Logistic regression and Ising networks: Prediction and estimation when violating lasso assumptions

Waldorp, L.; Marsman, M.; Maris, G.

DOI
10.1007/s41237-018-0061-0

Publication date
2019

Document Version
Final published version

Published in
Behaviormetrika

Behaviormetrika (2019) 46:49–72
https://doi.org/10.1007/s41237-018-0061-0

ORIGINAL PAPER

Logistic regression and Ising networks: prediction and estimation when violating lasso assumptions

Lourens Waldorp[1] · **Maarten Marsman**[1] · **Gunter Maris**[1,2]

Received: 23 July 2018 / Accepted: 31 July 2018 / Published online: 11 August 2018

Abstract
The Ising model was originally developed to model magnetisation of solids in statistical physics. As a network of binary variables with the probability of becoming 'active' depending only on direct neighbours, the Ising model appears appropriate for many other processes. For instance, it was recently applied in psychology to model co-occurrences of mental disorders. It has been shown that the connections between the variables (nodes) in the Ising network can be estimated with a series of logistic regressions. This naturally leads to questions of how well such a model predicts new observations and how well parameters of the Ising model can be estimated using logistic regressions. Here we focus on the high-dimensional setting with more parameters than observations and consider violations of assumptions of the lasso. In particular, we determine the consequences for both prediction and estimation when the sparsity and restricted eigenvalue assumptions are not satisfied. We explain using the idea of connected copies (extreme multicollinearity) the fact that prediction becomes better when either sparsity or multicollinearity is not satisfied. We illustrate these results with simulations.

Keywords Ising model · Bernoulli graphs · Variational inference

1 Introduction

Logistic regression models the ratio of success versus failure for binary variables. These models are convenient and useful in many situations. To name one example, in genome wide association scans (GWAS), a combination of alleles

Communicated by Joe Suzuki.

✉ Lourens Waldorp
waldorp@uva.nl

[1] University of Amsterdam, Psychological Methods, Nieuwe Achtergracht 129-B, 1001 NK Amsterdam, The Netherlands

[2] ACT Next, 500 ACT Drive, Iowa City, IA 52245, USA

on single nucleotide polymorphisms (SNPs) is either present or not (Cantor et al. 2010). Recently, it became clear that logistic regression can also be used to obtain estimates of connections in a binary network (e.g. Ravikumar et al. 2010; Bühlmann and van de Geer 2011). A particular version of a binary network is the Ising model, in which the probability of a node being 'active' is determined only by its direct neighbours (pairwise interactions only). The Ising model originated in statistical physics and was used to model magnetisation of solids (Kindermann et al. 1980; Cipra 1987; Baxter 2007), and was investigated extensively by Besag (1974) and Cressie (1993) and recently by Marsman et al. (2017), amongst others, in a statistical modelling context. Recently, the Ising model has also been applied to modelling networks of mental disorders (Borsboom et al. 2011; van Borkulo et al. 2014). The objective in models of psychopathology is to both explain and predict certain observations such as co-occurrences of disorders (comorbidity).

Here we focus on violations of the assumptions of lasso in logistic regression with high-dimensional data (more parameters than observations, $p > n$). In particular, we consider the consequences for both prediction and estimation when violating the assumptions of sparsity and restricted eigenvalues (multicollinearity). For sparse models and $p > n$, it has been shown that statistical guarantees about the underlying network and its coefficients can be obtained with certain assumptions for Gaussian data (Meinshausen and Bühlmann 2006; Bickel et al. 2009; Hastie et al. 2015), for discrete data (Loh and Wainwright 2012), and for exponental family distributions (Bühlmann and van de Geer 2011). Specifically, Ravikumar et al. (2010) show that, under strong regularity conditions, using a series of regressions for the conditional probability in the Ising model (logistic regression), the correct structure (topology) of a network can be obtained in the high-dimensional setting.

In many practical settings, it is uncertain whether the assumptions of the lasso for accurate network estimation hold. Specifically, the assumptions of sparsity and the restricted eigenvalues (Bickel et al. 2009) are in many situations untestable. We therefore investigate here how estimation and prediction in Ising networks are affected by violating the sparsity and restricted eigenvalue assumptions. The setting of logistic regression and nodewise estimation of the Ising model parameters allows us to clearly determine how and why prediction and estimation are affected. We use the idea of connected nodes in a graph that are identical in the observations (and call them connected copies) to show why prediction is better for graph structures that violate the restricted eigenvalue or sparsity assumption. These connected copies represent the idea of extreme multicollinearity. One way to view connected copies is obtaining edge weights that lead to a network with perfect correlations between nodes (variables). We therefore compare in terms of prediction and estimation different situations where we violate the restricted eigenvalue or sparsity assumption based on different data generating processes. An example of a setting where near connected copies in networks are found is in high resolution functional magnetic resonance imaging. Here time series of contiguous voxels that are connected (also physically, see e.g. Johansen-Berg et al. 2004), are near exact copies of one another. The concept of connected copies allows us to determine the consequences for prediction loss, using the fact that subsets of connected copies do not change the risk

or ℓ_1 norm. We also show that prediction loss is a lower bound for estimation error (in ℓ_1) and so by consequence, if prediction loss increases, so does estimation error.

We first provide some background in Sect. 2 on the Ising model and its relation to logistic regression. To show the consequences of violating the assumptions of multicollinearity and sparsity, we discuss these assumptions at length in Sect. 3. We also show how they provide the statistical guarantees for the lasso (e.g. Negahban et al. 2012; Bühlmann and van de Geer 2011; Ravikumar et al. 2010). Then armed with these intuitions, we give in Sect. 4 some insight into the consequences for prediction and estimation when the sparsity or restricted eigenvalue assumption is violated. We also provide some simulations to confirm our results.

2 Logistic regression and the Ising model

The Ising model is part of the exponential family of distributions (see, e.g. Brown 1986; Young and Smith 2005; Wainwright and Jordan 2008). Let $X = (X_1, X_2, \ldots, X_p)$ be a random variable with values in $\{0, 1\}^p$. The Ising model can then be defined as follows. Let G be a graph consisting of nodes in $V = \{1, 2, \ldots, p\}$ and edges (s, t) in $E \subseteq V \times V$. To each node $s \in V$, a random variable X_s is associated with values in $\{0, 1\}$. The probability of each configuration x depends on a main effect (external field) and pairwise interactions. It is sometimes referred to as the auto logistic function (Besag 1974), or a pairwise Markov random field to emphasise that the parameter and sufficient statistic space are limited to pairwise interactions (Wainwright and Jordan 2008). Each $x_s \in \{0, 1\}$ has conditional on all remaining variables (nodes) $X_{\setminus s}$ probability of success $\pi_s := \mathbb{P}(X_s = 1 \mid x_{\setminus s})$, where $x_{\setminus s}$ contains all nodes except s. Let $\xi = (m, A)$ contain all parameters, where the $p \times p$ matrix A contains the pairwise interaction strengths and the p vector m is the main effects (external field). The distribution for configuration x of the Ising model is then

$$\mathbb{P}(x) = \frac{1}{Z(\xi)} \exp\left(\sum_{s \in V} m_s x_s + \sum_{(s,t) \in E} A_{st} x_s x_t \right). \tag{1}$$

In general, the normalisation $Z(\xi)$ is intractable, because the sum consists of 2^p possible configurations for $x \in \{0, 1\}^p$; for example, for $p = 30$, we obtain over 1 million configurations to evaluate in the sum in $Z(\xi)$ (see, e.g. Wainwright and Jordan (2008) for lattice (Bethe) approximations).

Alternatively, the conditional distribution does not contain the normalisation constant $Z(\xi)$ and so is more amenable to analysis. The conditional distribution is again an Ising model (Besag 1974; Kolaczyk 2009)

$$\pi_s = \mathbb{P}(X_s = 1 \mid x_{\setminus s}) = \frac{\exp\left(m_s + \sum_{t:(s,t) \in E} A_{st} x_t\right)}{1 + \exp\left(m_s + \sum_{t:(s,t) \in E} A_{st} x_t\right)}. \tag{2}$$

It immediately follows that the log odds (Besag 1974) is

$$\mu_s(x_{\setminus s}) = \log\left(\frac{\pi_s}{1-\pi_s}\right) = m_s + \sum_{t:(s,t)\in E} A_{st}x_t. \tag{3}$$

For each node $s \in V$, we collect the p parameters m_s and $(A_{st}, t \in V\setminus\{s\})$ in the vector θ. Note that the log odds $\theta \mapsto \mu_\theta$ is a linear function, and so if $x = (1, x_{\setminus s})$ then $\mu_\theta = x^\top\theta$. The theory of generalised linear models (GLM) can therefore immediately be applied to yield consistent and efficient estimators of θ when sufficient observations are available, i.e. $p < n$ (Nelder and Wedderburn 1972; Demidenko 2004). To obtain an estimate of θ when $p > n$, we require regularisation or another method (Hastie et al. 2015; Bühlmann et al. 2013).

2.1 Nodewise logistic regression

Meinshausen and Bühlmann (2006) showed that for sparse models the true neighbourhood of a graph can be obtained with high probability by performing a series of conditional regressions with Gaussian random variables. For each node $s \in V$, the set of nodes with nonzero A_{st} are determined, culminating in a neighbourhood for each node. Combining these results leads to the complete graph, even when the number of nodes p is much larger than the number of observations n. This is called neighbourhood selection, or nodewise regression. This idea was extended to Bernoulli (Ising) graphs by Ravikumar et al. (2010), but see also van de Geer (2011, chapters 6 and 13). Nodewise regression allows us to use standard logistic regression to determine the neighbourhood for each node. This framework, of course, comes at a cost, and two strong assumptions are required. We discuss these assumptions in Sect. 3.

To estimate the coefficients, Meinshausen and Bühlmann (2006) used a sequential regression procedure for Gaussian data where each node in turn is treated as the dependent variable and the remaining ones as independent variables. By repeating this analysis for all nodes in V, a total of $p-1$ neighbourhood estimates of nonzero parameters are obtained for all nodes $s \in V$. Since each node is considered twice, the estimates are often combined by either an *and*-rule, where an edge is obtained if $\hat{A}_{st} \neq 0$ and $\hat{A}_{ts} \neq 0$, or an *or*-rule, where either parameter estimate can be nonzero (Meinshausen and Bühlmann 2006).

Ravikumar et al. (2010) translated this procedure to binary variables using pseudo-likelihoods. Recall that $\theta \mapsto \mu_\theta$ is the linear function $\mu_{\theta_s}(x_{\setminus s}) = m_s + \sum_{t\in V\setminus s} A_{st}x_t$ of the conditional Ising model obtained from the log odds (3). The parameters in the p dimensional vector θ are m_s for the intercept (external field) and $(A_{st}, t \in V\setminus\{s\})$, representing the connectivity parameters for node s based on all remaining nodes $V\setminus\{s\}$. Let the $n \times p$ matrix $X_{\setminus s} = (1_n, X_1, \ldots, X_p)$ be the matrix with the vector of 1s in 1_n and the remaining variables without X_s. We write y_i for the observation $x_{i,s}$ of node s and $x_i = (1, x_{i,\setminus s})$ and $\mu_i := \mu_{i,\theta_s}(x_{i,\setminus s})$, basically leaving out the subscript s to index the node, and only use the node index s whenever circumstances demand

it. Let the loss function be the negative log of the conditional probability π in (2), known as a pseudo log-likelihood (Besag 1974)

$$\psi(y_i, \mu_i) := -\log \mathbb{P}(y_i \mid x_i) = -y_i\mu_i + \log\left(1 + \exp(\mu_i)\right). \tag{4}$$

For logistic loss ψ, the theoretical risk is defined as

$$R_\psi(\mu) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\psi(y_i, \mu_i). \tag{5}$$

The value that optimises the risk is $\theta^* = \arg\inf_{\theta\in\mathbb{R}^p} R_\psi(\mu)$; given the choice of logistic loss we can do no better than θ^* in terms of the population. Of course we do not have the theoretical risk and so we use an empirical version

$$R_{n,\psi}(\mu) = \frac{1}{n}\sum_{i=1}^{n} \psi(y_i, \mu_i). \tag{6}$$

Define $\mu^* := \mu_{\theta^*}(x)$, which uses the optimal value θ^* under theoretical risk. For sparse estimation, the ℓ_1 (lasso) minimisation is given by

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^{n} \psi(y_i, \mu_i) + \lambda||\theta||_1 \right\}, \tag{7}$$

where $||\theta||_1 = \sum_{t\in V\setminus\{s\}} |\theta_t|$ is the ℓ_1 norm, λ is a fixed penalty parameter. Since ψ is convex and $||\theta||_1$ is convex, the objective function $R_{n,\psi} + \lambda||\theta||_1$ in (7) is convex, which allows us to apply convex optimisation. We discuss how to obtain the parameters with the coordinate descent algorithm in Sect. 4.

Once the parameters are obtained it turns out that inference on network parameters is in general difficult with ℓ_1 regularisation (Pötscher and Leeb 2009). One solution is to desparsify it by adding a projection of the residuals (van de Geer et al. 2013; Javanmard and Montanari 2014; Zhang and Zhang 2014; Waldorp 2015), which is sometimes referred to as the desparsified lasso. Another type of inference is one where clusters of nodes obtained from the lasso are interpreted instead of individual nodes (Lockhart et al. 2014).

To illustrate the result of an implementation of logistic regression for the Ising model, consider Fig. 1. We generated a random Erdös–Renyi graph (left panel) with $p = 100$ nodes and probability of an edge 0.05, resulting in 258 edges. The *igraph* package in R was used with *erdos.renyi.game* (Csardi and Nepusz 2006). To generate data ($n = 50$ observations of the $p = 100$ nodes) from the Ising model, the package *IsingSampler* was used, and to obtain estimates of the parameters the package *IsingFit* was used (by Epskamp, see van Borkulo et al. 2014) in combination with the *and* rule.

The recall (true positive rate) for this example was 0.69 and the precision (positive predictive value) was 0.42. So we see that about 30% of the true edges are

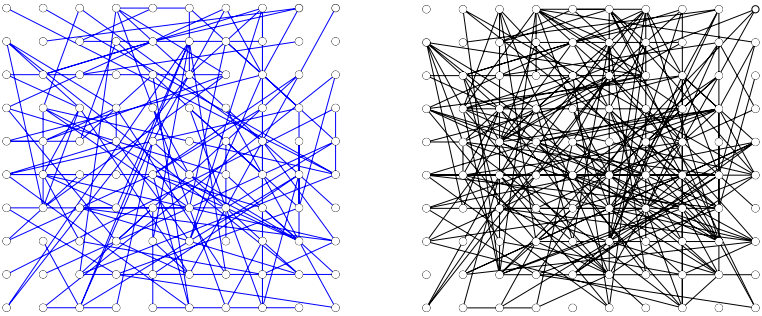

Fig. 1 Ising networks with $p = 100$ nodes. Left panel: true network used to generate data. Right panel: estimated Ising model with nodewise logistic regression from $n = 50$ generated observations

missing and about 60% of the estimated edges is incorrect. This is not surprising given that we have 4950 possible edges to determine and only 50 observations. (More details on the simulation are in Sect. 4.2.)

3 Assumptions for prediction and estimation

To determine the consequences of violating the assumptions of the lasso in logistic regression, we discuss the assumptions for accurate prediction and estimation. Both prediction and estimation require that the solution is sparse; informally, that the number of non-zero edges in the graph is relatively small (see Assumption 1 below). For accurate estimation we also require an assumption on the covariance between the nodes in the graph. Several types of assumptions have been proposed (see van de Geer et al. 2009, for an excellent over view and additional results on obtaining the lasso solution), but here we focus on the restricted eigenvalue assumption because of its direct connection to multicollinearity.

3.1 Sparsity

Central to lasso estimation is the assumption that the underlying problem is low dimensional (Bühlmann and van de Geer 2011; Giraud 2014). This is the assumption of sparsity. This is essential because whenever $p > n$ there is no unique solution to the empirical risk $R_{n,\psi}(\mu)$ defined in (6) (Wainwright 2009). Sparsity can be defined in different ways. The most common is a restriction on the number of nonzero edges, sometimes referred to as coordinate sparsity (Giraud 2014). Let S_0 denote the support containing the indices of the nonzero coefficients, i.e., $S_0 := \{j : \theta_j \neq 0\}$ and its size $s_0 = |S_0|$.

Assumption 1 (Coordinate sparsity) The size s_0 of the set of nonzero coefficients S_0 in θ^* is of order $o(\sqrt{n/\log p})$.

There are other forms of sparsity, such as the fused sparsity, where the support is defined as $\{j : \theta_j - \theta_{j-1} \neq 0\}$. This ensures that there are relatively few jumps in, for instance, a piecewise continuous function (see Giraud 2014, for more details). Another form of sparsity is where the ℓ_1 size of the parameter vector θ is restricted. We use this to show that prediction (classification) in logistic regression is accurate.

Assumption 2 (ℓ_1-sparsity) The ℓ_1 norm of the coefficients θ^* is of order $o(\sqrt{n/\log p})$, i.e. $||\theta^*||_1 = o(\sqrt{n/\log p})$.

In logistic regression, there is a natural classifier that predicts whether y_i is 1 or 0. We simply check whether the probability of a 1 is greater than 1/2, that is, whether $\pi_i > 1/2$. Because $\mu_i > 0$ if and only if $\pi_i > 1/2$, we obtain the natural classifier

$$C(y_i) = \mathbb{1}\{\mu_i > 0\}, \tag{8}$$

where $\mathbb{1}$ is the indicator function. This is called 0–1 loss or sometimes Bayes loss (Hastie et al. 2015). Instead of 0–1 loss we use logistic loss (4) to determine how well we predict individual observations y_i to which class they belong, 0 or 1. Define the prediction loss (sometimes called excess risk) with logistic loss ψ as

$$\mathcal{L}_\psi(\mu) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big(\psi(y_i, \mu_i) - \psi(y_i, \mu_i^*)\big). \tag{9}$$

Note that by definition of θ^* that $\mathcal{L}_\psi(\mu) \geq 0$ for any $\theta \mapsto \mu_\theta$. A similar definition is possible for 0–1 loss using C, which is $\mathcal{L}_C(\mu)$. (Bartlett et al. (2003), Theorem 3.3) show that $\mathcal{L}_\psi \to 0$ implies $\mathcal{L}_C \to 0$ as $n \to \infty$. In other words, using logistic loss finally results in the optimal 0–1 (Bayes) loss, and so nothing is lost in using logistic loss as a proxy for 0–1 loss.

Prediction has been shown to be accurate using Assumption 2. Suppose that the regularisation parameter λ is of order $O(\sqrt{\log p/n})$, then the prediction loss is bounded above (Bühlmann and van de Geer 2011)

$$\mathcal{L}_\psi(\hat{\mu}) + \lambda||\hat{\theta}||_1 \leq 2\lambda||\theta^*||_1. \tag{10}$$

If in addition Assumption 2 holds, where $||\theta^*||_1$ is of order $o(\sqrt{n/\log p})$, then $\mathcal{L}_\psi(\hat{\mu}) = o(1)$. This result is in the Appendix as Lemma 3 and corresponds to that in Ravikumar et al. (2010), but see also (Bühlmann and van de Geer (2011), Sect. 14.8) for stronger results. The requirement that the regularisation parameter is of order $O(\sqrt{\log p/n})$ is obtained because the stochastic part in the prediction loss has to be negligible (see Lemma 2 in the Appendix for details). If we choose λ sufficiently large, we are guaranteed with probability at least $1 - 2\exp(-nt^2)$ for some $t > 0$ that the prediction loss is bounded by the ℓ_1 norm of the parameter of interest θ^* as in (10).

It follows directly from (10) that the lasso estimation error is larger than prediction loss, and so prediction is easier than estimation (see also Hastie et al. 2001). From (10), we get an upper bound on prediction loss

$$\mathcal{L}_{\psi}(\hat{\mu}) \leq 2\lambda\big(||\hat{\theta} - \theta^*||_1\big), \tag{11}$$

where we used the reverse triangle inequality (see Lemma 4 in the Appendix for details). This shows that lasso estimation error is larger than prediction error.

3.2 Restricted eigenvalues

Next to sparsity, the second assumption for the lasso is related to the problem that when $p > n$ the empirical risk $R_{n,\psi}$ is not strongly convex and hence no unique solution is available. It turns out that we need to consider a subset of lasso estimation errors $\delta = \hat{\theta} - \theta^*$ such that strong convexity holds for that subset (Negahban et al. 2012).

Because we have $p > n$ we cannot obtain strong convexity in general, and we need to relax the assumption. This is how we get to the restricted eigenvalue assumption. Let $\nabla_j \psi(y_i, x_i^{\mathsf{T}}\theta)$ be the first derivative with respect θ_j and $\nabla^2_{jj}\psi(y_i, x_i^{\mathsf{T}}\theta)$ the second derivative with respect to θ_j. Then demanding strong convexity means that if we consider the $s_0 \times s_0$ submatrix $\nabla^2_{S_0}\psi_n(\theta)$ then we need that $\nabla^2_{S_0}\psi_n(\theta) \geq \gamma I$, where I is the identity matrix and we used $\psi(\theta)$ instead of $\psi(y, \mu)$ to emphasise dependence on θ (and $\mu = x^{\mathsf{T}}\theta$). This we can never get (see the Appendix for more details on strong convexity). But from (10) it follows that if the directions of the lasso error $\delta = \hat{\theta} - \theta^*$ follow a cone shaped region with $||\delta_{S_0^c}||_1 \leq \alpha||\delta_{S_0}||_1$ (see Theorem 1 in the Appendix), then within these directions strong convexity holds. We refer to this set as $\mathbb{C}_\alpha = \{\delta \in \mathbb{R}^p : ||\delta_{S_0^c}||_1 \leq \alpha||\delta_{S_0}||_1\}$. In the directions where the cone shape holds so that $\delta \in \mathbb{C}_\alpha$, the loss function is strictly larger than 0, except at $\delta = 0$, but is flat and can be 0 if $\delta \notin \mathbb{C}_\alpha$ (see Negahban et al. (2012) or Hastie et al. (2015) for an excellent discussion). This assumption is called the restricted eigenvalue assumption.

The second derivative or Fisher information matrix is

$$\nabla^2\psi(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\pi(\mu_i)\pi(-\mu_i)x_i x_i^{\mathsf{T}}. \tag{12}$$

We assume that this population level matrix is positive definite. Then by strong convexity we have for $\gamma > 0$ that $\nabla^2\psi \geq \gamma I$, and so

$$\mathcal{L}_{\psi}(\hat{\mu}) \geq \frac{1}{2}\delta^{\mathsf{T}}\nabla^2\psi(\hat{\theta})\delta \geq \frac{\gamma}{2}||\delta||_2^2,$$

which allows us to relate the lasso estimation error to prediction loss such that we can conclude consistency because of the bound on prediction error in (10) (see Lemma 3 in the Appendix). The problem is that we work with the empirical $p \times p$ matrix $\nabla^2 \psi_n(\theta)$ which is necessarily singular since $p > n$. The empirical Fisher information is

$$\nabla^2 \psi_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \pi(\mu_i)\pi(-\mu_i) x_i x_i^{\mathsf{T}}, \tag{13}$$

which has zero eigenvalues because it is positive semidefinite whenever $p > n$. Bickel et al. (2009) suggested the restricted eigenvalue assumption that is sufficient to guarantee that $\nabla^2 \psi_n(\theta)$ has positive eigenvalues for lasso errors $\delta \in \mathbb{C}_\alpha$. Here we require in the setting of nodewise logistic regression that for all nodes s simultaneously the lower bound $\gamma_G > 0$ is sufficient and $\alpha = 1$. We emphasise the nodewise estimation of all edges in E using ψ_s and δ_s.

Assumption 3 (Restricted eigenvalue) The population Fisher information matrix $\nabla^2 \psi_s$ of dimensions $p \times p$ is nonsingular and $\max_j \nabla^2_{jj} \psi_s(\theta) < K$, for some $K > 0$ and for all $s \in V$. The empirical matrix $\nabla^2 \psi_{n,s}(\theta)$ satisfies the restricted eigenvalue (RE) assumption if for some $\gamma_G > 0$ it holds that

$$\min_{s \in V} \frac{\delta_s^{\mathsf{T}} \nabla^2 \psi_{n,s}(\theta) \delta_s}{||\delta_s||_2^2} \geq \gamma_G \qquad \text{for all} \quad 0 \neq \delta_s \in \mathbb{C}_1. \tag{14}$$

The restricted eigenvalue assumption has been investigated in the context of Gaussian data (Bickel et al. 2009; Wainwright 2009; Raskutti et al. 2010; Hastie et al. 2015, chapter11), in the setting of the Ising model (Ravikumar et al. 2010, Lemma 3), and in generalised linear models (Van de Geer 2008; Bühlmann and van de Geer 2011, chapter 6). The original restricted eigenvalue assumption as presented in Bickel et al. (2009) is slightly stronger than the compatibility assumption of van de Geer et al. (2009). See van de Geer et al. (2009) for more details on the compatibility and other assumptions to bound estimation error in the lasso. Here we use the RE assumption because of its connection to multicollinearity, discussed in Sect. 4.

Let θ_S be the vector where for each $t \in V$ we have $\theta_t \mathbb{1}\{t \in S\}$. It follows that $\theta = \theta_S + \theta_{S^c}$, where S^c is the complement of S. The RE assumption implies that the $s_0 \times s_0$ submatrix $\nabla^2_{S_0} \psi_n(\theta)$ indexed by S_0 has smallest eigenvalue > 0. This can be seen as follows. RE implies that there is a δ such that $\delta_{S_0} \neq 0$, $\delta_{S_0^c} = 0$, implying $||\delta_{S_0^c}||_1 \leq ||\delta_{S_0}||_1$, and $\delta^{\mathsf{T}}(\nabla^2 \psi_n)\delta > 0$. This implies that for some $\gamma_G > 0$

$$\nabla^2_{S_0} \psi_n(\theta) \geq \gamma_G I$$

and so we have restricted strong convexity for this $\delta \in \mathbb{C}_1$. The two Assumptions 1 on coordinate sparsity and 3 on restricted eigenvalues make it possible to derive the ℓ_1 estimation error bound (see Theorem 1 in the Appendix for details)

$$\max_{s \in V} ||\delta_s||_1 = \max_{s \in V} ||\hat{\theta}_s - \theta_s^*||_1 \leq \frac{16}{\gamma_G} s_0 \lambda. \tag{15}$$

The bound corresponds to the one given in Negahban et al. (2012, Corollary 2, discussed in Sect. 4.4), and the one in Bühlmann and van de Geer (2011, Lemma 6.8). Because we require the smallest γ such that the RE assumption holds, we have that this bound holds simultaneously for all nodes in the Ising graph.

The bounds on prediction and estimation are important to know the circumstances for the statistical guarantees. However, in many practical situations, we cannot be certain of the assumptions of sparsity and restricted eigenvalues. These assumptions cannot be checked. And so it becomes relevant to know what the consequences for prediction and estimation are when the assumptions are not satisfied. This is what we investigate next.

4 Violation of sparsity and restricted eigenvalues

If we violate either the sparsity or restricted eigenvalue assumption, then we would expect that lasso estimation error becomes worse, and indeed this happens. However, this is not so clear for prediction. In fact, it turns out that prediction becomes better for non-sparse models that violate the restricted eigenvalue (RE) assumption. Our main result is that violating the RE or sparsity assumption leads to a decrease in empirical risk, and hence in loss. The RE assumption is violated by an extreme case of multicollinearity, namely where some nodes are copies of other nodes. When such copies are connected we call them connected copies. In connected copies, the coefficients are proportional to the original ones, such that we do not arbitrarily change the data generating process. One way to view connected copies is to find multiplicative constants for the edge weights that lead to a network with perfect correlations between nodes. We therefore compare prediction and estimation of different situations where we violate the RE or sparsity assumption based on different data generating processes. Proposition 1 shows that the number of connected copies co-determines the decrease in empirical risk, and hence, violating the RE assumption leads to a decrease in risk. Next, in Corollary 1, we show that violating the sparsity assumption leads to either a decrease or increase of empirical risk depending on whether the set of coefficients in the different subsets of nodes is positive or negative, respectively. We illustrate the theoretical results with some simulations in Sect. 4.2.

4.1 Connected copies

Suppose that for some nodes $s, t \in V$ we have that the observations are identical, that is, $x_{i,s} = x_{i,t}$ for all i. Then the coefficients obtained with the lasso using the quadratic approximation to the logistic loss in coordinate descent will be identical, i.e. $\hat{\theta}_s = \hat{\theta}_t$ (Hastie et al. 2015, see also the Appendix for a discussion of the coordinate descent algorithm). This can be seen from the following considerations. By (13) we have that element (s, s) of the second derivative matrix is

$$\nabla^2_{ss}\psi_n = \frac{1}{n}\sum_{i=1}^{n} \pi(\mu_i)\pi(-\mu_i)x^2_{i,s},$$

and this is the same for element (t, t) since $x_s = x_t$. Similarly, for the sth element $\nabla_s\psi_n$, we obtain

$$\nabla_s\psi_n = \frac{1}{n}\sum_{i=1}^{n}(-y_i + \pi(\mu_i))x_{i,s},$$

which equals $\nabla_t\psi_n$ because $x_s = x_t$. In the coordinate descent algorithm, the updating scheme using the quadratic approximation (see the Appendix for details) is at time $q + 1$

$$\theta_j^{q+1} = \theta_j^q - \begin{cases} (\nabla^2_{jj}\psi^q)^{-1}\nabla_j\psi^q - \lambda & \text{if} \quad (\nabla^2_{jj}\psi^q)^{-1}\nabla_j\psi^q > \lambda \\ (\nabla^2_{jj}\psi^q)^{-1}\nabla_j\psi^q + \lambda & \text{if} \quad (\nabla^2_{jj}\psi^q)^{-1}\nabla_j\psi^q < -\lambda \\ 0 & \text{if} \quad |(\nabla^2_{jj}\psi^q)^{-1}\nabla_j\psi^q| \leq \lambda, \end{cases} \tag{16}$$

where $(\nabla^2_{jj}\psi^q)^{-1}$ is element (j, j) of the inverse of the second order derivative matrix $\nabla^2\psi^q$ for step q in the coordinate descent algorithm. Then we obtain in the coordinate descent algorithm $(\nabla^2_{ss}\psi^q_n)^{-1}\nabla_s\psi^q_n$ at each step q for both nodes s and t, implying that the coefficients are the same. So for each node in the nodewise regressions, we obtain a Fisher matrix where column s is the same as column t. Now if both s and t are in S_0, then the smallest eigenvalue of $\nabla^2\psi_{n,S_0}$ is 0, and hence, the RE assumption is violated. We will use this idea of identical nodes to explain why prediction loss becomes better when we violate the RE assumption.

We call a node t in the subset $L \subset V$ a connected copy of $s \in K = V\backslash L$ if $(s, t) \in E$ and $x_t = x_s$. This says that two directly connected nodes are identical to each other for all n observations. Note that the coefficient between a connected copy and its original must be positive; if the coefficient was negative, then the connected copy would also have to be the reverse of its original, which cannot be true because the variables are defined to be identical. We know from estimation that if a node is a connected copy then the lasso solution is no longer unique (Hastie et al. 2015). In fact, if t is a connected copy of s, then all solutions with

$\alpha\hat{\theta}_s$ and $(1-\alpha)\hat{\theta}_t$, with $0 \leq \alpha \leq 1$ and $\hat{\theta}_s$, $\hat{\theta}_t$ are estimates of the parameters of nodes s and t, respectively, result in identical empirical risk $R_{n,\psi}$ as when those connected copies have been deleted. Similarly, we will have the same ℓ_1 norm as when the connected copies have been deleted. As a consequence, we cannot distinguish between the situation with or without the connected copy in ℓ_1 optimisation. We denote by L_t the set of all connected copies $s \in L_t$ of $t \in K$, which defines an equivalence relation on L, such that $L_t \cap L_s = \emptyset$ and $\cup L_t = L$. We denote the parameter vector where the connected copies in L have been deleted by $\theta_{\setminus L}$ and correspondingly $\mu_{\setminus L} = x_{\setminus L}^{\mathsf{T}} \theta_{\setminus L}$.

Lemma 1 In the Ising graph $G = (V, E)$ suppose nodes in $L \subset V$ are connected copies of nodes in $K = V \setminus L$. Furthermore, the nodewise lasso solutions $\hat{\theta}$ are obtained with (7) where for each connected copy $t \in L_t$ of node $s \in K$, with $\alpha_t \hat{\theta}_t$, we have that $\sum_{t \in L_t} \alpha_t = 1$. Then the empirical risk $R_{n,\psi}(\hat{\mu})$ and ℓ_1 norm of $\hat{\theta}$ are the same as when the connected copies in L are deleted, i.e. $R_{n,\psi}(\hat{\mu}) = R_{n,\psi}(\hat{\mu}_{\setminus L})$ and $||\hat{\theta}||_1 = ||\hat{\theta}_{\setminus L}||_1$.

So we have that the non-uniqueness of the lasso in case of a connected copy, results in the exact same value for the empirical risk whether we delete it or take any one of the weighted versions such that the coefficients sum to 1. Note that we do not change the underlying process in any arbitrary way; the nodes are connected and the coefficients remain proportional to the original ones. We immediately obtain that the size $|L|$ of the set of connected copies co-determines the prediction loss. We obtain this result because the coefficients of the connected copies with respect to their originals are positive.

Proposition 1 For the Ising graph, let L_1 and L_2 be subsets of connected copies of nodes in $V \setminus L_1 \cup L_2$ such that $L_1 \subset L_2$ and hence $|L_1| < |L_2|$. Then we have for the prediction loss that the sum of coefficients in $L_1^c \cap L_2$ is > 0, and the risk $R_{n,\psi}(\hat{\mu}_{\setminus L_1}) \geq R_{n,\psi}(\hat{\mu}_{\setminus L_2})$.

This follows from Lemma 1 directly, since there we saw that the prediction loss including connected copies is equal to the prediction error when those connected copies are deleted. This idea explains why the empirical risk decreases as a function of an increasing number of connected copies.

The same idea can be used to determine why prediction becomes better for non-sparse sets. Proposition 1 can be altered such that a similar result holds for sparsity, where we do not need the connected copies. The only requirement is that we know what the sum of the coefficients is that are in the larger set of connected nodes, because the nodes need not be connected in this case. Let the S_a be a set of nodes with a possibly non-sparse set of nonzero edges in the sense that $|S_a| > O(\sqrt{n/\log p})$. Suppose that $S_0 \subset S_a$ so that $|S_0| < |S_a|$.

Corollary 1 In the Ising graph $G = (V, E)$ suppose that we have a particular, not necessarily sparse, node set with nonzero edges in S_a, and define the subset $S_0 \subset S_a$. Then we obtain for the empirical risk $R_{n,\psi}$ that

(1) if the sum of coefficients in $S_0^c \cap S_a$ is > 0, then $R_{n,\psi}(\hat{\mu}_{\setminus S_0}) \geq R_{n,\psi}(\hat{\mu}_{\setminus S_a})$;
(2) if the sum of coefficients in $S_0^c \cap S_a$ is < 0, then $R_{n,\psi}(\hat{\mu}_{\setminus S_0}) \leq R_{n,\psi}(\hat{\mu}_{\setminus S_a})$.

We see that by eliminating the requirement of connectedness, we find that prediction loss decreases given that the coefficients in the remaining set of non-zero coefficients are positive.

We focus here on prediction loss because by (11) we have that the ℓ_1 estimation error is larger than prediction loss (given that the penalty parameter λ is of the right order), and hence if we find that prediction loss becomes higher, it follows that ℓ_1 estimation error becomes larger.

The above presented ideas of violating the sparsity assumption or restricted eigenvalue assumption are confirmed by some numerical illustrations.

4.2 Numerical illustration

To show the effects of non-sparse underlying representations and violation of the restricted eigenvalue assumption (multicollinearity), we performed some simulation studies. Here 0–1 data were generated by a Metropolis–Hastings algorithm, implemented in the R package *IsingSampler* (van Borkulo et al. 2014), according to a random graph (Erdös–Renyi) with $p = 100$ nodes and $n = 50$ observations. All edge coefficients were positive, so that we expect the prediction error to improve with increasing collinearity. Sparsity of the graph was varied by the probability of an edge from $p_e = 0.025$, which complies with the sparsity assumption, to the probability of an edge of $p_e = 0.2$, which does not comply with the sparsity assumption. For interpretation we defined sparsity in these simulations as $1 - p_e$, so that high sparsity means few non-zero edges. Multicollinearity was induced by equating two columns of the data X if there was an edge in the edge set of the true graph for a percentage α, ranging from 0 to 0.6. This ensured that the smallest αs_0 eigenvalues of the submatrix $\nabla^2 \psi_{n,S_0}$ are 0, thereby violating the RE assumption.

The parameters for the nodes m and for interactions in A were estimated by node-wise logistic regressions, implemented in *IsingFit* (by Epskamp, see van Borkulo et al. 2014). Here the extended Bayesian information criterion (EBIC) is used to determine the optimal λ for each logistic regression separately (Foygel and Drton 2013). This procedure was run 100 times and the averages across these runs (and nodes) are presented. We evaluated estimation accuracy by recall ($|\hat{S} \cap S_0|/|S_0|$) and precision ($|\hat{S} \cap S_0|/|\hat{S}|$). We also used a scaled ℓ_1 norm for the estimation error $||\delta||_1/u$, where $\delta = \hat{\theta} - \theta^*$ and u is the maximal value obtained. Prediction was evaluated by logistic loss ψ and Bayes loss C. We determine loss for data z_i independent from data y_i, upon which the estimate $\hat{\theta}$ is based (predictive risk).

Figure 2b shows that recovery of parameters is accurate when sparsity is high (few non-zero edges), but recovery becomes poor when sparsity does not hold; from sparsity 0.95 and lower. This is seen in all three measures: recall, precision and the scaled ℓ_1 norm. In contrast, the 0–1 loss from (8) and the logistic loss in (4) actually become better (the loss decreases) when the data generating process is no longer sparse, as can be seen in Fig. 2a. This corresponds to Corollary 1, which shows

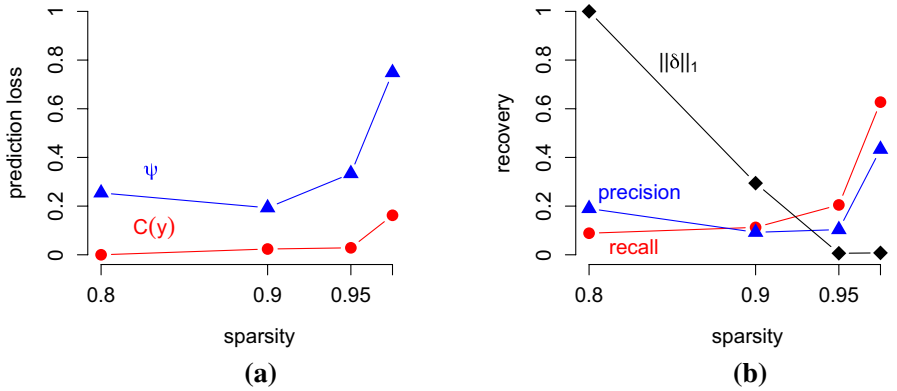

Fig. 2 Performance measures of constructing networks with the lasso as a function of sparsity, where sparsity is defined as $1 - p_e$, the reverse of the edge probability. In **a**, Bayes loss (redcircle) and logistic loss (bluetriangle), and in **b**, recovery in terms of recall (redcircle) and precision (bluetriangle) and the scaled ℓ_1 norm of the error (blackdiamond)

that sparsity is not necessary for accurate prediction. We do require that the penalty parameter λ is of the appropriate order (i.e. $\lambda = O(\sqrt{\log p/n})$); here λ was selected by the EBIC (Foygel and Drton 2013) which ensured such a penalty. The EBIC has an additional hyperparameter γ to control the impact of the size of the search domain; we set γ to 0.25 in line with the reasonable performance obtained in Foygel and Drton (2013). Prediction loss is high at high sparsity because in the simulation there are only about 2–3 edges, which means that prediction of other nodes is extremely difficult.

In Fig. 3, the results can be seen when multicollinearity is varied. As expected, Fig. 3b shows that increasing multicollinearity reduced recovery; both recall and

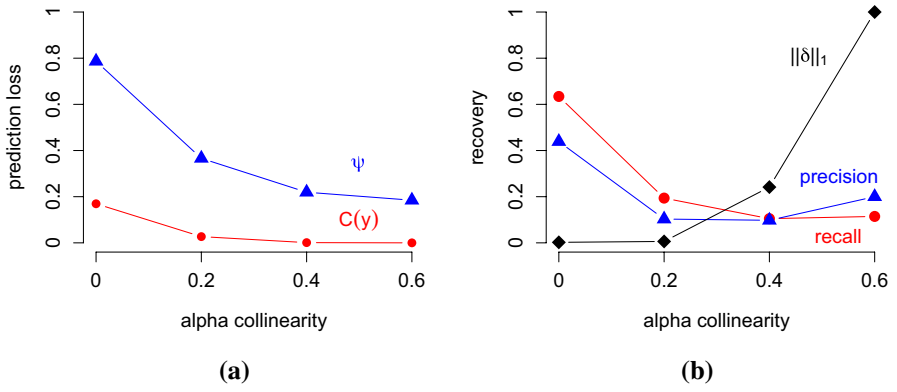

Fig. 3 Performance measures of constructing networks with the lasso as a function of collinearity (α); collinearity is defined as the probability of identical observations for two nodes whenever these nodes are connected. In **a**, Bayes loss (redcircle) and logistic loss (bluetriangle), and in **b**, recovery in terms of recall (redcircle) and precision (bluetriangle) and the scaled ℓ_1 norm of the error (blackdiamond)

precision decreased to around 10%. Prediction loss, on the other hand, becomes smaller as shown in Fig. 3a, indicating better prediction for multicollinear data. This is in line with Proposition 1. We can also think of it in the following way. With increasing multicollinearity α, more equal columns in X are present for connected nodes. This leads to more similar behaviour of connected nodes in the Ising network and hence to better prediction.

These results demonstrate that when either the sparsity assumption or multicollinearity (RE) assumption is violated, the prediction loss decreases, making prediction better. But also that estimation error increases. Hence, the estimated network that predicts well, will not be similar to the true underlying network. On the other hand, if the assumptions of sparsity and RE hold, then many of the edges in the Ising model are estimated correctly but because of the high-dimensional setting many true edges are also missed. And since in sparse settings fewer edges are present that determine the prediction, prediction is poorer.

5 Discussion

Logistic regression is an appropriate tool for prediction and estimation of parameters of the Ising model. Statistical guarantees have been given for prediction and estimation of the parameters of the Ising model using a sequence of logistic regressions whenever at least the assumptions of sparsity and restricted eigenvalues hold. Here we focused on violations of these assumptions and showed why prediction becomes better whenever sparsity or restricted eigenvalues do not hold. Intuitively, for predicting the underlying structure of the graph is irrelevant and when nodes behave similarly, prediction becomes easier. To confirm these intuitions we showed, using connected copies, that prediction loss can decrease as a function of multicollinearity and sparsity. When multicollinearity increases or sparsity decreases, then prediction loss decreases. By consequence of the fact that prediction loss can be considered as a lower bound for estimation error (albeit not a tight bound), estimation error is seen to become worse (increase) as multicollinearity increases or sparsity decreases. Our simulations support these findings and additionally show that recovery in terms of precision and recall becomes worse when violating the assumption of sparsity and multicollinearity.

The concept of connected copies used here is of course an idealisation of reality. Connected copies can be seen as a way to compare prediction and estimation for different structures (topologies) of graphs, where a connected copy is an extreme case in which the correlation between two variables is 1. We required this idealisation to obtain the analytical results. In practice, we will not encounter $x_s = x_t$ but $x_s \approx x_t$. This case is much more difficult to treat analytically. In the case where $x_s \approx x_t$ then the parameter estimates will not be equal and the result would depend on the exact differences in estimates. But if we suppose that the sign of all the coefficients is positive, say, then we would expect similar behaviour of the empirical risk based on the results of Proposition 1 and Corollary 1.

We showed here the consequences of violating the restricted eigenvalue and sparsity assumptions in the Ising model using logistic regression. The next step

is obviously to generalise these results to exponential family distributions. This will require additional restrictions such as the margin condition. The margin condition bridges the gap between estimation error and prediction loss. Because for logistic regression we have the linear functional $\mu = \theta^{\mathsf{T}} x$, we obtain a quadratic margin. For logistic regression, the margin condition then implies that $||\hat{\mu} - \mu^*||_2^2 \geq \gamma ||\delta||_2^2$, where $\delta = \hat{\theta} - \theta^*$ and using strong convexity on $\frac{1}{n}\sum_{i=1}^n x_i x_i^{\mathsf{T}}$. But the margin condition does not hold in general and so requires additional assumptions (see Bühlmann and van de Geer 2011) to apply the current analysis of the consequences of violating RE and sparsity on estimation and prediction.

Compliance with ethical standards

Conflict of interest statement On behalf of all authors, I declare that none of the authors has a conflict of interest.

Appendix

Strong convexity The function ψ is strongly convex if for some $\gamma > 0$ and all $\theta \in \mathbb{R}^p$ it holds that

$$\psi(y, \mu) - \psi(y, \mu^*) \geq \nabla\psi(y, \mu^*)^{\mathsf{T}}(\theta - \theta^*) + \frac{\gamma}{2}||\theta - \theta^*||_2^2 \tag{17}$$

with derivative $\nabla\psi = \partial\psi/\partial\theta$ (Boyd and Vandenberghe 2004a), which for the logistic function is $(-y + \pi(\mu))x$. Because logistic loss is also L-Lipschitz continuous, we have that each element of the derivative $\nabla\psi$ is bounded above by L, and so

$$\psi(y, \mu) - \psi(y, \mu^*) \geq \frac{\gamma}{2}||\theta - \theta^*||_2^2. \tag{18}$$

This is equivalent to requiring that the second derivative $\nabla^2\psi$ has smallest eigenvalue γ, because if we assume $\nabla^2\psi \geq \gamma I$, where I is the identity matrix, then we have

$$\psi(y, \mu) - \psi(y, \mu^*) \geq \frac{\gamma}{2}(\theta - \theta^*)^{\mathsf{T}}\nabla^2\psi(\theta - \theta^*) \geq \frac{\gamma}{2}||\theta - \theta^*||_2^2. \tag{19}$$

Lemma 2 Let $\mathbb{D}_{n,s}(\theta_s)$ be as defined in (23) for each node $s \in V$, and choose $\lambda_0 = c\sqrt{t^2 + \log p/n}$ such that for some $c > 0$, $\mathbb{D}_{i,s}(\theta_s) \leq c$ for all i and $s \in V$. Then for all $t > 0$ we have that

$$\mathbb{P}\left(\max_{s \in V} \left|\mathbb{D}_{n,s}(\theta_s)\right| \leq \lambda_0\right) \geq 1 - 2\exp\left(-nt^2\right).$$

Proof (Lemma 2) To obtain a bound on the prediction loss, we need to bound the stochastic part in the empirical risk $R_{n,\psi}(\mu)$. We have by definition of $\hat{\theta}$ that

$$\frac{1}{n}\sum_{i=1}^{n}\psi(y_i,\hat{\mu}_i)+\lambda||\hat{\theta}||_1 \leq \frac{1}{n}\sum_{i=1}^{n}\psi(y_i,\mu_i^*)+\lambda||\theta^*||_1. \tag{20}$$

Let

$$\mathbb{G}_n(\theta)=\frac{1}{n}\sum_{i=1}^{n}\left(\psi(y_i,\mu_i)-\mathbb{E}\psi(y_i,\mu_i)\right) \tag{21}$$

be the empirical process of the logistic loss indexed by $\theta=(m_s,(A_{st},t\in V\backslash\{s\}))$ through $\mu_\theta(x)$. We can rewrite the left hand side of the above Eq. (20), by subtracting and adding the theoretical risk, as

$$\mathbb{G}_n(\hat{\theta})+\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\psi(y_i,\hat{\mu}_i)+\lambda||\hat{\theta}||_1$$

and similarly for the right hand side of (20) with θ^*, where we obtain $\mathbb{G}_n(\theta^*)+\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\psi(y_i,\mu_i^*)+\lambda||\theta^*||_1$. Then plugging these in (20), we obtain

$$\mathcal{L}_\psi(\hat{\mu})+\lambda||\hat{\theta}||_1 \leq -\left(\mathbb{G}_n(\hat{\theta})-\mathbb{G}_n(\theta^*)\right)+\lambda||\theta^*||_1. \tag{22}$$

Define the increments of the empirical process by

$$\mathbb{D}_n(\hat{\theta})=\mathbb{G}_n(\hat{\theta})-\mathbb{G}_n(\theta^*). \tag{23}$$

If we can bound $|\mathbb{D}_n(\theta)|$ such that the influence of the increment is negligible, then we see from (22) that we can bound prediction loss by the ℓ_1 norm of θ^*.

We define $\mathbb{D}_{i,s}(\theta)=(\psi_{i,s}-\psi_{i,s}^*)-(\mathbb{E}\psi_{i,s}-\mathbb{E}\psi_{i,s}^*)$, where $\psi_{i,s}$ is the logistic function for observation i and node $s\in V$. Note that $\frac{1}{n}\sum_{i=1}^{n}\mathbb{D}_{i,s}(\theta)=\mathbb{D}_{n,s}(\theta)$ as defined in (23), and $\mathbb{E}(\mathbb{D}_{i,s}(\theta))=0$ for all $i=1,\ldots,n$. By assumption there is $c>0$ such that $|\mathbb{D}_{i,s}(\theta)|\leq c$ for all i and s. For some $\lambda_0>0$ by Hoeffding's lemma (e.g. Bousquet et al. 2004; Venkatesh 2013), we have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{D}_{i,s}(\theta)\right|>\lambda_0\right) \leq 2\exp\left(-\frac{2n\lambda_0^2}{c^2}\right).$$

And with $\lambda_0 = c\sqrt{t^2 + \log p/n}$ for some $t > 0$

$$\mathbb{P}\left(\max_{s \in V} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{D}_{i,s}(\theta) \right| > \lambda_0 \right) \leq 2p \exp\left(-\frac{2nt^2}{2} - \frac{2n \log p}{2n} \right),$$

where in the last inequality, we have used the union bound to account for all nodes $p = |V|$. □

Lemma 3 Let $\theta \mapsto \mu_\theta(x)$ be the linear function for the Ising model (3) and $\hat{\theta}$ is the lasso estimate (7) obtained with $\lambda \geq 2\lambda_0$. Let θ^* be the optimum of the theoretical risk $\theta^* = \arg\inf_\theta R_\psi(\mu_\theta)$. Then for all nodes in V such that $|\mathbb{D}_n(\theta)| \leq \lambda_0$, we have with probability $1 - 2\exp(-nt^2/2)$

$$\mathcal{L}_\psi(\hat{\mu}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(\psi(y_i, \hat{\mu}_i) - \psi(y_i, \mu_i^*)) \leq 2\lambda ||\theta^*||_1. \tag{24}$$

If λ is of order $O(\sqrt{\log p/n})$ and if $||\theta^*||_1 = o(\sqrt{n/\log p})$, then $\mathcal{L}_\psi(\hat{\mu}) = o(1)$.

Proof (Lemma 3) By assumption $|\mathbb{D}_n(\hat{\theta})| = |\mathbb{G}_n(\hat{\theta}) - \mathbb{G}_n(\theta^*)| \leq \lambda_0$ with high probability (Lemma 2) for all θ. Then if we choose $\lambda \geq 2\lambda_0$, we have from (22) with probability $1 - 2\exp(-nt^2/2)$ that

$$\mathcal{L}_\psi(\hat{\mu}) + \lambda ||\hat{\theta}||_1 \leq 2\lambda ||\theta^*||_1$$

from which the result follows. If λ is of order $O(\sqrt{\log p/n})$ and if $||\theta^*||_1 = o(\sqrt{n/\log p})$, then $\mathcal{L}_\psi(\hat{\mu}) = o(1)$. □

Lemma 4 Choose λ_0 as in Lemma 2 and assume the $\mathbb{D}_{i,s}(\theta_s)$ are uniformly bounded by c as in Lemma 2. Then the prediction loss is bounded by the estimation error as

$$\mathcal{L}_\psi(\hat{\mu}) \leq 2\lambda(||\hat{\theta} - \theta^*||_1).$$

Proof By (22) we have with high probability that $\mathcal{L}_\psi(\hat{\mu}) + \lambda ||\hat{\theta}||_1 \leq \lambda ||\theta^*||_1$. This implies that $\mathcal{L}_\psi(\hat{\mu}) \leq \lambda(||\theta^*||_1 - ||\hat{\theta}||_1)$. Also from (22) we obtain that $||\theta^*||_1 - ||\hat{\theta}||_1 \geq 0$; and by the reverse triangle inequality we find that $||\theta^*||_1 - ||\hat{\theta}||_1 \leq ||\theta^* - \hat{\theta}||_1 = ||\hat{\theta} - \theta^*||_1$. This completes the proof. □

Theorem 1 For each node s in the Ising graph $G(V, E)$, let $\psi(y, \mu_\theta)$ be the logistic loss in (4) with positive definite second derivative matrix $\nabla^2\psi(\theta)$, and let $\hat{\theta}$ be the lasso estimate (7) obtained with $\lambda \geq 2\lambda_0 = O(\sqrt{\log p/n})$. Suppose that the sparsity Assumption 1 holds with $s_0 = o(\sqrt{n/\log p})$ for all nodes and that the RE Assumption 3 holds with $\gamma_G > 0$. Then for $\delta_s \in \mathbb{C}_1$

$$\max_{s\in V} ||\delta_s||_1 = \max_{s\in V} ||\hat{\theta}_s - \theta^*_s||_1 \leq \frac{16}{\gamma_G} s_0 \lambda$$

and $\hat{\theta}$ is consistent for θ^*.

Proof (Theorem 1) Using the ℓ_1 norm, we obtain for each node $s \in V$ the lasso error $\delta = \hat{\theta} - \theta^*$

$$||\delta||_1 = ||\delta_{S_0}||_1 + ||\delta_{S_0^c}||_1.$$

Choosing $\lambda > 2\lambda_0$ as in Lemma 2, we have from (22) that with probability at least $1 - 2\exp(-nt^2)$

$$\mathcal{L}_\psi(\hat{\mu}) + \lambda||\hat{\theta}_{S_0}||_1 + \lambda||\hat{\theta}_{S_0^c}||_1 \leq \lambda||\theta^*_{S_0}||_1 + \lambda||\theta^*_{S_0^c}||_1.$$

Note that $||\theta^*_{S_0}||_1 - ||\hat{\theta}_{S_0}||_1 \leq ||\hat{\theta}_{S_0} - \theta^*_{S_0}||_1$ and $\theta^*_{S_0^c} = 0$. By rearranging the above equation, for $\lambda > 2\sqrt{\log p/n}$, we have by Lemma 3 that

$$||\delta_{S_0^c}||_1 \leq ||\delta_{S_0}||_1 \tag{25}$$

with high probability. We therefore have that $\delta \in \mathbb{C}_1$. We can then bound ℓ_1 estimation error and obtain

$$||\delta||_1 \leq 2||\delta_{S_0}||_1 \leq 2\sqrt{s_0}||\delta_{S_0}||_2, \tag{26}$$

where we used the inequality $||\nu||_1 \leq \sqrt{k}||\nu||_2$ for $\nu \in \mathbb{R}^k$. We can connect the above to prediction loss by considering strong convexity for the restricted setting for $\delta \in \mathbb{C}_1$. We have by the RE assumption that for $\gamma_G > 0$ for all nodes in the graph G that

$$\frac{1}{n}\sum_{i=1}^{n} \psi(y_i, \hat{\mu}_{i,S_0}) - \psi(y_i, \mu^*_{i,S_0}) \geq \frac{\gamma_G}{2}||\delta_{S_0}||_2^2,$$

where $\mu_{i,S_0} = x_{i,S_0}^{\mathsf{T}} \theta_{S_0}$. Using the empirical process $\mathbb{G}_n(\theta)$ defined in (21) and the increments $\mathbb{D}_n(\theta)$ defined in (23), we obtain

$$\frac{1}{n}\sum_{i=1}^{n} \psi(y_i, \hat{\mu}_{i,S_0}) - \psi(y_i, \mu^*_{i,S_0}) = \mathbb{D}_n(\hat{\theta}_{S_0}) + \mathcal{L}_\psi(\hat{\mu}_{S_0}).$$

If $\lambda > 2\lambda_0$, then by Lemma 2 the increments $\mathbb{D}_n(\theta) < c$ with probability $1 - 2\exp(-nt^2)$, and so

$$2\lambda||\delta||_1 \geq \mathcal{L}_\psi(\hat{\mu}) \geq \mathcal{L}_\psi(\hat{\mu}_{S_0}) \geq \frac{\gamma_G}{2}||\delta_{S_0}||_2^2 \geq \frac{\gamma_G}{2s_0}||\delta_{S_0}||_1^2,$$

where we used the fact that for each subset S of nodes the prediction loss $\mathcal{L}_\psi(\mu_S) \geq 0$. Rearranging and using that $\delta \in \mathbb{C}_1$, gives

$$\frac{16}{\gamma_G}\lambda s_0||\delta||_1 \geq 4||\delta_{S_0}||_1^2 \geq ||\delta||_1^2$$

as claimed. □

Coordinate descent An optimisation algorithm for the convex lasso problem (7) can be characterised by the Karush–Kuhn–Tucker (KKT) conditions (see e.g. Boyd and Vandenberghe 2004b) . The function $\theta_j \mapsto |\theta_j|$ is not differentiable at 0, and so we require a subdifferential $\partial|\theta_j|$ for each j. The KKT condition is the subgradient

$$\frac{1}{n}\sum_{i=1}^{n}(-y_i + \pi(\hat{\mu}))x_i = \lambda\partial||\hat{\theta}||_1, \tag{27}$$

where $\hat{\mu} = \mu_{\hat{\theta}}(x_i)$, and the subdifferential vector $\partial||\hat{\theta}||_1$ is

$$\partial|\hat{\theta}_j| \in \begin{cases} \{-1\} & \text{if} \quad \hat{\theta}_j < 0, \\ \{1\} & \text{if} \quad \hat{\theta}_j > 0, \\ [-1, 1] & \text{if} \quad \hat{\theta}_j = 0. \end{cases}$$

The left hand side of the KKT condition (27) is the first derivative of the empirical risk function $R_{n,\psi}$. The KKT condition with the subdifferential makes clear that a solution $\hat{\theta}_j$ for all j will be shrunken towards 0 by the penally λ. If $|\frac{1}{n}\sum_{i=1}^{n}(-y_i + \pi(\hat{\mu}))x_{i,j}| \leq \lambda$ then it will be set to 0, otherwise it will be shrunken towards 0 by λ.

A solution $\hat{\theta}$ that satisfies the KKT condition can be obtained by, for instance, a subgradient or coordinate descent algorithm (Hastie et al. 2015). The advantage of a convex program is that any local optimum is in fact a global optimum (see e.g. Bertsimas and Tsitsiklis 1997). In a coordinate descent algorithm, each θ_j is obtained by minimisation using the KKT condition and a quadratic approximation to ψ and updated in turn for all p parameters. The parameters θ_j can be updated in turn because the empirical risk $R_{n,\psi}$ is twice differentiable and convex and the ℓ_1 penalty is a sum of convex functions and hence convex (see, e.g. Hastie et al. 2015, chapter 5). For logistic regression

optimisation remains slow since logistic loss needs to be optimised iteratively. Optimisation can be speeded up using a quadratic approximation to logistic loss to obtain for each coordinate of θ separately an update. Let θ^t be the update obtained at step t and let $\delta^t = \theta - \theta^t$. Then we obtain for the quadratic approximation the KKT condition

$$\frac{1}{n}\sum_{i=1}^{n} \nabla_j \psi\left(y_i, x_i^{\mathsf{T}}\theta^t\right) + \nabla_{jj}^2 \psi\left(y_i, x_i^{\mathsf{T}}\theta^t\right)\delta_j^t = \lambda \partial |\delta_j^t|, \tag{28}$$

which leads to an estimate for each coordinate j in $1, \ldots, p$. We define

$$(\nabla_{jj}^2 \psi^t)^{-1} \nabla_j \psi^t = \frac{1}{n}\sum_{i=1}^{n} \left(\nabla_{jj}^2 \psi(y_i, x_i^{\mathsf{T}}\theta^t)\right)^{-1} \nabla_j \psi\left(y_i, x_i^{\mathsf{T}}\theta^t\right). \tag{29}$$

At time $t+1$ this gives the update

$$\theta_j^{t+1} = \theta_j^t - \begin{cases} (\nabla_{jj}^2 \psi^t)^{-1} \nabla_j \psi^t - \lambda & \text{if} \quad (\nabla_{jj}^2 \psi^t)^{-1} \nabla_j \psi^t > \lambda \\ (\nabla_{jj}^2 \psi^t)^{-1} \nabla_j \psi^t + \lambda & \text{if} \quad (\nabla_{jj}^2 \psi^t)^{-1} \nabla_j \psi^t < -\lambda \\ 0 & \text{if} \quad |(\nabla_{jj}^2 \psi^t)^{-1} \nabla_j \psi^t| \leq \lambda. \end{cases} \tag{30}$$

This last equation clearly shows how the threshold of the lasso works: If the update is within λ of 0, then it is put to 0 exactly, otherwise it is shrunk to 0 by λ. Nodewise optimisation for the Ising graph is implemented in the R package *IsingFit* (van Borkulo et al. 2014) using *glmnet* (Friedman et al. 2010) which employs a coordinate descent algorithm.

Proof (Lemma 1) Suppose $t \in L_t = \{t\}$ is a connected copy of $s \in K$. Recall that logistic loss $\psi(y_i, \mu_i) = -y_i\mu_i + \log(1 + \exp(\mu_i))$ depends on the parameter θ only in $\mu_i = x_i^{\mathsf{T}}\theta$. Because $x_s = x_t$ we have the estimate $\hat{\theta}_s = \hat{\theta}_t$. Then a solution $\hat{\theta}$ obtained with the lasso in (7) with $\alpha\hat{\theta}_s$ and $(1-\alpha)\hat{\theta}_s$ yields for all i

$$\mu_i = \sum_{j \in V \setminus \{s,t\}} x_{i,j}\hat{\theta}_j + x_{i,s}\alpha\hat{\theta}_s + x_{i,t}(1-\alpha)\hat{\theta}_t = \sum_{j \in V \setminus \{t\}} x_{i,j}\hat{\theta}_j,$$

which equals the version where node t is deleted from the data. Hence, for any $0 \leq \alpha \leq 1$, the parameter in logistic loss is the same for all i for such connected copies. It follows that the empirical risk $R_{n,\psi}$ is the same for all such copies. Similarly, the ℓ_1 norm gives

$$||\hat{\theta}||_1 = \sum_{j \in V \setminus \{s,t\}} |\hat{\theta}_j| + \alpha|\hat{\theta}_s| + (1-\alpha)|\hat{\theta}_t| = \sum_{j \in V \setminus \{t\}} |\hat{\theta}_j|$$

as claimed. If there are multiple connected copies in L_t, then choose $0 \leq \alpha_j \leq 1$ such that $\sum_{j=1}^{|L_t|} \alpha_j = 1$. Then μ_i is again a sum over $V \backslash L_t$ for any t, implying the same value for ψ_i and hence for $R_{n,\psi}$ as when the nodes in L_t were deleted. The same holds for the ℓ_1 norm. □

Proof (Proposition 1) By Lemma 1 we have that

$$\mu_{i,\backslash L_2} = \mu_{i,\backslash L_1} + \sum_{j \in L_1^c \cap L_2} x_{i,j}\theta_j$$

since $L_1 \subset L_2$ by assumption. It is easily seen that

$$\psi(y_i, \mu_{i,\backslash L_1}) - \psi(y_i, \mu_{i,\backslash L_2}) = y_i(\mu_{i,\backslash L_2} - \mu_{i,\backslash L_1}) + \log\left(\frac{1+\exp(\mu_{i,\backslash L_1})}{1+\exp(\mu_{i,\backslash L_2})}\right).$$

Recall that $\log(1+\exp(a)) \geq a$. Because y_i and $x_{i,j}$ are either 0 or 1, we obtain by the assumption of connected copies that $\sum_{j \in L_1^c \cap L_2} x_{i,j}\theta_j > 0$, and so $\psi(y_i, \mu_{i,\backslash L_1}) - \psi(y_i, \mu_{i,\backslash L_2}) \geq 0$. □

Proof (Corollary 1) From Proposition 1, we have the first half that shows that if $\sum_{j \in S_0^c \cap S_a} x_{i,j}\theta_j > 0$, then $\psi(y_i, \mu_{i,\backslash S_0}) - \psi(y_i, \mu_{i,\backslash S_a}) \geq 0$. The other way around is similar: if $\sum_{j \in S_0^c \cap S_a} x_{i,j}\theta_j < 0$, then $\psi(y_i, \mu_{i,\backslash S_0}) - \psi(y_i, \mu_{i,\backslash S_a}) \leq 0$, which completes the proof. □

References

Bartlett PL, Jordan MI, McAuliffe JD (2003) Large margin classifiers: convex loss, low noise, and convergence rates. In: NIPS

Baxter RJ (2007) Exactly solved models in statistical mechanics. Courier corporation

Bertsimas D, Tsitsiklis J (1997) Introduction to linear optimization. Athena Scientific and Dynamic Ideas, Belmont

Besag J (1974) Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc Ser B (Methodol) 36(2):192–236

Bickel PJ, Ritov Y, Tsybakov AB (2009) Simultaneous analysis of lasso and dantzig selector. Ann Stat 37:1705–1732

Borsboom D, Cramer AOJ, Schmittmann VD, Epskamp S, Waldorp LJ (2011) The small world of psychopathology. PLoS One 6(11):e27407

Bousquet O, Boucheron S, Lugosi G (2004) Introduction to statistical learning theory. Advanced lectures on machine learning. Springer, Berlin, pp 169–207

Boyd S, Vandenberghe L (2004a) Convex optimization. Cambridge University Press, Cambridge

Boyd S, Vandenberghe L (2004b) Convex optimization. Cambridge University Press, Cambridge

Brown L (1986) Fundamentals of statistical exponential families. Inst of Math Stat

Bühlmann P, van de Geer S (2011) Statistics for high-dimensional data: methods. Springer, Theory and Applications, Berlin

Bühlmann P et al (2013) Statistical significance in high-dimensional linear models. Bernoulli 19(4):1212–1242
Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing gwas results: a review of statistical methods and recommendations for their application. Am J Human Genet 86(1):6–22
Cipra B (1987) An introduction to the ising model. Am Math Mon 94(10):937–959
Cressie N (1993) Statistics for spatial data. Wiley, Hoboken
Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Systems 1695(5):1–9 http://igraph.org
Demidenko E (2004) Mixed models: Theory and applications. Wiley, Hoboken
Foygel R, Drton M (2013) Bayesian model choice and information criteria in sparse generalized linear models. University of Chicago, Tech. rep., Chicago
Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22
Giraud C (2014) Introduction to high-dimensional statistics, vol 138. CRC Press, Boca Raton
Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer-Verlag, New York
Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity: the lasso and generalizations. CRC Press, Boca Raton
Javanmard A, Montanari A (2014) Confidence intervals and hypothesis testing for high-dimensional regression. Tech. rep., arXiv:1306.317
Johansen-Berg H, Behrens TEJ, Robson MD, Drobnjak I, Rushworth MFS, Brady JM, Smith SM, Higham DJ, Matthews PM (2004) Changes in connectivity profiles define functionally distinct regions in human medial frontal cortex. Proc Nat Acad Sci Am 101(36):13335–13340
Kindermann R, Snell JL et al (1980) Markov random fields and their applications, vol 1. American Mathematical Society Providence, Providence
Kolaczyk ED (2009) Statistical analysis of network data: methods and models. Springer, New York
Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R et al (2014) A significance test for the lasso. Ann Stat 42(2):413–468
Loh P-L, Wainwright M (2012) High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. Ann Stat 40(3):1637–1664
Marsman M, Waldorp L, Maris G (2017) A note on large-scale logistic prediction: Using an approximate graphical model to deal with collinearity and missing data. Behaviormetrika 44(2):513–534
Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the lasso. Ann Stat 34(3):1436–1462
Negahban SN, Ravikumar P, Wainwright MJ, Yu B (2012) A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Stat Sci 27(4):538–557
Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc Ser A (Gen) 135(3):370–384
Pötscher BM, Leeb H (2009) On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. J Multivar Anal 100(9):2065–2082
Raskutti G, Wainwright MJ, Yu B (2010) Restricted eigenvalue properties for correlated gaussian designs. J Mach Learn Res 11:2241–2259
Ravikumar P, Wainwright M, Lafferty J (2010) High-dimensional ising model selection using ℓ_1-regularized logistic regression. Ann Stati 38(3):1287–1319
van Borkulo CD, Borsboom D, Epskamp S, Blanken TF, Boschloo L, Schoevers RA, Waldorp LJ (2014) A new method for constructing networks from binary data. Scientific reports 4
van de Geer S, Bühlmann P, Ritov Y (2013) On asymptotically optimal confidence regions and tests for high-dimensional models. arXiv preprint arXiv:1303.0518
Van de Geer SA (2008) High-dimensional generalized linear models and the lasso. Ann Stat 36:614–645
van de Geer SA, Bühlmann P et al (2009) On the conditions used to prove oracle results for the lasso. Electron J Stat 3:1360–1392
Venkatesh S (2013) The theory of probability. Cambridge University Press, Cambridge
Wainwright MJ (2009) Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). Inform Theory IEEE Trans 55(5):2183–2202
Wainwright MJ, Jordan MI (2008) Graphical models, exponential families, and variational inference. Found Trends Mach Learn 1(1–2):1–305
Waldorp L (2015) Testing for graph differences using the desparsified lasso in high-dimensional data. (submitted)

Young G, Smith R (2005) Essentials of statistical inference. Cambridge University Press, Cambridge
Zhang C-H, Zhang SS (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. J R Stat Soc 76(1):217–242