Autologistic Network Model on Binary Data for Disease Progression Study

Yei Eun Shin[1,*], Huiyan Sang[2], Dawei Liu[3], Toby A. Ferguson[3] and Peter X. K. Song[4]

[1]Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, U.S

[2]Department of Statistics, Texas A&M University, College Station, Texas, U.S.A.

[3]Biogen, Cambridge, Massachusetts, U.S.A.

[4]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, U.S.A.

**email:* yei-eun.shin@nih.gov

Summary: This paper focuses on analysis of spatio-temporal binary data with absorbing states. The research was motivated by a clinical study on amyotrophic lateral sclerosis (ALS), a neurological disease marked by gradual loss of muscle strength over time on multiple body regions. We propose an autologistic regression model to capture complex spatial and temporal dependencies in muscle strength among different muscles. As it is not clear how the disease spreads from one muscle to another, it may not be reasonable to define a neighborhood structure based on spatial proximity. Relaxing the requirement for pre-specification of spatial neighborhoods as in existing models, our method can learn underlying network structure of disease spreading pattern directly from observed data. The model also allows the network autoregressive effects to differ in accordance with muscles' previous status. Based on joint distribution derived from this autologistic model specification, joint transition probabilities of responses over locations can be estimated and the one-time ahead disease status can be predicted. Model parameters are estimated through maximization of penalized pseudo-likelihood. Post-model selection inference was conducted via a bias-correction method, for which the asymptotic distributions were derived. Simulation studies were conducted to evaluate the performance of the proposed method. The method was applied to the analysis of muscle strength loss from ALS clinical study.

Key words: Absorbing states; Amyotrophic lateral sclerosis disease; Bias-corrected lasso; Network; Penalized peudo-likelihood estimation; Spatio-temporal dependence.

1 Introduction

This research was motivated by a clinical study on amyotrophic lateral sclerosis (ALS). ALS, also known as Lou Gehrig's disease, is a neurological disease that mainly affects the nerve cells in the brain and the spinal cord that are responsible for controlling voluntary muscle movements. As the disease progresses, a patient's brain gradually loses the ability to signal and control muscle movements, which leads to muscle weakness, impaired physical functionality and finally death. Currently there is no treatment for the disease. The symptom typically starts from a particular muscle group and then spreads to other muscles as the disease progresses. In other words, muscles at different locations are interconnected so that a "normal" muscle can become diseased due to another "diseased" muscle. The spreading pattern, however, remains unknown. Such a disease spreading can arise from many other occasions; for example, a pathogen in a farmland or a virus in a computer network.

Our research interest is to characterize how the disease spreads over space and time. There are several features that pose a challenge to statistical modeling and inference. First, neighborhood is not clearly defined, because spatial closeness may not reflect the underlying disease spreading pattern. For example, disease in one location can spread to another distant location rather than any nearby locations. Therefore, actual dependence over space is determined by some sophisticated but unknown structure in a latent space. Second, outcome of interest is irreversible over time. For example, in ALS once a muscle becomes "diseased", it can never return to "normal" as there is no treatment. Thus, data generation mechanism has an absorbing state. Lastly, the strength of temporal association depends strongly on previous disease statuses. For example, recently diseased muscles pose a higher risk than muscles that are diseased a while ago.

The autologistic model, first proposed by Besag (1974), is one of the most widely used modeling methods for spatial binary data. Being closely related to a joint Markov random field for binary responses, this model is known to be advantageous for its direct modeling of spatial dependence over other existing models via latent variables for dependence. Caragea and Kaiser (2009) propose a centered autologistic model to allow for more interpretable parameters, and Hughes et al. (2011) conduct comparative studies to evaluate the performance of several computation methods for fitting the autologistic model. In these papers, a pre-specified neighborhood structure is often required to establish spatial dependence. To relax this requirement, there has been a surge of recent work (Höfling and Tibshirani, 2009; Ravikumar et al., 2010; Xue et al., 2012) in the context of the Ising models, a special case of autologistic regression, where sparse regularization techniques is invoked to learn sparse network associations among nodes. These regularization methods, however, focus mostly on spatial data and hence are not directly applicable to the evaluation of spreading patterns over space and time, as needed in the case of ALS disease. For spatio-temporal binary data, Zhu et al. (2005) develop a model via joint distributions to first estimate spatial correlation then prediction. To incorporate absorbing states, Kaiser et al. (2014) formulate a model in which sufficient support conditions and specify to construct well-defined joint distributions of all observations. Both approaches rely on pre-specified neighbor structures on lattices. Agaskar and Lu (2013) consider a binary vector autologistic regressive model in time and use regularization estimation methods to study sparse network. However, they neither model absorbing states nor consider simultaneous spatial dependence.

The main contribution of our paper is to develop an autologistic network model in space and time that addresses the aforementioned challenges: the model learns a spatial network from data without the need to pre-specify a neighborhood structure; accounts for absorbing states; and considers varying impacts depending on previous status. Also, it has centered autocovariates to capture the residual dependence structure from the large-scale structure (Caragea and Kaiser, 2009; Hughes and Haran, 2013), and consequently alleviates spatial confounding issues and enhance parameter interpretability. While the model bases on the conditional probability of a single location, we derive a valid joint distribution of all locations to establish one-time transition probability, which is essential for spreading pattern analysis.

For the estimation of model parameters, we invoke a pseudo-likelihood inference (Besag, 1975). Since the proposed model has an excessive number of parameters representing all possible pairwise spatial associations, penalization is employed with the least absolute shrinkage and selection operator (LASSO) penalty as suggested by Tibshirani (1996). The connection of this pseudo-likelihood with the standard framework of generalized linear model (GLM) estimation is established. Also, as is well known, since the LASSO estimator is biased and does not have a tractable limiting distribution, we propose a bias-correction for penalized pseudo-likelihood estimator and establish its asymptotic distribution, following the ideas of post-selection inference (Van de Geer et al., 2014; Tang et al., 2016).

The remainder of the paper is organized as follows. In Section 2, we propose an autologistic network model for binary data observed in space and time. In Section 3, we derive a valid joint distribution for the proposed model and formulate transition probabilities. In Section 4, we discuss a bias-corrected penalized pseudo-likelihood estimator with an iterative algorithm and a large-sample theorem. In Section 5, we present simulation studies to assess how accurately our proposed approach performs statistical inferences. In Section 6, the application to the motivating ALS clinical study is illustrated. Finally, we summarize the research findings and suggest future studies in Section 7.

2 Autologistic Network Model with Absorbing States

Denote a binary random variable such that $Y_m(s_j, t)$ is 1 if a location s_j is diseased at time t for a subject m, and 0 if normal. Let M be the number of subjects, N_s the number of locations that are fixed over subjects, and $T(m)$ the number of times that may vary over subjects. We define two index sets,

$$\begin{aligned}\mathcal{P}^0_{mt} &= \{j : Y_m(s_j, t-1) = 0, j = 1, \ldots, N_s\}; \qquad (1)\\ \mathcal{P}^1_{mt} &= \{j : Y_m(s_j, t-1) = 1, j = 1, \ldots, N_s\},\end{aligned}$$

where $\mathcal{P}^0_{mt}$ is an *active* set consisting of locations previously normal, in which they have a chance to change their status at the next time, and $\mathcal{P}^1_{mt}$ is an *absorbing* set consisting of remaining locations, previously diseased. The vector of independent variables including an intercept and time t as well as other covariates is denoted by $\boldsymbol{X}_m$.

We specify the conditional probability of presence of a progressive disease, given independent variables and other locations' status at previous and current times,

$$\mathrm{P}[Y_m(s_j, t) = 1 | \boldsymbol{X}_m, Y_m(s_k, t-1), Y_m(s_k, t) \text{ for } \forall k \in \{1, \ldots, N_s\} \setminus \{j\}] = p_m(s_j, t), \qquad (2)$$

where $A \setminus B$ denotes the set A excluding B. This conditional probability is assumed to be Markovian over time, while subjects are sampled independently. We propose an autologistic network model, for $j \in \mathcal{P}^0_{mt}$,

$$\mathrm{logit}\{p_m(s_j, t)\} = \boldsymbol{X}_m^T \boldsymbol{\beta} + \sum_{k \in \mathcal{P}^0_{mt} \setminus \{j\}} \eta_{0jk} \{Y_m(s_k, t) - \kappa_m\} + \sum_{k \in \mathcal{P}^1_{mt} \setminus \{j\}} \eta_{1jk} \{Y_m(s_k, t) - \kappa_m\} \qquad (3)$$

$$\text{subject to } \eta_{0jk} = \eta_{0kj} \text{ and } \eta_{1jk} = \eta_{1kj} \text{ for all } j \neq k$$

where $\mathrm{logit}(p) = \log\{p/(1-p)\}$ and $\kappa_m = \exp(\boldsymbol{X}_m^T \boldsymbol{\beta})/\{1 + \exp(\boldsymbol{X}_m^T \boldsymbol{\beta})\}$, $m = 1, \ldots, M$. Note that $p_m(s_j, t) = 1$ for $j \in \mathcal{P}^1_{mt}$ because of absorbing features.

We center the autocovariate terms by κ_m to reduce bias and make the better interpretations on η-parameters. Without centering, $p_m(s_j, t)$ is completely biased toward 1-valued autocovariates and impossible to obtain effect from 0-valued autocovariates; $p_m(s_j, t)$ would never decrease in time. The centering constant κ_m is the expectation of $\mathrm{logit}\{p_m(s_j, t)\}$ under an *independence* model without autocovariates. When $\boldsymbol{\beta} = 0$, for example, 0 and 1 of autocovariates are distinguished by -0.5 and 0.5, respectively. See Caragea and Kaiser (2009) for more discussions.

We divide the autocovariates into the active and absorbing set to allow normal and diseased locations having different levels of contributing risk through η_{0jk} and η_{1jk}, respectively. The parameter η_{0jk} indicates an impact of the location s_k on s_j when s_k is previously normal. Similarly, η_{1jk} is an impact of s_k on s_j when s_k is previously diseased. These parameters characterize associations between s_j and s_k for any $j \neq k$, which allow us to learn a network structure concerning all possible connections of any two locations.

Also, we restrict symmetricity on η-parameters to ensure both a valid joint distribution (as shown in Section 3) and nondirectional correlations. In infectious factor spreading pattern studies such as ALS data analysis (in Section 6), these η-parameters can be further restricted to take only non-negative values from practical considerations; it is unusual that a normal location makes others more likely to be diseased.

[Table 1 about here.]

Table 1 specifies a simple example on η-parameters. Consider two locations s_1 and s_2, and assume that the effect of s_2 on s_1 is of interest when s_1 is normal at $t-1$. The probability of s_1 being diseased at t, $p_m(s_1,t)$, follows the proposed model (3) given the status of s_2 and other locations at $t-1$ and t. Model parameters η_{012} and η_{112} will characterize the effect of s_2 on s_1 as follows. If s_2 were both previously and currently normal (Case 1), $\text{logit}\{p_m(s_1,t)\}$ would decrease as much as $\eta_{012}(0-\kappa_m)$. If s_2 were previously normal but currently diseased (Case 2), $\text{logit}\{p_m(s_1,t)\}$ would increase as much as $\eta_{012}(1-\kappa_m)$. These two cases imply that strongly linked locations with high value of η_{012} are more likely to stay healthy (or be diseased) simultaneously. On the other hand, s_1 would always be ill-affected if s_2 were diseased at the previous time (Case 3); $\text{logit}\{p_m(s_1,t)\}$ will increase as much as $\eta_{112}(1-\kappa_m)$. There are no other cases for s_2 such as being previously diseased but currently normal because of absorbing feature. Likewise, η_{0jk} and η_{1jk} characterize the impacts of s_k on s_j for $j \neq k$.

3 Joint Distribution and Transition Probability

We show that the conditional probabilities modeled by (3) uniquely determine a valid joint distribution of spatio-temporal binary responses. For simplicity, let $\boldsymbol{\eta}_0 = \{\eta_{0jk}\}_{j<k}$ and $\boldsymbol{\eta}_1 = \{\eta_{1jk}\}_{j<k}$ be the vectorized autoregressive coefficients of size $N_s(N_s-1)/2$ where $j, k \in \{1, \ldots, N_s\}$, and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1)^T$, all coefficients of size p.

We first consider the spatial joint distribution for a given t and a given m: the distribution of a random vector $\boldsymbol{Y}_{mt} = \big\{Y_m(s_1,t), \ldots, Y_m(s_{N_s},t)\big\}^T$. Since the realizations of $\boldsymbol{Y}_{mt}$ against absorbing states have no chances to occur, we restrict its domain. Let $\mathcal{S}$ be all possible joint responses at N_s locations so that 2^{N_s} elements are in $\mathcal{S}$. Define $\mathcal{S}_{mt}$ as a subset of $\mathcal{S}$ including all available joint responses, $\mathcal{S}_{mt} = \big\{\boldsymbol{Y}_{mt} \in \mathcal{S} | Y_m(s_j,t) = 1 \text{ for } s_j \text{ s.t. } Y_m(s_j, t-1) = 1\big\}$, which includes $2^{|\mathcal{P}^0_{mt}|}$ elements, where $|\mathcal{P}^0_{mt}|$ is the number of active locations at t. According to Theorem 3 in Kaiser and Cressie (2000), we have a valid *spatial joint distribution* of $\boldsymbol{Y}_{mt} \in \mathcal{S}_{mt}$,

$$f(\boldsymbol{Y}_{mt}|\boldsymbol{\theta}) = \frac{\exp\{Q(\boldsymbol{Y}_{mt}|\boldsymbol{\theta})\}}{\sum_{\boldsymbol{Y}_{mt}\in\mathcal{S}_{mt}} \exp\{Q(\boldsymbol{Y}_{mt}|\boldsymbol{\theta})\}} \tag{4}$$

for a fixed subject m and time t, where

$$\begin{aligned} Q(\boldsymbol{Y}_{mt}|\boldsymbol{\theta}) = & \sum_{j\in\mathcal{P}^0_{mt}} Y_m(s_j,t)\Big\{\boldsymbol{X}_m^T\boldsymbol{\beta} - \sum_{k\in\mathcal{P}^0_{mt}\setminus\{j\}} \eta_{0jk}\kappa_m - \sum_{k\in\mathcal{P}^1_{mt}\setminus\{j\}} \eta_{1jk}\kappa_m\Big\} \\ & + \frac{1}{2}\sum_{j\in\mathcal{P}^0_{mt}} \Big\{ \sum_{k\in\mathcal{P}^0_{mt}\setminus\{j\}} \eta_{0jk} Y_m(s_j,t)Y_m(s_k,t) + \sum_{k\in\mathcal{P}^1_{mt}\setminus\{j\}} \eta_{1jk} Y_m(s_j,t)Y_m(s_k,t)\Big\}. \end{aligned}$$

The details for deriving (4) are deferred to Appendix A.

We can now write the *full joint distribution* of responses over all times, locations and subjects, denoted by $\boldsymbol{Y} = \{Y_m(s_j,t)|m = 1,\ldots,M; j = 1,\ldots,N_s; t = 1,\ldots,T(m)\}$, as $\mathbb{P}(\boldsymbol{Y}|\boldsymbol{\theta}) = \prod_{m=1}^M \prod_{t=1}^{T(m)} f(\boldsymbol{Y}_{mt}|\boldsymbol{\theta})$. This follows the independent assumption among responses across subjects, based on an inductive method with a valid Markov random field model at an initial time $t = 0$ (Kaiser et al., 2014).

Here, the spatial joint distribution in (4) can be viewed as conditional distribution at t given

$t-1$, because $\boldsymbol{Y}_{m(t-1)}$ determines both $\mathcal{P}^0_{mt}$ and $\mathcal{S}_{mt}$. Therefore, when making an inference on responses at next time, $\boldsymbol{Y}_{mt}$ given $\boldsymbol{Y}_{m(t-1)}$, we calculate *one-time transition probability*,

$$f(\boldsymbol{Y}_{mt}|\boldsymbol{\theta}) = \mathrm{P}(\boldsymbol{Y}_{mt}|\boldsymbol{Y}_{m(t-1)};\boldsymbol{\theta}), \tag{5}$$

which is essential for infectious factor spreading pattern studies. By plugging in estimates $\hat{\boldsymbol{\theta}}$ to (5), one can infer one-time ahead disease status or predict at which locations the disease is most (or least) likely to occur next. See Section 5 for numerical studies on this.

4 Penalized Maximum Pseudo Likelihood Estimation with Bias-correction

4.1 The Penalized Maximum Pseudo Likelihood Estimation

Estimating the autologistic model parameters in (3) is challenging since the normalizing constant in its joint likelihood function, the denominator of (4), is computationally costly when $\mathcal{S}_{mt}$ is a large set. For efficient computation, we replace the full likelihood with a product of conditional likelihoods, a pseudo-likelihood approach (Besag, 1974).

For simplicity, we stack all active responses into a longitudinal vector with a single index i as $\boldsymbol{\mathcal{Y}} = \{\mathcal{Y}_i | i = 1,\ldots,n\} = \{Y_m(s_j,t)|m = 1,\ldots,M;\ s_j \in \mathcal{P}^0_{mt};\ t = 1,\ldots,T(m)\}$ where $n = \sum_{m=1}^{M}\sum_{t=1}^{T(m)}\sum_{j=1}^{N_s} I(s_j \in \mathcal{P}^0_{mt})$. In other words, each i is uniquely assigned to an index combination (m,s_j,t). With this notation, $p_m(s_j,t)$ is expressed by the probability of a binary response $\mathcal{Y}_i$ from a logistic linear regression model; model (3) is equivalent to

$$\mathrm{logit}\{P(\mathcal{Y}_i = 1)\} = \boldsymbol{\mathcal{X}}_i(\kappa_i)^T\boldsymbol{\theta}, \tag{6}$$

where a centering parameter κ_i is determined for some i with respect to (m,s_j,t). The design matrix $\boldsymbol{\mathcal{X}} = \{\boldsymbol{\mathcal{X}}_1(\kappa_1),\ldots,\boldsymbol{\mathcal{X}}_n(\kappa_n)\}^T$ is a set of $\boldsymbol{\mathcal{X}}_i(\kappa_i) = \{\boldsymbol{X}_m, (\boldsymbol{Y}^{-j}_{mt} - \boldsymbol{\kappa}_m)I(\boldsymbol{Y}^{-j}_{m(t-1)} = 0), (\boldsymbol{Y}^{-j}_{mt} - \boldsymbol{\kappa}_m)I(\boldsymbol{Y}^{-j}_{m(t-1)} = 1)\}^T$ with $\boldsymbol{Y}^{-j}_{mt} = \boldsymbol{Y}_{mt} \setminus \{Y_m(s_j,t)\}$, $\boldsymbol{\kappa}_m = \kappa_m \mathbf{1}_{N_s-1}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta},\boldsymbol{\eta}_0,\boldsymbol{\eta}_1)^T$. To ensure the model identifiability, we assume that the design matrix $\boldsymbol{\mathcal{X}}_i(\kappa_i)$ is of full rank. Let $\boldsymbol{\theta}^*$ be the true parameter. This equivalent transformation is often used for autoregressive models (Wang, 2012). Consequently, the pseudo log-likelihood of original binary spatio-temporal data $\{Y_m(s_j,t)\}$ is reformulated as the log-likelihood of longitudinal binary vectors $\{\mathcal{Y}_i\}$,

$$\boldsymbol{\mathcal{L}}_c(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{c,i}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\Big(\mathcal{Y}_i\boldsymbol{\mathcal{X}}_i(\kappa_i)^T\boldsymbol{\theta} - \log\big[1+\exp\{\boldsymbol{\mathcal{X}}_i(\kappa_i)^T\boldsymbol{\theta}\}\big]\Big), \tag{7}$$

which can be maximized by the standard framework of generalized linear model (GLM) estimation for a fixed κ_i.

We regulate (7) using the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996). The sparsity on $\boldsymbol{\theta}$ is needed not only because of a large number of regressors but also in order to reflect the reality that a specific covariate or location possibly has negligible effect. Specifically, we maximize the ℓ_1-penalized pseudo log-likelihood,

$$F_\lambda(\boldsymbol{\theta}) = \boldsymbol{\mathcal{L}}_c(\boldsymbol{\theta}) - \lambda\|\boldsymbol{\theta}\|_1, \tag{8}$$

where $\|\cdot\|_1$ is the ℓ_1-norm and $\lambda > 0$ is a tuning parameter for regularization. Other regularization approaches for sparsity can be employed, such as the adaptive LASSO (Zou, 2006), the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), and the maximum a posteriori (MAP) estimation.

4.2 The Bias-corrected Estimator and Inference

The consistency of LASSO estimators for GLM has been approved under appropriate regularity conditions (Van de Geer, 2008). However, they do not have a tractable limiting distribution to make statistical inference. Along the lines of Van de Geer et al. (2014) and Tang et al. (2016), we find a bias-corrected LASSO estimator which asymptotically behaves as a maximum pseudo-likelihood estimator under the assumption that the nonzero set of true parameters $\boldsymbol{\theta}^*$ is known in advance.

Let $\hat{\boldsymbol{\theta}}_\lambda = (\hat{\boldsymbol{\beta}}_\lambda, \hat{\boldsymbol{\eta}}_{0\lambda}, \hat{\boldsymbol{\eta}}_{1\lambda})^T$ be the regularized estimator at λ, that is $\hat{\boldsymbol{\theta}}_\lambda = \arg\max_{\boldsymbol{\theta}} F_\lambda(\boldsymbol{\theta})$. By the Karush-Kuhn-Tucker (KKT) optimality conditions (Kuhn and Tucker, 2014), i.e. the subgradient of the objective in (8) is 0, the regularized pseudo likelihood estimator satisfies

$$S_n(\hat{\boldsymbol{\theta}}_\lambda) - \lambda \hat{\boldsymbol{Z}} = 0, \tag{9}$$

where $S_n(\hat{\boldsymbol{\theta}}_\lambda) = \dfrac{1}{n}\sum_{i=1}^n \dot{\mathcal{L}}_{c,i}(\hat{\boldsymbol{\theta}}_\lambda)$ is a pseudo-score function of (7) and $\hat{\boldsymbol{Z}} = (\hat{Z}_1, \ldots, \hat{Z}_p)^T$ is a subdifferential satisfying $\hat{Z}_j = \text{sign}(\hat{\theta}_j)$ if $\hat{\theta}_j \neq 0$ and $\hat{Z}_j \in \{-1, 1\}$ otherwise, for $j = 1, \ldots, p$. The first-order Taylor expansion of $S_n(\hat{\boldsymbol{\theta}}_\lambda)$ in (9) leads to $S_n(\boldsymbol{\theta}^*) + \dot{S}_n(\boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*) - \lambda\hat{\boldsymbol{Z}} \approx 0$. Multiplying both sides by $\{\dot{S}_n(\boldsymbol{\theta}^*)\}^{-1}$ and reordering terms, we have

$$\hat{\boldsymbol{\theta}}_\lambda + \{-\dot{S}_n(\boldsymbol{\theta}^*)\}^{-1}\lambda\hat{\boldsymbol{Z}} - \boldsymbol{\theta}^* + \{\dot{S}_n(\boldsymbol{\theta}^*)\}^{-1}S_n(\boldsymbol{\theta}^*) \approx 0. \tag{10}$$

Combine the first two terms and define $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_\lambda + \{-\dot{S}_n(\boldsymbol{\theta}^*)\}^{-1}\lambda\hat{\boldsymbol{Z}}$. From (10), we have $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \approx \{-\dot{S}_n(\boldsymbol{\theta}^*)\}^{-1}S_n(\boldsymbol{\theta}^*)$, a property also satisfied by the maximum pseudo-likelihood estimator asymptotically. This motivates us to use $\tilde{\boldsymbol{\theta}}$ as a bias-corrected estimator. In practice, $\boldsymbol{\theta}^*$ is unknown, and $-\dot{S}_n(\boldsymbol{\theta}^*)$ is estimated by an observed Hessian matrix, $\hat{\boldsymbol{H}} = -\dot{S}_n(\hat{\boldsymbol{\theta}}_\lambda)$. Since $\lambda\hat{\boldsymbol{Z}} = S_n(\hat{\boldsymbol{\theta}}_\lambda)$ from (9), the *bias-corrected LASSO estimator* is therefore

$$\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + \hat{\boldsymbol{H}}^{-1}S_n(\hat{\boldsymbol{\theta}}_\lambda). \tag{11}$$

For the log-likelihood in (7), $S_n(\hat{\boldsymbol{\theta}}_\lambda) = \dfrac{1}{n}\boldsymbol{\mathcal{X}}^T(\boldsymbol{\mathcal{Y}} - \hat{\boldsymbol{\pi}}_\lambda)$ and $\hat{\boldsymbol{H}} = \dfrac{1}{n}\boldsymbol{\mathcal{X}}^T\hat{\boldsymbol{\mathcal{V}}}_\lambda\boldsymbol{\mathcal{X}}$, where $\hat{\boldsymbol{\pi}}_\lambda = (\hat{\pi}_{1\lambda}, \ldots, \hat{\pi}_{n\lambda})^T$ and $\hat{\boldsymbol{\mathcal{V}}}_\lambda = \text{diag}\{\hat{\pi}_{1\lambda}(1-\hat{\pi}_{1\lambda}), \ldots, \hat{\pi}_{n\lambda}(1-\hat{\pi}_{n\lambda})\}$ with $\hat{\pi}_{i\lambda} = \text{logit}^{-1}\{\boldsymbol{\mathcal{X}}_i(\kappa_i)^T\hat{\boldsymbol{\theta}}_\lambda\}$.

We introduce notations for the asymptotic framework. Recall $n = \sum_{m=1}^M \sum_{t=1}^{T(m)} \sum_{j=1}^{N_s} I(s_j \in \mathcal{P}^0_{mt})$ and $p = N_s(N_s - 1) + p_x$, where p_x is the number of other covariates. We let the number of subjects M and the number of locations N_s go to infinity while fixing $T(m)$ and assuming $p < n$. Given two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \asymp b_n$ means $-\infty < \liminf(a_n/b_n) \leqslant \limsup(a_n/b_n) < \infty$; and $a_n = \mathcal{O}_p(b_n)$ means $0 < \liminf(a_n/b_n) \leqslant \limsup(a_n/b_n) < \infty$. Denote $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$ as the ℓ_1, ℓ_2 and the maximum norm of a vector or a matrix, respectively. We will make use of the following regularity conditions.

(C1) Suppose that $\|\boldsymbol{\mathcal{X}}_i\|_\infty = \mathcal{O}_p(1)$, for $i = 1, \ldots, n$. Also assume $\Lambda_{\min}(\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}}/n) = \mathcal{O}(1)$ and

$\Lambda_{\min}(\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}}/n) = \mathcal{O}(1)$, where $\Lambda_{\min}(\mathcal{M})$ and $\Lambda_{\max}(\mathcal{M})$ denote the minimum and maximum eigen values of a matrix $\mathcal{M}$, respectively.

(C2) There exists $\delta > 0$ such that in a neighborhood around a true value $\boldsymbol{\theta}^*$, denoted as $N_\delta(\boldsymbol{\theta}^*)$, it holds for some constant $0 < \epsilon_0 < 1$, $\epsilon_0 < \text{logit}^{-1}(\boldsymbol{\mathcal{X}}_i^T\boldsymbol{\theta}) < 1-\epsilon_0$, for $\forall\boldsymbol{\theta} \in N_\delta(\boldsymbol{\theta}^*)$.

(C3) It holds $s^* = o(\sqrt{n/(p\log p)})$ and $\lambda \asymp \sqrt{\log p/n}$, where $s^* \ll p$ is the number of true signals.

We establish the consistency and asymptotic normality of the bias-corrected estimator in the following theorem.

THEOREM 1: *Suppose the conditions (C1)-(C3) hold, the bias-corrected estimator* $\tilde{\boldsymbol{\theta}}$ *defined in* (11) *is consistent. For a fixed* r*, let* $\mathcal{A}_r = \{\boldsymbol{A} \in \mathbb{R}^{r\times p} : 0 < \Lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^T) \leqslant \Lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^T) < \infty\}$*. For any* $\boldsymbol{A} \in \mathcal{A}_r$*, we have the following asymptotic normality result*

$$n^{1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\xrightarrow{d}\mathcal{N}_r(\boldsymbol{0}, \boldsymbol{I}_r)$$

where $\boldsymbol{\Sigma} = \boldsymbol{A}\{\boldsymbol{H}^*\}^{-1}\boldsymbol{J}^*\{\boldsymbol{H}^*\}^{-1}\boldsymbol{A}^T$, $\boldsymbol{H}^* = E\{-\ddot{\boldsymbol{\mathcal{L}}}_c(\boldsymbol{\theta}^*)\}$ *and* $\boldsymbol{J}^* = var\{\dot{\boldsymbol{\mathcal{L}}}_c(\boldsymbol{\theta}^*)\}$.

The proof is provided in Appendix B. This asymptotic normality result enables us to establish statistical inferences for the biased corrected estimator, such as hypothesis tests or confidence interval constructions. Theorem 1 is general because the result is applicable to cases beyond the regularized pseudo-likelihood considered in this paper as long as $p < n$. For other ℓ_1-norm regularized composite likelihood estimators that are built upon a weighted product of a collection of component likelihoods such as low dimensional conditional or marginal densities (Varin et al., 2011), the asymptotic results still hold under appropriate regularity conditions.

4.3 The Iterative Algorithm for Estimation

The common algorithms for a logistic regression such as Newton's method cannot apply because κ-parameter in (3) is a nonlinear function of $\boldsymbol{\beta}$. Instead, we first estimate the parameters which are linear, with the bias-correction, and then update the nonlinear portion κ_m's (or κ_i's equivalently) iteratively until all converges, as follows.

1. Fit the independence model, the model (3) with $\forall \boldsymbol{\eta} = 0$, and set as $\hat{\boldsymbol{\beta}}^{(0)}$;
2. At the l-th iteration, for a fixed $\hat{\kappa}_i^{(l-1)} = \text{logit}^{-1}\{\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}^{(l-1)}\}$ and given λ (see below), fit the penalized logistic model (6) by maximizing (8), and update estimates $\hat{\boldsymbol{\theta}}^{(l)}_{\lambda} = \{\hat{\boldsymbol{\beta}}^{(l)}_{\lambda}, \hat{\boldsymbol{\eta}}^{(l)}_{0\lambda}, \hat{\boldsymbol{\eta}}^{(l)}_{1\lambda}\}^T$;
3. Calculate the bias-corrected estimates by (11), $\tilde{\boldsymbol{\theta}}^{(l)} = \hat{\boldsymbol{\theta}}^{(l)}_{\lambda} + \{\hat{\boldsymbol{H}}(\hat{\boldsymbol{\theta}}^{(l)}_{\lambda})\}^{-1} S_n(\hat{\boldsymbol{\theta}}^{(l)}_{\lambda})$;
4. Update the centering parameters $\hat{\kappa}_i^{(l)} = \text{logit}^{-1}\{\boldsymbol{X}_i^T \tilde{\boldsymbol{\beta}}^{(l)}\}$;
5. Return to step 2 until all converges.

In step 1 and step 2, the standard logistic regression (Hastie and Pregibon, 1992) and the GLM with regularization (Friedman et al., 2010) are used, respectively. For example, `glm` and `glmnet` in R software can apply.

The algorithm involves the selection of a tuning parameter λ which controls the sparsity of network connectivities. In practice, an optimal λ can be determined via some data-dependent model selection criteria, such as generalized cross-validation (GCV) (Golub et al., 1979), Bayesian information criterion (BIC) (Schwarz et al., 1978) and extended Bayesian information criterion (EBIC) (Chen and Chen, 2008). Alternatively, one can manually choose λ that meets a desired degree of sparsity from certain domain knowledge. We suggest to use any technique at the first iteration, $l = 1$, and fix it for the remaining to save computation.

5 Simulation Studies

Simulation studies were conducted to investigate, first, how well the proposed model estimation and inference work, and second, how better the proposed prediction via transition probabilities performs than a simple Markov model.

We set $N_s = 8$ locations, $(s_1, s_2, \ldots, s_8)$, and the dimension of each $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ is $8 \times (8-1) = 56$. We assigned different values to $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ that are symmetric and moderately sparse, as illustrated in Figure 1. For simplicity, only an intercept was included in the independent model; $\boldsymbol{X}_m = 1$ with a coefficient $\boldsymbol{\beta} = -2$. This leaded $\kappa_m \approx 0.12$ for all $m = 1, \ldots, M$, and 0 and 1 of autocovariates are transformed to -0.12 and 0.88, respectively. In other words, we let the influence of diseased status be stronger than normal status.

[Figure 1 about here.]

Recall that $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1)^T$ determines one-time transition probability from a joint outcome to another by the formula (5). Since there are 2^8 joint outcomes with $N_s = 8$, it is not feasible to present all transition probabilities regarding our $\boldsymbol{\theta}$. Alternatively, we illustrate an example to provide better insights on relationships between $\boldsymbol{\theta}$ and $\mathrm{P}(\boldsymbol{Y}_{mt}|\boldsymbol{Y}_{m(t-1)}; \boldsymbol{\theta})$. Suppose $\boldsymbol{Y}_{m0} = (1, 0, 0, 0, 1, 0, 1, 1)$, then the most probable next outcome is $\boldsymbol{Y}_{m1} = (1, 0, 0, 0, 1, 1, 1, 1)$ with probability 0.21; that is, s_6 is the most likely to be newly diseased. This can be partly explained by $\eta_{156} (= \eta_{165}) = 2.15$, a strong positive contribution to switch the status of s_6 from 0 to 1 when $Y_m(s_5, 0) = Y_m(s_5, 1) = 1$. Likewise, $\boldsymbol{Y}_{m1} = (1, 1, 0, 0, 1, 0, 1, 1)$ has the second highest probability of 0.18 with s_2 being newly diseased next, by $\eta_{112} (= \eta_{121}) = 1.69$.

The initial status at 8 locations, $\boldsymbol{Y}_{m0}$, was generated from a Bernoulli distribution with a probability 0.25; $Y(s_j, 0) \sim \text{Bernoulli}(0.25)$, for $j \in \{1, \ldots, 8\}$. The next status was generated from the true transition probabilities, $\mathrm{P}(\boldsymbol{Y}_{mt}|\boldsymbol{Y}_{m(t-1)}; \boldsymbol{\theta}^*)$, and we repeated this until all is diseased. Such sample sequences were independently generated for $M = 500$ subjects to fit the model. The iterative algorithm described in Section 4.3 was applied to estimate model parameters. The sparsity parameter was tuned at the first iteration by the method of cross-validation via function in R (`cv.glmnet`) and withheld for the subsequent iterations to speed up algorithmic convergence where high-quality initial values are always desirable. For non-negativity constraint, the minimum value of η is set to be zero using the option (`lower.limits`). Most runs have converged in 10 or less iterations. We ran $B = 100$ rounds of simulations.

Figure 2 illustrates the simulation results with the means and standard deviations of the estimated parameters and the means of the estimated asymptotic standard deviations, denoted by $\hat{\boldsymbol{\eta}}$, $\mathsf{SD}_{\hat{\boldsymbol{\eta}}}$ and $\overline{\mathsf{SE}}_{\hat{\boldsymbol{\eta}}}$, respectively. In the panels (a) and (b), the point estimates verified that our estimation procedure is, in general, able to recover the network structures and discriminates autoregressive effects among different pairs of locations. For example, from the true parameters we set in Figure 1, the impact of s_7 on s_8 (or s_8 on s_7) when the previous state was 0 is stronger than that of s_5 on s_7 (or s_7 on s_5), that is, $\eta_{078}(=\eta_{087}) = 2.98 > \eta_{057}(= \eta_{075}) = 1.21$. These impacts were estimated as $\hat{\eta}_{078}(=\hat{\eta}_{087}) = 2.99 > \hat{\eta}_{057}(=\hat{\eta}_{075}) = 1.18$, which are very close to the true values. Moreover, the panels (c) and (d) demonstrated the asymptotic distribution derived by Theorem 1. The good correspondence between the empirical and estimated variances convinced that the covariance matrix in Theorem 1 is a proper estimator for asymptotic variances of parameters in the proposed model.

[Figure 2 about here.]

We also compared the autologistic network model (3) with a simple Markov model,

$$\text{logit}\{p_m(s_j,t)\} = \boldsymbol{X}_m^T\boldsymbol{\beta} + \sum_{k\neq j}\eta_{jk}\{Y_m(s_k,t-1)-\kappa_m\} \tag{12}$$

$$\text{subject to } \eta_{jk} = \eta_{kj} \text{ for all } j \neq k,$$

which has both centered autocovariates and symmetric η-parameters. For each model, we estimated parameters and transition probabilities, and computed individual root-mean-square errors at every active transition, defined as

$$\text{RMSE}^2_{mt} = \frac{1}{B}\sum_{b=1}^{B}\{\text{P}(\boldsymbol{Y}_{mt}|\boldsymbol{Y}_{m(t-1)};\hat{\boldsymbol{\theta}}_{(b)}) - \text{P}(\boldsymbol{Y}_{mt}|\boldsymbol{Y}_{m(t-1)};\boldsymbol{\theta}^*)\}^2$$

where $\hat{\boldsymbol{\theta}}_{(b)}$ denotes the estimated model parameters at the b-th simulation run. Since it is quite infeasible to show the RMSEs at all active transitions, we instead showed their summaries in Table 2. Note that the number of active transitions is $\sum_{k=0}^{N_s} 2^k \binom{N_s}{N_s-k}$, which is 6561 for

$N_s = 8$. From Table 2, the proposed model (3) estimated the transition probabilities more precisely than the simple Markov model (12).

[Table 2 about here.]

Despite of the overall reasonable estimation performance, a theoretically guaranteed recovery of the true sparsity by the LASSO (i.e. variable selection consistency) requires *the irrepresentable condition* (Zhao and Yu, 2006); relevant variables (signal with non-zero η) are not strongly correlated with the irrelevant variables (noise with zero η). This condition is generally considered too stringent to hold in practice. In our application, we expect some dependence in the design matrix consisting of autocovariates; for example, the empirical correlation between the two columns corresponding η_{167} and η_{168} is about 0.4 while their true values are $\eta_{167} = 0$ and $\eta_{168} = 0.68$. However, we focused parameter estimation consistency shown in Theorem 1 under the restricted eigen condition (C1) and Donoho and Johnstone (1994)'s hard threshold rate $\sqrt{\log p/n}$ in condition (C3). The simulation numerical results also demonstrate that our estimator is in general a good approximation of true parameter values.

6 Application to ALS Patients Data

Data used in this research came from the EMPOWER study, a double-blind, placebo-controlled phase III clinical trial on dexpramipexole in patients with ALS (Cudkowicz et al., 2013). Participants were 18 to 80 years old, with first symptom onset 24 months or less at study entry and an upright slow vital capacity of at least 65% of the predicted value for age, height, and sex at screening. A total of 942 patients were enrolled from 81 academic medical centers in 11 countries. Sixteen muscles (eight bilaterally) were tested at study entry and every two months thereafter for up to 12 months. As shown in Figure 3, the sixteen muscles are: the left and right of shoulder flexion, elbow flexion, hip flexion, knee flexion, elbow extension, knee entension, wrist extension and ankle dorsiflexion.

[Figure 3 about here.]

For the ALS disease, the association in muscle strength between different muscles is not merely determined by the spatial proximity of muscles at different body locations, it can also be affected by the proximity of nerves controlling muscles in the spinal cord. For example, when a patient's right wrist muscle loses strength, the left wrist muscle, although far away from the right one, can be affected before the right elbow, which is physically closer to the right wrist. Absorbing features also need to be considered, because once a muscle becomes diseased it can never recover. Moreover, in the spread of muscle weakness by ALS, a newly diseased muscle may have different effects on a muscle compared to the others that are diseased earlier. Our model is thus suitable for the ALS disease spreading pattern study.

In this study, the raw muscle strength data were dichotomized using the regression equations in National Isometric Muscle Strength Database Consortium (1996) and Bohannon (1997), which established the predictive strength of each muscle for healthy people based on their gender, age, height, and weight. The predicted strengths were used as a benchmark to determine whether muscles were diseased or not. Specifically, a muscle was declared as impaired ($= 1$) if its measured strength is 40% less than the predicted strength, or healthy ($= 0$) otherwise. Once a muscle was declared as impaired at a time, it would remain so from that time point on. We fitted the model (3) to these data with independent variables $\boldsymbol{X} = \{1, t, \text{symptom onset site}, \text{symptom duration}\}$ and estimated the model parameters using the regularized pseudo-likelihood in (8). The tuning parameter, λ, was chosen by 10-fold cross validation. In the iterative algorithm for parameter estimation given in Section 4.3, we stopped the iteration when every updated estimate falls within 1% difference from the old estimate.

Figure 4 shows the heat maps of estimated $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$, respectively. The horizontal connections between the right and left side of the same muscle were mostly stronger than other connections for both previously healthy and diseased patients. This implies that a muscle is likely to remain at the same status as its opposite side, no matter what status a muscle was at the previous visit. Also, the estimates of $\boldsymbol{\eta}_1$ were sparser than those of $\boldsymbol{\eta}_0$, under the same degree of regularization (at the same value of λ). This implies that newly impaired muscles have different impacts on others, even vertically between upper and lower body muscles, while muscles impaired far in the past were mostly associated only with physically neighboring muscles or their counterparts. Moreover, the fairly strong connection was observed between elbow and knee in $\boldsymbol{\eta}_1$, which can be a clue to a biological link of spreading path between upper and lower body locations.

[Figure 4 about here.]

By computing the transition probabilities from (5), we could make prediction on one-time ahead disease progression. One of clinical interests is to single out which muscles have the most likelihood of being impaired before long. We could also track the most susceptible muscles continuously and sequentially until all muscles are impaired. This resulting pathway of muscle impairment provides a simple yet informative prediction of disease spreading. A more rigorous way is perhaps to make inference on a probable path according to a transition probability matrix; however, the dimension of this path space is too large for practical use.

Suppose a male patient visits a clinic for the first time when only one muscle is impaired by ALS, say 21.6 months ago, and his symptom is not bulbar onset. Figure 5 illustrates two examples of probable disease progression paths for this hypothetical patient. In Figure 5 (a), the left wrist extension got impaired first, spread to the right wrist extension, and then followed by the knee muscles. This progression path is in conjunction with the implication of the estimated parameters in Figure 4; spreading directions occurred between the left and right sides. Also, the knee muscles, which are highly linked to lower body muscles, were likely to get impaired first among lower body sites, and so were the elbow muscles among upper body sites. The other spreading path, with the left ankle flexor initially impaired, exhibited quite similar progression in Figure 5 (b); the transitions between muscles of right and left sides and between elbow and knee muscles were remarkable.

[Figure 5 about here.]

Table 3 summarizes the bias-corrected estimates of regression coefficients $\boldsymbol{\beta}$; the confidence intervals and p-values were based on Theorem 1. The coefficients $\hat{\boldsymbol{\beta}}$ can be better interpreted in terms of the estimated centering parameter $\hat{\kappa} = \text{logit}^{-1}(\boldsymbol{X}^T\hat{\boldsymbol{\beta}})$. For example, the estimated overall probability of disease progression with no contributions from independent covariates or autocovariates is $\hat{\kappa} = \text{logit}^{-1}(3.57) \approx 0.03$, which is reasonably low. In that manner, this probability would decrease over time because $\hat{\beta} = -0.05$ for the visiting time is negative with $p < .0001$. In other words, individual muscles would be likely to stay healthy if there were no inter-muscle spatial dependency and other risk factors. Also, the negative effect of the symptom duration ($\hat{\beta} = -0.01$ with $p = 0.0071$) suggests that a patient having a longer symptom duration tends to have a lower probability of progression, in contrast, a patient who had a most recently onset tend to have higher probability of progression.

[Table 3 about here.]

7 Discussion

We proposed an autologistic network model for spatio-temporal binary data with absorbing state. The major contributions are: we relaxed the need of pre-specification on neighborhood structure; we considered absorbing state of binary processes by partitioning the inferrable active set and non-inferrable absorbing set; the model incorporated previously diseased and normal locations with their different profiles; the model can apply to other applications with a similar data structure, such as an epidemic, a pathogen, a virus, and so on. Furthermore, we established a valid joint distribution from the proposed conditional probability model and derived the transition probabilities for useful to characterize spreading patterns of a disease.

For the estimation, the LASSO-penalized pseudo-likelihood maximization was invoked to enforce sparsity on network associations. We proposed an efficient iterative algorithm for model implementation by converting the optimization into the ordinary penalized GLM problem. In addition, a bias-correction method was applied to obtain the asymptotic normality. Note that since the asymptotic properties are proved at a fixed level of sparsity, their validity would hold when the tuning parameter, which is chosen according to some data-driven criteria such as the cross validation, satisfies the condition given in (C3). This technical work is worth future exploration.

Our simulation study affirmed that the proposed estimation approach is valid for inference on model parameters. We also studied in simulation that the proposed model can estimate transition probabilities more precisely than a simple Markov model, which has neither simultaneous spatial dependency or different impacts depending on previous status. Meanwhile, the application to the ALS data demonstrates that our model offers further insights into the spreading mechanism of muscle weakness by ALS disease.

Future research could focus on ordered categorical data or mixed data of continuous and discrete measures, rather than dichotomized data, to retain more information. Moreover, instead of using ℓ_1-penalty, other regularization approaches could be employed; ℓ_0-penalty could be appealing as it does not lead to estimation bias. It would be of interest to consider methods that combine both dimensionality reduction and sparsity. Lastly, three-way association, rather than two-way, would be worthwhile to consider, which is useful in brain imaging data analysis.

Acknowledgements

We gratefully acknowledge that this research is financially supported by Biogen. Also, Yei Eun's work was done while she was a PhD student at the Department of Statistics in Texas A&M University as a part of dissertation. Huiyan Sang's research was partly supported by *NSF DMS-1622433* and *NSF DMS-1737885*.

References

Agaskar, A. and Lu, Y. M. (2013). Alarm: A logistic auto-regressive model for binary processes on networks. In *Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE*, pages 305–308. IEEE.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 192–236.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician* pages 179–195.

Bohannon, R. W. (1997). Reference values for extremity muscle strength obtained by hand-held dynamometry from adults aged 20 to 79 years. *Archives of Physical Medicine and Rehabilitation* **78,** 26–32.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media.

Caragea, P. C. and Kaiser, M. S. (2009). Autologistic models with interpretable parameters. *Journal of Agricultural, Biological, and Environmental Statistics* **14,** 281–300.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika* **95,** 759–771.

Cudkowicz, M. E., van den Berg, L. H., Shefner, J. M., Mitsumoto, H., Mora, J. S., Ludolph, A., Hardiman, O., Bozik, M. E., Ingersoll, E. W., Archibald, D., et al. (2013). Dexpramipexole versus placebo for patients with amyotrophic lateral sclerosis

(EMPOWER): a randomised, double-blind, phase 3 trial. *The Lancet Neurology* **12,** 1059–1067.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika* **81,** 425–455.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** 1348–1360.

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

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics* **31,** 1208–1211.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21,** 215–223.

Hastie, T. J. and Pregibon, D. (1992). Statistical models in s, chapter generalized linear models. *Wadsworth & Brooks/Cole* **51,**.

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

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75,** 139–159.

Hughes, J., Haran, M., and Caragea, P. C. (2011). Autologistic models for binary data on a lattice. *Environmetrics* **22,** 857–871.

Kaiser, M. S. and Cressie, N. (2000). The construction of multivariate distributions from markov random fields. *Journal of Multivariate Analysis* **73,** 199–220.

Kaiser, M. S., Pazdernik, K. T., Lock, A. B., and Nutter, F. W. (2014). Modeling the spread of plant disease using a sequence of binary random fields with absorbing states. *Spatial*

Statistics **9,** 38–50.

Kuhn, H. W. and Tucker, A. W. (2014). Nonlinear programming. In *Traces and emergence of nonlinear programming*, pages 247–258. Springer.

National Isometric Muscle Strength Database Consortium, N. (1996). Muscular weakness assessment: use of normal isometric strength data. *Archives of Physical Medicine and Rehabilitation* **77,** 1251–1255.

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

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

Song, P. X.-K. (2007). *Correlated data analysis: modeling, analytics, and applications.* Springer Science & Business Media.

Tang, L., Zhou, L., and Song, P. X.-K. (2016). Method of divide-and-combine in regularised generalised linear models for big data. *arXiv preprint arXiv:1611.06208* .

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 267–288.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics* **42,** 1166–1202.

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics* pages 614–645.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* pages 5–42.

Wang, Z. (2012). *Analysis of Binary Data via Spatial-Temporal Autologistic Regression Models*. University of Kentucky.

Xue, L., Zou, H., Cai, T., et al. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *The Annals of Statistics* **40,** 1403–1429.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research* **7,** 2541–2563.

Zhu, J., Huang, H.-C., and Wu, J. (2005). Modeling spatial-temporal binary data using markov random fields. *Journal of Agricultural, Biological, and Environmental Statistics* **10,** 212–225.

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

Received October 2007. *Revised February* 2008. *Accepted March* 2008.

Appendix

(A): Derivation of Joint Distribution in Section 3

We build a negpotenial function Q following Besag (1974) and Kaiser and Cressie (2000) to derive a valid joint distribution from conditionals. For a fixed subject m and time t whose previous state is zero (i.e. $Y_m(s_j,t-1)=0$ and so $Y_m(s_j,t)\in\mathcal{P}^0_{mt}$), the conditional density of a response $Y_m(s_j,t)$ at y is

$$f_j\{y|\boldsymbol{X}_m, Y_m(s_k,t-1), Y_m(s_k,t) \text{ for } \forall k\neq j;\boldsymbol{\theta}\} = p_m(s_j,t)^y\{1-p_m(s_j,t)\}^{1-y}.$$

From the model specification in (3), we have its log-conditional density as

$$\begin{aligned}&\log f_j\{Y_m(s_j,t)|\boldsymbol{X}_m, Y_m(s_k,t-1), Y_m(s_k,t) \text{ for } \forall k\neq j;\boldsymbol{\theta}\}\\ &\quad= Y_m(s_j,t)\Big[\boldsymbol{X}_m^T\boldsymbol{\beta}+\sum_{k\in\mathcal{P}^0_{mt}\backslash\{j\}}\eta_{0jk}\{Y_m(s_k,t)-\kappa_m\}+\sum_{k\in\mathcal{P}^1_{mt}\backslash\{j\}}\eta_{1jk}\{Y_m(s_k,t)-\kappa_m\}\Big]\\ &\qquad-\log\left(1+\exp\Big[\boldsymbol{X}_m^T\boldsymbol{\beta}+\sum_{k\in\mathcal{P}^0_{mt}\backslash\{j\}}\eta_{0jk}\{Y_m(s_k,t)-\kappa_m\}+\sum_{k\in\mathcal{P}^1_{mt}\backslash\{j\}}\eta_{1jk}\{Y_m(s_k,t)-\kappa_m\}\Big]\right)\end{aligned}$$

for all $Y_m(s_j,t)$ in the active set $\mathcal{P}^0_{mt}$. As the above conditionals indicate only pairwise dependencies, the negpotential function of all responses in the active set has only the first and second order of cliques, so it has the following permutation invariance form,

$$Q(\boldsymbol{Y}_{mt}|\boldsymbol{\theta}) = \sum_{j:\ Y_m(s_j,t)\in\mathcal{A}_{mt}} H_j\{Y_m(s_j,t)\} + \sum_{\substack{j:\, Y_m(s_j,t)\in\mathcal{A}_{mt}\\ k:\, j<k\leqslant N_s}} H_{j,k}\{Y_m(s_j,t), Y_m(s_k,t)\}.$$

To derive $H_j\{Y_m(s_j,t)\}$ and $H_{j,k}\{Y_m(s_j,t), Y_m(s_k,t)\}$, we follow Besag (1974) and define

$$H_j\{Y_m(s_j,t)\} = \log\frac{f_j\{Y_m(s_j,t)|Y_m^*(s_{-j},t)\}}{f_j\{Y_m^*(s_j,t)|Y_m^*(s_{-j},t)\}}$$

$$H_{j,k}\{Y_m(s_j,t), Y_m(s_k,t)\} = \log\frac{f_j\{Y_m(s_j,t)|Y_m(s_k,t), Y_m^*(s_{-j,-k},t)\}f_j\{Y_m^*(s_j,t)|Y_m^*(s_{-j},t)\}}{f_j\{Y_m^*(s_j,t)|Y_m(s_k,t), Y_m^*(s_{-j,-k},t)\}f_j\{Y_m(s_j,t)|Y_m^*(s_{-j},t)\}}$$

Choosing $Y_m^*(s_j,t)=0$ for each j in active set $\mathcal{P}_{mt}^0$, we obtain

$$H_j\{Y_m(s_j,t)\} = Y_m(s_j,t)\Big\{\boldsymbol{X}_m^T\boldsymbol{\beta} - \sum_{k\in\mathcal{P}_{mt}^0\backslash\{j\}}\eta_{0jk}\kappa_m - \sum_{k\in\mathcal{P}_{mt}^1\backslash\{j\}}\eta_{1jk}\kappa_m\Big\}$$

$$H_{j,k}\{Y_m(s_j,t), Y_m(s_k,t)\} = \sum_{k\in\mathcal{P}_{mt}^0\backslash\{j\}}\eta_{0jk}Y_m(s_j,t)Y_m(s_k,t) + \sum_{k\in\mathcal{P}_{mt}^1\backslash\{j\}}\eta_{1jk}Y_m(s_j,t)Y_m(s_k,t).$$

The negpotential function then takes the form in (4) and finally, the joint distribution of $\boldsymbol{Y}_{mt}$ in the support set $\mathcal{S}_{mt}$ given a complete set of conditional distributions, denoted by f, can be specified up to a normalizing constant by Theorem 3 in Kaiser and Cressie (2000).

(B): Proof of Theorem 1 in Section 4

Proof. We first introduce some notations to simplify mathematical expressions. For a function $\rho:\mathbb{X}\times\mathbb{Y}\to\mathbb{R}$, write $\mathcal{P}_n\rho = \sum_{i=1}^n \rho_i/n$, and $\mathcal{P}\rho = E(\mathcal{P}_n\rho)$. Also define a function $\rho(\alpha,y) = y\alpha - \log\{1+\exp(\alpha)\}$. For the binary logistic regression model, we can rewrite the pseudo loglikelihood function as $\ell_c(\boldsymbol{\theta}) = \mathcal{P}_n\rho(\boldsymbol{\mathcal{X}}_i\boldsymbol{\theta}, Y_i)$.

When condition (C2) holds, the second derivative of $\rho(\alpha,y)$ with respect to α is $\ddot{\rho}(\alpha,y) = \mathrm{logit}^{-1}(\alpha)\{1-\mathrm{logit}^{-1}(\alpha)\}$, which is positive and bounded away from zero. It indicates that $\rho(\alpha,y)$ behaves quadratically near $\alpha^* = \boldsymbol{\mathcal{X}}_i\boldsymbol{\theta}^*$ and hence the quadratic margin condition holds (see, e.g., Section 6.4 of Bühlmann and Van De Geer (2011)), i.e., $\mathcal{P}_n\{\rho(\boldsymbol{\mathcal{X}}_i\hat{\boldsymbol{\theta}}_\lambda, Y_i) - \rho(\boldsymbol{\mathcal{X}}_i\boldsymbol{\theta}^*, Y_i)\} \geqslant c\|\boldsymbol{\mathcal{X}}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)\|_2^2/n$ for some constant c.

Furthermore, the restricted eigenvalue condition (C1) implies that the compatibility condition required in Theorem 6.4 in Bühlmann and Van De Geer (2011) holds. Combining these two conditions together, the oracle inequality of the LASSO estimator can be established as $c\|\boldsymbol{\mathcal{X}}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)\|_2^2/n + \lambda\|\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*\|_1 \leqslant \mathcal{P}_n\{\rho(\boldsymbol{\mathcal{X}}_i\hat{\boldsymbol{\theta}}_\lambda, Y_i) - \rho(\boldsymbol{\mathcal{X}}_i\boldsymbol{\theta}^*, Y_i)\} + \lambda\|\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*\|_1 = \mathcal{O}(s_0\lambda^2)$, which provides asymptotic bounds for both the prediction error and the ℓ_1 error, i.e., $\|\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*\|_1 = \mathcal{O}_{\mathcal{P}}(s_0\lambda)$, $\|\boldsymbol{\mathcal{X}}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)\|_2^2/n = \mathcal{O}(s_0\lambda^2)$. Under condition (C3), $\hat{\boldsymbol{\theta}}_\lambda$ is a consistent estimator of $\boldsymbol{\theta}$.

Notice that $\hat{\boldsymbol{H}} = \frac{1}{n}\boldsymbol{\mathcal{X}}^T\mathrm{diag}\big[\hat{\pi}_1(\hat{\boldsymbol{\theta}}_\lambda)\{1 - \hat{\pi}_1(\hat{\boldsymbol{\theta}}_\lambda)\}, \ldots, \hat{\pi}_n(\hat{\boldsymbol{\theta}}_\lambda)\{1 - \hat{\pi}_n(\hat{\boldsymbol{\theta}}_\lambda\}\big]\boldsymbol{\mathcal{X}}$. By conditions (C1) and (C2), $\Lambda_{\min}\{\hat{\boldsymbol{H}}\} = \min_{\|u\|_2=1} u^T(\frac{1}{n}\boldsymbol{\mathcal{X}}^T\mathrm{diag}\big[\hat{\pi}_1(\hat{\boldsymbol{\theta}})\{1 - \hat{\pi}_1(\hat{\boldsymbol{\theta}})\}, \ldots, \hat{\pi}_n(\hat{\boldsymbol{\theta}})\{1 - \hat{\pi}_n(\hat{\boldsymbol{\theta}}\}\big]\boldsymbol{\mathcal{X}})u = \mathcal{O}\{\min_{\|u\|_2=1} u^T(\frac{1}{n}\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}})u\} = \mathcal{O}(\Lambda_{\min}\{\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}}/n)\}$. Similarly, $\Lambda_{\max}\{\hat{\boldsymbol{H}}\} = \mathcal{O}\{\Lambda_{\max}(\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{X}}/n)\}$, which indicates $\hat{\boldsymbol{H}}$ is strictly positive definite. Consider the inverse matrix of $\hat{\boldsymbol{H}}$, and define it as $\hat{\Theta}\hat{\boldsymbol{H}} = I$. Recall $\Lambda_{\max}(\Theta) = 1/\Lambda_{\min}(\hat{\boldsymbol{H}})$, and $\Lambda_{\min}(\Theta) = 1/\Lambda_{\max}(\boldsymbol{H})$, which suggests that $\hat{\Theta}$ is also strictly positive definite with bounded eigenvalues and hence $\|\hat{\Theta}\|_2 = \Lambda_{\max}(\hat{\Theta}) = \mathcal{O}(1)$. Combing this fact with $S_n(\hat{\boldsymbol{\theta}}_\lambda) = \lambda\hat{\kappa}(\hat{\boldsymbol{\theta}}_\lambda) \to 0$ and $\hat{\boldsymbol{\theta}}_\lambda \to \boldsymbol{\theta}^*$, we prove the consistency of $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_\lambda + \hat{\Theta}S_n(\hat{\boldsymbol{\theta}}_\lambda)$, i.e. $\tilde{\boldsymbol{\theta}} \to \boldsymbol{\theta}^{\boldsymbol{*}}$ as $n \to \infty$.

We next show the asymptotic normality of the biased-corrected estimator. When condition (C2) holds, the third derivative of $\rho(\alpha, y)$ with respect to α exists and its absolute value is bounded by 1, which ensures that the second derivative of $\rho(\alpha, y)$ with respect to α is locally Lipschitz with a universal constant.

From the Taylor expansion of $\dot{\rho}(\boldsymbol{\mathcal{X}}_i\boldsymbol{\theta}, y)$ and the Lipschitz conditions on $\ddot{\rho}(\boldsymbol{\mathcal{X}}_i\boldsymbol{\theta}, y)$ for $\forall\theta \in N_\delta(\boldsymbol{\theta}^*)$, we have $\dot{\rho}(\boldsymbol{\mathcal{X}}_i^T\hat{\boldsymbol{\theta}}_\lambda, Y_i) = \dot{\rho}(\boldsymbol{\mathcal{X}}_i^T\boldsymbol{\theta}^*, Y_i) + \ddot{\rho}(\boldsymbol{\mathcal{X}}_i^T\hat{\boldsymbol{\theta}}_\lambda, Y_i)\boldsymbol{\mathcal{X}}_i^T(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*) + \mathcal{O}(|\boldsymbol{\mathcal{X}}_i^T(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)|^2)$. Therefore,

$$\begin{aligned}
&\hat{\boldsymbol{\theta}}_\lambda + \hat{\Theta}S_n(\hat{\boldsymbol{\theta}}_\lambda) - \boldsymbol{\theta}^* \\
=&\hat{\boldsymbol{\theta}}_\lambda + \hat{\Theta}\mathcal{P}_n\{\dot{\rho}(\boldsymbol{\mathcal{X}}_i^T\hat{\boldsymbol{\theta}}_\lambda, Y_i)\boldsymbol{\mathcal{X}}_i\} - \boldsymbol{\theta}^* \\
=&\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* + \hat{\Theta}\mathcal{P}_n\{\dot{\rho}(\boldsymbol{\mathcal{X}}_i^T\boldsymbol{\theta}^*, Y_i)\boldsymbol{\mathcal{X}}_i + \boldsymbol{\mathcal{X}}_i\ddot{\rho}(\boldsymbol{\mathcal{X}}_i^T\hat{\boldsymbol{\theta}}_\lambda, Y_i)\boldsymbol{\mathcal{X}}_i^T(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*) + \boldsymbol{\mathcal{X}}_i\mathcal{O}(|\boldsymbol{\mathcal{X}}_i^T(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)|^2)\} \\
=&\hat{\Theta}\mathcal{P}_n\{\dot{\rho}(\boldsymbol{\mathcal{X}}_i^T\boldsymbol{\theta}^*, Y_i)\boldsymbol{\mathcal{X}}_i + \boldsymbol{\mathcal{X}}_i\mathcal{O}(|\boldsymbol{\mathcal{X}}_i^T(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)|^2)\} + [\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* + \hat{\Theta}\mathcal{P}_n\{\boldsymbol{\mathcal{X}}_i\ddot{\rho}(\boldsymbol{\mathcal{X}}_i^T\hat{\boldsymbol{\theta}}_\lambda, Y_i)\boldsymbol{\mathcal{X}}_i^T\}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)] \\
=&\hat{\Theta}\mathcal{P}_n\{\dot{\rho}(\boldsymbol{\mathcal{X}}_i^T\boldsymbol{\theta}^*, Y_i)\boldsymbol{\mathcal{X}}_i + \boldsymbol{\mathcal{X}}_i\mathcal{O}(|\boldsymbol{\mathcal{X}}_i^T(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)|^2)\} + [\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* + \hat{\Theta}\hat{\boldsymbol{H}}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)]
\end{aligned}$$

$$
\begin{aligned}
&=\hat{\boldsymbol{\Theta}}\mathcal{P}_n\{\dot{\rho}(\boldsymbol{\mathcal{X}}_i^T\boldsymbol{\theta}^*, Y_i)\boldsymbol{\mathcal{X}}_i + \boldsymbol{\mathcal{X}}_i\mathcal{O}(|\boldsymbol{\mathcal{X}}_i^T(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)|^2)\} + [\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* + \hat{\boldsymbol{\Theta}}\hat{\boldsymbol{H}}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)] \\
&= \underbrace{\boldsymbol{\Theta}^*\mathcal{P}_n\{\dot{\rho}(\boldsymbol{\mathcal{X}}_i^T\boldsymbol{\theta}^*, Y_i)\boldsymbol{\mathcal{X}}_i\}}_{\boldsymbol{T}_1} + \underbrace{\hat{\boldsymbol{\Theta}}\mathcal{P}_n\{\boldsymbol{\mathcal{X}}_i\mathcal{O}(|\boldsymbol{\mathcal{X}}_i^T(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)|^2)\}}_{\boldsymbol{T}_2} + \underbrace{(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*)\mathcal{P}_n\{\dot{\rho}(\boldsymbol{\mathcal{X}}_i^T\boldsymbol{\theta}^*, Y_i)\boldsymbol{\mathcal{X}}_i\}}_{\boldsymbol{T}_3}
\end{aligned}
\tag{A.1}
$$

When conditions (C1)-(C3) hold, by Hölder's inequality, the second term in (A.2) is

$$
\begin{aligned}
\|\boldsymbol{T}_2\|_\infty &= \|\hat{\boldsymbol{\Theta}}\mathcal{P}_n\{\boldsymbol{\mathcal{X}}_i\mathcal{O}(|\boldsymbol{\mathcal{X}}_i^T(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)|^2)\}\|_\infty \leqslant \mathcal{O}(\mathcal{P}_n\{\|\hat{\boldsymbol{\Theta}}\boldsymbol{\mathcal{X}}_i\|_\infty|\boldsymbol{\mathcal{X}}_i^T(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)|^2\}) \\
&\leqslant \mathcal{O}(\mathcal{P}_n\{\|\hat{\boldsymbol{\Theta}}\|_1\|\boldsymbol{\mathcal{X}}_i\|_\infty|\boldsymbol{\mathcal{X}}_i^T(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)|^2\}) \\
&\leqslant \mathcal{O}(\mathcal{P}_n\{\sqrt{p}\|\hat{\boldsymbol{\Theta}}\|_2\|\boldsymbol{\mathcal{X}}_i\|_\infty|\boldsymbol{\mathcal{X}}_i^T(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)|^2\}) \\
&= \mathcal{O}(\sqrt{p}\|\boldsymbol{\mathcal{X}}^T(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)\|_2^2/n) = \mathcal{O}(\sqrt{p}s^*\lambda^2) = o_p(1/\sqrt{n})
\end{aligned}
$$

Notice that we consider the case with $p < n$, and hence $\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_1 = \mathcal{O}(1/\sqrt{n})$. By Hölder's inequality, the third term in (A.2) is as follows

$$
\begin{aligned}
\|\boldsymbol{T}_3\|_\infty &\leqslant \|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_1\|\mathcal{P}_n\{\dot{\rho}(\boldsymbol{\mathcal{X}}_i^T\boldsymbol{\theta}^*, Y_i)\boldsymbol{\mathcal{X}}_i\}\|_\infty = \|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_1\|\{Y_i - \text{logit}^{-1}(\boldsymbol{\mathcal{X}}_i^T\boldsymbol{\theta}^*)\}\boldsymbol{\mathcal{X}}_i\|_\infty \\
&= \|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_1\|\mathcal{P}_n\{Y_i - \text{logit}^{-1}(\boldsymbol{\mathcal{X}}_i^T\boldsymbol{\theta}^*)\}\|_1\|\mathcal{P}_n\boldsymbol{\mathcal{X}}_i\|_\infty \leqslant o_p(1/\sqrt{n})
\end{aligned}
$$

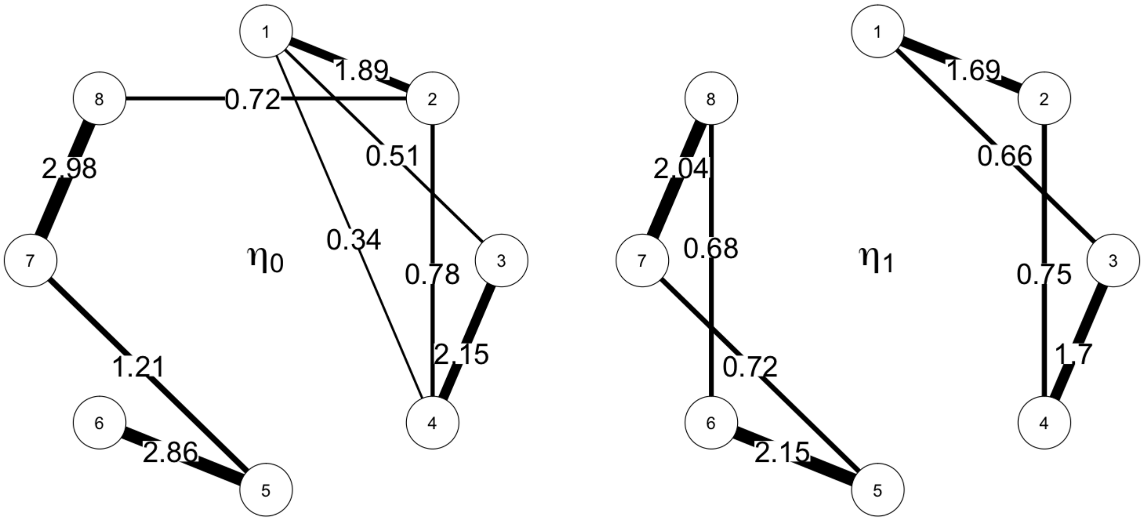
We now consider $n^{1/2}\boldsymbol{A}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = n^{1/2}\boldsymbol{A}\boldsymbol{T}_1 + n^{1/2}\boldsymbol{A}(\boldsymbol{T}_2 + \boldsymbol{T}_3)$. For any $\boldsymbol{A} \in \mathcal{A}_r$ with fixed r, we have $\|n^{1/2}\boldsymbol{A}(\boldsymbol{T}_2 + \boldsymbol{T}_3)\|_\infty \leqslant \|\boldsymbol{A}\|_1\|\sqrt{n}(\boldsymbol{T}_2 + \boldsymbol{T}_3)\|_\infty = o_r(1)$

Also recall the Fisher information for the logistic regression is $\boldsymbol{J}^* = \text{var}\{\dfrac{1}{n}\boldsymbol{\mathcal{X}}^T(\boldsymbol{\mathcal{Y}} - \boldsymbol{\pi}^*)\}$, and the Hessian information is $\boldsymbol{H}^* = \frac{1}{n}\boldsymbol{\mathcal{X}}^T\text{diag}\big[\pi_1^*(1 - \pi_1^*), \ldots, \pi_n^*(1 - \pi_n^*)\big]\boldsymbol{\mathcal{X}}$, where $\pi_i^* = \text{logit}^{-1}\{\boldsymbol{\mathcal{X}}_i(\kappa_i)^T\boldsymbol{\theta}^*\}$. From conditions (C1) and (C2), both $\boldsymbol{J}^*$ and $\{\boldsymbol{H}^*\}^{-1}$ exist. When $\boldsymbol{A}\boldsymbol{A}^T$ is positive definite with bounded eigen values, $\boldsymbol{\Sigma}^{-1/2}$ exists.

From the central limit theorem and the theory of unbiased estimating equation theory (see, e.g., Chapter 3 of Song (2007)), we have $n^{1/2}\boldsymbol{A}\boldsymbol{T}_1 \xrightarrow{d} \mathcal{N}_r(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Theta}^*\boldsymbol{J}^*\boldsymbol{\Theta}^*\boldsymbol{A}^T)$.

Finally, we prove that

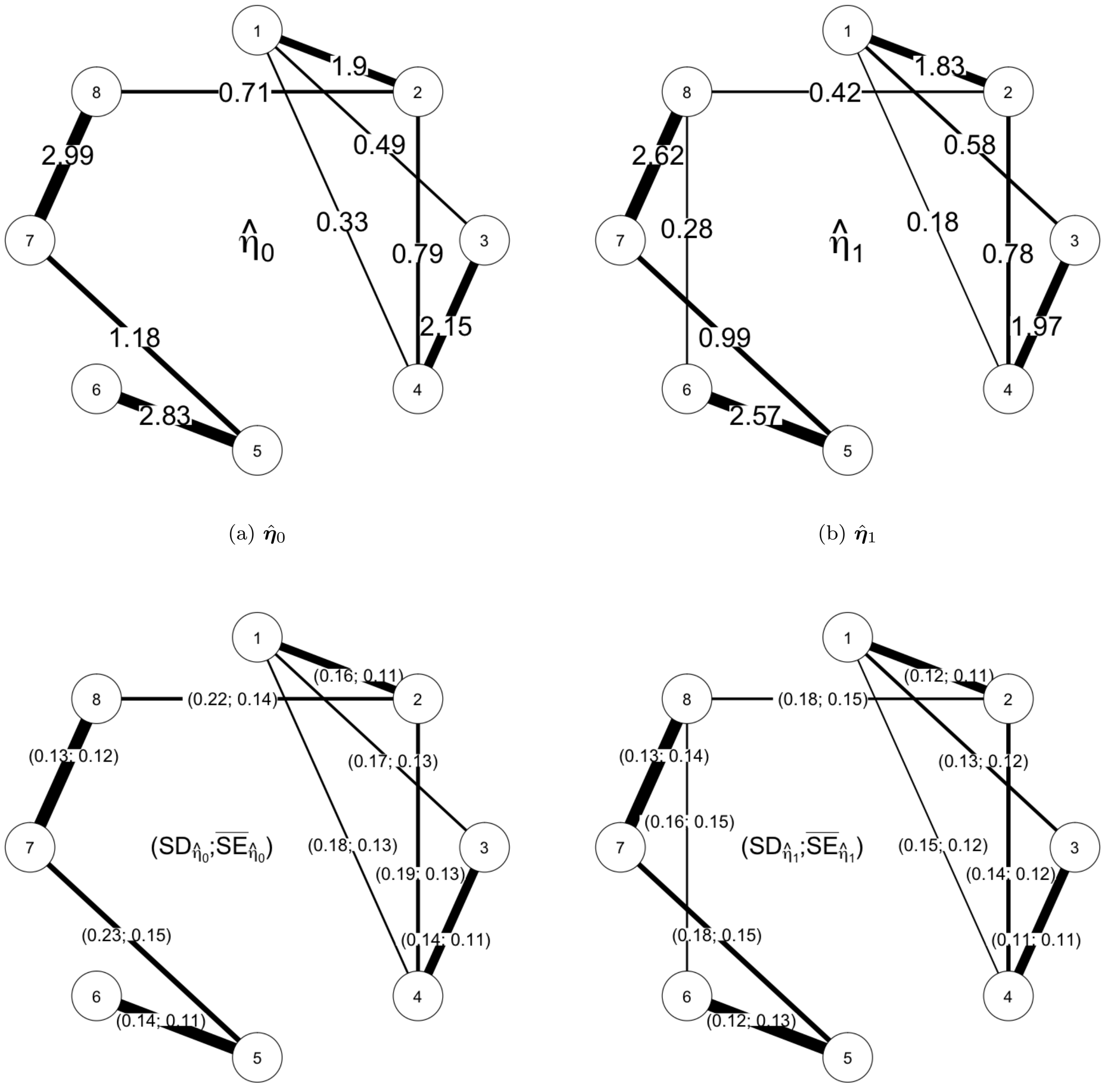
$$
n^{1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{A}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}_r(\boldsymbol{0}, \boldsymbol{I}_r)
$$



(a) $\boldsymbol{\eta}_0$

(b) $\boldsymbol{\eta}_1$

Figure 1: Illustration of true parameters designated for simulation; node indices $j \in \{1, 2, \ldots, 8\}$ are written in circles; nonzero values of $\eta_{jk}(= \eta_{kj})$ are labeled on edges in (a) for $\boldsymbol{\eta}_0$ and in (b) for $\boldsymbol{\eta}_1$, while zeros have void edges; the width of edges represents the strength of conditional dependence between two nodes.



(a) $\hat{\boldsymbol{\eta}}_0$

(b) $\hat{\boldsymbol{\eta}}_1$

(c) SD and $\overline{\mathsf{SE}}$ of $\hat{\boldsymbol{\eta}}_0$

(d) SD and $\overline{\mathsf{SE}}$ of $\hat{\boldsymbol{\eta}}_1$

Figure 2: Simulation results; (a)(b) mean of estimates; (c)(d) standard deviation of estimates, SD, and mean of asymptotic standard deviation (standard error) of estimates, $\overline{\mathsf{SE}}$.

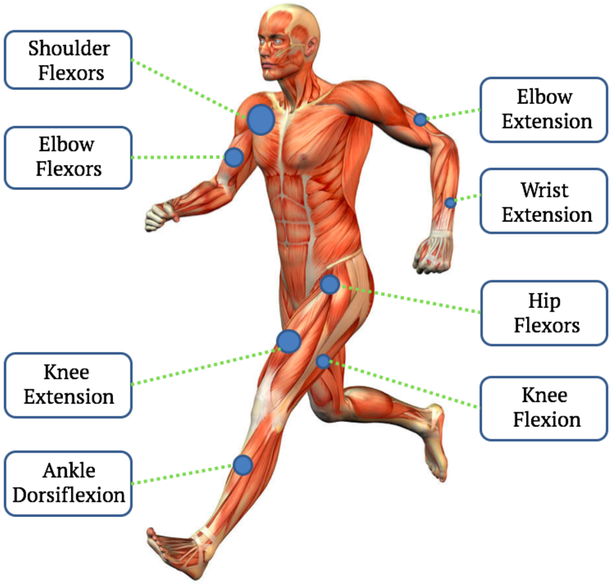
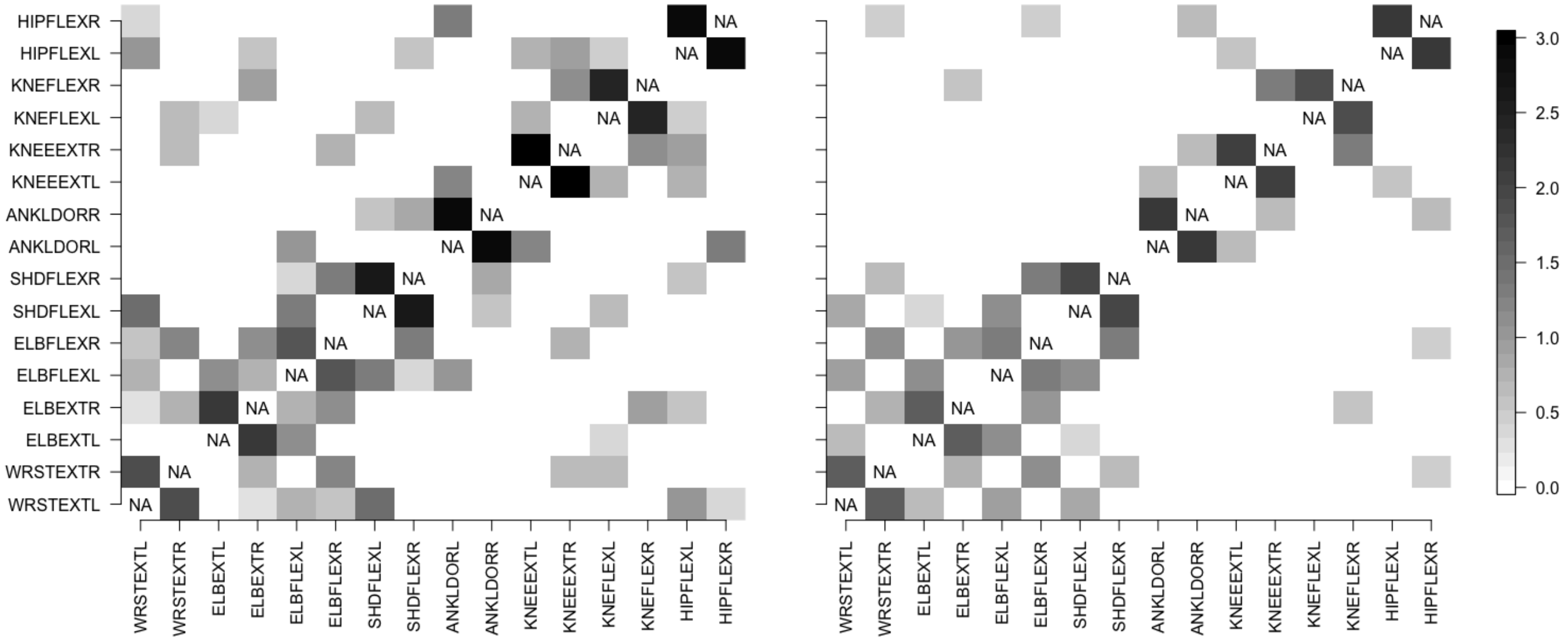


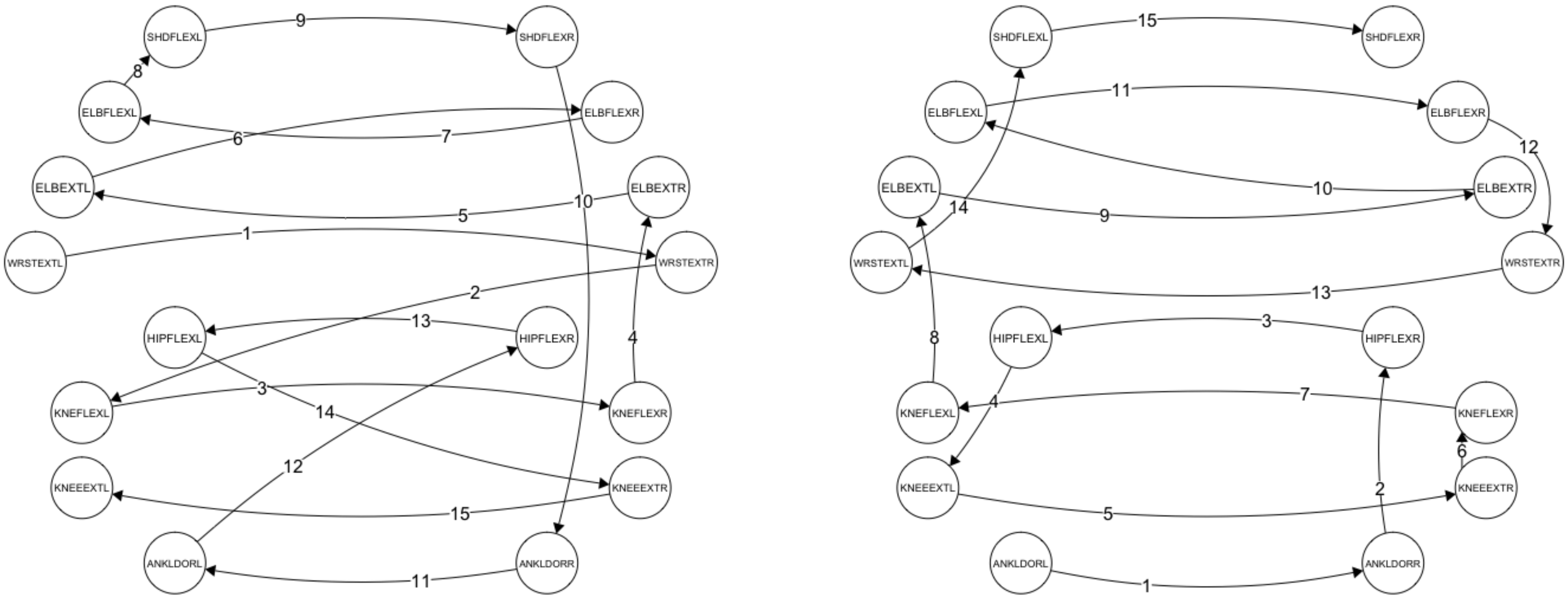
Figure 3: Measured muscles on a human body map; right and left sides of eight pairs of muscle groups, totally sixteen number of muscles (16 nodes), are examined.



(a) $\boldsymbol{\eta}_0$

(b) $\boldsymbol{\eta}_1$

Figure 4: Illustration of estimates; muscles are labeled by their abbreviated letters followed by ‘R’(right) or ‘L’(left); color depth represents the strength of conditional association between two muscles; no coefficients for the same muscle, denoted by ‘NA’.



(a) when a left wrist (WRSTEXTL) is diseased at $t = 0$

(b) when a left ankle (ANKLDORL) is diseased at $t = 0$

Figure 5: Probable paths of ALS progression for a hypothetical patient; circle nodes with abbreviated muscle names are drawn on a concised human body map; times are labeled at arrows; each pathway is the most probable one based on transition probability.

Table 1: Demonstration of η-parameters describing the effect of s_2 on the probability of s_1 being diseased, $p_m(s_1, t)$, which depends on the status of s_2 at previous and current times.

	$Y_m(s_2, t-1)$	$Y_m(s_2, t)$	change in logit$\{p_m(s_1, t)\}$
Case 1	0	0	$\eta_{012}(0 - \kappa_m)$
Case 2	0	1	$\eta_{012}(1 - \kappa_m)$
Case 3	1	1	$\eta_{112}(1 - \kappa_m)$

Table 2: Summary of RMSEs of transition probabilities at active transitions

	model	median	mean	max
(3)	autologistic network	0.0040	0.0087	0.0925
(12)	simple Markov	0.0081	0.0335	0.4461

Table 3: Summary for the bias-corrected estimates of $\boldsymbol{\beta}$ from EMPOWER study

covariate	estimate	95% confidence iinterval	p-value
Intercept	-3.5695	$(-3.7758, -3.3633)$	< 0.0001
Visiting time (t)	-0.0512	$(-0.0656, -0.0367)$	< 0.0001
Onset site	-0.0011	$(-0.1012, \quad 0.0991)$	0.6470
Symptom duration	-0.0100	$(-0.0182, -0.0017)$	0.0071