{BIC}_\gamma(J_0) \;\geq\; |J \backslash J_0| \cdot \big(\log(n) + 2\big[\gamma - (1+\epsilon)(1+\nu)\big] \log(p) \big) .$$

This is positive by the assumed inequality from (2.6). □

Based on Theorem 2.3, we can identify asymptotic scenarios under which BIC_γ 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

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

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

Corollary 2.1. *Suppose that $p_n \to \infty$ as $n \to \infty$ with $p_n \leq n^\kappa$ for some $\kappa \in (0, \infty]$ and $\log(p_n) \leq n^\tau$ for some $0 < \tau < 1$. Suppose further that $q_n \leq n^\psi$ for some $0 \leq \psi < \frac{1}{3}(1-\tau)$, and that $\beta_{\min}(n) \geq n^{-\phi/2}$ for some $0 \leq \phi <$*

$1-\psi-\tau$. *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_γ *is consistent in the sense that the event*

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

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

Proof. Since $p_n \leq n^\kappa$, 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 ϵ and ν sufficiently small. Fix a suitable choice of (ϵ, ν) for the rest of the argument.

Our scaling assumptions for p_n, q_n and $\beta_{\min}(n)$ 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. □

Remark 2.2. Corollary 2.1 requires $p_n \leq n^\kappa$ and $\log(p_n) \leq n^\tau$, for $\kappa \in (0, \infty]$ and $\tau \in (0, 1)$. For $\kappa < \infty$, this means that p_n grows polynomially with n. In this case, τ can be chosen arbitrarily close to 0, and the conditions on ψ and ϕ become $0 \leq \psi < 1/3$ and $0 \leq \phi < 1 - \psi$. For $\kappa = \infty$, the growth of p_n can be faster than polynomial; the remaining condition $\log(p_n) \leq n^\tau$ allows for subexponential growth. In this latter case, since $\kappa = \infty$, we require $\gamma > 1$ in order to ensure consistency of BIC_γ.

3. Consistency of extended BIC for Ising models

Turning to neighborhood selection for Ising models, let $Z_1, \ldots, Z_n$ be an i.i.d. sample, where each $Z_i = (Z_{i1}, \ldots, Z_{ip})$ is a vector of binary random variables with values in $\{-1, 1\}$. Suppose the Z_i follow 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 $\{v, w\} \in E$. Assume that G is minimal in that $\{v, w\} \in E$ if and only $\theta_{vw} \neq 0$.

We will consider selection of G by means of variable selection in the p logistic regression models, where the v-th regression problem has response variable Z_v and the $p-1$ covariates Z_w, $w \in [p] \setminus \{v\}$. We write $\mathrm{BIC}_\gamma(J, v)$ for the BIC score from (1.3) evaluated for the logistic regression model with response Z_v and covariates Z_w, $w \in J$, with $J \subseteq [p] \setminus \{v\}$. Correct inference of G is achieved if, for each $v \in [p]$, the neighborhood

$$\mathrm{ne}(v) = \{w \in [p] \setminus \{v\} \,:\, \theta_{vw} \neq 0\} = \{w \in [p] \setminus \{v\} \,:\, \{v, w\} \in E\}$$

(uniquely) minimizes $\mathrm{BIC}_\gamma(\cdot, v)$.

Using $Z_1 = (Z_{11}, \ldots, Z_{1p})^\top$ 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 between a variable and its neighborhood is bounded as

$$\sqrt{\sum_{w \in \mathrm{ne}(v)} \theta_{vw}^2} \leq b_0 \quad \text{for all} \quad v \in [p].$$

(B2) For any q-sparse unit vector u, $\mathbb{E}[(Z_1^\top u)^2] \geq b_1$.
(B3) For any q-sparse unit vector u, $\mathbb{E}[|Z_1^\top u|^3] \leq b_2$.

As explained in Santhanam and Wainwright (2012), the graph selection problem is ill-posed without some upper bound on the interaction between a variable and its neighborhood, as we impose in (B1). Assumption (B2) constitutes a lower bound on the eigenvalues of the $q \times q$ principal submatrices of the covariance matrix $\mathbb{E}[Z_1 Z_1^\top]$ and is akin to requirements in Ravikumar, Wainwright and Lafferty (2010) and Loh and Wainwright (2013). As we clarify at the end of this section, condition (B2) is implied by (B1) for asymptotic scenarios in which all neighborhoods $\mathrm{ne}(v)$ 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 neighbor 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 neighorhoods $\mathrm{ne}_n(v)$ and edge set $E(n)$. Let d_n be the maximum cardinality of any neighborhood $\mathrm{ne}_n(v)$, $v \in [p_n]$, and let

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

be the non-zero interaction of smallest magnitude.

Theorem 3.1. *Suppose that $p_n \to \infty$ as $n \to \infty$ with $p_n \leq n^\kappa$ for some $\kappa \in (0, \infty]$ and $\log(p_n) \leq n^\tau$ for some $0 < \tau < 1$. Suppose further that $q_n \leq n^\psi$ for some $0 \leq \psi < \frac{1}{3}(1-\tau)$, and that $\theta_{\min}(n) \geq n^{-\phi/2}$ for some $0 \leq \phi < 1-\psi-\tau$. 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 BIC_γ is consistent in the sense that the event that, simultaneously for all $v \in [p_n]$,*

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

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

Remark 3.1. As in Corollary 2.1, this result allows for subexponential rather than polynomial growth of p_n relative to n, by setting $\kappa = \infty$.

Proof of Theorem 3.1. 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 $p_n \leq n^\kappa$, condition (2.6) in Theorem 2.3 holds for all n if

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

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

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

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

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

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 $|\mathrm{ne}(v)| \leq q$ for all $v \in [p]$. If the interaction parameters θ_{vw} for Z satisfy (B1) then it holds for any q-sparse unit vector u that*

$$\mathbb{E}\left[(Z^\top u)^2\right] \ \geq \ \frac{4}{q} \cdot \frac{e^{2b_0\sqrt{q}}}{\left(1 + e^{2b_0\sqrt{q}}\right)^2}.$$

Proof. Without loss of generality, we consider a q-sparse unit vector u that has $\mathrm{supp}(u) = \{1, \ldots, q\}$ 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,

$$\mathrm{Var}[X] = \min_{a \in \mathbb{R}} \mathbb{E}\left[(X - a)^2\right].$$

Therefore,

$$\mathbb{E}\left[(Z^\top u)^2 \,|\, Z_{-1}\right] \;\geq\; \mathrm{Var}\left[Z_1 u_1 \,|\, Z_{-1}\right] \;\geq\; \frac{1}{q}\,\mathrm{Var}\left[Z_1 \,|\, Z_{-1}\right].$$

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

$$\frac{\exp\left\{2\sum_{w\in\mathrm{ne}(1)}\theta_{1w}Z_w\right\}}{1+\exp\left\{2\sum_{w\in\mathrm{ne}(1)}\theta_{1w}Z_w\right\}};$$

recall (1.2). We obtain that

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

By assumption (B1),

$$-b_0\sqrt{q} \;\leq\; \sum_{w\in\mathrm{ne}(1)}\theta_{1w}Z_w \;\leq\; b_0\sqrt{q}.$$

It follows that

$$\mathbb{E}\left[(Z^\top u)^2\right] \;=\; \mathbb{E}\left[\mathbb{E}\left[(Z^\top u)^2 \,|\, Z_{-1}\right]\right] \;\geq\; \frac{4}{q}\cdot\frac{e^{2b_0\sqrt{q}}}{\left(1+e^{2b_0\sqrt{q}}\right)^2}. \qquad \square$$

4. Practical considerations when applying information criteria

Theorem 3.1 shows that, with sufficient data, application of BIC_γ 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_γ, 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 large.

(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_γ to select from this list. For each node, we use an ℓ_1-penalized logistic likelihood (Ravikumar, Wainwright and

Lafferty, 2010) with varying levels of penalization ρ to produce the candidate models:

$$\hat{\beta}_v^{(\rho)} = \arg \min_{\beta \in \mathbb{R}^{V \setminus \{v\}}} \left\{ -\sum_{i=1}^{n} \log \text{Prob}\left(Z_{iv} \,\middle|\, \sum_{w \neq v} Z_{iw}\beta_w \right) + \rho \|\beta\|_1 \right\} \tag{4.1}$$

where the probability term is given by the logistic model, i.e.

$$\log \text{Prob}\left(Z_{iv} \,\middle|\, \sum_{w \neq v} Z_{iw}\beta_w \right) = Z_{iv} \cdot \sum_{w \neq v} Z_{iw}\beta_w - \text{b}\Big(\sum_{w \neq v} Z_{iw}\beta_w \Big).$$

As in Section 3, Z_{iv} refers to the v-th coordinate of the binary vector $Z_i = (Z_{i1}, \ldots, Z_{ip})$, which is the i-th vector in a sample $Z_1, \ldots, Z_n$.

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 only that either $\hat{\beta}_{vw} \neq 0$ 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

We study the performance of the extended BIC on both simulated and real data. The real data consists of precipitation measurements from weather stations across the midwest, where we aim to recover a graph that is consistent with the true geographical layout of the weather stations. For this data set, we compare BIC_γ (with a range of values for the parameter γ) with cross-validation as well as with the stability selection method of Meinshausen and Bühlmann (2010). Our simulations replicate those of Ravikumar, Wainwright and Lafferty (2010), including three different sparse graph structures. For the simulated data, we compare different values of the γ parameter for BIC_γ.

5.1. Simulated data

5.1.1. Data and methods for model selection

We generate data from sparse Ising models associated to lattice graphs and star graphs on p nodes for $p \in \{64, 100, 225\}$. For each graph structure, the sample size n is chosen based on the settings that produced moderately high success rates in the simulations of Ravikumar, Wainwright and Lafferty (2010). We consider the following three graph types:

4-*nearest neighbor lattice:* Arranging the nodes in a lattice of size $\sqrt{p} \times \sqrt{p}$, each node is connected to the nodes directly above, below, left or right, giving maximal degree $d = 4$. For adjacent nodes v and w, we either set $\theta_{vw} = 0.5$ (attractive couplings) or draw θ_{vw} at random from $\{+0.5, -0.5\}$ (random couplings). The sample size is $n = \lceil 15 d \log(p) \rceil$.

8-*nearest neighbor lattice:* Analogous to the above but also connecting nodes along diagonals. The maximal degree is $d = 8$. For edges $\{v, w\}$, we either set $\theta_{vw} = 0.25$ (attractive couplings), or draw θ_{vw} at random from $\{+0.25, -0.25\}$ (random couplings). The sample size is $n = \lceil 25 d \log(p) \rceil$.

Star graph: Edges are drawn from a designated "hub" node to q other nodes, where either $q = \lceil \log(p) \rceil$ (logarithmic sparsity) or $q = \lceil 0.1p \rceil$ (linear sparsity). For edges $\{v, w\}$, we set $\theta_{vw} = +0.25$. The sample size is $n = \lceil 10 d \log(p) \rceil$, where $d = q$ is the maximal degree of the graph.

These three graph structures are illustrated in Figure 1 of Ravikumar, Wainwright and Lafferty (2010).

For each of the three settings, we simulate 100 data sets. Each time, we perform nodewise ℓ_1-penalized logistic regressions as in (4.1), where we consider a wide range of penalty parameters ρ in order to produce a 'path' of candidate models for that node. To this end, we used the `glmnet` package for `R` (Friedman, Hastie and Tibshirani, 2010). For each node, we then optimize BIC_γ in order to select a model from the path. Evaluating BIC_γ involves refitting each candidate model without ℓ_1-penalization, which was done using the function `glm` in `R`. We then symmetrized the neighborhoods inferred by applying the OR rule. The resulting graph is compared to the underlying true graph. This procedure was carried out for five choices for γ, namely, $\gamma \in \{0, 0.25, 0.5, 0.75, 1\}$.

We note that the AND rule for symmetrization led to qualitatively similar conclusions, and we do not report the results here.

5.1.2. Results

Results for the 4- and 8-nearest neighbor lattices as well as the star graph are shown in Figures 1, 2, and 3, respectively. For each scenario, we plot the positive selection rate (proportion of true edges that are identified) and the false discovery rate (the proportion of selected edges that are false positives). In each case, we observe a tradeoff between positive selection rate and false discovery rate as the parameter γ for BIC_γ varies. Most notably, for nearly every setting considered, we see that increasing γ from 0 to a positive value can significantly reduce the false discovery rate without much detriment to the positive selection rate, demonstrating a clear benefit to using the extended BIC with $\gamma > 0$ as opposed to the ordinary $\mathrm{BIC} = \mathrm{BIC}_0$ for this high-dimensional setting.

5.2. Real data: Regional weather patterns

5.2.1. Data and methods for model selection

We apply BIC_γ, 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

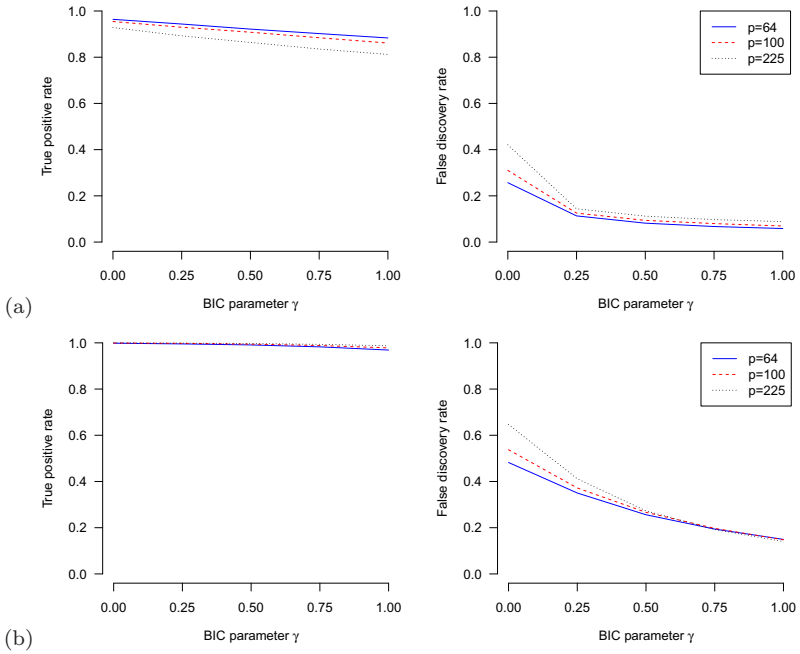

FIG 1. *Results for the* 4*-nearest neighbor graph with (a) attractive couplings and (b) random couplings.*

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). Following the same steps as in our simulation study, we compute a set of candidate models for each node using the ℓ_1-penalized logistic regression and then select a model from the set using either the ordinary BIC $=$ BIC$_0$ or BIC$_\gamma$ with $\gamma \in \{0.25, 0.5\}$. In addition, we considered cross-validation and stability selection (Meinshausen and Bühlmann, 2010). For cross-validation, we select the model that minimizes average error on test sets over 10 folds. For stability selection, we used the `stabsel` function in the `mboost` package for `R` (Hothorn et al., 2013), setting the expected support size to 10.[2] As noted by Meinshausen and Bühlmann (2010), changing the settings within

[2]Parameters for the `stabsel` function were set at `q` $= 10$, the expected support size, and `cutoff` $= 0.75$, the midpoint of the suggested range.

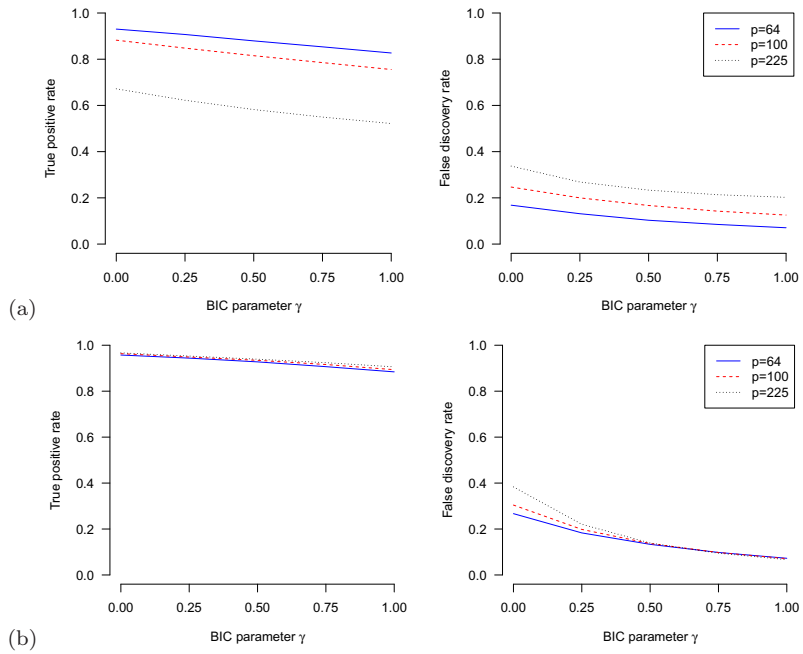

FIG 2. *Results for the* 8*-nearest neighbor graph with (a) attractive couplings and (b) random couplings.*

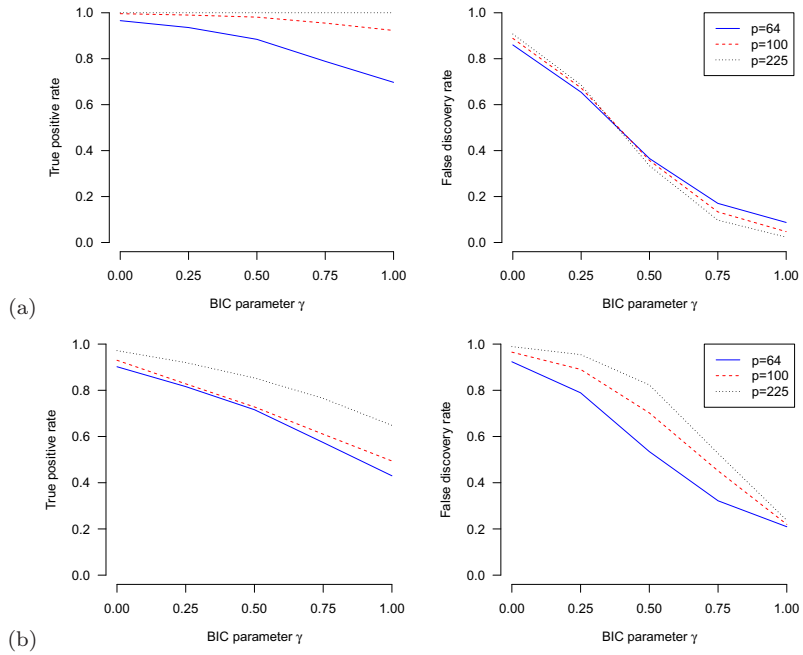

FIG 3. *Results for the star neighbor graph with (a) linear sparsity and (b) logarithmic sparsity.*

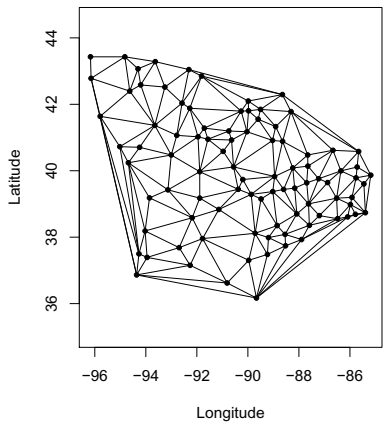

FIG 4. *Delaunay triangulation for 92 weather stations in Illinois, Indiana, Iowa, and Missouri.*

a reasonable range did not have a large effect on the output. For each of the mentioned methods, 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).[3] 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 precipation are not as pronounced in the 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 4 shows a map of the 92 stations, along with an undirected graph representing the Delaunay triangulation of the 92 locations.

5.2.2. 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

[3]Available at http://cdiac.ornl.gov/ftp/ushcn_daily/.

TABLE 1
Positive selection rate (%) and false discovery rate (%) in the weather data experiment, where the true graph is defined via the Delaunay triangulation

	AND rule		OR rule	
	PSR	FDR	PSR	FDR
BIC_0	41.98	32.93	55.73	46.72
$\text{BIC}_{0.25}$	37.40	27.94	52.67	42.02
$\text{BIC}_{0.5}$	34.73	26.61	50.38	38.89
Cross-validation	59.16	57.65	71.37	75.65
Stability selection	45.04	38.54	53.05	45.28

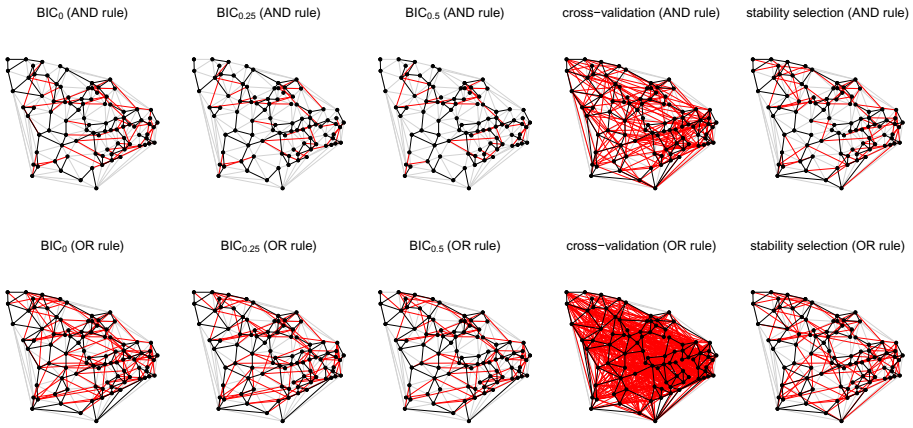

FIG 5. *Graphs recovered under each method. Black edges indicate true positives, red edges indicate false positives, and light gray edges indicate false negatives, i.e. true edges that were not recovered by the method, where the true graph is defined via the Delaunay triangulation.*

triangulation graph. Figure 5 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 performs similarly to BIC_0, but is computationally more expensive.

While it does not seem unreasonable to assume that the edges of the Delaunay 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 6, 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 (under either the OR or the AND rule) does not decay to zero as distance

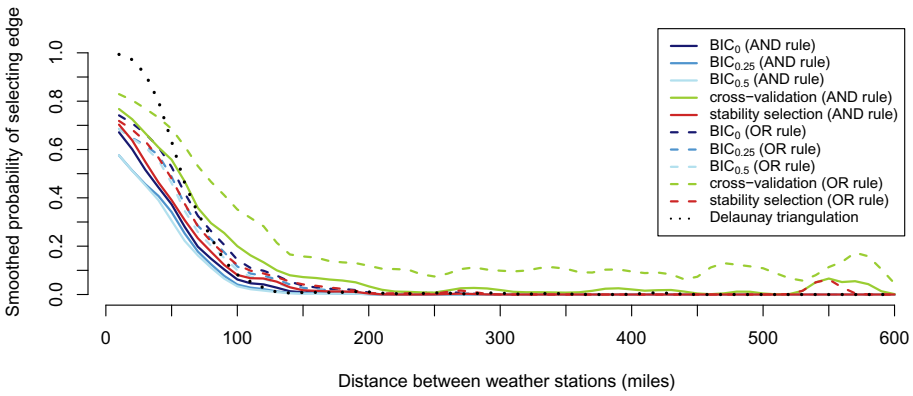

FIG 6. *Smoothed probability of selecting edges as a function of distance, for each method under the* OR *rule and the* AND *rule.*

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. 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 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 information 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 BIC_γ 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).

Acknowledgments

This work was supported by the U.S. National Science Foundation under Postdoctoral Fellowship No. DMS-1203762 and Grant No. DMS-1305154. We would like to thank Kean Ming Tan for 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, $\mathbb{E}[(X_1^\top u)^2] \le a_2'$.

However, we now show that (A2′) is not sufficient for the desired results in any asymptotic scenario where q grows with n, no matter how slow the 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 $\mathbf{1}_q = (1, \dots, 1)^\top$. Then define a random vector X by setting

$$X = \begin{cases} \mathbf{1}_q & \text{with prob. } \frac{1}{2q}, \\ -\mathbf{1}_q & \text{with prob. } \frac{1}{2q}, \\ Z & \text{with prob. } 1 - \frac{1}{q}. \end{cases}$$

Clearly, $\mathbb{E}[Z] = \mathbb{E}[X] = 0$, and $\mathbb{E}[ZZ^\top] = \mathbf{I}_q$. Therefore,

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

has minimal and maximal eigenvalue equal to

$$\lambda_{\min}(\mathbb{E}\left[XX^\top\right]) = 1 - \frac{1}{q}, \qquad \lambda_{\max}(\mathbb{E}\left[XX^\top\right]) = 2 - \frac{1}{q},$$

respectively. We observe that the eigenvalues are bounded above and below by positive constants that are independent of q (for all $q \ge 2$), as required by (A1) and (A2′).

Now take the unit vector $u = \frac{1}{\sqrt{q}}\mathbf{1}_q$. We see that $|X^\top u| = \sqrt{q}$ with probability at least $1/q$. For independent random vectors $X_1, \dots, X_n$ that all have the same distribution as X, it follows that $\#\{i : |X_i^\top u| = \sqrt{q}\}$ is at least as large as a Binomial$(n, 1/q)$ random variable. 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 $\frac{1}{2}$,

$$\#\{i : |X_i^\top u| = \sqrt{q}\} \geq \frac{n}{q}.$$

In the remainder of this section, we prove that this property contradicts the inequalities in (2.2), which for $\beta_0 = 0$ state that

$$H(\beta) \preceq (1+\epsilon)H(0) \text{ for all } \|\beta\|_2 \leq \delta.$$

More precisely, for any $\epsilon > 0$ there should be some $\delta = \delta(\epsilon) > 0$ such that the statement holds, and the relationship between δ and ϵ (given by $\delta = \delta(\epsilon)$) should not depend on the dimensions of the problem. Note that since we have simplified the problem by setting $p = q$, we do not need to make reference to submatrices of $H(0)$.

Next take $\beta = u \cdot \frac{1}{\sqrt{q}} = \frac{1}{q}\mathbf{1}_q$; then $\|\beta\|_2 = \frac{1}{\sqrt{q}}$. Since $\mathrm{b}''(0) \geq \mathrm{b}''(z)$ and $\mathrm{b}''(z) = \mathrm{b}''(-z)$ for all $z \in \mathbb{R}$, we have

$$\begin{aligned} u^\top \left(H(0) - H(\beta)\right) u \;&=\; \sum_{i=1}^n (X_i^\top u)^2 \left(\mathrm{b}''(0) - \mathrm{b}''(X_i^\top \beta)\right) \\ &\geq \sum_{i:|X_i^\top u| = \sqrt{q}} (X_i^\top u)^2 \left(\mathrm{b}''(0) - \mathrm{b}''(X_i^\top \beta)\right) = \sum_{i:|X_i^\top u| = \sqrt{q}} q\left(\mathrm{b}''(0) - \mathrm{b}''(1)\right) \\ &\geq n\left(\mathrm{b}''(0) - \mathrm{b}''(1)\right) > 0.05n, \end{aligned}$$

where the next-to-last step holds with probability at least $\frac{1}{2}$ by the work above.

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 $H(0) \preceq n \cdot c_2 \mathbf{I}_q$, where c_2 is a constant that is independent of the dimensions (n, q) of the problem. Then for the choice $\epsilon = 0.05/c_2$, we require that there exists some $\delta > 0$, not depending on the dimensions (n, q) of the problem, such that

$$H(\beta) \;\succeq\; (1-\epsilon)H(0) \;\succeq\; H(0) - n \cdot \epsilon c_2 \mathbf{I}_q,$$

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 $\|\beta\|_2 \leq \delta$,

$$0.05n = n \cdot \epsilon c_2 \;\geq\; u^\top \left(H(0) - H(\beta)\right) u. \tag{A.1}$$

In particular, this must be true for $\beta = u \cdot \delta'$ for any $\delta' \leq \delta$. But from the work above, the bound (A.1) is not true for $\beta = u \cdot \frac{1}{\sqrt{q}}$, and so we must have $\delta < \frac{1}{\sqrt{q}}$. This contradicts the requirement that the relationship between ϵ and δ should not depend on the dimensions of the problem.

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 β.

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 β_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 $\tau_r, \tilde{\tau}_r > 0$ be defined via

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

and

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

respectively. Assume that

$$\tau_r \leq \frac{\epsilon' \sqrt{n\, c_{\text{lower}} (\|\beta_0\|_2)^3}}{(1-\epsilon') c_{\text{change}}} \tag{B.1}$$

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

$$\left\| \text{Proj}_{\mathcal{S}_J^\perp} \left(H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) \right) \right\|_2 \leq \tilde{\tau}_{|J \setminus J_0|}, \tag{B.3}$$

where the projection is onto the orthogonal complement of the subspace

$$\mathcal{S}_J = \left\{ H_J(\beta_0)^{\frac{1}{2}} z \,:\, z \in \mathbb{R}^{J_0} \right\} \subset \mathbb{R}^J.$$

To be clear, in the definition of $\mathcal{S}_J$, 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 ϵ-net argument based on the fact that for any vector $z \in \mathbb{R}^p$,

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

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

$$\begin{aligned}\text{Prob}\left\{u^\top H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) > \tau \middle| X\right\} \\ \leq \mathbb{E}\left[\exp\left\{\tau \cdot u^\top H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) - \tau^2\right\} \middle| X\right]. \quad \text{(B.4)}\end{aligned}$$

By definition,

$$s_J(\beta_0) = \sum_{i=1}^n X_{iJ}(Y_i - \mathrm{b}'(X_i^\top \beta_0)), \quad \text{(B.5)}$$

and since the conditional distribution of Y_i given X_i belongs to an exponential family, we have

$$\mathbb{E}\left[\exp\{sY_i\}|X_i\right] = \exp\left\{\mathrm{b}(X_i^\top \beta_0 + s) - \mathrm{b}(X_i^\top \beta_0)\right\}. \quad \text{(B.6)}$$

Plugging (B.5) into (B.4) and using (B.6), we obtain that

$$\begin{aligned}&\log \text{Prob}\left\{u^\top H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) > \tau \middle| X\right\} \\ &\leq \sum_{i=1}^n \left[\mathrm{b}\left(X_i^\top(\beta_0 + \tau H_J(\beta_0)^{-\frac{1}{2}} u)\right) - \mathrm{b}\left(X_i^\top \beta_0\right)\right] \\ &\qquad - \sum_{i=1}^n \left[\mathrm{b}'(X_i^\top \beta_0) \cdot \tau X_{iJ}^\top H_J(\beta_0)^{-\frac{1}{2}} u\right] \; - \; \tau^2 \\ &= \frac{1}{2}\sum_{i=1}^n \left[\mathrm{b}''\left(X_i^\top(\beta_0 + \xi \cdot \tau H_J(\beta_0)^{-\frac{1}{2}} u)\right) \cdot \left(\tau X_{iJ}^\top H_J(\beta_0)^{-\frac{1}{2}} u\right)^2\right] \; - \; \tau^2,\end{aligned}$$

where the last equation is a 2nd-order Taylor expansion with $\xi \in [0,1]$. We may rewrite the inequality just obtained as

$$\begin{aligned}&\log \text{Prob}\left\{u^\top H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) > \tau \middle| X\right\} \\ &\qquad \leq \frac{\tau^2}{2} u^\top H_J(\beta_0)^{-\frac{1}{2}} H_J\left(\beta_0 + \xi \cdot \tau H_J(\beta_0)^{-\frac{1}{2}} u\right) H_J(\beta_0)^{-\frac{1}{2}} u \; - \; \tau^2.\end{aligned}$$

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

$$\left\|\xi \cdot \tau_r' H_J(\beta_0)^{-\frac{1}{2}} u\right\|_2 \leq \tau_r(1-\epsilon') \cdot \sqrt{\frac{1}{n\, c_{\text{lower}}(\|\beta_0\|_2)}} \leq \epsilon' \cdot \frac{c_{\text{lower}}(\|\beta_0\|_2)}{c_{\text{change}}};$$

recall (B.1). Via (2.2) and (2.5), the assumed Hessian conditions imply that

$$\begin{aligned}&H_J(\beta_0)^{-\frac{1}{2}} H_J\left(\beta_0 + \xi \cdot \tau_r' \cdot H_J(\beta_0)^{-\frac{1}{2}} u\right) H_J(\beta_0)^{-\frac{1}{2}} \\ &\qquad \preceq \; H_J(\beta_0)^{-\frac{1}{2}}\left[(1+\epsilon')H_J(\beta_0)\right] H_J(\beta_0)^{-\frac{1}{2}} \; = \; (1+\epsilon') \cdot \mathbf{I}_J,\end{aligned}$$

and thus

$$\begin{aligned}\text{Prob}\left\{u^\top H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) > \tau_r' \middle| X\right\} \\ \leq \exp\left\{\frac{\tau_r'^2}{2}(1+\epsilon') - \tau_r'^2\right\} = \exp\left\{-\frac{\tau_r'^2}{2}(1-\epsilon')\right\}. \quad \text{(B.7)}\end{aligned}$$

Next, let $\mathcal{U}_J$ be an ϵ'-net for the unit sphere in $\mathbb{R}^J$ with respect to the Euclidean norm, that is, $\mathcal{U}_J$ is a subset of the sphere such that for any unit vector v there exists a (unit) vector $u \in \mathcal{U}_J$ such that $\|u - v\|_2 < \epsilon'$. In particular, for the unit vector

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

and corresponding $u \in \mathcal{U}_J$ 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

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

We can take the ϵ-net such that

$$|\mathcal{U}_J| \leq \left(1 + \frac{2}{\epsilon'}\right)^{|J|} \leq \left(\frac{3}{\epsilon'}\right)^{|J|}; \quad \text{(B.9)}$$

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

$$\begin{aligned}&\text{Prob}\left\{\left\|H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0)\right\|_2 > \tau_r\right\} \\ &\quad \leq \text{Prob}\left\{u^\top H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) \geq \tau_r' \text{ for some } u \in \mathcal{U}_J\right\} \\ &\quad \leq |\mathcal{U}_J| \cdot \text{Prob}\left\{u^\top H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0) \geq \tau_r' \text{ for any single } u \in \mathcal{U}_J\right\}.\end{aligned}$$

Applying inequalities (B.7) and (B.9), and plugging in the definition of τ_r', we obtain that

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

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

$$\text{Prob}\left\{\left\|H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0)\right\|_2 > \tau_{|J \setminus J_0|} \text{ for some } J \supseteq J_0, |J| \leq q\right\}$$

$$\leq \sum_{r=0}^{q-|J_0|} \text{Prob}\left\{\left\|H_J(\beta_0)^{-\frac{1}{2}} s_J(\beta_0)\right\|_2 \geq \tau_r \text{ for some } J \supseteq J_0 \text{ with } |J\backslash J_0| = r\right\}.$$

Using the fact that there are at most $\binom{p}{r} \leq p^r$ sets $J \supseteq J_0$ with $|J\backslash J_0| = r$, inequality (B.10) and another union bound imply that

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

To prove the analogous statement about the projection operator, we instead take $\mathcal{U}_J$ to be an ϵ'-net of the unit sphere in the orthogonal complement $\mathcal{S}_J^\perp \subset \mathbb{R}^J$, which has dimension $|J\backslash J_0|$. Consequently, we have $|\mathcal{U}_J| \leq (3/\epsilon')^{|J\backslash J_0|}$. The rest of the argument proceeds identically with a bound of $1/(2p^\nu)$ for the probability of the considered event. A union bound over the two cases gives the claimed bound of $1/p^\nu$ 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 β_0. The statement being true uniformly for vectors with $\|\beta_0\|_2$ bounded by a_0 follows from the monotonicity of the functions c_{lower} and c_{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

$$\tilde{\gamma} = H_J(\beta_0)^{-\frac{1}{2}} \cdot \text{Proj}_{\mathcal{S}_J}\left(H_J(\beta_0)^{\frac{1}{2}}\gamma\right) \in \mathbb{R}^{J_0},$$

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

$$\left\|H_J(\beta_0)^{\frac{1}{2}}\gamma\right\|_2^2 = \left\|H_J(\beta_0)^{\frac{1}{2}}\tilde{\gamma}\right\|_2^2 + \left\|H_J(\beta_0)^{\frac{1}{2}}(\gamma - \tilde{\gamma})\right\|_2^2. \tag{B.11}$$

Using (2.3), we obtain that

$$\begin{aligned}\|\tilde{\gamma}\|_2 &\leq \frac{1}{\sqrt{nc_{\text{lower}}(\|\beta_0\|_2)}} \left\|\text{Proj}_{\mathcal{S}_J}\left(H_J(\beta_0)^{\frac{1}{2}}\gamma\right)\right\|_2 \\ &\leq \frac{1}{\sqrt{nc_{\text{lower}}(\|\beta_0\|_2)}} \left\|H_J(\beta_0)^{\frac{1}{2}}\gamma\right\|_2 \\ &\leq \sqrt{\frac{c_{\text{upper}}(\|\beta_0\|_2)}{c_{\text{lower}}(\|\beta_0\|_2)}}\, \|\gamma\|_2. \end{aligned} \tag{B.12}$$

We now compare the values of the log-likelihood function at β_0, $\beta_0+\gamma$, and $\beta_0+\tilde{\gamma}$, using Taylor-expansions. Using Proposition 2.1, we calculate

$$\begin{aligned}\log \mathrm{L}(\beta_0+\gamma) - \log \mathrm{L}(\beta_0) &= s_J(\beta_0)^\top \gamma - \frac{1}{2}\gamma^\top H_J(\beta_0+\xi\cdot\gamma)\gamma \\ &\leq s_J(\beta_0)^\top \gamma - \frac{1}{2}\left(1 - \frac{c_{\text{change}}}{c_{\text{lower}}(\|\beta_0\|_2)}\cdot\|\gamma\|_2\right)\gamma^\top H_J(\beta_0)\gamma \quad \text{(B.13)}\end{aligned}$$

and

$$\begin{aligned}\log \mathrm{L}(\beta_0+\tilde{\gamma}) - \log \mathrm{L}(\beta_0) &= s_J(\beta_0)^\top \tilde{\gamma} - \frac{1}{2}\tilde{\gamma}^\top H_J(\beta_0+\tilde{\xi}\cdot\tilde{\gamma})\tilde{\gamma} \\ &\geq s_J(\beta_0)^\top \tilde{\gamma} - \frac{1}{2}\left(1 + \frac{c_{\text{change}}}{c_{\text{lower}}(\|\beta_0\|_2)}\cdot\|\tilde{\gamma}\|_2\right)\tilde{\gamma}^\top H_J(\beta_0)\tilde{\gamma}, \quad \text{(B.14)}\end{aligned}$$

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

$$\begin{aligned}\log \mathrm{L}(\beta_0+\gamma) - \log \mathrm{L}(\beta_0+\tilde{\gamma}) \;&\leq\; s_J(\beta_0)^\top(\gamma-\tilde{\gamma}) \\ -\frac{1}{2}\left\|H_J(\beta_0)^{\frac{1}{2}}(\gamma-\tilde{\gamma})\right\|_2^2 &+ \frac{c_{\text{change}}\left(\|\gamma\|_2\gamma^\top H_J(\beta_0)\gamma + \|\tilde{\gamma}\|_2\tilde{\gamma}^\top H_J(\beta_0)\tilde{\gamma}\right)}{2c_{\text{lower}}(\|\beta_0\|_2)}.\end{aligned}$$

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

$$\begin{aligned}\log \mathrm{L}(\beta_0+\gamma) - \log \mathrm{L}(\beta_0+\tilde{\gamma}) \;&\leq\; s_J(\beta_0)^\top(\gamma-\tilde{\gamma}) \\ -\frac{1}{2}\left\|H_J(\beta_0)^{\frac{1}{2}}(\gamma-\tilde{\gamma})\right\|_2^2 &+ n\cdot c_{\text{change}}\left(\frac{c_{\text{upper}}(\|\beta_0\|_2)}{c_{\text{lower}}(\|\beta_0\|_2)}\right)^{\frac{3}{2}}\|\gamma\|_2^3. \quad \text{(B.15)}\end{aligned}$$

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

$$\begin{aligned}s_J(\beta_0)^\top(\gamma-\tilde{\gamma}) - \frac{1}{2}\left\|H_J(\beta_0)^{\frac{1}{2}}(\gamma-\tilde{\gamma})\right\|_2^2 &\leq \sup_{z\in\mathcal{S}_J^\perp}\left(H_J(\beta_0)^{-\frac{1}{2}}s_J(\beta_0)\right)^\top z - \frac{1}{2}\|z\|_2^2 \\ &= \frac{1}{2}\left\|\mathrm{Proj}_{\mathcal{S}_J^\perp}\left(H_J(\beta_0)^{-\frac{1}{2}}s_J(\beta_0)\right)\right\|_2^2,\end{aligned}$$

which is at most $\tilde{\tau}^2_{|J\setminus J_0|}/2$ by the assumed inequality (B.3).

Consider now the MLE $\beta = \widehat{\beta}_J = \beta_0+\gamma$, and define $\tilde{\gamma} \in \mathbb{R}^{J_0}$ as before. Then $\log \mathrm{L}(\widehat{\beta}_{J_0}) \geq \log \mathrm{L}(\beta_0+\tilde{\gamma})$ because $\beta_0+\tilde{\gamma} \in \mathbb{R}^{J_0}$, and so applying the calculations above, we have

$$\log \mathrm{L}(\widehat{\beta}_J) - \log \mathrm{L}(\widehat{\beta}_{J_0}) \leq \log \mathrm{L}(\widehat{\beta}_J) - \log \mathrm{L}(\beta_0+\tilde{\gamma})$$

$$\leq \frac{1}{2}\tilde{\tau}^2_{|J\setminus J_0|} + n \cdot c_{\text{change}} \left(\frac{c_{\text{upper}}(\|\beta_0\|_2)}{c_{\text{lower}}(\|\beta_0\|_2)} \right)^{\frac{3}{2}} \cdot \|\widehat{\beta}_J - \beta_0\|_2^3. \tag{B.16}$$

We can thus bound the difference between the maxima of the log-likelihood functions if we can bound the distance $\|\widehat{\beta}_J - \beta_0\|_2$.

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

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

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

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

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

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

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

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

Since $\log \mathrm{L}(\hat{\beta}_J) - \log \mathrm{L}(\beta_0) \geq 0$, this shows that

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

as long as we assume that

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

Taking up (B.16), we get

$$\log \mathrm{L}(\widehat{\beta}_J) - \log \mathrm{L}(\widehat{\beta}_{J_0}) \;\leq \frac{1}{2}\tilde{\tau}^2_{|J\backslash J_0|} + n \cdot c_{\mathrm{change}} \left(\frac{c_{\mathrm{upper}}(\|\beta_0\|_2)}{c_{\mathrm{lower}}(\|\beta_0\|_2)} \right)^{\frac{3}{2}} \left(\frac{4\sqrt{c_{\mathrm{upper}}(\|\beta_0\|_2)} \cdot \tau_{|J\backslash J_0|}}{\sqrt{n} c_{\mathrm{lower}}(\|\beta_0\|_2)} \right)^3 .$$

If

$$\sqrt{n} \geq \frac{2c_{\mathrm{change}}}{\epsilon'} \left(\frac{c_{\mathrm{upper}}(\|\beta_0\|_2)}{c_{\mathrm{lower}}(\|\beta_0\|_2)} \right)^{\frac{3}{2}} \left(\frac{4\sqrt{c_{\mathrm{upper}}(\|\beta_0\|_2)}}{c_{\mathrm{lower}}(\|\beta_0\|_2)} \right)^3 \cdot \frac{\tau^3_{|J\backslash J_0|}}{\tilde{\tau}^2_{|J\backslash J_0|}}, \tag{B.18}$$

then we get

$$\begin{aligned} \log \mathrm{L}(\widehat{\beta}_J) - \log \mathrm{L}(\widehat{\beta}_{J_0}) &\leq \frac{1}{2}\tilde{\tau}^2_{|J\backslash J_0|} \cdot (1+\epsilon') \\ &= \frac{(1+\epsilon')}{(1-\epsilon')^3} \cdot \Big(|J\backslash J_0| \log\left(\frac{6p}{\epsilon'}\right) + \log(4p^\nu) \Big). \end{aligned} \tag{B.19}$$

Hence, this inequality holds whenever (B.18) holds. Now, to determine a simpler lower bound on n, we calculate

$$\frac{\tau^2_{|J\backslash J_0|}}{\tilde{\tau}^2_{|J\backslash J_0|}} = \frac{\frac{2}{(1-\epsilon')^3} \cdot \left[|J| \log\left(\frac{3}{\epsilon'}\right) + \log(4p^\nu) + |J\backslash J_0| \log(2p) \right]}{\frac{2}{(1-\epsilon')^3} \cdot \left[|J\backslash J_0| \log\left(\frac{3}{\epsilon'}\right) + \log(4p^\nu) + |J\backslash J_0| \log(2p) \right]} \leq \frac{|J|}{|J\backslash J_0|} \leq q.$$

Hence, (B.18) holds as long as n exceeds a constant multiple of $q^2\tau^2_{|J\backslash J_0|}$. For p large enough, which we can ensure by choice of the constant C_{dim}, we have that $\tau^2_{|J\backslash J_0|}$ is no larger than a constant times $q\log(p)$. So, by choosing the constant $C_{\mathrm{sample},1}$ appropriately, (B.18) holds as long as

$$n \geq C_{\mathrm{sample},1} \cdot q^3 \log(p).$$

Now fix $\epsilon' \in (0, \epsilon)$ such that

$$\frac{(1+\epsilon')^2}{(1-\epsilon')^3} < 1 + \epsilon. \tag{B.20}$$

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

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

which implies, by (B.19) and (B.20), that

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

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

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

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

$$\log \mathrm{L}(\widehat{\beta}_J) - \log \mathrm{L}(\beta_0) \leq \min_{j \in J_0} |(\beta_0)_j| \left(\sqrt{n c_{\mathrm{upper}}(\|\beta_0\|_2)} \cdot \tau_{|J \setminus J_0|} - \frac{n c_{\mathrm{lower}}(\|\beta_0\|_2)}{4} \min \left\{ \min_{j \in J_0} |(\beta_0)_j|, \frac{c_{\mathrm{lower}}(\|\beta_0\|_2)}{2 c_{\mathrm{change}}} \right\} \right),$$

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 $C_{\mathrm{sample},2}$. In particular, for appropriate choice of $C_{\mathrm{sample},2}$, we get

$$\log \mathrm{L}(\widehat{\beta}_J) - \log \mathrm{L}(\beta_0) \leq - \min_{j \in J_0} |(\beta_0)_j| \frac{n c_{\mathrm{lower}}(\|\beta_0\|_2)}{8} \min \left\{ \min_{j \in J_0} |(\beta_0)_j|, \frac{c_{\mathrm{lower}}(\|\beta_0\|_2)}{2 c_{\mathrm{change}}} \right\}.$$

Since $\min_{j \in J_0} |(\beta_0)_j|$ is also upper bounded by a constant, namely $\min_{j \in J_0} |(\beta_0)_j| \leq \|\beta_0\|_2 \leq a_0$, this is sufficient to prove claim (b) of Theorem 2.2.

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

$$c_{\mathrm{change}} = 1 + a_2 + 12\sqrt{2}\, a_3^3. \tag{C.1}$$

We claim that if $n \geq q^3 \log(2p)$, then with probability at least

$$1 - \exp\left\{ -\frac{n}{2 a_3^6 q^3} \right\}, \tag{C.2}$$

we have

$$\sup_{J \subseteq [p], |J| \leq q} \sup_{\beta \neq \beta' \in \mathbb{R}^J} \frac{\|H_J(\beta) - H_J(\beta')\|}{\|\beta - \beta'\|_2} \leq c_{\mathrm{change}} \cdot n.$$

To show this claim, take any set J with $|J| \leq q$, any unit vector $u \in \mathbb{R}^J$ and any pair of distinct vectors $\beta \neq \beta' \in \mathbb{R}^J$. Then we have

$$\begin{aligned}\left|u^\top \left(H(\beta) - H(\beta')\right) u\right| &\leq \sum_{i=1}^n (X_i^\top u)^2 \cdot \left|\mathrm{b}''(X_i^\top \beta) - \mathrm{b}''(X_i^\top \beta')\right| \\ &\leq \sum_{i=1}^n (X_i^\top u)^2 \cdot \left|X_i^\top \beta - X_i^\top \beta'\right| \cdot \max_{t \in [0,1]} \left|\mathrm{b}'''(X_i^\top (t\beta + (1-t)\beta')\right|.\end{aligned}$$

Define the unit vector $v = \frac{\beta - \beta'}{\|\beta - \beta'\|_2} \in \mathbb{R}^J$. In logistic regression, $|\mathrm{b}'''(z)| \leq 1$ for all $z \in \mathbb{R}$. Using this fact[4], we obtain that

$$\begin{aligned}\left|u^\top \left(H(\beta) - H(\beta')\right) u\right| &\leq \|\beta - \beta'\|_2 \cdot n \left(\frac{1}{n}\sum_{i=1}^n (X_i^\top u)^2 \cdot \left|X_i^\top v\right|\right) \\ &\leq \|\beta - \beta'\|_2 \cdot n \left(\frac{1}{n}\sum_{i=1}^n |X_i^\top u|^3\right)^{\frac{2}{3}} \left(\frac{1}{n}\sum_{i=1}^n |X_i^\top v|^3\right)^{\frac{1}{3}} \\ &\leq \|\beta - \beta'\|_2 \cdot n \cdot \left(\sup_{q\text{-sparse unit } w} \frac{1}{n}\sum_{i=1}^n |X_i^\top w|^3\right).\end{aligned}$$

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

$$\sup_{q\text{-sparse unit } w} \frac{1}{n}\sum_{i=1}^n |X_i^\top w|^3 \;\leq\; 1 + a_2 + 12\sqrt{2}\, a_3^3 = c_{\text{change}},$$

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

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

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

C.2. Upper bound on Hessian

In this subsection, we prove that if inequality (2.4) holds, then with probability at least

$$1 - \exp\left\{-\frac{n}{2a_3^4 q^2}\right\}, \tag{C.3}$$

it also holds that

$$H_J(\beta) \preceq n \cdot c_{\text{upper}}(\|\beta\|_2) \cdot \mathbf{I}_J$$

[4]For other exponential families, one could bound the $\mathrm{b}'''(\cdot)$ term by taking q to be constant and only considering β and β' of bounded norm.

for all J with $|J| \le q$ and all $\beta \in \mathbb{R}^J$. Here, we define

$$c_{\text{upper}}(r) \coloneqq \mathrm{b}''(0) \cdot \left(1 + a_2 + 8\sqrt{2}\, a_3^2\right) + c_{\text{change}} \cdot r$$

where c_{change} is the constant from (C.1) and $\mathrm{b}''(0) = 1/4$ for logistic regression. The idea for our proof is to show that, on a suitable event, $\sup_{|J|\le q}\|H_J(0)\| = \mathbf{O}(n)$. 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

$$\mathbb{E}\left[(X^\top u)^2\right] \le \mathbb{E}\left[|X^\top u|^3\right]^{\frac{2}{3}} \le a_2^{\frac{2}{3}}.$$

Then we have, with at least the probability in (C.3),

$$\begin{aligned}
\sup_{|J|\le q} \|H_J(0)\| &= \sup_{|J|\le q} \left\| \sum_{i=1}^n X_{iJ} X_{iJ}^\top \mathrm{b}''(X_i^\top 0) \right\| \\
&= \mathrm{b}''(0) \cdot \sup_{|J|\le q} \left\| \sum_{i=1}^n X_{iJ} X_{iJ}^\top \right\| \\
&= \mathrm{b}''(0) \cdot \sup_{|J|\le q,\ \text{unit } u\in\mathbb{R}^J} \sum_{i=1}^n (X_i^\top u)^2 \\
&\le \mathrm{b}''(0) \cdot n \left(1 + a_2 + 8\sqrt{2}\, a_3^2\right),
\end{aligned}$$

where for the last step we apply Corollary D.1 with $k = 2$, using the assumption that $n \ge q^2 \log(2p)$. The bounded change condition from (2.4) now implies the desired conclusion, namely, that for all J with $|J| \le q$ and all $\beta \in \mathbb{R}^J$,

$$\|H_J(\beta)\| \le \|H_J(0)\| + \|H_J(0) - H_J(\beta)\| \le n \cdot c_{\text{upper}}(\|\beta\|_2).$$

C.3. Lower bound on Hessian

Finally, we prove that with probability at least

$$1 - 2\exp\left\{-\frac{n}{2} \cdot \left(\frac{a_1^3}{512 a_2^2}\right)^2\right\}, \tag{C.4}$$

it holds for all $|J| \le q$, for all $|J| \le q$ and for all $\beta \in \mathbb{R}^J$ that

$$H_J(\beta) \succeq n \cdot c_{\text{lower}}(\|\beta\|_2) \cdot \mathbf{I}_J,$$

where

$$c_{\text{lower}}(r) \coloneqq \frac{a_1^4}{2048 a_2^2} \cdot \min_{|z| \le r \cdot 2\sqrt[3]{256} a_2/a_1} \mathrm{b}''(z)$$

$$= \frac{a_1^4}{2048 a_2^2} \cdot \frac{\exp\{r \cdot 2\sqrt[3]{256} a_2/a_1\}}{\left(1 + \exp\{r \cdot 2\sqrt[3]{256} a_2/a_1\}\right)^2}$$

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) below.

For a proof, since

$$H_J(\beta) = \sum_{i=1}^{n} X_{iJ} X_{iJ}^\top \, \mathrm{b}''(X_i^\top \beta),$$

we consider the quantity

$$\sum_{i=1}^{n} (X_i^\top u)^2 \mathrm{b}''(X_i^\top \beta)$$

where $u \in \mathbb{R}^J$ is a unit vector. For any choice of $w_1, w_2 \geq 0$, we have

$$\sum_{i=1}^{n} (X_i^\top u)^2 \mathrm{b}''(X_i^\top \beta) \;\geq\; \sum_{i=1}^{n} w_1^2 \min_{|z| \leq \|\beta\|_2 w_2} \mathrm{b}''(z) \cdot \mathbb{1}_{\{|X_i^\top u| \geq w_1, |X_i^\top \beta| \leq \|\beta\|_2 w_2\}}.$$

Using the symmetry and monotonicity of b'' for logistic regression we find

$$\sum_{i=1}^{n} (X_i^\top u)^2 \mathrm{b}''(X_i^\top \beta) \;\geq\; n \cdot w_1^2 \, \mathrm{b}''(\|\beta\|_2 w_2) \times \left(1 - \frac{\#\{i : |X_i^\top u| < w_1\}}{n} - \frac{\#\left\{i : |X_i^\top \beta|/\|\beta\|_2 > w_2\right\}}{n}\right). \quad \text{(C.5)}$$

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 β.

By Lemma D.3, for any $t > 0$, with probability at least $1 - 2e^{-nt^2/2}$, for all q-sparse unit vectors u,

$$\frac{\#\{i : |X_i^\top u| < w_1\}}{n} \;\leq\; \mathrm{Prob}\left\{|X_1^\top u| < w_1 + \frac{1}{t}\sqrt{\frac{32 a_3^2 q \log(2p)}{n}}\right\} + 2t \quad \text{(C.6)}$$

and

$$\frac{\#\{i : |X_i^\top u| > w_2\}}{n} \;\leq\; \mathrm{Prob}\left\{|X_1^\top u| > w_2 - \frac{1}{t}\sqrt{\frac{32 a_3^2 q \log(2p)}{n}}\right\} + 2t. \quad \text{(C.7)}$$

Now set

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

and assume

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

Then, for shorter notation, define the two scalars

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

Assume now that (C.6) and (C.7) hold. We begin by simplifying the bound in (C.6). By (C.8), ${w_1'}^2 \le \frac{a_1}{2}$, and so applying Lemma D.5 with $Z = (X_1^\top u)^2$, $h(Z) = \sqrt{Z}$ and $a = {w_1'}^2$ yields that for all q-sparse unit vectors u,

$$\begin{aligned}\text{Prob}\left\{|X_1^\top u| < w_1'\right\} &\le 1 - \frac{\mathbb{E}\left[(X_1^\top u)^2\right] - {w_1'}^2}{\inf\left\{x \ge 0 : \sqrt{x} > \frac{\mathbb{E}\left[|X_1^\top u|^3\right]}{\mathbb{E}\left[(X_1^\top u)^2\right] - {w_1'}^2}\right\}} \\ &\le 1 - \frac{a_1 - {w_1'}^2}{\inf\left\{x \ge 0 : \sqrt{x} > \frac{\mathbb{E}\left[|X_1^\top u|^3\right]}{\mathbb{E}\left[(X_1^\top u)^2\right] - {w_1'}^2}\right\}}.\end{aligned}$$

The term involving the supremum satisfies

$$\frac{a_1 - {w_1'}^2}{\inf\left\{x \ge 0 : \sqrt{x} > \frac{a_2}{a_1 - {w_1'}^2}\right\}} = \frac{a_1 - {w_1'}^2}{\left(\frac{a_2}{a_1 - {w_1'}^2}\right)^2} = \frac{(a_1 - {w_1'}^2)^3}{8a_2^2} \ge \frac{a_1^3}{64a_2^2},$$

and so

$$\sup_{q\text{-sparse unit } u} \frac{\#\{i : |X_i^\top u| < w_1\}}{n} \;\le\; \left(1 - \frac{a_1^3}{64a_2^2}\right) + 2t \;\le\; 1 - \frac{3a_1^3}{256a_2^2}.$$

Next, we simplify the bound in (C.7). By (C.8), for any u,

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

By Markov's inequality,

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

We obtain that

$$\sup_{q\text{-sparse unit } u} \frac{\#\{i : |X_i^\top u| > w_2\}}{n} \;\le\; \frac{a_1^3}{256a_2^2} + 2t \;\le\; \frac{a_1^3}{128a_2^2}.$$

Returning to (C.5), we conclude that, with at least the probability from (C.4), for all $|J| \le q$ and all unit $u \in \mathbb{R}^J$ and all $\beta \in \mathbb{R}^J$,

$$\sum_{i=1}^n (X_i^\top u)^2\, \mathrm{b}''(X_i^\top \beta) \ge n \cdot w_1^2 \mathrm{b}''(\|\beta\|_2 w_2) \cdot \left[1 - \left(1 - \frac{3a_1^3}{256a_2^2}\right) - \frac{a_1^3}{128a_2^2}\}\right]$$

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

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 $1 - e^{-t^2/2}$,*

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

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). Recall that a random variable Z is σ-subgaussian if, for all $t \in \mathbb{R}$,

$$\mathbb{E}\left[e^{tZ} \right] \leq e^{t^2 \sigma^2 / 2}.$$

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

$$\mathbb{E}\left[\max_{1 \leq i \leq m} |Z_i| \right] \leq \sigma \cdot \sqrt{2 \log(2m)}.$$

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 $1 - e^{-t^2 n/2}$, for all q-sparse unit vectors u,*

$$\frac{1}{n}\#\{i : |X_i^\top u| < a\} \;\le\; \text{Prob}\left\{|X^\top u| < a + \frac{1}{t}\sqrt{\frac{32a_3^2 q\log(2p)}{n}}\right\} + 2t.$$

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

$$\frac{1}{n}\#\{i : |X_i^\top u| > a\} \;\le\; \text{Prob}\left\{|X^\top u| > a - \frac{1}{t}\sqrt{\frac{32a_3^2 q\log(2p)}{n}}\right\} + 2t.$$

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

$$c \coloneqq t^{-1}\sqrt{\frac{32a_3^2 q\log(2p)}{n}},$$

and define the piece-wise linear function

$$g(z) = \begin{cases} 1 & \text{if } |z| \le a, \\ 0 & \text{if } |z| \ge a + c, \\ (a + c - |z|)/c & \text{if } a \le |z| \le a + c. \end{cases}$$

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

$$\mathbb{1}_{\{|z|<a\}} \le g(z) \le \mathbb{1}_{\{|z|<a+c\}}.$$

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

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

Assume from now on that this event occurs.

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

$$\mathbb{E}_{\nu,X}\left[\sup_{q\text{-sparse unit } u}\left|\sum_{i=1}^n \nu_i X_i^\top u\right|\right] \le \mathbb{E}_{\nu,X}\left[\sup_{|J|\le q}\left\|\sum_{i=1}^n \nu_i X_{iJ}\right\|_2\right] \le \sqrt{q}\,\mathbb{E}_{\nu,X}\left[\sup_{1\le j\le p}\left|\sum_{i=1}^n \nu_i X_{ij}\right|\right]. \quad \text{(D.2)}$$

By assumption (A3), $|X_{ij}| \le a_3$ for all i, j, and therefore $\sum_{i=1}^n \nu_i X_{ij}$ is $(a_3\sqrt{n})$-subgaussian because, by independence of the X_i's,

$$\mathbb{E}\left[e^{t\sum_{i=1}^n \nu_i X_{ij}}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{t\nu_i X_{ij}}\right] \le \prod_{i=1}^n \mathbb{E}\left[\mathbb{E}\left[e^{t\nu_i X_{ij}} \mid X_{ij}\right]\right]$$
$$= \prod_{i=1}^n \frac{1}{2}\mathbb{E}\left[e^{tX_{ij}} + e^{-tX_{ij}}\right] \le \prod_{i=1}^n \mathbb{E}\left[e^{t^2 X_{ij}^2/2}\right] \le e^{nt^2 a_3^2/2}.$$

Applying Lemma D.2 to (D.2), we obtain that

$$\mathbb{E}_{\nu,X}\left[\sup_{q\text{-sparse unit } u}\left|\sum_{i=1}^n \nu_i X_i^\top u\right|\right] \;\le\; \sqrt{q}\cdot a_3\sqrt{n}\cdot\sqrt{2\log(2p)}. \tag{D.3}$$

We deduce that

$$\sup_{q\text{-sparse unit } u}\left|\sum_{i=1}^n g(X_i^\top u) - \mathbb{E}\left[g(X^\top u)\right]\right|$$
$$\le\; 4c^{-1}\sqrt{q}\cdot a_3\sqrt{n}\cdot\sqrt{2\log(2p)} + nt \;=\; 2nt,$$

by our choice of c. Returning to (D.1), we have shown that, as desired,

$$\sup_{q\text{-sparse unit } u} \#\{i : |X_i^\top u| < a\} \;\le\; \sup_{q\text{-sparse unit } u}\sum_{i=1}^n g(X_i^\top u)$$
$$\le\; n\mathbb{E}\left[g(X^\top u)\right] + 2nt \;\le\; n\,\mathrm{Prob}\{|X^\top u| < a + c\} + 2nt.$$

□

D.3. Bounding functions of sparse unit linear combinations

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

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

for all $0 \le z, z' \le a_3\sqrt{q}$. If $\mathbb{E}[f(|X^\top u|)] \le a_2$ for all q-sparse unit vectors u, then under assumption (A3), it holds with probability at least $1 - e^{-n/(2M^2)}$ that

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

Proof. Define $h(z) = f(|z|)$ for $z \in \mathbb{R}$. By our assumptions on f, the function h is M-bounded and L-Lipschitz over $z \in [-a_3\sqrt{q}, a_3\sqrt{q}] \subset \mathbb{R}$.

Applying the concentration bound from Lemma D.1 to h, with $t = \sqrt{n}/M$, we obtain that with probability at least $1 - e^{-n/(2M^2)}$,

$$\sup_{|J|\le q,\ \text{unit } u\in\mathbb{R}^J}\left|\sum_{i=1}^n \left(h(X_i^\top u) - \mathbb{E}\left[h(X^\top u)\right]\right)\right|$$

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

where the second inequality follows from (D.3) using the same reasoning as in the proof of Lemma D.3. Since $h(z) = f(|z|)$ for all $z \in \mathbb{R}$, we have

$$\mathbb{E}\left[h(X^\top u)\right] = \mathbb{E}\left[f(|X_i^\top u|)\right] \leq a_2.$$

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

$$\sup_{|J|\leq q,\ \text{unit } u\in\mathbb{R}^J}\sum_{i=1}^n h(X_i^\top u) \ \leq\ n\left(1 + a_2 + 4La_3\sqrt{\frac{2q\log(2p)}{n}}\right). \tag{D.4}$$

□

We obtain the following corollary about moments of sparse unit linear combinations.

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

$$\sup_{|J|\leq q,\ \textit{unit } u\in\mathbb{R}^J}\sum_{i=1}^n |X_i^\top u|^k \leq n\left(1 + a_2 + 4\sqrt{2}\cdot k a_3^k\sqrt{\frac{q^k\log(2p)}{n}}\right)$$

holds with probability at least

$$1 - \exp\left\{-\frac{n}{2a_3^{2k}q^k}\right\}.$$

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

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

and collecting terms to find the upper bound. □

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

Lemma D.5. *Let $h : [0,\infty) \to [0,\infty)$ be a continuous and nondecreasing function such that $z \mapsto z\cdot h(z)$ is convex. Let $Z \geq 0$ be a random variable with $\mathbb{E}[Z] < \infty$ and $\mathbb{E}[Z\cdot h(Z)] < \infty$. Then for any $a \leq \mathbb{E}[Z]$,*

$$\text{Prob}\{Z > a\} \geq \frac{\mathbb{E}[Z] - a}{\inf\left\{x \geq 0 : h(x) > \frac{\mathbb{E}[Z\cdot h(Z)]}{\mathbb{E}[Z]-a}\right\}}.$$

Remark D.1. This inequality is an extension of the Paley-Zygmund inequality (Shorack, 2000, Inequality 4.9), which makes the same statement for $h(z) = z$.

Proof. First, we have

$$\mathbb{E}\left[Z \cdot \mathbb{1}_{Z>a}\right] = \mathbb{E}\left[Z\right] - \mathbb{E}\left[Z \cdot \mathbb{1}_{Z\leq a}\right] \geq \mathbb{E}\left[Z\right] - a. \tag{D.5}$$

Now, let Y be a random variable whose distribution is equal to the distribution of Z conditional on $Z > a$, that is,

$$\text{Prob}\left\{Y \leq y\right\} = \begin{cases} 0, & y \leq a, \\ \frac{\text{Prob}\{a<Z\leq y\}}{\text{Prob}\{Z>a\}}, & y > a. \end{cases}$$

Next, write $g(z) = z \cdot h(z)$, which by assumption is convex, continuous, and strictly increasing, with $g(0) = 0$. We can then define the inverse $g^{-1} : \mathbb{R}_+ \mapsto \mathbb{R}_+$, which is also strictly increasing. Then, by Jensen's inequality,

$$\begin{aligned} \mathbb{E}\left[Z \cdot \mathbb{1}_{Z>a}\right] &= \mathbb{E}\left[Y\right] \cdot \text{Prob}\left\{Z > a\right\} \\ &\leq g^{-1}\left(\mathbb{E}\left[g(Y)\right]\right) \cdot \text{Prob}\left\{Z > a\right\} \\ &= g^{-1}\left(\mathbb{E}\left[g(Z) \mid Z > a\right]\right) \cdot \text{Prob}\left\{Z > a\right\} \\ &= g^{-1}\left(\frac{\mathbb{E}\left[g(Z) \cdot \mathbb{1}_{Z>a}\right]}{\text{Prob}\left\{Z > a\right\}}\right) \cdot \text{Prob}\left\{Z > a\right\} \\ &\leq g^{-1}\left(\frac{\mathbb{E}\left[g(Z)\right]}{\text{Prob}\left\{Z > a\right\}}\right) \cdot \text{Prob}\left\{Z > a\right\} \end{aligned}$$

and so

$$g^{-1}\left(\frac{\mathbb{E}\left[g(Z)\right]}{\text{Prob}\left\{Z > a\right\}}\right) \geq \frac{\mathbb{E}\left[Z \cdot \mathbb{1}_{Z>a}\right]}{\text{Prob}\left\{Z > a\right\}} \geq \frac{\mathbb{E}\left[Z\right] - a}{\text{Prob}\left\{Z > a\right\}},$$

which implies

$$\frac{\mathbb{E}\left[g(Z)\right]}{\text{Prob}\left\{Z > a\right\}} \geq g\left(\frac{\mathbb{E}\left[Z\right] - a}{\text{Prob}\left\{Z > a\right\}}\right).$$

Rewriting this last conclusion in terms of h gives

$$\frac{\mathbb{E}\left[Z \cdot h(Z)\right]}{\text{Prob}\left\{Z > a\right\}} \geq \frac{\mathbb{E}\left[Z\right] - a}{\text{Prob}\left\{Z > a\right\}} \cdot h\left(\frac{\mathbb{E}\left[Z\right] - a}{\text{Prob}\left\{Z > a\right\}}\right)$$

and thus

$$\frac{\mathbb{E}\left[Z \cdot h(Z)\right]}{\mathbb{E}\left[Z\right] - a} \geq h\left(\frac{\mathbb{E}\left[Z\right] - a}{\text{Prob}\left\{Z > a\right\}}\right).$$

We conclude that for any x such that

$$h(x) > \frac{\mathbb{E}\left[Z \cdot h(Z)\right]}{\mathbb{E}\left[Z\right] - a},$$

we must have $\frac{\mathbb{E}[Z]-a}{\text{Prob}\{Z>a\}} \leq x$ because h is nondecreasing. This proves that

$$\text{Prob}\left\{Z > a\right\} \geq \frac{\mathbb{E}\left[Z\right] - a}{\inf\left\{x \geq 0 : h(x) > \frac{\mathbb{E}[Z \cdot h(Z)]}{\mathbb{E}[Z]-a}\right\}}.$$

□

References

Anandkumar, A., Tan, V. Y. F., Huang, F. and Willsky, A. S. (2012). High-dimensional structure estimation in Ising models: Local separation criterion. *Ann. Statist.* **40** 1346–1375. MR3015028

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

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *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. *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. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 641–656. MR1979381

Bühlmann, P. and van de Geer, S. (2011). *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. *Biometrika* **95** 759–771. MR2443189

Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *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.

Foygel, R. and Drton, M. (2011). Bayesian model choice and information criteria in sparse generalized linear models. *ArXiv e-prints* 1112.5635.

Foygel, R. and Drton, M. (2014). High-dimensional Ising model selection with Bayesian information criteria. *ArXiv e-prints* 1403.3374v1.

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. *R package version 2.2-3*.

Jalali, A., Johnson, C. C. and Ravikumar, P. K. (2011). On learning discrete graphical models using greedy methods. In *Advances in Neural Information Processing Systems* 1935–1943.

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. (2013). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Ann. Statist.* **41** 3022–3049. MR3161456

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. *Stat. Interface* **6** 275–284. MR3066691

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. MR2758523

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 ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343 (2011d:62066)

Roudi, Y., Aurell, E. and Hertz, J. A. (2009). Statistical physics of pairwise probability models. *Frontiers in Computational Neuroscience* **3**.

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. MR2943079

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)

Żak-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)