Statistica Sinica Preprint No: SS-2024-0021	
Title	An Ising Similarity Regression Model for Modeling Multivariate Binary Data
Manuscript ID	SS-2024-0021
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0021
Complete List of Authors	Zhi Yang Tho, Francis K. C. Hui and Tao Zou
Corresponding Authors	Zhi Yang Tho
E-mails	zhiyang.tho@anu.edu.au
Notice: Accepted author version.	

An Ising Similarity Regression Model for Modeling Multivariate Binary Data

Zhi Yang Tho*, Francis K. C. Hui, and Tao Zou*

The Australian National University

Abstract: Understanding the dependence structure between response variables is an important component in the analysis of correlated multivariate data. This article focuses on modeling dependence structures in multivariate binary data, motivated by a study aiming to understand how patterns in different U.S. senators' votes are determined by similarities (or lack thereof) in their attributes, e.g., political parties and social network profiles. To address such a research question, we propose a new Ising similarity regression model which regresses pairwise interaction coefficients in the Ising model against a set of similarity measures available/constructed from covariates. Model selection approaches are further developed through regularizing the pseudo-likelihood function with an adaptive lasso penalty to enable the selection of relevant similarity measures. We establish estimation and selection consistency of the proposed estimator under a general setting where the number of similarity measures and responses tend to infinity. Simulation study demonstrates the strong finite sample performance of the proposed estimator, particularly compared with several existing Ising model estimators in estimating the matrix of pairwise interaction coefficients. Applying the Ising similarity regression model to a dataset of roll call voting

*Corresponding author.

records of 100 U.S. senators, we are able to quantify how similarities in senators' parties, businessman occupations and social network profiles drive their voting associations.

Key words and phrases: Conditional dependence, Ising model, Lasso, Model selection, Multivariate data, Pseudo-likelihood

1. Introduction

The study of the dependence structure in correlated multivariate data has drawn much attention in recent years, due to the increasing accessibility of multi-response data across many disciplines such as finance (Tsay, 2013), economics (McElroy and Trimbur, 2023) and ecology (Hui, Müller, and Welsh, 2023). For a multivariate response vector $\boldsymbol{y} = (y_1, \cdots, y_p)^\top$, the dependencies across any set of two components y_j and $y_{j'}$ for $j, j' = 1, \cdots, p$ can be modeled by specifying a joint probabilistic distribution for $\boldsymbol{y}$. Two prominent examples are Gaussian graphical models (Whittaker, 1990) and Ising models (Ising, 1925), which specify a probability density function (pdf) and a probability mass function (pmf) for a continuous and binary response vector $\boldsymbol{y}$, respectively, where the pdf and pmf of these two models both contain a component of the form $\sum_{1\le j<j'\le p} \exp(\theta_{jj'} y_j y_{j'})$ with $\theta_{jj'}$ denoting a coefficient for the interaction term between y_j and $y_{j'}$. Accordingly, the set of $\{\theta_{jj'} : j, j' = 1, \cdots, p\}$ are referred to as pairwise interaction coefficients in these joint models. Moreover,

in both models each $\theta_{jj'}$ is directly tied to the conditional dependence relationship between y_j and $y_{j'}$ given the remaining elements in the response vector $\boldsymbol{y}$ (Hammersley-Clifford equivalence, Hammersley and Clifford, 1971): if $\theta_{jj'}$ is close to/far away from zero, the conditional dependence between y_j and $y_{j'}$ is weak/strong, with the limiting case of $\theta_{jj'} = 0$ implying conditional independence. This interpretation of the pairwise interaction coefficient $\theta_{jj'}$ has contributed to the popularity of Gaussian graphical models and Ising models for correlated multivariate data analysis; see for instance Meinshausen and Bühlmann (2006), Yuan and Lin (2007) and Friedman, Hastie, and Tibshirani (2007) for the Gaussian graphical models, and Majewski, Li, and Ott (2001), Guo et al. (2015) and Bhattacharya and Mukherjee (2018) for the Ising models, among many others.

In this article, we focus on modeling the dependence relationships for a multivariate binary response vector $\boldsymbol{y}$ through, or equivalently, modeling the pairwise interaction coefficients $\theta_{jj'}$ in, the Ising model. As demonstrated above, the conditional dependence of multivariate binary responses in the Ising model is captured by the pairwise interaction coefficients $\theta_{jj'}$, which can be collected into a symmetric interaction matrix $\boldsymbol{\Theta} = (\theta_{jj'})_{p\times p}$ where $\theta_{j'j} = \theta_{jj'}$. A variety of approaches have been proposed to estimate and regularize $\boldsymbol{\Theta}$. For example, Banerjee, El Ghaoui, and d' Aspremont (2008) proposed a block-coordinate de-

scent algorithm to solve an approximate sparse maximum likelihood problem for the estimation of $\mathbf{\Theta}$. Ravikumar, Wainwright, and Lafferty (2010) considered a neighborhood estimation method based on fitting a separate regularized logistic regression with the lasso penalty to each binary response, while Höfling and Tibshirani (2009) and Guo et al. (2010) used a pseudo-likelihood function coupled with a lasso-type penalty to simultaneously estimate and regularize all pairwise interaction coefficients in $\mathbf{\Theta}$. Further penalized likelihood approaches for sparse $\mathbf{\Theta}$ estimation have been developed by Lee, Ganapathi, and Koller (2006) and Xue, Zou, and Cai (2012), among others.

In contrast to estimating and regularizing the elements of the interaction matrix $\mathbf{\Theta}$, this article proposes a novel Ising similarity regression model which regresses the pairwise interaction coefficients $\theta_{jj'}$ against a set of pairwise similarity measures $w_{jj'}^{(k)}$; that is,

$$\theta_{jj'} = \sum_{k=1}^{K} \alpha_k w_{jj'}^{(k)}, \text{ for } j \neq j', \tag{1.1}$$

where α_k denotes the regression coefficient associated with the k-th similarity measure $w_{jj'}^{(k)}$ for $k = 1, \cdots, K$. It is worth noting $w_{jj'}^{(k)}$ measures the similarity between j and j', which can be either observed directly as part of the data collection process, or induced from available auxiliary information. As a motivating

example, we consider U.S. Senate roll call voting data where the binary response y_j represents the j-th senator's voting record (Yea or Nay) to a particular piece of legislation or bill, and $w_{jj'}^{(k)}$ are constructed based on additional attributes of the j-th and j'-th senators, such as political parties and occupations. In such an example, the regression coefficients α_k in model (1.1) offer a clear, explicit quantification of how the k-th similarity measure between the j-th and j'-th senators affects their conditional dependence when it comes to voting patterns.

At this point, it is important to acknowledge the work of Cheng et al. (2014), who proposed an Ising regression model to regress the pairwise interaction coefficients $\theta_{jj'}$ on a covariate vector $\boldsymbol{x}$. Such a model with $\theta_{jj'} = \theta_{jj'}(\boldsymbol{x})$ does not provide the same interpretation for our motivating example however, as their interaction coefficients $\theta_{jj'}$ for different pairs (j, j') only depend on the same covariate vector $\boldsymbol{x}$ instead of the pairwise similarity measures $w_{jj'}^{(k)}$ considered in model (1.1). Also, in a parallel line of research, several studies have considered using covariates to model the precision matrix encoding the dependence structure in Gaussian graphical models. Liu et al. (2010) partitioned the covariate space based on classification and regression trees into multiple subspaces that can have different precision matrices, while Lee and Xue (2018) proposed a nonparametric mixture of Gaussian graphical models with mixture probabilities and precision matrices that are both covariate-dependent. More recently, Wang

et al. (2022) and Ni, Stingo, and Baladandayuthapani (2022) proposed Bayesian approaches to model the elements of the precision matrix as linear functions of covariates. As these studies focus on precision matrices of Gaussian graphical models, they are not directly applicable to the multivariate binary response setting which we focus on in this article.

Apart from the aforementioned example in political science, the proposed Ising similarity regression model has a multitude of applications across finance and ecology. For example, in the study of capital markets, y_j can be an indicator denoting whether the j-th firm has distributed dividends to its shareholders, and $w_{jj'}^{(k)}$ can be the similarity between the j-th and j'-th firms' financial fundamentals such as market value, cash flow, and leverage. Also, in ecology, y_j can be a binary variable indicating the presence or absence of the j-th species, and $w_{jj'}^{(k)}$ can be the similarity between the j-th and j'-th species' trait values; see an application of the proposed model to such an ecology dataset in Section S9 of the supplementary material, to illustrate the wide applicability of the proposed method. More broadly, the use of similarity measures in regression is also motivated by recent developments in covariance regression modeling (Zou et al., 2017, 2020, 2022), where the covariance between continuous responses is modeled as a linear combination of similarity measures. Indeed, the proposed model can be written in a matrix regression form, and this connects to the wider litera-

ture linking (functions of) matrix parameters to a linear combination of matrices; see for example Anderson (1973), Pourahmadi (1999) and Bonat and Jørgensen (2016). To the best of our knowledge though, the proposed Ising similarity regression model is the first to establish such an idea for the Ising model specifically, by explicitly linking similarity matrices $\boldsymbol{W}_k = (w_{jj'}^{(k)})_{p\times p}$ to the interaction matrix $\boldsymbol{\Theta}$ that captures the conditional dependence of binary responses.

To estimate the proposed model, we study a regularized pseudo-likelihood approach which augments the pseudo-likelihood function of model (1.1) with an adaptive lasso penalty (Zou, 2006). Doing so induces sparse estimation of the regression coefficients $\{\alpha_k : k = 1, \cdots, K\}$, which allows us to recover similarity measures that are truly relevant in explaining the conditional dependence relationships between the binary responses. It is important here to highlight that this article differs from the aforementioned studies on Ising model estimation, as our main focus is to induce sparsity on the regression coefficients α_k as opposed to the similarity measures $w_{jj'}^{(k)}$ and the resulting interaction matrix $\boldsymbol{\Theta}$. Put another way, we aim to identify important drivers of the conditional dependence relationships i.e., similarity selection rather than edge selection, by treating the similarity measures as given covariates.

Under a setting where the number of regression coefficients K and responses p tend to infinity, we establish estimation and selection consistency for the regu-

larized pseudo-likelihood estimator. To select the tuning parameter in the adaptive lasso penalty, we employ a cross-validation approach which preserves the dependencies between the elements of $\boldsymbol{y}$. Simulation results support the theoretical findings of the proposed estimator, demonstrating its strong finite sample estimation and model selection performance. Specifically, the proposed estimator not only outperforms other estimators such as the unpenalized estimator and the lasso-penalized estimator in estimating the parameters of the Ising similarity regression model, but also performs much better than the traditional Ising model estimators that ignore the additional information from similarity measures in estimating the Ising model interaction matrix $\boldsymbol{\Theta}$. Additionally, we carry out simulation studies to compare the similarity selection performance of our cross-validation approach to the use of AIC (Akaike, 1998) and BIC (Schwarz, 1978) criteria for choosing the tuning parameter, with results showing that BIC has a comparable performance to cross-validation approach while AIC tends to suffer from overfitting. We apply the Ising similarity regression model to roll call voting records of 100 U.S. senators from the 117-th Congress, with results demonstrating how similarities of senators' attributes and social network activities drive the association between their voting patterns. In particular, aside from the expected findings such as senators from the same state or party being more likely to vote similarly, we find that senators who are businessmen or share

certain follower-followee relationships on Twitter tend to exhibit more similar voting patterns.

The rest of this article is organized as follows. Section 2 introduces the Ising similarity regression model along with the proposed regularized pseudo-likelihood estimator. Section 3 discusses the theoretical properties of the regularized estimator. Section 4 presents simulation studies, while an application to U.S. roll call voting dataset is provided in Section 5. Section 6 offers some concluding remarks. All theoretical proofs of the theorems developed in this article, along with detailed empirical comparisons to other estimation approaches, as well as an additional application to Scotland Carabidae ground beetle dataset, are presented in the supplementary material.

2. An Ising Similarity Regression Model

2.1 Model Set-Up

Let $\boldsymbol{u} = (u_1, \cdots, u_p)^\top$ be any vector in the space $\{0,1\}^p$. The Ising model (Ising, 1925) specifies the following pmf for the p-dimensional binary response vector $\boldsymbol{y}$,

$$f(\boldsymbol{u};\boldsymbol{\theta}) = P(\boldsymbol{y}=\boldsymbol{u};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\left(\sum_{j=1}^{p} \theta_{jj} u_j + \sum_{1\le j<j'\le p} \theta_{jj'} u_j u_{j'}\right), \quad (2.2)$$

where $\boldsymbol{\theta} = (\theta_{11}, \cdots, \theta_{1p}, \cdots, \theta_{p-1,p-1}, \theta_{p-1,p}, \theta_{pp})^\top$ is a $p(p+1)/2$-dimensional parameter vector, and the partition function $Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{u}\in\{0,1\}^p} \exp(\sum_{j=1}^p \theta_{jj} u_j + \sum_{1\le j<j'\le p} \theta_{jj'} u_j u_{j'})$ is an intractable normalization constant in the pmf. The main effect parameters in (2.2) are given by θ_{jj} for $j = 1, \cdots, p$, while the pairwise interaction coefficients in the pmf are given by $\theta_{jj'}$ for $j, j' = 1, \cdots, p$ and $\theta_{j'j} = \theta_{jj'}$. As discussed in Section 1, the symmetric interaction matrix $\boldsymbol{\Theta} = (\theta_{jj'})_{p\times p}$ is useful for studying the conditional dependence structure among binary responses.

Suppose now that, in addition to observing the binary response vector $\boldsymbol{y}$, we also record a set of $p \times p$ symmetric similarity matrices $\boldsymbol{W}_k = (w_{jj'}^{(k)})_{p\times p}$ for $k = 1, \cdots, K$, which may be available directly as part of the data collection process or constructed from auxiliary information variables $z_{j1}, \cdots, z_{jK}$ associated to the j-th response for $j = 1, \cdots, p$. In the latter, each element $w_{jj'}^{(k)}$ of the similarity matrix $\boldsymbol{W}_k$ measures the similarity between z_{jk} and $z_{j'k}$ for $j \neq j'$. For instance, if z_{jk} is quantitative, then we can set $w_{jj'}^{(k)} = \exp(-|z_{jk} - z_{j'k}|^2)$, whereas if z_{jk} is qualitative then we set $w_{jj'}^{(k)} = 1$ if z_{jk} and $z_{j'k}$ have the same categorical level, and $w_{jj'}^{(k)} = 0$ otherwise (see also Johnson and Wichern, 1992). For completeness and reasons of parameter identifiability, the diagonals $w_{jj}^{(k)}$ are set to be zeros for all $j = 1, \cdots, p$ and $k = 1, \cdots, K$.

Given a set of similarity matrices $\boldsymbol{W}_k$ for $k = 1, \cdots, K$, the Ising similarity

regression model as introduced in equation (1.1) can be equivalently formulated as modeling the interaction matrix $\boldsymbol{\Theta}$ via the form

$$\boldsymbol{\Theta} = \sum_{j=1}^{p} \theta_{jj} \boldsymbol{\Delta}_{jj} + \sum_{k=1}^{K} \alpha_k \boldsymbol{W}_k, \tag{2.3}$$

where $\boldsymbol{\Delta}_{jj}$ is a $p \times p$ matrix with the (j, j)-th element being one and other elements being zeros for $j = 1, \cdots, p$. This model re-parameterizes the vector $\boldsymbol{\theta}$ in equation (2.2) by a new parameter vector $\boldsymbol{\vartheta} = (\theta_{11}, \cdots, \theta_{pp}, \boldsymbol{\alpha}^\top)^\top$ with $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)^\top$, such that the re-parameterized pmf for $\boldsymbol{y}$ can be written as

$$f(\boldsymbol{u}; \boldsymbol{\vartheta}) = \frac{1}{Z(\boldsymbol{\vartheta})} \exp\left\{ \sum_{j=1}^{p} \theta_{jj} u_j + \sum_{1 \le j < j' \le p} \left(\sum_{k=1}^{K} \alpha_k w_{jj'}^{(k)} \right) u_j u_{j'} \right\}, \tag{2.4}$$

where $Z(\boldsymbol{\vartheta}) = \sum_{\boldsymbol{u} \in \{0,1\}^p} \exp\{ \sum_{j=1}^{p} \theta_{jj} u_j + \sum_{1 \le j < j' \le p} (\sum_{k=1}^{K} \alpha_k w_{jj'}^{(k)}) u_j u_{j'} \}$.

In equation (2.3), the vector of regression coefficients $\boldsymbol{\alpha}$ describes how similarities directly affect the pairwise interaction coefficients (conditional dependence relationships) of the binary responses. Note also by utilizing similarity matrices to model the interaction matrix $\boldsymbol{\Theta}$, model (2.4) involves a substantially smaller number of parameters ($p+K$ parameters) compared to the standard Ising model in (2.2), which has $(p+1)p/2$ parameters. The proposed Ising similarity regression model is thus particularly useful when the dimension of the binary re-

sponse vector p is large: even when the number of similarity matrices K grows at the same rate as the dimension p, the number of parameters in model (2.4) only grows linearly in p compared to $O(p^2)$ parameters in model (2.2).

2.2 Estimation and Similarity Selection

Suppose we have observations of p-dimensional response vectors $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{ip})^\top \in \{0,1\}^p$ for $i = 1, \cdots, n$, where $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n \overset{i.i.d.}{\sim} f(\cdot; \boldsymbol{\vartheta})$ follows the Ising similarity regression model in equation (2.4), and $i.i.d.$ denotes independent and identically distributed. Recall the pmf $f(\cdot; \boldsymbol{\vartheta})$ involves the normalization constant $Z(\boldsymbol{\vartheta})$, which is a sum of 2^p terms. As a result, maximum likelihood estimation based on $\prod_{i=1}^n f(\boldsymbol{y}_i; \boldsymbol{\vartheta})$ is computationally not feasible when the dimension p is large. To overcome this problem, we adapt the existing literature (e.g., Höfling and Tibshirani, 2009; Ravikumar, Wainwright, and Lafferty, 2010) and propose a pseudo-likelihood estimation approach to fit the Ising similarity regression model. Let $\boldsymbol{y}_{i\setminus j} = (y_{i1}, \cdots, y_{i,j-1}, y_{i,j+1}, \cdots, y_{ip})^\top$ denote the i-th observed response vector without the j-th element, and $f_j(\cdot|\boldsymbol{y}_{i\setminus j}; \boldsymbol{\vartheta})$ denote the corresponding conditional pmf of y_{ij} given $\boldsymbol{y}_{i\setminus j}$ for $i = 1, \cdots, n$ and $j = 1, \cdots, p$. Then the (unregularized) pseudo-likelihood estimator can be obtained by maximizing the pseudo-likelihood function $\prod_{i=1}^n \prod_{j=1}^p f_j(y_{ij}|\boldsymbol{y}_{i\setminus j}; \boldsymbol{\vartheta})$. In particular, the conditional pmf $f_j(\cdot|\boldsymbol{y}_{i\setminus j}; \boldsymbol{\vartheta})$ can be derived based on $f(\cdot; \boldsymbol{\vartheta})$ in

equation (2.4), and takes the simple form below,

$$f_j(u|\boldsymbol{y}_{i\setminus j};\boldsymbol{\vartheta}) = \frac{\exp\left\{u\left(\theta_{jj} + \sum_{k=1}^{K}\alpha_k \sum_{j'\neq j} w_{jj'}^{(k)} y_{ij'}\right)\right\}}{1+\exp\left(\theta_{jj} + \sum_{k=1}^{K}\alpha_k \sum_{j'\neq j} w_{jj'}^{(k)} y_{ij'}\right)}, \tag{2.5}$$

for $u \in \{0,1\}$, $i = 1,\cdots,n$ and $j = 1,\cdots,p$. It follows that the conditional log-odds is given by

$$\log\left\{\frac{\mathrm{P}(y_{ij}=1|\boldsymbol{y}_{i\setminus j};\boldsymbol{\vartheta})}{1-\mathrm{P}(y_{ij}=1|\boldsymbol{y}_{i\setminus j};\boldsymbol{\vartheta})}\right\} = \theta_{jj} + \sum_{k=1}^{K}\alpha_k \sum_{j'\neq j} w_{jj'}^{(k)} y_{ij'}, \tag{2.6}$$

for $i = 1,\cdots,n$ and $j = 1,\cdots,p$, from which we observe that the conditional pmf $f_j(\cdot|\boldsymbol{y}_{i\setminus j};\boldsymbol{\vartheta})$ and thus pseudo-likelihood estimation no longer involve the intractable normalization constant $Z(\boldsymbol{\vartheta})$. In fact, equation (2.6) bears a similar form to fitting a logistic regression for the conditional log-odds of $y_{ij} = 1$ against the set of K covariates $\{\sum_{j'\neq j} w_{jj'}^{(k)} y_{ij'} = \boldsymbol{W}_{j\cdot}^{(k)\top}\boldsymbol{y}_i : k = 1,\cdots,K\}$, with an intercept term θ_{jj} and regression coefficients $\boldsymbol{\alpha} = (\alpha_1,\cdots,\alpha_K)^\top$, where $\boldsymbol{W}_{j\cdot}^{(k)\top} \in \mathbb{R}^{1\times p}$ denotes the j-th row of $\boldsymbol{W}_k$.

We can further augment the above pseudo-likelihood estimator with a penalty to perform variable selection on the elements of $\boldsymbol{\alpha}$. This is useful in practice when there are a non-negligible number of similarity matrices available for data analysis, but only a subset of them are anticipated to be relevant in equation (2.3).

To perform simultaneous estimation and regularization on the coefficient vector $\boldsymbol{\alpha}$, we augment the log pseudo-likelihood function based on equation (2.5) with an adaptive lasso penalty (Zou, 2006), resulting in a regularized pseudo-likelihood estimator that minimizes the objective function

$$\begin{aligned} & -\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\log\{f_j(y_{ij}|\boldsymbol{y}_{i\backslash j};\boldsymbol{\vartheta})\}+\lambda\sum_{k=1}^{K}w_k\left|\alpha_k\right| \\ =& -\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left[y_{ij}\left(\theta_{jj}+\sum_{k=1}^{K}\alpha_k\boldsymbol{W}_{j\cdot}^{(k)\top}\boldsymbol{y}_i\right)\right. \\ & \left.-\log\left\{1+\exp\left(\theta_{jj}+\sum_{k=1}^{K}\alpha_k\boldsymbol{W}_{j\cdot}^{(k)\top}\boldsymbol{y}_i\right)\right\}\right]+\lambda\sum_{k=1}^{K}w_k\left|\alpha_k\right|, \end{aligned} \tag{2.7}$$

given a tuning parameter $\lambda > 0$. Following Zou (2006) and Huang, Ma, and Zhang (2008) among others, we set the adaptive weights as $w_k = 1/|\bar{\alpha}_k|$ for $k = 1, \cdots, K$, where $\bar{\boldsymbol{\alpha}} = (\bar{\alpha}_1, \cdots, \bar{\alpha}_K)^\top$ denotes the unregularized pseudo-likelihood estimator i.e., the estimator of $\boldsymbol{\alpha}$ which minimizes the objective function (2.7) with $\lambda = 0$. It is worth noting that θ_{jj}, the j-specific intercepts for $j = 1, \cdots, p$, are not regularized; this is similar to the literature of sparse Ising models (Guo et al., 2010; Ravikumar, Wainwright, and Lafferty, 2010) as well as other regularized regression settings in general (e.g, Hui, Müller, and Welsh, 2017b, 2018). Regarding the choice of the penalty function, in this article we focus our developments on the adaptive lasso penalty $\lambda\sum_{k=1}^{K}w_k\left|\alpha_k\right|$; see Section

6 for a discussion on alternative penalty functions. In particular, with adaptive weights there is only a single regularization parameter λ in the penalty, and the whole objective function (2.7) remains convex for optimization. Moreover, the incorporation of adaptive weights w_k allows for varying degrees of regularization on the coefficients α_k across $k = 1, \cdots, K$, and facilitates selection consistency for a sparse coefficient vector $\boldsymbol{\alpha}$ which we theoretically examine in the next section. We also emphasize that the adaptive lasso penalty here is used for the selection of regression coefficients associated with similarity matrices i.e., similarity selection, and not for the selection of edges in $\boldsymbol{\Theta}$. This is apparent when we see that the penalty induces sparsity in the regression coefficients, but not the similarity matrices and the resulting $\boldsymbol{\Theta}$. The details of implementing the regularized pseudo-likelihood estimation by minimizing (2.7) are discussed in Section 4.

3. Theoretical Results

In this section, we establish asymptotic properties for the regularized pseudo-likelihood estimator of the Ising similarity regression model (2.4). Since our main interest lies in estimating similarity regression coefficients $\boldsymbol{\alpha}$ that only depend on the pairwise interaction coefficients in equation (1.1), then we focus on

a variant of the model with no main effects, giving rise to the criterion

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\{ -l(\boldsymbol{\alpha}) + \lambda \sum_{k=1}^{K} w_k \left| \alpha_k \right| \right\}, \tag{3.8}$$

where $l(\boldsymbol{\alpha}) = \sum_{i=1}^{n} l_i(\boldsymbol{\alpha})/(np)$, $l_i(\boldsymbol{\alpha}) = \sum_{j=1}^{p}[y_{ij}(\sum_{k=1}^{K} \alpha_k \boldsymbol{W}_{j\cdot}^{(k)\top} \boldsymbol{y}_i) - \log\{1 + \exp(\sum_{k=1}^{K} \alpha_k \, \boldsymbol{W}_{j\cdot}^{(k)\top} \boldsymbol{y}_i)\}]$, $w_k = 1/|\bar{\alpha}_k|$ and $\bar{\boldsymbol{\alpha}} = (\bar{\alpha}_1, \cdots, \bar{\alpha}_K)^\top = \arg\min_{\boldsymbol{\alpha}}\{ - l(\boldsymbol{\alpha})\}$; see Guo et al. (2010, 2015) for a similar theoretical treatment. It is worth noting that while the pseudo-likelihood function is used in criterion (3.8) to overcome the issue of intractable normalization constant, theoretical results in this section are still obtained under the joint probability distribution of the Ising similarity regression model.

Let $\boldsymbol{\alpha}^{(0)} = (\alpha_1^{(0)}, \cdots, \alpha_K^{(0)})^\top$ denote the true value of the coefficient vector $\boldsymbol{\alpha}$, $S = \{k : \alpha_k^{(0)} \neq 0, \text{ for } k = 1, \cdots, K\}$ denote the set indexing all truly non-zero coefficients, $S^c = \{1, \cdots, K\} \backslash S$, and let the cardinality of S be denoted by $|S| = K_0$, where K_0 is assumed to be finite. Furthermore, we construct

$$\boldsymbol{\mathcal{X}}^{(i)} = \begin{pmatrix} \boldsymbol{\mathcal{X}}^{(i,1)\top} \\ \vdots \\ \boldsymbol{\mathcal{X}}^{(i,p)\top} \end{pmatrix} = \begin{pmatrix} \boldsymbol{W}_{1\cdot}^{(1)\top} \boldsymbol{y}_i & \cdots & \boldsymbol{W}_{1\cdot}^{(K)\top} \boldsymbol{y}_i \\ \vdots & \ddots & \vdots \\ \boldsymbol{W}_{p\cdot}^{(1)\top} \boldsymbol{y}_i & \cdots & \boldsymbol{W}_{p\cdot}^{(K)\top} \boldsymbol{y}_i \end{pmatrix}, \text{ for } i = 1, \cdots, n, \tag{3.9}$$

and define the $K \times K$ matrices $\boldsymbol{U}^0 = E\{\sum_{i=1}^{n} \boldsymbol{\mathcal{X}}^{(i)\top} \boldsymbol{\mathcal{X}}^{(i)}/(np)\}$ and $\boldsymbol{M}^0 =$

$E\{-\nabla^2 l(\boldsymbol{\alpha}^{(0)})\}$. Finally, for a generic $m \times r$ matrix $\boldsymbol{H} = (h_{tv})_{m\times r}$, and subsets of row and column indices $\mathcal{T} \subseteq \{1, \cdots, m\}$ and $\mathcal{V} \subseteq \{1, \cdots, r\}$, let $\boldsymbol{H}_{\mathcal{T},\mathcal{V}}$ denote the submatrix of $\boldsymbol{H}$ consisting of rows and columns indexed by $\mathcal{T}$ and $\mathcal{V}$, respectively, $\|\boldsymbol{H}\|_1 = \max_{1\le v\le r}\{\sum_{t=1}^{m} |h_{tv}|\}$ denote the matrix 1-norm of $\boldsymbol{H}$, and $\Lambda_{\min}(\boldsymbol{H})$ and $\Lambda_{\max}(\boldsymbol{H})$ denote the smallest and largest eigenvalues of $\boldsymbol{H}$, respectively, when $\boldsymbol{H}$ is a square matrix (i.e., $m = r$). We also refer the reader to Table S1 in Section S1 of the supplementary material for a list of important notations used throughout this article along with their definitions.

We introduce the following technical conditions.

Condition 1. *There exist finite positive constants* $C_{\min}$ *and* $C_{\max}$ *such that* $\Lambda_{\min}(\boldsymbol{M}^0) \ge C_{\min}$ *and* $\Lambda_{\max}(\boldsymbol{U}^0) \le C_{\max}$, *where* $\boldsymbol{M}^0$ *and* $\boldsymbol{U}^0$ *are defined below equation* (3.9).

Condition 2. *There exists a finite positive constant* C_W *such that* $\|\boldsymbol{W}_k\|_1 \le C_W$ *for all symmetric similarity matrices* $\{\boldsymbol{W}_k : k = 1, \cdots, K\}$.

By the Cauchy Interlacing Theorem (Parlett, 1980, p. 186) and Condition 1, similar conditions hold for the submatrices $\boldsymbol{M}^0_{S,S}$ and $\boldsymbol{U}^0_{S,S}$; in particular, $\Lambda_{\min}(\boldsymbol{M}^0_{S,S}) \ge C_{\min}$ and $\Lambda_{\max}(\boldsymbol{U}^0_{S,S}) \le C_{\max}$. Condition 2 is a bounded matrix norm assumption which is similar to the conditions imposed on the similarity matrices $\boldsymbol{W}_k$ in the literature (see for instance, Condition C8 in Zou et al., 2022). For example, when $\boldsymbol{W}_k$ is a symmetric adjacency matrix with elements

$w_{jj'}^{(k)} \in \{0, 1\}$ capturing the neighborhood relationship among p nodes, the condition is equivalent to a column-sparsity (and row-sparsity since $\boldsymbol{W}_k$ are symmetric) condition that requires the number of neighbors for each node to be finite even when the total number of nodes p diverges. Other classes of similarity matrices such as those with bounded elements $|w_{jj'}^{(k)}| \leq C$ and s-sparse column (and row) vectors (e.g., Wainwright, 2019, p. 156) i.e., $\sum_{j=1}^{p} 1_{\{|w_{jj'}^{(k)}|>0\}} \leq s$, where C and s are finite positive constants and $1_{\{\cdot\}}$ is the indicator function, also satisfy Condition 2 with $C_W = sC$. Together with the above assumed sparsity on the true coefficient vector $\boldsymbol{\alpha}^{(0)}$, Condition 2 implies similar bounded matrix norm condition for the true interaction matrix $\boldsymbol{\Theta}^{(0)} = \sum_{k=1}^{K} \alpha_k^{(0)} \boldsymbol{W}_k$; that is, $\|\boldsymbol{\Theta}^{(0)}\|_1 \leq C_W \sum_{k \in S} |\alpha_k^{(0)}|$.

We first establish the estimation consistency for the unregularized pseudo-likelihood estimator that is used to construct the adaptive weights.

Theorem 1. *Assume Conditions 1 – 2 are satisfied. If* $K\sqrt{\log(p)/n} = o(1)$ *and there exists a finite positive constant* C_∇ *such that* $K = o(p^{C_\nabla^2/(8C_W^2)})$ *as* $n, p \to \infty$*, then with probability tending to one it holds that* $\|\bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}\|_2 \leq \bar{M}\sqrt{K \log(p)/n}$*, for a finite positive constant* $\bar{M} > 4C_\nabla/C_{\min}$.

The proofs of Theorem 1 along with all other theoretical results are provided in the supplementary material. The above convergence rate bears a similar form to the convergence rate of many lasso-penalized estimators (e.g., Guo

et al., 2010; Nghiem et al., 2022), where the factor of K, instead of K_0, in the numerator is expected given this is the unregularized estimator. In addition, Theorem 1 is valid under both cases of fixed K and diverging K, so long as $K\sqrt{\log(p)/n} \to 0$ and $K/p^{C_\nabla^2/(8C_W^2)} \to 0$.

To study the asymptotic properties of the regularized pseudo-likelihood estimator, we introduce two additional technical conditions.

Condition 3. *There exists a constant* $C_M \in (0,1)$ *such that* $\|\boldsymbol{M}^0_{S^c,S}(\boldsymbol{M}^0_{S,S})^{-1}\|_\infty \leq 1 - C_M$.

Condition 4. $\lambda\sqrt{n}/\{\min\{|\alpha_k^{(0)}| : k \in S\}\sqrt{\log(p)}\} \to 0$ *and* $\lambda n/\{\sqrt{K}\log(p)\} \to \infty$.

Condition 3 is commonly known as the mutual incoherence or irrepresentability condition (Hastie, Tibshirani, and Wainwright, 2015), and, together with the conditions on $\boldsymbol{M}^0_{S,S}$ and $\boldsymbol{U}^0_{S,S}$ implied by Condition 1, are similar to those assumed by Meinshausen and Bühlmann (2006) and Guo et al. (2010) among others. Condition 4 is similar to existing conditions in the literature regarding the rates of the tuning parameters for adaptive lasso regression (e.g., Wang, Li, and Leng, 2009; Hui, Müller, and Welsh, 2018).

We now state the main results of this paper for the regularized pseudo-likelihood estimator $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_1, \cdots, \hat{\alpha}_K)^\top$ of the Ising similarity regression model.

Theorem 2. *Assume Conditions 1 – 4 are satisfied. If* $K\sqrt{\log(p)/n} = o(1)$ *and there exists a finite positive constant* C_∇ *such that* $K = o(p^{C_\nabla^2/(8C_W^2)})$ *as* $n, p \to \infty$*, then with probability tending to one it holds that*

(a) $\left\|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^{(0)}\right\|_2 \leq M\sqrt{K_0 \log(p)/n}$ *for a finite positive constant* $M > 4/C_{\min}$*;*

(b) $\hat{\alpha}_k \neq 0$ *for all* $k \in S$ *and* $\hat{\alpha}_k = 0$ *for all* $k \in S^c$*.*

Theorem 2(a) establishes estimation consistency of the regularized pseudo-likelihood estimator, noting its convergence rate can be faster than that of the unregularized estimator. Theorem 2(b) establishes selection consistency of the regularized pseudo-likelihood estimator i.e., it can recover the underlying sparsity pattern of $\boldsymbol{\alpha}^{(0)}$. This is attractive when the Ising similarity regression model is applied to datasets where the dimension of responses and number of similarity matrices may be large.

To conclude this section, we remark Condition 4 provides requirements on the rate of the tuning parameter λ based on the smallest truly non-zero coefficient $\min\{|\alpha_k^{(0)}| : k \in S\}$, as the basis for establishing the results in Theorem 2. For instance, if $\min\{|\alpha_k^{(0)}| : k \in S\}$ is bounded away from zero, then it can be verified that $\lambda = \{n/\log(p)\}^{-t}K^v$ for some $t \in (1/2, 1)$ and $v \in [2t - 3/2, 2t - 1]$ suffices for Condition 4 to hold. If this is relaxed and we permit $\min\{|\alpha_k^{(0)}| : k \in S\}$ to tend to zero at a rate satisfying $\min\{|\alpha_k^{(0)}| : k \in S\}\{n/\log(p)\}^m \to$

∞ for some $m \in (0, 1/4]$, then $\lambda = O(\{\log(p)/n\}^q)$ for some $q \in [m+1/2, 3/4]$ will satisfy the requirements. In practice, since $\min\{|\alpha_k^{(0)}| : k \in S\}$ is unknown, we adopt a data-driven approach to select λ as discussed in the next section.

4. Simulation Study

We conduct a numerical study to evaluate the finite sample performance of the proposed regularized pseudo-likelihood estimator for the Ising similarity regression model (2.4). Briefly, we consider sample sizes $n \in \{50, 100, 200, 400\}$, numbers of binary responses $p \in \{10, 25, 50, 100, 200\}$, and $K = 20$ similarity matrices with only the first $K_0 = 5$ of them having non-zero true regression coefficients i.e., $S = \{1, \cdots, 5\}$. A total of 1000 datasets are simulated from the Ising similarity regression model (2.4) for each combination of n and p. Details of the simulation settings, including the true values of the main effect parameters $\{\theta_{jj}^{(0)} : j = 1, \cdots, p\}$ and regression coefficients $\{\alpha_k^{(0)} : k = 1, \cdots, K\}$ are provided in Section S7.1 of the supplementary material.

We employ a two-step algorithm to compute the regularized pseudo-likelihood estimator. First, from the discussion below equations (2.6) – (2.7), the unregularized estimator is obtained by fitting a logistic regression model with $\boldsymbol{y}_{1:n} = (y_{11}, \cdots, y_{n1}, \cdots, y_{1p}, \cdots, y_{np})^\top$ as the response and $(\boldsymbol{I}_p \otimes \mathbf{1}_n, \boldsymbol{\mathcal{X}})$ as the model matrix, where $\mathbf{1}_n$ is an n-dimensional vector of ones, $\otimes$ is the Kronecker prod-

uct operator, $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}^{(1,1)}, \cdots, \boldsymbol{\mathcal{X}}^{(n,1)}, \cdots, \boldsymbol{\mathcal{X}}^{(1,p)}, \cdots, \boldsymbol{\mathcal{X}}^{(n,p)})^\top$, and $\boldsymbol{\mathcal{X}}^{(i,j)}$ denotes the j-th row of the matrix $\boldsymbol{\mathcal{X}}^{(i)}$ given in equation (3.9). In the second step, we compute the regularized pseudo-likelihood estimator by fitting an adaptive lasso regularized logistic regression to $\boldsymbol{y}_{1:n}$ and $(\boldsymbol{I}_p \otimes \mathbf{1}_n, \boldsymbol{\mathcal{X}})$ via the `R` package `glmnet` (Friedman, Hastie, and Tibshirani, 2010), where the adaptive regularization weights in (2.7) are constructed based on the unregularized estimator.

To select the tuning parameter λ in (2.7), we utilize a ten-fold cross-validation approach, where the observations in each dataset are grouped at the level indexed by $i = 1, \cdots, n$. That is, we split the data $\{(\boldsymbol{y}_1, \boldsymbol{\mathcal{X}}^{(1)}), \cdots, (\boldsymbol{y}_n, \boldsymbol{\mathcal{X}}^{(n)})\}$ randomly into ten folds and select an optimal λ that gives the largest mean pseudo-likelihood (averaged across ten test sets), noting that the same set of adaptive weights constructed based on the unregularized estimator from the full dataset is used throughout the cross-validation. The grouping of the observations is done to preserve the dependence structure of the data, since y_{ij} are independent in the index i but dependent in the response index j through the Ising model (see also Warton, Thibaut, and Wang, 2017).

We compare our proposed regularized pseudo-likelihood estimator (Regularized) with three basic approaches: the lasso-regularized estimator (Lasso) based on (2.7) with $w_k = 1$ for $k = 1, \cdots, K$ and λ selected using a similar ten-fold cross-validation approach, the unregularized estimator (Unregularized)

based on (2.7) with $\lambda = 0$, and the oracle estimator (Oracle) based on minimizing an alternative to (2.7) where $\lambda = 0$ and α_k are set to be zeros for $k \in S^c$. The lasso-regularized estimator is considered to study the effect of penalty choice on the empirical performance of the estimator, the unregularized estimator is included to examine whether regularization could lead to better overall estimation performance, and the oracle estimator is used as a benchmark since it incorporates additional information regarding the index set $S = \{1, \cdots, 5\}$. We compare point estimation performance for all four estimators of the regression coefficients and main effect parameters using their mean square error (MSE), MSE_α and MSE_θ, respectively, and the model selection performance of the proposed estimator and lasso-regularized estimator of the regression coefficients using the true positive rate (TPR) and false positive rate (FPR); see Section S7.1 of the supplementary material for details on the computation of these performance measures.

Table 1 shows the MSE_α for all four estimators decrease as n and p increase, while their MSE_θ exhibit a clear decreasing trend when n increases but no obvious patterns when p increases; this is to be expected given the main effect parameter θ_{jj} is specific to each of the responses in the Ising similarity regression model for $j = 1, \cdots, p$. The MSE_α and MSE_θ for the proposed regularized pseudo-likelihood estimators are typically much smaller than the unregularized

Table 1: MSE for the Oracle, Regularized, Lasso, and Unregularized estimators of the regression coefficients (MSE_α) and main effect parameters (MSE_θ) in the simulation study involving the Ising similarity regression model. MSE_α is multiplied by 1000 for clarity.

		$1000 \times \text{MSE}_\alpha$				MSE_θ			
p	n	Oracle	Regularized	Lasso	Unregularized	Oracle	Regularized	Lasso	Unregularized
10	50	45.791	68.420	37.524	351.622	0.350	0.402	0.346	0.896
	100	21.544	44.121	28.569	147.225	0.167	0.207	0.193	0.370
	200	10.755	27.506	19.048	68.262	0.082	0.111	0.099	0.172
	400	5.390	16.298	11.717	33.021	0.038	0.057	0.052	0.080
25	50	5.858	14.717	13.474	25.985	0.298	0.391	0.351	0.546
	100	2.846	6.817	6.992	12.089	0.137	0.184	0.171	0.242
	200	1.434	3.025	3.546	5.996	0.068	0.088	0.087	0.118
	400	0.710	1.356	1.780	2.950	0.033	0.041	0.043	0.057
50	50	2.100	5.901	5.908	8.368	0.550	0.788	0.719	0.737
	100	0.969	1.924	2.537	3.967	0.120	0.156	0.166	0.205
	200	0.469	0.860	1.266	1.935	0.053	0.067	0.074	0.093
	400	0.232	0.386	0.625	0.939	0.027	0.033	0.038	0.047
100	50	0.765	4.517	3.282	3.065	2.314	1.551	2.201	2.444
	100	0.362	1.390	1.609	1.441	0.500	0.559	0.512	0.567
	200	0.175	0.367	0.510	0.709	0.099	0.111	0.113	0.132
	400	0.089	0.119	0.227	0.355	0.033	0.034	0.041	0.049
200	50	0.351	5.493	2.406	1.282	1.971	2.180	2.420	2.073
	100	0.169	0.376	0.422	0.614	0.290	0.350	0.395	0.343
	200	0.079	0.108	0.197	0.295	0.073	0.076	0.089	0.102
	400	0.039	0.061	0.096	0.148	0.030	0.031	0.035	0.044

estimators, and in fact similar to the oracle estimators especially when n and p are large. This is consistent with the notion that the proposed estimators take advantage of the underlying sparsity in the model, resulting in better overall estimation performance than the unregularized estimators. Furthermore, the MSEs of the proposed estimators tend to be smaller than the lasso-regularized estimators when n and p are large.

Turning to the model selection performance of the proposed estimator in Table 2, while there is slight underfitting as reflected by the comparably low TPR for the case of smallest $p = 10$, both the TPR and FPR tend to one and zero, respectively, as n and p increase. This is consistent with Theorem 2(b), empiri-

Table 2: TPR and FPR for the Regularized (top panel) and Lasso (bottom panel) estimators of the regression coefficients in the simulation study involving the Ising similarity regression model.

Regularized								
	TPR				FPR			
$p \backslash n$	50	100	200	400	50	100	200	400
10	0.337	0.490	0.654	0.824	0.167	0.202	0.225	0.235
25	0.835	0.970	0.999	1	0.218	0.218	0.197	0.167
50	0.954	0.999	1	1	0.186	0.174	0.154	0.102
100	0.976	0.998	0.999	1	0.043	0.105	0.076	0.009
200	0.925	0.998	1	1	0.043	0.058	0.005	0
Lasso								
	TPR				FPR			
$p \backslash n$	50	100	200	400	50	100	200	400
10	0.307	0.500	0.711	0.888	0.124	0.185	0.237	0.290
25	0.859	0.980	1	1	0.298	0.356	0.392	0.395
50	0.982	0.999	1	1	0.403	0.412	0.426	0.428
100	0.999	1	1	1	0.294	0.327	0.405	0.429
200	0.996	1	1	1	0.254	0.354	0.392	0.433

cally demonstrating that the proposed method is able to recover the underlying sparsity in the Ising similarity regression model. The lasso-regularized estimator suffers from overfitting as seen from its FPR not converging towards zero, thus supporting the use of the adaptive lasso penalty in our proposed estimator.

Results from comparing our ten-fold cross-validation approach to the use of AIC and BIC criteria for choosing the tuning parameter are provided in Table S2 of the supplementary material. Overall, results show the BIC performs similarly well as our cross-validation approach, while the AIC suffers from clear overfitting when p is small due to its weaker model complexity penalty. These results provide empirical support for using the cross-validation approach to choose λ.

Next, we compare the proposed estimator to other traditional estimators of the Ising model in the literature that involve direct estimation of the Ising model

interaction matrix $\boldsymbol{\Theta}$, without considering the similarity matrices $\boldsymbol{W}_k$ (e.g., Section 3 of Höfling and Tibshirani, 2009). From Table S3 in the supplementary material, we see that the proposed estimator greatly outperforms the traditional Ising model estimators in recovering the true $\boldsymbol{\Theta}$ matrix, since it incorporates the additional information from the similarity measures.

Finally, we conduct simulation studies with varying number of similarity matrices $K \in \{10, 20, 40, 80, 200\}$ while keeping the same number of truly non-zero regression coefficients $K_0 = 5$, as a further investigation on the effect of K on the empirical performance of various estimators. Unsurprisingly, the estimation and model selection performance of all methods decline as more irrelevant similarity matrices are being added, noting the proposed estimator still performs reasonably well relative to other estimators under different settings of K; see Tables S4 – S6 in the supplementary material for full results.

5. Application to U.S. Senate Roll Call Voting Data

We apply the Ising similarity regression model (2.4) to roll call voting data from the U.S. Senate as part of the 117-th Congress, covering the period from 6 January 2021 to 20 May 2021 (date of data collection). Roll call voting data has previously been studied using a variety of statistical techniques including undirected graphical models (Banerjee, El Ghaoui, and d' Aspremont, 2008).

Here, we use our model to study how voting associations between senators are associated with their similarities in various demographic attributes and social network profiles. The dataset is obtained from the U.S. Senate's website (`https://www.senate.gov/`), which originally consists of binary voting records, coded as one for 'Yea' and zero for 'Nay', on 199 bills by 100 senators of the 117-th Congress. After performing some preliminary data wrangling procedures (see Section S8 of the supplementary material), the final dataset analyzed consists of $n = 138$ bills voted by $p = 100$ senators, where all bills are treated as independent (e.g., Banerjee, El Ghaoui, and d' Aspremont, 2008; Guo et al., 2010).

We obtain several attributes for each senator, including their state, political party, class, and age from the U.S. Senate website, along with gender and occupation from Wikipedia. Additionally, each senator's Twitter handle is obtained from the 'us-senate' GitHub project of 'CivilServiceUSA', and used to compute the number of tweets and number of followers for each senator. Each of these attributes is then converted into a similarity matrix $\boldsymbol{W}_k$ following the procedure described in Section 2.1, depending on whether it is a qualitative (state, party, class, gender and occupation) or quantitative (age, number of tweets, number of Twitter followers) attribute. We refer the reader to Section S8 of the supplementary material for detailed construction of these similarity matrices, as well as a

descriptive analysis for the above attributes and the binary votes of the senators. Note that a symmetric adjacency matrix $\boldsymbol{W}_k$ is also constructed to summarize the Twitter follower-followee relationship among the senators, such that $w_{jj'}^{(k)} = 1$ if the j-th senator follows the j'-th senator on Twitter or vice versa, and zero otherwise. As a result, the analysis consists of $K = 15$ similarity matrices.

We fit the Ising similarity regression model using the regularized pseudo-likelihood estimator to identify and quantify truly important attributes driving the voting associations between senators, where λ is selected using ten-fold cross-validation with the groupings being done at the bill level. After performing variable selection and obtaining the estimated non-zero regression coefficients, we construct 95% Wald confidence intervals for each coefficient based on standard errors from the empirical sandwich covariance matrix obtained by deriving the score and Hessian matrix of the log pseudo-likelihood for each observation with respect to the set of chosen similarity measures; see Section S6 of the supplementary material for details of its derivation.

Table 3 shows only 7 of the 15 regression coefficients are estimated to be non-zero using the adaptive lasso penalty; estimation results for the main effect parameters are given in the supplementary material. Of these, unsurprisingly there is statistically clear evidence that senators from the same state and/or party are more likely to vote similarly on the bills (analogous state and party effects

Table 3: Point estimates and 95% confidence intervals (in parentheses) for the regression coefficients corresponding to the $K = 15$ similarity matrices, based on fitting the Ising similarity regression model (2.4) to the U.S. Senate roll call voting data using regularized pseudo-likelihood estimation. Estimates whose corresponding confidence interval excludes zero are bolded.

Estimation of α_k					
State	Party	Class	Age	Gender	Lawyer
2.342	**0.167**	0.021	0	0	0.018
(1.846,2.838)	**(0.153,0.181)**	(-0.034,0.075)	0	0	(-0.047,0.083)
Executive	Businessman	Farmer	Army	Teacher	Professor
0	**0.451**	0	0	0	0
0	**(0.018,0.884)**	0	0	0	0
Tweets	Followers	Twitter Follower-Followee Relationship			
-0.091	0	**0.144**			
(-0.136,-0.046)	0	**(0.110,0.177)**			

are found in Banerjee, El Ghaoui, and d' Aspremont, 2008; Guo et al., 2010, among others), although the former exhibits a much stronger effect. Most occupations are found to be uninformative for the conditional dependence relationships between senator voting patterns, except for businessman and lawyer, although the confidence interval corresponding to the effect of lawyer contains zero. The presence of a positive effect for the businessman similarity matrix could be attributed to senators who are businessmen tending to vote similarly on bills related to the economy.

There is statistically clear evidence of a positive association between the Twitter follower-followee relationship adjacency matrix and senator voting patterns. A possible explanation is that one senator who follows the other senator on Twitter has more exposure to their advocated ideologies, and hence is more likely to vote similarly. It is also possible that the senator who follows the other senator on Twitter already agrees with their ideologies in the first place, while

their interactions on Twitter further reinforce such agreement, leading to positive associations in their voting patterns. The similarity in terms of senators' popularity on Twitter, as measured by the number of followers, does not have any effect on the association between senators' votes. Interestingly, although the effect of the senators' number of tweets is found to be negative, when we run a separate analysis based on fitting an Ising similarity regression model with only this similarity matrix, its associated coefficient becomes positive: $\hat{\alpha}_{\text{Tweets}} = 0.083$ with 95% confidence interval being $(0.080, 0.087)$. Therefore, the negative coefficient found in Table 3 is conjectured to be due to a large amount of information contained within this similarity matrix that could be explained by other similarity matrices i.e., a form of matrix multicollinearity.

Figure 1 presents graphs of the weighted similarity/adjacency matrices $\hat{\alpha}_k \boldsymbol{W}_k$ for selected attributes, together with the estimated interaction matrix $\hat{\boldsymbol{\Theta}}$ describing the voting associations for a subset of 20 senators, where $\hat{\boldsymbol{\Theta}} = \sum_{j=1}^{100} \hat{\theta}_{jj} \boldsymbol{\Delta}_{jj} + \sum_{k=1}^{15} \hat{\alpha}_k \boldsymbol{W}_k$ based on equation (2.3), and $\hat{\alpha}_k$ and $\hat{\theta}_{jj}$ are the regularized pseudo-likelihood estimators. The choice of senators to be included is made by beginning with an empty set, and sequentially adding pairs of senators who have the largest off-diagonal elements $\hat{\theta}_{jj'}$ in $\hat{\boldsymbol{\Theta}}$ until the set contains the top 20 senators. To further improve clarity of the graph based on the estimated $\hat{\boldsymbol{\Theta}}$, we remove all edges with $\hat{\theta}_{jj'} \leq 0.1297$ (the median of $\{\hat{\theta}_{jj'} : j \neq j'\}$); this explains the visu-

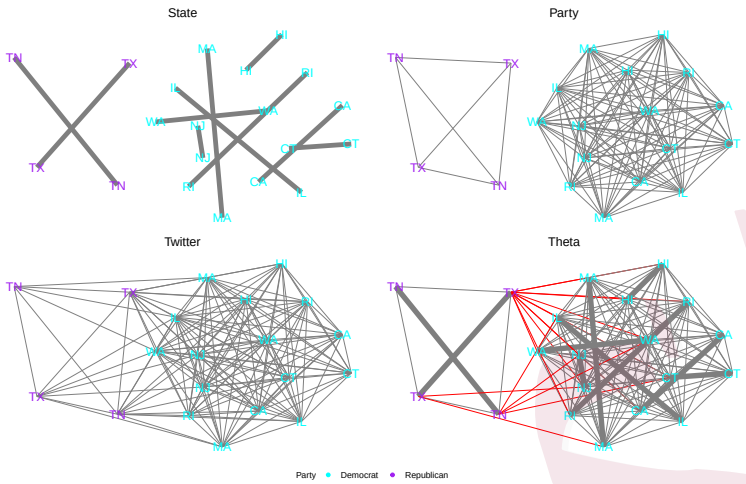

Figure 1: Graphs of weighted similarity/adjacency matrices $\hat{\alpha}_k \boldsymbol{W}_k$ associated with state, party and Twitter follower-followee relationship for a subset of 20 senators, where the edge width is proportional to the estimated $\hat{\alpha}_k$. The bottom right plot presents the weighted graph based on the estimated interaction matrix $\hat{\boldsymbol{\Theta}}$, where the edge width is proportional to its associated $\hat{\theta}_{jj'}$ and edges between senators with different states and parties are colored red. Nodes are labeled with the state abbreviation of each senator and the color represents senator's party.

ally denser graph of Twitter follower-followee relationship compared to that of $\hat{\boldsymbol{\Theta}}$ in Figure 1. We emphasize that the removal of edges here is purely to improve the visualization and interpretability of the graph, noting our focus is on similarity selection and not edge selection as discussed earlier in Sections 1 and 2.2. The corresponding graph without such removal of edges can be found in Figure S5 in the supplementary material, and provides qualitatively similar conclusions as above. It can be seen that the estimated strong voting associations (i.e., large estimated values of $\theta_{jj'}$) between senators are being driven heavily not only by similarities in senators' state and party affiliations, but also their relationships on Twitter as indicated by the red-colored edges in the graph of $\hat{\boldsymbol{\Theta}}$. Indeed, the fit-

ted Ising similarity regression model yields a log pseudo-likelihood of -1125.28, compared to -9441.01 for a null model with only the main effect parameters. This corresponds to a pseudo R^2 (Cohen et al., 2013) of approximately 0.88, indicating the selected model provides a much better fit than the null model through the inclusion of the available $\boldsymbol{W}_k$ matrices to explain the dependence structure.

6. Conclusion

We develop an Ising similarity regression model to study the effect of predictor information, arising in the form of similarity matrices, on conditional dependence relationships among binary responses. A computationally efficient regularized pseudo-likelihood estimator using the adaptive lasso penalty is proposed, which we demonstrate to be estimation and model selection consistent under a general setting where the number of similarity matrices K and responses p grow with sample size n. Simulations demonstrate the strong finite sample point estimation and selection performance of the proposed estimator, especially compared with several traditional Ising model estimators in recovering the interaction matrix, as it incorporates additional information from relevant similarity matrices. The cross-validation approach for choosing the tuning parameter is shown to have comparable performance to the BIC in terms of model selec-

tion. Applying the proposed model to the U.S. Senate roll call voting data not only identifies well-documented state and party effects in driving senator voting patterns, but also establishes new insights into the importance of senators' similarities in businessman occupation and social network relationships on their voting dependence. These findings are new and differ from those obtained from the graphical structure analyses using standard Ising models (Guo et al., 2010, 2015). In particular, while graphical structure analyses aim to reveal dependency structures among responses (senators) through edge selection, our method seeks to recover and quantify how different auxiliary information (senator's attributes) influence this structure via similarity selection. While various methods have been developed for the former problem, our approach is one of the first to have been developed for addressing the latter.

A logical next step would be to extend the proposed model to incorporate predictor information from both observations and responses, thus forming a sort of "double Ising similarity regression model". This may involve relaxing the assumption of $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n \overset{i.i.d.}{\sim} f(\cdot; \boldsymbol{\vartheta})$ to consider correlated data such as spatial and/or temporal data, where similarity matrices could be built up based on knowledge of spatial or temporal distance (e.g., Bonat and Jørgensen, 2016) to account for spatial or temporal dependence between observations. The assumption could also be relaxed by extending the Ising similarity regression model to

allow for heterogeneous interaction matrices $\boldsymbol{\Theta}^{(i)}$ for different $\boldsymbol{y}_i$ vectors, e.g. by considering $\boldsymbol{W}_k^{(i)}$ that are heterogeneous across $i = 1, \cdots, n$; see for instance, Zou et al. (2022) who model heterogeneous covariance matrices of continuous response vectors based on linear combination of heterogeneous similarity matrices.

While this article focuses on the estimation and selection of the regression coefficients $\boldsymbol{\alpha}$, future research would involve rigorous theoretical study of the inferential component of our method e.g., by establishing asymptotic normality results for the regularized pseudo-likelihood estimator. Given the comparable model selection performance of the BIC to our cross-validation method, it would also be useful to extend this investigation to consider other types of information criterion in selecting λ (e.g., Zhang, Li, and Tsai, 2010; Fan and Tang, 2013). On a related note, other penalty functions could be used to replace the adaptive lasso penalty for similarity selection, including the use of group or fusion penalties when the set of similarity matrices exhibits some sort of hierarchy or ordering (Hui, Müller, and Welsh, 2017a; Zhang et al., 2023). Finally, while this work focuses on binary responses using the Ising model, similar modeling idea could be extended to quadratic exponential families (Gourieroux, Monfort, and Trognon, 1984) to allow for other response types.

Supplementary Materials

The Supplementary Material contains sample versions of Conditions 1 and 3, proofs of the theorems, inference method, additional simulation results, along with supplementary details of application to the U.S. Senate roll call voting data, as well as an additional application to the Scotland Carabidae ground beetle data.

Acknowledgements

Zhi Yang Tho was supported by an Australian Government Research Training Program scholarship. Francis KC Hui was supported by an Australian Research Council Discovery Project DP230101908. Thanks to Alan Welsh for useful discussions.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pp. 199–213. Springer.

Anderson, T. W. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics* **1**, 135–141.

Banerjee, O., L. El Ghaoui, and A. d' Aspremont (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research* **9**, 485–516.

REFERENCES

Bhattacharya, B. B. and S. Mukherjee (2018). Inference in Ising models. *Bernoulli* **24**, 493–525.

Bonat, W. H. and B. Jørgensen (2016). Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65**, 649–675.

Cheng, J., E. Levina, P. Wang, and J. Zhu (2014). A sparse Ising model with covariates. *Biometrics* **70**, 943–953.

Cohen, J., P. Cohen, S. West, and L. Aiken (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*. Taylor & Francis.

Fan, Y. and C. Y. Tang (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* ***75***, 531–552.

Friedman, J., T. Hastie, and R. Tibshirani (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**, 1–22.

Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Theory. *Econometrica* **52**, 681–700.

Guo, J., J. Cheng, E. Levina, G. Michailidis, and J. Zhu (2015). Estimating heterogeneous graphical models for discrete data with an application to roll call voting. *The Annals of Applied Statistics* **9**, 821–848.

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2010). Joint structure estimation for categorical Markov networks.

REFERENCES

Hammersley, J. M. and P. Clifford (1971). Markov fields on finite graphs and lattices.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: The lasso and generalizations*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

Höfling, H. and R. Tibshirani (2009). Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *Journal of Machine Learning Research* ***10***, 883–906.

Huang, J., S. Ma, and C.-H. Zhang (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica* **18**, 1603–1618.

Hui, F. K. C., S. Müller, and A. H. Welsh (2017a). Hierarchical selection of fixed and random effects in generalized linear mixed models. *Statistica Sinica* **27**, 501–518.

Hui, F. K. C., S. Müller, and A. H. Welsh (2017b). Joint selection in mixed models using regularized PQL. *Journal of the American Statistical Association* **112**, 1323–1333.

Hui, F. K. C., S. Müller, and A. H. Welsh (2018). Sparse pairwise likelihood estimation for multivariate longitudinal mixed models. *Journal of the American Statistical Association* **113**, 1759–1769.

Hui, F. K. C., S. Müller, and A. H. Welsh (2023). GEE-assisted variable selection for latent variable models with multivariate binary data. *Journal of the American Statistical Association* **118**, 1252–1263.

Ising, E. (1925). Beitrag zur theorie der ferromagnetismus. *Zeitschrift für Physik* **31**, 253–258.

Johnson, R. A. and D. W. Wichern (1992). *Applied multivariate statistical analysis*. Prentice Hall.

Lee, K. H. and L. Xue (2018). Nonparametric finite mixture of Gaussian graphical models. *Technometrics* **60**, 511–521.

REFERENCES

Lee, S.-I., V. Ganapathi, and D. Koller (2006). Efficient structure learning of Markov networks using L_1-regularization. In B. Schölkopf, J. Platt, and T. Hoffman (Eds.), *Advances in Neural Information Processing Systems*, Volume **19**. MIT Press.

Liu, H., X. Chen, L. Wasserman, and J. Lafferty (2010). Graph-valued regression. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (Eds.), *Advances in Neural Information Processing Systems*, Volume **23**. Curran Associates, Inc.

Majewski, J., H. Li, and J. Ott (2001). The Ising model in physics and statistical genetics. *The American Journal of Human Genetics* **69**, 853–862.

McElroy, T. S. and T. Trimbur (2023). Variable targeting and reduction in large vector autoregressions with applications to workforce indicators. *Journal of Applied Statistics* **50**, 1515–1537.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**, 1436–1462.

Nghiem, L. H., F. K. C. Hui, S. Müller, and A. H. Welsh (2022). Sparse sliced inverse regression via Cholesky matrix penalization. *Statistica Sinica* **32**, 2431–2453.

Ni, Y., F. C. Stingo, and V. Baladandayuthapani (2022). Bayesian covariate-dependent Gaussian graphical models with varying structure. *Journal of Machine Learning Research* **23**, 1–29.

Parlett, B. (1980). *The symmetric eigenvalue problem*. Prentice-Hall International Series in Computer Science. Prentice-Hall.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Uncon-

strained parameterisation. *Biometrika* **86**, 677–690.

Ravikumar, P., M. J. Wainwright, and J. D. Lafferty (2010). High-dimensional Ising model selection using l_1-regularized logistic regression. *The Annals of Statistics* **38**, 1287–1319.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Tsay, R. (2013). *Multivariate time series analysis: With R and financial applications*. Wiley Series in Probability and Statistics. Wiley.

Wainwright, M. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, H., B. Li, and C. Leng (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 671–683.

Wang, Z., V. Baladandayuthapani, A. O. Kaseb, H. M. Amin, M. M. Hassan, W. Wang, and J. S. Morris (2022). Bayesian edge regression in undirected graphical models to characterize interpatient heterogeneity in cancer. *Journal of the American Statistical Association* **117**, 533–546.

Warton, D. I., L. Thibaut, and Y. A. Wang (2017). The PIT-trap – A "model-free" bootstrap procedure for inference about regression models with discrete, multivariate responses. *PloS one* **12**, e0181790.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley Series in Probability and Statistics. Wiley.

Xue, L., H. Zou, and T. Cai (2012). Nonconcave penalized composite conditional likelihood estimator of sparse Ising models. *The Annals of Statistics* **40**, 1403–1429.

REFERENCES

Yuan, M. and Y. Lin (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* **94**, 19–35.

Zhang, X., F. Huang, F. K. C. Hui, and S. Haberman (2023). Cause-of-death mortality forecasting using adaptive penalized tensor decompositions. *Insurance: Mathematics and Economics* **111**, 193–213.

Zhang, Y., R. Li, and C.-L. Tsai (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* ***105***, 312–323.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Zou, T., W. Lan, R. Li, and C.-L. Tsai (2022). Inference on covariance-mean regression. *Journal of Econometrics* **230**, 318–338.

Zou, T., W. Lan, H. Wang, and C.-L. Tsai (2017). Covariance regression analysis. *Journal of the American Statistical Association* **112**, 266–281.

Zou, T., R. Luo, W. Lan, and C.-L. Tsai (2020). Covariance regression model for non-normal data. In C. F. Lee and J. C. Lee (Eds.), *Handbook of Financial Econometrics, Mathematics, Statistics, and Machine Learning*, Chapter 113, pp. 3933–3945. World Scientific.

Zhi Yang Tho

The Australian National University, Canberra, ACT 2600, Australia.

E-mail: zhiyang.tho@anu.edu.au

Francis K.C. Hui

REFERENCES

The Australian National University, Canberra, ACT 2600, Australia.

E-mail: francis.hui@anu.edu.au

Tao Zou

The Australian National University, Canberra, ACT 2600, Australia.

E-mail: tao.zou@anu.edu.au