Shin Yei Eun ORCID iD: 0000-0002-0739-1281
Song Peter ORCID iD: 0000-0001-7881-7182

Author Manuscript

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.A.
[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 identifies an underlying network structure empirically to describe the pattern of spreading disease. The model also allows the network autoregressive effects to vary depending on the muscles previous status. Based on the joint distribution derived from this autologistic model, joint transition probabilities of responses among locations can be estimated and disease status can be predicted in the next time interval. 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 the ALS clinical study.

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

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi:10.1111/biom.13111

This article is protected by copyright. All rights reserved.

1 Introduction

This research was motivated by a clinical study on amyotrophic lateral sclerosis (ALS). ALS, also known as Lou Gehrigs 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 patients 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 symptoms typically start 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.

Our research interest is to characterize how disease spreads over space and time, and to address challenges to statistical modeling and inference. First, neighborhood is not clearly defined, because spatial closeness may not reflect underlying disease spreading patterns. For example, disease in one location can spread to another distant location rather than any nearby locations so that actual dependence over space is determined by some complex but unknown mechanism. Second, the outcome of interest is irreversible over time. In ALS, a diseased muscle can never return to normal as there is no treatment. Thus, the data generation mechanism has an absorbing state. Lastly, the strength of temporal association depends strongly on the previous disease statuses of various muscles. For example, recently diseased muscles pose a higher risk than muscles that were diseased in the distant past.

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 better than latent variable models for modeling spatial dependence. Caragea and Kaiser (2009) propose a centered autologistic model to allow for more interpretable parameters, and Hughes et al. (2011) conducted comparative studies to evaluate the performance of several computational 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) on Ising models, a special case of autologistic regression; sparse regularization techniques are used to identify sparse network associations among nodes. These regularization methods, however, focus mostly on spatial data and hence are not directly applicable to the evaluation of ALS spreading patterns over space and

This article is protected by copyright. All rights reserved.

Author Manuscript

Author Manuscript

time. For spatio-temporal binary data, Zhu et al. (2005) developed a model via joint distributions to first estimate spatial correlation and them temporal spread prediction. To incorporate absorbing states, Kaiser et al. (2014) formulated a model with sufficient support conditions to construct a well-defined joint distribution 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 methods to estimate a 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 estimates a spatial network from data without the need to pre-specify a neighborhood structure; accounts for absorbing states; and allows for varying effects depending on muscles' 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). This feature alleviates spatial confounding and enhances parameter interpretability. While the model is based on the conditional probability at a single location, we derive a valid joint distribution for all locations to establish the transition probabilities needed to project disease progression.

We use pseudo-likelihood (Besag, 1975) to estimate model parameters. We employ penalization with the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) to deal with the large number of pairwise spatial associations. We adapt the pseudo-likelihood to the generalized linear model (GLM) framework. Because the LASSO estimator is biased and does not have a tractable limiting distribution, we propose a bias-correction for the penalized pseudo-likelihood estimator and establish its asymptotic distribution, following methods for 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 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 our proposed approach. In Section 6, we apply these methods to the motivating ALS clinical study. Finally, we summarize the research findings and suggest future studies in Section 7.

This article is protected by copyright. All rights reserved.

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, $\mathcal{P}^0_{mt} = \{j : Y_m(s_j,t-1) = 0, j = 1,\ldots,N_s\}$ and $\mathcal{P}^1_{mt} = \{j : Y_m(s_j,t-1) = 1, j = 1,\ldots,N_s\}$; $\mathcal{P}^0_{mt}$ is an *active* set consisting of locations previously normal that have a chance to change their status at the next time, and $\mathcal{P}^1_{mt}$ is an *absorbing* set of previously diseased locations. The vector of independent variables includes time and other covariates, 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, $\mathrm{P}[Y_m(s_j,t) = 1|\boldsymbol{X}_m, Y_m(s_k,t-1), Y_m(s_k,t) \text{ for } \forall k \in \{1,\ldots,N_s\}\backslash\{j\}] = p_m(s_j,t)$, where $A\backslash B$ indicates the set A excluding B. This probability is assumed to be first order Markovian over time. Subjects are independent. We propose an autologistic network model, for $j \in \mathcal{P}^0_{mt}$,

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

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

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

We center the autocovariates by κ_m to reduce bias and facilitate interpretation of η-parameters (Caragea and Kaiser, 2009). Without centering, the odds of $Y_m(s_j,t) = 1$ in model (1) increases for any nonzero neighbors, and can never decrease even when most its neighbors are zeros. The centering constant κ_m is the expectation of $\mathrm{logit}\{p_m(s_j,t)\}$ under an independence model without autocovariates. When $\boldsymbol{\beta} = 0$, the centered autocovariates, $\{Y_m(s_k,t)-\kappa_m\}$, are -0.5 and 0.5 for previously normal and diseased muscles respectively.

We divide the autocovariates into the active and absorbing set to allow normal and diseased locations to have different effects η_{0jk} and η_{1jk}. The parameter η_{0jk} indicates the effect of location s_k on s_j if s_k was previously normal, and η_{1jk} is the effect of s_k on s_j if s_k was previously diseased. These parameters characterize associations between s_j and s_k for any $j \neq k$ and define a network structure for all two-way connections.

Also, we assume symmetry on η-parameters to ensure both a valid joint distribution (Section 3) and nondirectional correlations. For spreading pattern studies like ALS (Section 6), these η-parameters can be further restricted to be non-negative because

This article is protected by copyright. All rights reserved.

Author Manuscript

it is not plausible that a normal location increases the risk of disease at other sites. See also Section S1 in Supporting Information for details on interpretation of η-parameters.

3 Joint Distribution and Transition Probability

We show that the conditional probabilities modeled by (1) 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 let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\eta}_0^{\mathrm{T}}, \boldsymbol{\eta}_1^{\mathrm{T}})^{\mathrm{T}}$ be all coefficient vectors of dimension p.

We first consider the spatial joint distribution of $\boldsymbol{Y}_{mt} = \big\{Y_m(s_1,t), \ldots, Y_m(s_{N_s},t)\big\}^{\mathrm{T}}$. Since changed realizations of $\boldsymbol{Y}_{mt}$ from absorbing states have zero probability, we restrict its domain. Let $\mathcal{S}$ be all 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 has $2^{|\mathcal{P}^0_{mt}|}$ 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 $\boldsymbol{Y}_{mt} \in \mathcal{S}_{mt}$,

$$f(\boldsymbol{Y}_{mt}|\boldsymbol{\theta}) = \frac{\exp\{Q(\boldsymbol{Y}_{mt}|\boldsymbol{\theta})\}}{\sum_{\boldsymbol{Y}_{mt}\in\mathcal{S}_{mt}} \exp\{Q(\boldsymbol{Y}_{mt}|\boldsymbol{\theta})\}} \tag{2}$$

for a fixed subject m and time t, and $Q(\boldsymbol{Y}_{mt}|\boldsymbol{\theta}) = \sum_{j\in\mathcal{P}^0_{mt}} Y_m(s_j,t)\{\boldsymbol{X}_m^{\mathrm{T}}\boldsymbol{\beta} - \sum_{k\in\mathcal{P}^0_{mt}\setminus\{j\}} \eta_{0jk}\kappa_m - \sum_{k\in\mathcal{P}^1_{mt}\setminus\{j\}} \eta_{1jk}\kappa_m\} + 1/2 \sum_{j\in\mathcal{P}^0_{mt}}\{\sum_{k\in\mathcal{P}^0_{mt}\setminus\{j\}} \eta_{0jk} Y_m(s_j,t)Y_m(s_k,t) + \sum_{k\in\mathcal{P}^1_{mt}\setminus\{j\}} \eta_{1jk} Y_m(s_j,t)Y_m(s_k,t)\}$. We defer the derivation of the function, Q, in (2) to Section S2 in Supporting Information.

Importantly, the spatial joint distribution (2) can be viewed as a conditional distribution at t given status at $t-1$, because $\boldsymbol{Y}_{m(t-1)}$ determines both $\mathcal{P}^0_{mt}$ and $\mathcal{S}_{mt}$. As a result, we have an *one-time transition probability*, $\mathrm{P}(\boldsymbol{Y}_{mt}|\boldsymbol{Y}_{m(t-1)};\boldsymbol{\theta}) = f(\boldsymbol{Y}_{mt}|\boldsymbol{\theta})$, which is essential for inference on future responses; by plugging estimates $\widehat{\boldsymbol{\theta}}$ into (2), one can predict one-time ahead disease status or the locations at which the disease is most (or least) likely to occur next (see Section 5 for numerical studies).

The *full joint distribution* of responses for all times, locations and subjects is $\mathbb{P}(\boldsymbol{Y}|\boldsymbol{\theta}) = \prod_{m=1}^{M}\prod_{t=1}^{T_m} f(\boldsymbol{Y}_{mt}|\boldsymbol{\theta})$ for $\boldsymbol{Y} = \{Y_m(s_j,t) | m = 1, \ldots, M; j = 1, \ldots, N_s; t = 1, \ldots, T_m\}$. This follows from the independence of responses across subjects, based on an inductive method with a valid Markov random field model at $t = 0$ (Kaiser et al., 2014).

This article is protected by copyright. All rights reserved.

Author Manuscript

4 Penalized Maximum Pseudo Likelihood Estimation with Bias-correction

4.1 The Penalized Maximum Pseudo Likelihood Estimation

Estimating the autologistic model parameters in (1) is challenging since the normalizing constant in the denominator of (2) is computationally costly when $|\mathcal{S}_{mt}|$ is large. For efficient computation, we replace the full likelihood with the product of conditional likelihoods, the pseudo-likelihood (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 the probability of $\mathcal{Y}_i = 1$ and model (1) is equivalent to a logistic linear regression model as

$$\text{logit}\{P(\mathcal{Y}_i = 1)\} = \boldsymbol{\mathcal{X}}_i(\kappa_i)^T \boldsymbol{\theta}, \tag{3}$$

where a centering parameter κ_i corresponds to an index i with respect to (m, s_j, t). The design matrix $\boldsymbol{\mathcal{X}} = \{\boldsymbol{\mathcal{X}}_1(\kappa_1), \ldots, \boldsymbol{\mathcal{X}}_n(\kappa_n)\}^{\mathrm{T}}$ is a set of

$$\boldsymbol{\mathcal{X}}_i(\kappa_i) = \{\boldsymbol{X}_m^{\mathrm{T}}, \widetilde{\boldsymbol{Y}}^1_{mt}(0)^{\mathrm{T}}, \ldots, \widetilde{\boldsymbol{Y}}^{(N_s-1)}_{mt}(0)^{\mathrm{T}}, \widetilde{\boldsymbol{Y}}^1_{mt}(1)^{\mathrm{T}}, \ldots, \widetilde{\boldsymbol{Y}}^{(N_s-1)}_{mt}(1)^{\mathrm{T}}\}^{\mathrm{T}}$$

where $\widetilde{\boldsymbol{Y}}^k_{mt}(\delta)$ denotes a $(N_s - k) \times 1$ vector for a given s_j such that

$$\widetilde{\boldsymbol{Y}}^k_{mt}(\delta) = \begin{cases} [\mathbf{0}_{j-1}, \{Y_m(s_j, t) - \kappa_m\} I\{Y_m(s_j, t) = \delta\}, \mathbf{0}_{N_s-j-k}]^{\mathrm{T}} & k < j \\ [\{Y_m(s_{j+1}, t) - \kappa_m\} I\{Y_m(s_{j+1}, t) = \delta\}, \ldots, \{Y_m(s_{N_s}, t) - \kappa_m\} I\{Y_m(s_{N_s}, t) = \delta\}]^{\mathrm{T}} & k = j \\ \mathbf{0}_{N_s-k} & k > j \end{cases}$$

for $k = 1, \ldots, (N_s - 1)$ and $\delta = 0, 1$. Here $\mathbf{0}_k$ denotes a k length vector of zeros. The elements of the design matrix correspond to each of $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\eta}_0^{\mathrm{T}}, \boldsymbol{\eta}_1^{\mathrm{T}})^{\mathrm{T}}$. To ensure 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,

$$\boldsymbol{\mathcal{L}}_c(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_{c,i}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \Big(\mathcal{Y}_i \boldsymbol{\mathcal{X}}_i(\kappa_i)^{\mathrm{T}} \boldsymbol{\theta} - \log\big[1 + \exp\{\boldsymbol{\mathcal{X}}_i(\kappa_i)^{\mathrm{T}} \boldsymbol{\theta}\}\big] \Big), \tag{4}$$

which can be maximized using generalized linear model (GLM) estimation for a fixed κ_i.

We penalize (4) with 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 because a specific covariate or location possibly may have negligible effect.

This article is protected by copyright. All rights reserved.

Author Manuscript

Specifically, we maximize the ℓ_1-penalized pseudo log-likelihood, $F_\lambda(\boldsymbol{\theta}) = \boldsymbol{\mathcal{L}}_c(\boldsymbol{\theta}) - \lambda\|\boldsymbol{\theta}\|_1$, 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 adaptive LASSO (Zou, 2006), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), and maximum a posteriori (MAP) estimation.

4.2 The Bias-corrected Estimator and Inference

The consistency of LASSO estimators for GLM has been proved 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 $\widehat{\boldsymbol{\theta}}_\lambda = (\widehat{\boldsymbol{\beta}}_\lambda, \widehat{\boldsymbol{\eta}}_{0\lambda}, \widehat{\boldsymbol{\eta}}_{1\lambda})^{\mathrm{T}}$ be the regularized estimator at λ, that is $\widehat{\boldsymbol{\theta}}_\lambda = \arg\max_{\boldsymbol{\theta}} F_\lambda(\boldsymbol{\theta})$. By the Karush-Kuhn-Tucker (KKT) optimality conditions (Kuhn and Tucker, 2014), i.e. the subgradient of the objective, $F_\lambda(\boldsymbol{\theta})$, is 0, the regularized pseudo likelihood estimator satisfies

$$S_n(\widehat{\boldsymbol{\theta}}_\lambda) - \lambda\widehat{\boldsymbol{Z}} = 0, \tag{5}$$

where $S_n(\widehat{\boldsymbol{\theta}}_\lambda) = \frac{1}{n}\sum_{i=1}^n \dot{\boldsymbol{\mathcal{L}}}_{c,i}(\widehat{\boldsymbol{\theta}}_\lambda)$ is a pseudo-score function of (4) and $\widehat{\boldsymbol{Z}} = (\widehat{Z}_1, \ldots, \widehat{Z}_p)^{\mathrm{T}}$ is a subdifferential satisfying $\widehat{Z}_j = \mathrm{sign}(\widehat{\theta}_j)$ if $\widehat{\theta}_j \neq 0$ and $\widehat{Z}_j \in \{-1, 1\}$ otherwise, for $j = 1, \ldots, p$. The first-order Taylor expansion of $S_n(\widehat{\boldsymbol{\theta}}_\lambda)$ in (5) leads to $S_n(\boldsymbol{\theta}^*) + \dot{S}_n(\boldsymbol{\theta}^*)(\widehat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*) - \lambda\widehat{\boldsymbol{Z}} \approx 0$. Multiplying both sides by $\{\dot{S}_n(\boldsymbol{\theta}^*)\}^{-1}$ and reordering terms, we have

$$\widehat{\boldsymbol{\theta}}_\lambda + \{-\dot{S}_n(\boldsymbol{\theta}^*)\}^{-1}\lambda\widehat{\boldsymbol{Z}} - \boldsymbol{\theta}^* + \{\dot{S}_n(\boldsymbol{\theta}^*)\}^{-1}S_n(\boldsymbol{\theta}^*) \approx 0. \tag{6}$$

Combine the first two terms and define $\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_\lambda + \{-\dot{S}_n(\boldsymbol{\theta}^*)\}^{-1}\lambda\widehat{\boldsymbol{Z}}$. From (6), we have $\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \approx \{-\dot{S}_n(\boldsymbol{\theta}^*)\}^{-1}S_n(\boldsymbol{\theta}^*)$, a property also satisfied by the maximum pseudo-likelihood estimator asymptotically. This motivates us to use $\widetilde{\boldsymbol{\theta}}$ as a bias-corrected estimator. In practice, $\boldsymbol{\theta}^*$ is unknown, and $-\dot{S}_n(\boldsymbol{\theta}^*)$ is estimated by an observed Hessian matrix, $\widehat{\boldsymbol{H}} = -\dot{S}_n(\widehat{\boldsymbol{\theta}}_\lambda)$. Since $\lambda\widehat{\boldsymbol{Z}} = S_n(\widehat{\boldsymbol{\theta}}_\lambda)$ from (5), the *bias-corrected LASSO estimator* is therefore

$$\widetilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}} + \widehat{\boldsymbol{H}}^{-1}S_n(\widehat{\boldsymbol{\theta}}_\lambda). \tag{7}$$

This article is protected by copyright. All rights reserved.

Author Manuscript

For the log-likelihood in (4), $S_n(\widehat{\boldsymbol{\theta}}_\lambda) = \frac{1}{n}\boldsymbol{\mathcal{X}}^{\mathrm{T}}(\boldsymbol{\mathcal{Y}} - \widehat{\boldsymbol{\pi}}_\lambda)$ and $\widehat{\boldsymbol{H}} = \frac{1}{n}\boldsymbol{\mathcal{X}}^{\mathrm{T}}\widehat{\boldsymbol{\mathcal{V}}}_\lambda\boldsymbol{\mathcal{X}}$, where $\widehat{\boldsymbol{\pi}}_\lambda = (\widehat{\pi}_{1\lambda}, \ldots, \widehat{\pi}_{n\lambda})^{\mathrm{T}}$ and $\widehat{\boldsymbol{\mathcal{V}}}_\lambda = \mathrm{diag}\{\widehat{\pi}_{1\lambda}(1-\widehat{\pi}_{1\lambda}), \ldots, \widehat{\pi}_{n\lambda}(1-\widehat{\pi}_{n\lambda})\}$ with $\widehat{\pi}_{i\lambda} = \mathrm{logit}^{-1}\{\boldsymbol{\mathcal{X}}_i(\kappa_i)^{\mathrm{T}}\widehat{\boldsymbol{\theta}}_\lambda\}$. We introduce notation for the asymptotic framework. Recall $n = \sum_{m=1}^{M}\sum_{t=1}^{T_m}\sum_{j=1}^{N_s} I(s_j \in \mathcal{P}_{mt}^0)$ 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 $\Lambda_{\min}(\boldsymbol{\mathcal{X}}^{\mathrm{T}}\boldsymbol{\mathcal{X}}/n) = \mathcal{O}(1)$ and $\Lambda_{\min}(\boldsymbol{\mathcal{X}}^{\mathrm{T}}\boldsymbol{\mathcal{X}}/n) = \mathcal{O}(1)$, where $\Lambda_{\min}(\mathcal{M})$ and $\Lambda_{\max}(\mathcal{M})$ denote the minimum and maximum eigenvalues of a matrix $\mathcal{M}$, respectively.

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

(C3) The number of true signals $s^* \ll p$ satisfies $s^* = o(\sqrt{n/(p\log p)})$ and $\lambda \asymp \sqrt{\log p/n}$.

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

THEOREM 1: *Suppose the conditions (C1)-(C3) hold. Then the bias-corrected estimator* $\widetilde{\boldsymbol{\theta}}$ *defined in* (7) *is consistent. Moreover, for a fixed* r, *let* $\mathcal{A}_r = \{\boldsymbol{A} \in \mathbb{R}^{r\times p} : 0 < \Lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^T) \leqslant \Lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^T) < \infty\}$. *Then for any* $\boldsymbol{A} \in \mathcal{A}_r$, *we have*

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

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 Section S3 of Supporting Information. This asymptotic normality results enable statistical inference for the bias-corrected estimator, such as hypothesis tests or confidence interval constructions. Theorem 1 is general because it is applicable 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 results still hold under appropriate regularity conditions.

This article is protected by copyright. All rights reserved.

Author Manuscript

Author Manuscript

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 (1) 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 convergence, as follows.

1. Fit the independence model, the model (1) with $\forall \boldsymbol{\eta} = 0$, and set as $\widehat{\boldsymbol{\beta}}^{(0)}$;
2. At the l-th iteration, for a fixed $\widehat{\kappa}_i^{(l-1)} = \text{logit}^{-1}\{\boldsymbol{X}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{(l-1)}\}$ and given λ (see below), fit the model (3) by maximizing F_λ, and update estimates $\widehat{\boldsymbol{\theta}}_\lambda^{(l)} = \{\widehat{\boldsymbol{\beta}}_\lambda^{(l)}, \widehat{\boldsymbol{\eta}}_{0\lambda}^{(l)}, \widehat{\boldsymbol{\eta}}_{1\lambda}^{(l)}\}^{\mathrm{T}}$;
3. Calculate the bias-corrected estimates by (7), $\widetilde{\boldsymbol{\theta}}^{(l)} = \widehat{\boldsymbol{\theta}}_\lambda^{(l)} + \{\widehat{\boldsymbol{H}}(\widehat{\boldsymbol{\theta}}_\lambda^{(l)})\}^{-1} S_n(\widehat{\boldsymbol{\theta}}_\lambda^{(l)})$;
4. Update the centering parameters $\widehat{\kappa}_i^{(l)} = \text{logit}^{-1}\{\boldsymbol{X}_i^{\mathrm{T}}\widetilde{\boldsymbol{\beta}}^{(l)}\}$;
5. Return to step 2 until all steps converge.

In step 1 and step 2, the standard logistic regression and the GLM with regularization 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 by some data-dependent model selection criteria, such as generalized cross-validation (GCV) (Golub et al., 1979), Bayesian information criterion (BIC) (Schwarz, 1978) and extended Bayesian information criterion (EBIC) (Chen and Chen, 2008). Alternatively, one can manually choose λ to meet a desired degree of sparsity based on domain knowledge. We suggest to use any technique of choosing λ 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 the proposed prediction via transition probabilities performs compared to a simple Markov model.

We set $N_s = 8$ locations, $(s_1, s_2, \ldots, s_8)$, and the dimensions of $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ each to be $8 \times (8-1) = 56$. We assigned different values to $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ that are symmetric and moderately sparse, as illustrated in the left panels of Figure 1. For simplicity, only an intercept was included in the independence model; $\boldsymbol{X}_m = 1$ with $\boldsymbol{\beta} = -2$. This led to $\kappa_m \approx 0.12$ for $\forall m$, and zero and nonzero of $Y_m(s_j, t)$ are transformed to -0.12 and 0.88, respectively. In other words, we let the influence of diseased status be stronger than normal status.

Recall that $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta}_0, \boldsymbol{\eta}_1)^{\mathrm{T}}$ determines the one-time transition probability from one joint

This article is protected by copyright. All rights reserved.

outcome to another by the formula (2). Instead of presenting all transition probabilities for 2^8 joint outcomes, we provide a simple example to illustrate 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 was generated from a Bernoulli distribution with a probability 0.25; $Y(s_j,0) \sim$ 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 sites are diseased. Such sample sequences were independently generated for $M = 500$ subjects. 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 the `cv.glmnet` function in R and reused for the subsequent iterations to speed up convergence. To impose a non-negativity constraint, the minimum value of η is set to be zero using the option (`lower.limits`). Most runs converged in 10 or fewer iterations. We ran $B = 100$ rounds of simulations.

The center and right panels of Figure 1 illustrate the simulation results with the means and standard deviations of the estimates and the means of the estimated asymptotic standard deviations, denoted by $\widehat{\boldsymbol{\eta}}$, $\mathsf{SD}_{\widehat{\boldsymbol{\eta}}}$ and $\overline{\mathsf{SE}}_{\widehat{\boldsymbol{\eta}}}$, respectively. From the center panels, 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 values, the effect 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); $\eta_{078}(=\eta_{087}) = 2.98 > \eta_{057}(=\eta_{075}) = 1.21$. These effects were estimated as $\widehat{\eta}_{078}(=\widehat{\eta}_{087}) = 2.99 > \widehat{\eta}_{057}(=\widehat{\eta}_{075}) = 1.18$, which are close to the true values. The right panels demonstrate the asymptotic distribution derived by Theorem 1. The good correspondence between the empirical and estimated variances is evidence that the covariance matrix in Theorem 1 is a good estimator for the asymptotic variances of parameters in the proposed model.

We also compared the autologistic network model (1) with a simple Markov model,

$$\mathrm{logit}\{p_m(s_j,t)\} = \boldsymbol{X}_m^{\mathrm{T}}\boldsymbol{\beta} + \sum_{k\neq j}\eta_{jk}\{Y_m(s_k,t-1) - \kappa_m\}, \tag{8}$$

This article is protected by copyright. All rights reserved.

Author Manuscript

Author Manuscript

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

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

where $\widehat{\boldsymbol{\theta}}_{(b)}$ denotes the estimated model parameters at the b-th simulation run. Since it is not feasible to show the RMSEs at all active transitions, we instead report their summaries; the mean, median, max of RMSEs of transition probabilities are 0.0040, 0.0087, 0.0925 for the proposed model (1), and 0.0081, 0.0335, 0.4461 for the simple Markov model (8). Therefore the proposed model (1) more closely estimated the transition probabilities than the simple Markov model (8).

Despite 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 eigenvalue condition (C1) and Donoho and Johnstone (1994)'s hard threshold rate $\sqrt{\log p/n}$ in condition (C3). The numerical simulation results demonstrate that our estimator approximates true parameter values well.

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 before 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. Figure 2 shows the sixteen muscles: the left and right of shoulder flexion, elbow flexion, hip flexion, knee flexion, elbow extension, knee entension, wrist extension and ankle dorsiflexion.

This article is protected by copyright. All rights reserved.

Author Manuscript

For 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 patients 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 (1) to these data with independent variables $\boldsymbol{X} = \{1, t, \text{symptom onset site}, \text{symptom duration}\}$ and estimated the model parameters using regularized pseudo-likelihood. 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 was within 1% of the previous estimate.

Figure 3 shows the heat maps of estimated $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$. The horizontal connections between the right and left side of each muscle were mostly stronger than other connections for both previously healthy and diseased patients. This implies that a muscle is likely to remain in the same status as its opposite side, no matter what status a muscle was at the previous visit. 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 effects 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. The fairly strong connection was observed between elbow and knee in $\boldsymbol{\eta}_1$, which can be a clue to a biological link between upper and lower body locations.

This article is protected by copyright. All rights reserved.

Author Manuscript

By computing the transition probabilities from (2), we predicted disease progression in the next time interval. It is of clinical interest to single out which muscles have the most likelihood of being impaired before long. We also tracked the most susceptible muscles continuously and sequentially until all muscles are impaired. This pathway of muscle impairment provides a simple yet informative prediction of disease spread. Another approach is 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 4 illustrates two examples of probable disease progression paths for this hypothetical patient. In Figure 4 (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 3; 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 4 (b); the transitions between muscles of right and left sides and between elbow and knee muscles were remarkable.

Table 1 summarizes the bias-corrected estimates of regression coefficients $\boldsymbol{\beta}$; the confidence intervals and p-values were based on Theorem 1. The $\widehat{\boldsymbol{\beta}}$ can be better interpreted as centering parameters, $\widehat{\kappa} = \text{logit}^{-1}(\boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$. For example, the estimated overall probability of disease progression with no contributions from independent covariates or autocovariates is $\widehat{\kappa} = \text{logit}^{-1}(3.57) \approx 0.03$, which is reasonably low; this would decrease over time because $\widehat{\beta} = -0.05$ for the visiting time is negative with $p < .0001$. As a result, individual muscles would be likely to stay healthy if there were no inter-muscle spatial dependency and other risk factors. The negative effect of the symptom duration ($\widehat{\beta} = -0.01$ 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 with a recently onset tends to have higher probability of progression.

This article is protected by copyright. All rights reserved.

7 Discussion

We proposed an autologistic network model for spatio-temporal binary data with absorbing states. The major contributions are: we relaxed the need of pre-specification on neighborhood structure; we considered absorbing states 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 to characterize spreading patterns of a disease.

We used LASSO-penalized pseudo-likelihood maximization to enforce sparsity on network associations. We proposed an efficient iterative algorithm for model implementation by converting the optimization into an ordinary penalized GLM problem. We also applied a bias-correction method and showed the asymptotic normality of estimators. 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 requires future exploration.

Our simulation study confirmed that the proposed estimation approach is valid for inference on model parameters. We also showed in simulations based on spatio-temporal model that the proposed model estimated transition probabilities more precisely than a simple Markov model, which does not allow for simultaneous spatial dependency or different effects depending on previous status. The application to ALS data demonstrated that our model offers insights into the spreading patterns 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 is 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 for applications such as brain imaging.

Author Manuscript

This article is protected by copyright. All rights reserved.

Acknowledgements

We gratefully acknowledge that this research is financially supported by Biogen. 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*. The authors thank the Editor, Associate Editor and a reviewer for their constructive comments, and Mitchell Gail for his final proofreading.

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.

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., 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.

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.

Höfling, H. and Tibshirani, R. (2009). Estimation of sparse binary pairwise markov

This article is protected by copyright. All rights reserved.

Author Manuscript

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., and Lafferty, J. D. (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.

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., and Dezeure, R. (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*

This article is protected by copyright. All rights reserved.

Author Manuscript

Models. University of Kentucky.

Xue, L., Zou, H., and Cai, T. (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.

Author Manuscript

This article is protected by copyright. All rights reserved.

Author Manuscript

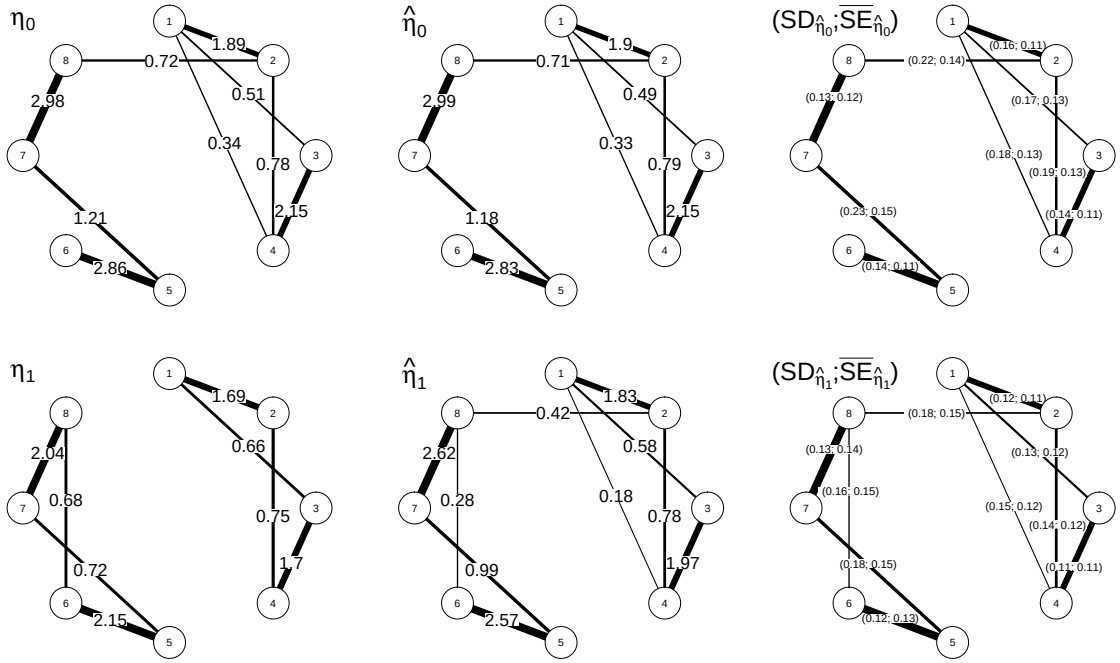


Figure 1: Simulation results for $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$. Left: true values. Center: mean of estimates. Right: standard deviation, SD, and mean of standard error, $\overline{\mathsf{SE}}$, of estimates.

This article is protected by copyright. All rights reserved.

Author Manuscript

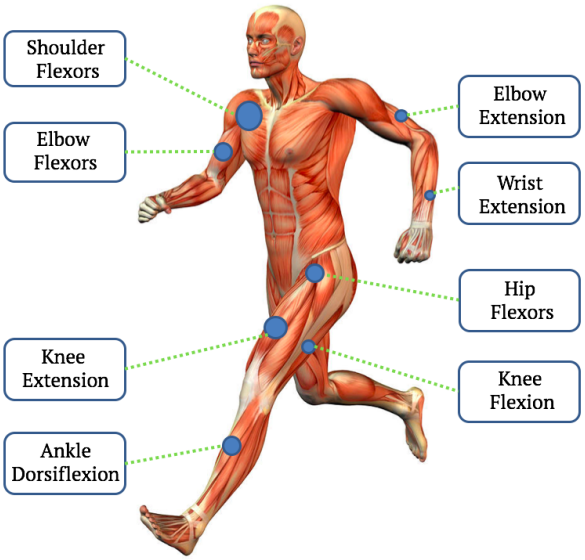
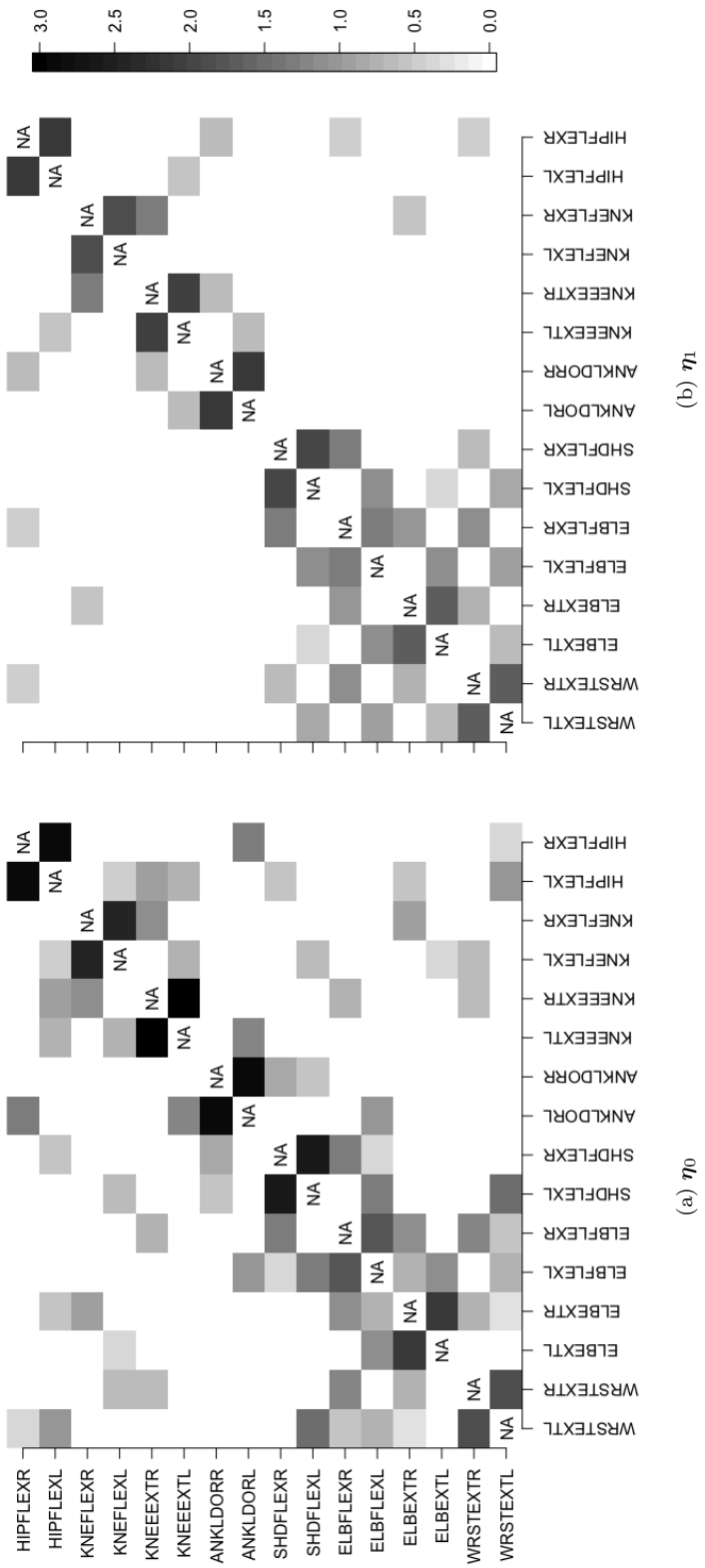


Figure 2: Measured muscles on a human body map; right and left sides of eight pairs of muscle groups. In total sixteen muscles (16 nodes) are examined. This figure appears in color in the electronic version of this article.

This article is protected by copyright. All rights reserved.

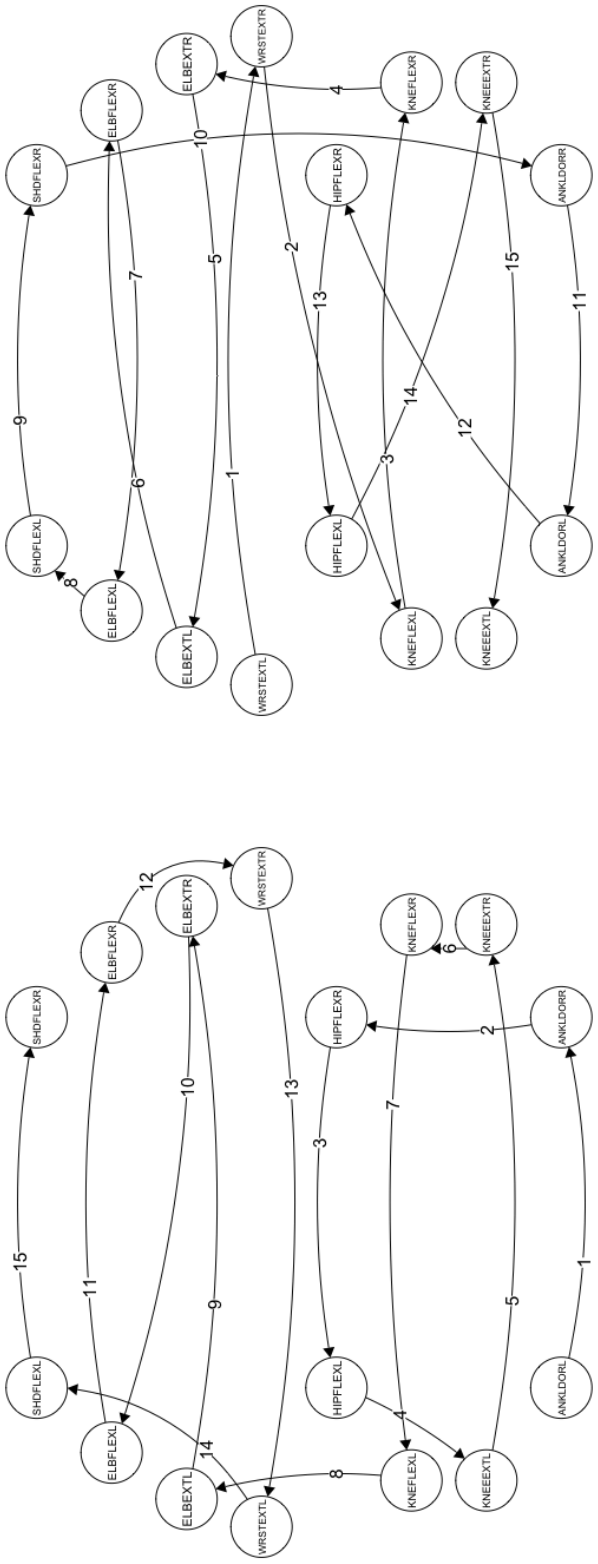


(a) $\boldsymbol{\eta}_0$ (b) $\boldsymbol{\eta}_1$

Figure 3: 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'.

Author Manuscript

This article is protected by copyright. All rights reserved.



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

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

Figure 4: Most probable paths of ALS progression for two initial statuses of a hypothetical patient; circle nodes with abbreviated muscle names are drawn in their relative positions on the human body map; times are labeled at arrows.

Author Manuscript

This article is protected by copyright. All rights reserved.

Table 1: 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

This article is protected by copyright. All rights reserved.