Received: 31 July 2020 | Revised: 2 March 2021 | Accepted: 3 March 2021

DOI: 10.1002/sim.8956

RESEARCH ARTICLE

Statistics in Medicine WILEY

A binary hidden Markov model on spatial network for amyotrophic lateral sclerosis disease spreading pattern analysis

Yei Eun Shin[1] | Dawei Liu[2] | Huiyan Sang[3] | Toby A. Ferguson[4] | Peter X. K. Song[5]

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

[2] Global Analytics and Data Sciences, Biogen, Cambridge, Massachusetts

[3] Department of Statistics, Texas A&M University, College Station, Texas

[4] Neurology Research and Early Clinical Development, Biogen, Cambridge, Massachusetts

[5] Department of Biostatistics, University of Michigan, Ann Arbor, Michigan

Correspondence
Yei Eun Shin, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA.
Email: yei-eun.shin@nih.gov

Amyotrophic lateral sclerosis (ALS) is a neurological disease that starts at a focal point and gradually spreads to other parts of the nervous system. One of the main clinical symptoms of ALS is muscle weakness. To study spreading patterns of muscle weakness, we analyze spatiotemporal binary muscle strength data, which indicates whether observed muscle strengths are impaired or healthy. We propose a hidden Markov model-based approach that assumes the observed disease status depends on two latent disease states. The model enables us to estimate the incidence rate of ALS disease and the probability of disease state transition. Specifically, the latter is modeled by a logistic autoregression in that the spatial network of susceptible muscles follows a Markov process. The proposed model is flexible to allow both historical muscle conditions and their spatial relationships to be included in the analysis. To estimate the model parameters, we provide an iterative algorithm to maximize sparse-penalized likelihood with bias correction, and use the Viterbi algorithm to label hidden disease states. We apply the proposed approach to analyze the ALS patients' data from EMPOWER Study.

KEYWORDS

autologistic regressive model, hidden Markov model, network, penalized likelihood, spatiotemporal, Viterbi algorithm

1 | INTRODUCTION

Amyotrophic lateral sclerosis (ALS), also known as Lou Gehrig's disease, is a neurological disease that affects nerve cells in the brain and the spinal cord which control voluntary muscle movements. Muscle weakness is a major symptom of ALS. The weakness typically starts from a particular muscle group and then spreads to other muscles. As the disease progresses, patients lose muscle movement in multiple muscles and finally die from the disease. Currently there is no treatment for the disease. How the muscle weakness spreads across body regions remains unknown. Motivated closely from the ALS patients' data from EMPOWER Study, in this article we develop a hidden Markov model (HMM) to explore the spreading pattern of the ALS disease progression. In particular, we model the dependence among susceptible muscles by extending the classical HMM in order to account for both spatial and temporal mechanisms of disease progression.

The motivating data consists of space-time binary disease states (1 for impaired and 0 for healthy) of 16 muscle groups of ALS patients during their longitudinal clinical visits. The observed disease states are determined in the clinical setting by thresholding the observed muscle strengths, leading to potential errors caused by measurement biases and inappropriate thresholds. Indeed, there exist some trajectories of observed states that show contradictory patterns of disease progression; for example, a muscle may appear to get healthier than before, which contradicts to the fact that ALS is a progressive disease. Also, because ALS is technically not a muscular disease but a neurological disease, their true disease states are difficult to observe precisely in practice. To overcome these difficulties, we adopt a hierarchical modeling framework for the disease spreading patterns, where patient's true muscular conditions, which are latent, drive the observed states recorded in the motivating data.

To model the spreading pattern of latent disease states, we notice that muscles are scattered over body locations and their conditions evolve in time. We develop a logistic autoregressive model that assume the current muscle states depend only on their previous states through the first-order Markov process. Specifically, for each muscle, the probability of one-time transition from healthy state to impaired state is governed by the previous states of the other muscles via lagged covariates. To reflect the fact that ALS is incurable as aforementioned, the transition probability from impaired state to healthy state is always zero. The literatures on spatiotemporal models[1-4] typically assume that spatially closer locations have stronger associations and associations between locations are symmetric. Such dependencies are indeed restrictive for ALS disease because human body parts are connected via a complex system whose neurological mechanisms remain largely unknown. In this article, we propose a flexible logistic autoregressive model to not only allow resilient network specification of neighboring muscle locations, but also allow autocovariates to have potential directional effects on the entire network of muscles under investigation. Moreover, the proposed model allows for muscles to have different effects to other muscles according to their previous states of being healthy or impaired.

Following the standard HMM-based approaches,[5] we develop an iterative statistical method involving three major steps for the proposed model in the study of ALS disease spreading patterns. In the first step, we derive the conditional distributions of observed states given hidden states, namely the emission probability, which helps calculate the observation likelihood in Section 2.1. In this step, the false positive or negative rates of disease diagnosis are estimated. In the second step, we compute the transition probability of hidden states based on the autoregressive model in Section 2.2. The model parameters are estimated by maximizing a L_1-penalized likelihood with bias correction, not only to obtain unbiased and sparse estimates but also to obtain their asymptotic distribution according the theory of post-selection inference.[6-8] For the ease of implementation, we show that such regularized estimation can be converted to an estimation procedure carried out in the generalized linear model (GLM) framework.[9] In the third step, we identify the optimal time sequence of hidden states for each muscle location using the Viterbi algorithm[10] in Section 2.3. These three steps and associated analysis goals are iterated, starting with a given initialization, until all estimates of the model parameters are converged. We apply the proposed model and estimation method to analyze the motivating ALS patients' data from the EMPOWER Study of Biogen[11] (Section 3).

2 | HIDDEN MARKOV MODEL FOR SPATIOTEMPORAL BINARY STATES

Let $Y_i(u_j, t)$ and $S_i(u_j, t)$ denote an observed and a hidden binary state, respectively; 1 if a muscle at location u_j is impaired at time t, and 0 otherwise for subject i, where $i = 1, \dots, N$, $j = 1, \dots, M$, and $t = 0, 1, \dots, n_i$. We assume that all subjects are completely observed for all locations, but may have short or different lengths of time sequences due to missing visits. Figure 1 shows the hierarchical structure of a HMM in that the observed states depend on the hidden states (vertical arrows) and the current hidden states depend only on the previous hidden states (horizontal arrows). In this article, we consider an HMM with two states corresponding to healthy and impaired muscle, respectively, at each location.

$$\begin{tikzcd}
{} \arrow[r] & S_i(u_j, t-1) \arrow[r, "\boldsymbol{A}_{ijt}"] \arrow[d, "\boldsymbol{E}_i"] & S_i(u_j, t) \arrow[r, "\boldsymbol{A}_{ij(t+1)}"] \arrow[d, "\boldsymbol{E}_i"] & S_i(u_j, t+1) \arrow[r] \arrow[d, "\boldsymbol{E}_i"] & {} \\
 & Y_i(u_j, t-1) & Y_i(u_j, t) & Y_i(u_j, t+1) &
\end{tikzcd}$$

FIGURE 1 The structure of a hidden Markov model where $S_i(u_j, t)$ and $Y_i(u_j, t)$ denote a hidden and an observed binary state, respectively, for subject i, location u_j and time t; $\boldsymbol{E}_i$ = emission probability from hidden state to observed state; $\boldsymbol{A}_{ijt}$ = transition probability from time $t - 1$ to t

2.1 | Emission probability

Emission probability refers to the conditional distribution of observations given hidden states. In the HMM, the observed states are conditionally independent given the latent states, and each observed muscle state $Y_i(u_j, t)$ is assumed to depend only on the corresponding latent muscle state $S_i(u_j, t)$, not on any other latent muscle states at different locations or time. Denote the emission probability as $e_\delta(\gamma) = P\left(Y_i(u_j, t) = \gamma | S_i(u_j, t) = \delta\right)$ for $\delta, \gamma \in \{0, 1\}$, which for simplicity is assumed to be homogeneous over all subjects, locations, and times. Obviously, $\sum_{\gamma=0}^{1} e_\delta(\gamma) = 1$, for $\delta \in \{0, 1\}$. Further, we write the four emission probabilities in the following 2×2 matrix:

$$\boldsymbol{E} = \begin{pmatrix} e_0(0) & e_0(1) \\ e_1(0) & e_1(1) \end{pmatrix} = \begin{pmatrix} 1 - e_0(1) & e_0(1) \\ e_1(0) & 1 - e_1(0) \end{pmatrix}, \tag{1}$$

where rows indexed by the subscript correspond to the hidden states and columns correspond to the observed states. Here, $e_0(1)$ and $e_1(0)$ are the probabilities of misclassification between the hidden and observed states, which can also be regarded as false positive rates and false negative rates, respectively.

If the hidden states $S_i(u_j, t)$ were known, $e_\delta(\gamma)$ could be estimated by the empirical frequencies of cross-classified observed states as

$$\widehat{e}_\delta(\gamma) = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{t=1}^{n_i} I\left(Y_i(u_j, t) = \gamma | S_i(u_j, t) = \delta\right) \Big/ \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{t=1}^{n_i} I\left(S_i(u_j, t) = \delta\right), \text{ for } \delta, \gamma \in \{0, 1\},$$

where $I(\cdot)$ denotes the indicator function. These proportion estimators above are consistent and asymptotically normally distributed by the classical statistical theories under the assumption of conditional independence.

2.2 | Transition probability

In the HMM, the binary latent process is assumed to be a discrete Markov chain of order 1, which is governed by the matrix of transition probabilities. A transition probability defines the probability law of lag-1 transition for a hidden state at time t for muscle j, denoted by $a_{\delta'\delta} = P\left(S_i(u_j, t) = \delta | S_i(u_j, t-1) = \delta'\right)$ for $\delta, \delta' \in \{0, 1\}$. For the ALS disease, we assume that the impaired state cannot revert to the healthy state; that is, $P(S_i(u_j, t) = 0 \mid S_i(u_j, t-1) = 1) = 0$. In other words, the impaired state is an absorbing state. Here we are particularly interested in estimating the transition probability of a muscle moving from its healthy state to the impaired state, namely $p_i(u_j, t) = P(S_i(u_j, t) = 1 \mid S_i(u_j, t-1) = 0)$. For this, we develop a statistical model to accommodate spatially the disease spreading patterns, in addition to the temporal Markov dynamics given by the classical HMM. The 2×2 transition probability matrix from time $t-1$ to t for location u_j of subject i is

$$\boldsymbol{A}_{ijt} = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix} = \begin{pmatrix} 1 - p_i(u_j, t) & p_i(u_j, t) \\ 0 & 1 \end{pmatrix}, \; t = 1, \ldots, n_i, j = 1, \ldots, M, i = 1, \ldots, N, \tag{2}$$

where each row corresponds to a hidden binary state occurring at time $t-1$, and each column corresponds to a hidden state occurring at time t, respectively. In this article, we assume that the transition probabilities can differ over locations and times (ie, a nonstationary Markov chain). For notational simplicity, we drop the index i, j and t from $a_{\delta'\delta}$ above and in the remaining article as long as the context is subject to no confusion.

2.2.1 | Logistic autoregression model

We propose a logistic autoregressive model for the transition probability $p_i(u_j, t)$ in that the current state is only dependent on the lag-1 historical states of neighbors:

$$\text{logit}\{p_i(u_j, t)\} = \boldsymbol{X}_i^T \boldsymbol{\beta} + \sum_{j' \in \mathcal{N}_{ijt}^0} \eta_{0jj'} S_i^*(u_{j'}, t-1) + \sum_{j' \in \mathcal{N}_{ijt}^1} \eta_{1jj'} S_i^*(u_{j'}, t-1) \;\text{ for }\; t > 0 \tag{3}$$

for logit $(x) = \log\{x/(1-x)\}$, where $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$ are the $p \times 1$ vectors of subject-specific independent variables (eg, demographic characteristics or clinical information) and their regression coefficients, respectively. Also $S_i^*(u_{j'}, t-1)$ for $\forall j' \neq j$ denotes the centered autocovariates of other muscle conditions at different locations j' from location j of interest at time $t-1$, defined by $S_i^*(u_{j'}, t-1) = S_i(u_{j'}, t-1) - 0.5$. This centering on the autocovariates will make their spatial influences balanced by converting 0 and 1 into −0.5 and 0.5, respectively. Without centering, only impaired neighboring muscle conditions (ie, $S_i(u_{j'}, t-1) = 1$) would affect the transition while healthy neighbors (ie, $S_i(u_{j'}, t-1) = 0$) would not contribute to the odds of transition. In reality, both healthy and impaired neighbors are expected to affect the transition probability. This strategy of centering autocovariates is commonly used in the literatures.[12,13]

For model (3), we separate the effects of autocovariates based on their previous states through two types of neighbors:

$$\mathcal{N}_{ijt}^0 = \{j' | S_i(u_{j'}, t-1) = 0 \text{ and } j' \in \mathcal{N}_j\},$$
$$\mathcal{N}_{ijt}^1 = \{j' | S_i(u_{j'}, t-1) = 1 \text{ and } j' \in \mathcal{N}_j\},$$

where $\mathcal{N}_j$ contains both types of neighbors of u_j, denoted by $\mathcal{N}_j = \{j' | u_{j'} \sim u_j\}$. Here $\mathcal{N}_j$ can be specified as a complete network among locations such that $\mathcal{N}_j = \{j' | u_{j'} \neq u_j\}$, where every pair of muscles are connected each other; or alternatively, it can be specified based on certain prior knowledge about their spatial dependencies among the states (see also Appendix A for the schematic diagram of example networks). For the ALS disease, we use the complete network because all muscles are collectively controlled by nerve cells in the brain and the spinal cord. This complete network results in two unknown autoregressive coefficient vectors $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ of size $M(M-1)$, consisting of parameters $\{\eta_{0jj'}\}_{j \neq j'}$ and $\{\eta_{1jj'}\}_{j \neq j'}$ respectively. Coefficient $\eta_{0jj'}$ represents the effect of a previously healthy neighbor $u_{j'}$ on u_j, and similarly, coefficient $\eta_{1jj'}$ represents the effect of a previously impaired neighbor $u_{j'}$ on u_j. Moreover, we do not restrict the autocovariates to have any symmetry effects but allow to have directed impacts on the transition probability in model (3);[14] for example, the effect of state u_j on the transition probability of $u_{j'}$ can differ from the effect of state $u_{j'}$ on the transition probability of state u_j; that is, $\eta_{0jj'} \neq \eta_{0jj'}$ or $\eta_{1jj'} \neq \eta_{1j'j}$ for any $j \neq j'$.

In addition, we model the initial state probability, $p_i(u_j, 0) = P(S_i(u_j, 0) = 1)$, using the standard logistic regression without autocovariates such that $\text{logit}\{p_i(u_j, 0)\} = \boldsymbol{X}_i^T \boldsymbol{\alpha}_j$ for each location j. Alternatively, any marginal proportion of impaired states can be used, for example, $p_i(u_j, 0) = \sum_{i=1}^N S_i(u_j, 0)/N$ for each j, or the most simply, one can set $p_i(u_j, 0) = 0.5$ for all i and j.

2.2.2 | Bias-corrected L_1-regularized likelihood estimation

Provided that the true states $S_i(u_j, t)$ are known, we seek to estimate the coefficients in model (3), denoted by $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\eta}_0^{\mathrm{T}}, \boldsymbol{\eta}_1^{\mathrm{T}})^{\mathrm{T}}$. Note that we begin with the complete network assuming all other locations have potential effects on a given location. It is desirable to impose sparsity regularization on both $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ such that we can identify those neighboring muscle groups from all available other muscle groups that are truly relevant to a given muscle group. By selecting variables properly, we avoid over-fitting in estimation and improve the interpretation of the model. Therefore, we propose to estimate the model parameters by maximizing the regularized likelihood function.

Under the assumption of HMM being a Markov process of order 1 in time, the hidden states with $t > 0$ are conditionally independent given their previous states of neighbors. Thus, for given initial hidden states $\{S_i(u_j, 0)\}$, the full log-likelihood of $\{S_i(u_j, t); t > 0\}$ is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{t=1}^{n_i} \log\{\mathcal{L}_{ijt}(\boldsymbol{\theta})\}\, I\left(S_i(u_j, t-1) = 0\right),$$

where $I(\cdot)$ denotes an indicator function and $\mathcal{L}_{ijt}(\boldsymbol{\theta})$ denotes a conditional density of $S_i(u_j, t)$ whose previous state was healthy (ie, $S_i(u_j, t-1) = 0$) as

$$\mathcal{L}_{ijt}(\boldsymbol{\theta}) = P\left(S_i(u_j, t) | \boldsymbol{\theta}; \boldsymbol{X}_i, \{S_i(u_{j'}, t-1)\ \forall j' \neq j\}\right) = p_i(u_j, t|\boldsymbol{\theta})^{S_i(u_j,t)} \{1 - p_i(u_j, t|\boldsymbol{\theta})\}^{1-S_i(u_j,t)}$$

with $p_i(u_j, t|\boldsymbol{\theta})$ being a function of $\boldsymbol{\theta}$ as in model (3). Note that any $S_i(u_j, t)$ whose previous state was impaired (ie, $S_i(u_j, t-1) = 1$) does not contribute to the likelihood due to the absorbing feature of disease.

We propose to penalize the coefficient estimation via the bias-corrected least absolute shrinkage and selection operator (LASSO), which enables to select an optimal subset of autocovariates from a large number of pairwise links between locations. Not only does this approach correct the bias of the classical LASSO estimator but also enables post-selection inference based on asymptotic normality.[7,8] To proceed, we first maximize the L_1-penalized log-likelihood,

$$F_\lambda(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \lambda \sum_{j \neq j'} (|\eta_{0jj'}| + |\eta_{1jj'}|), \tag{4}$$

where the solution $\widehat{\boldsymbol{\theta}}_\lambda = \arg_{\boldsymbol{\theta}} \max F_\lambda(\boldsymbol{\theta})$ is the classical LASSO estimator. The tuning parameter $\lambda > 0$, which encourages the amount of sparsity in the estimation solution, can be tuned by a data-dependent model selection criterion, such as generalized cross-validation (GCV),[15] Bayesian information criterion (BIC)[16] and extended Bayesian information criterion (EBIC).[17] We then correct the bias of the LASSO estimator $\widehat{\boldsymbol{\theta}}_\lambda$ as follows:

$$\tilde{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_\lambda + \left\{ -\ell''(\widehat{\boldsymbol{\theta}}_\lambda) \right\}^{-1} \ell'(\widehat{\boldsymbol{\theta}}_\lambda), \tag{5}$$

where $-\ell''(\boldsymbol{\theta}) = -\partial^2 \log \ell(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^2$ is the Hessian matrix and $\ell'(\boldsymbol{\theta}) = \partial \log \ell(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ is the vector of normal scores. It has been shown[7,8,18] that, under some regularity conditions, the bias-corrected estimator in (5), $\tilde{\boldsymbol{\theta}}$, asymptotically behaves as the oracle maximum likelihood estimator obtained by assuming the nonzero set of true parameters is known in advance, which is consistent and asymptotically normally distributed.

In implementation, under model (3), we form the outcome vector of previously healthy states and the design matrix of independent variables and autocovariates in suitable forms, so that we can make a direct use of standard software packages to facilitate easy computation, rather than designing fully new optimization algorithms for objective function (4) from scratch. Appendix C1 provides an R function that organizes independent variables and spatiotemporal binary data into, respectively, a set of the outcome vector (`y`) and the design matrix of covariates (`X`). The resulting data formats can then be used for the function `glmnet(x=X, y=y,…)` in the R package `glmnet`, where the option `penalty.factor` controls the penalization. For example, one can set `penalty.factor=c(`$\mathbf{0}_p^{\mathrm{T}}, \mathbf{1}_{M(M-1)}^{\mathrm{T}}, \mathbf{1}_{M(M-1)}^{\mathrm{T}}$`)` for the complete network $\mathcal{N}_j = \{j' | u_{j'} \sim u_j\}$ where 0 corresponds to the case of $\lambda = 0$, namely no penalty is imposed to $\boldsymbol{\beta}$, and 1 for L_1-penalty to all $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$. For a partial network, one can use `Inf` to avoid estimating certain $\eta_{0jj'}$ and $\eta_{1jj'}$. The function `cv.glmnet()` in

Algorithm 1. Estimation of the proposed hidden Markov model

```
Input independent variables $\boldsymbol{X}_i$, observed states $Y_i(u_j, t)$, and network structure $\mathcal{N}_j$ for $\forall i, j, t$
Set initial hidden states as $S_i^{(0)}(u_j, 0) = Y_i(u_j, 0)$ and $S_i^{(0)}(u_j, t) = \max\{Y_i(u_j, t), S_i^{(0)}(u_j, t-1)\}$ for $\forall i, j, t(>0)$
Start with $q = 0$
repeat
    $q \longleftarrow q + 1$
    Estimate emission probability matrix $\widehat{\boldsymbol{E}}^{(q)}$ as (1) of Section 2.1, using
        $\widehat{e}_\delta^{(q)}(\gamma) \longleftarrow \sum_{i,j,t} I\big(Y_i(u_j, t) = \gamma \mid S_i^{(q-1)}(u_j, t) = \delta\big) \big/ \sum_{i,j,t} I\big(S_i^{(q-1)}(u_j, t) = \delta\big)$ for $\delta, \gamma \in \{0, 1\}$
    Estimate transition probability matrix $\widetilde{\boldsymbol{A}}_{ijt}^{(q)}$ for each $i, j, t$ as (2) of Section 2.2 using
        $\tilde{p}_i^{(q)}(u_j, t) = \mathrm{logit}^{-1}\big\{\boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}}^{(q)} + \sum_{u_{j'} \in \mathcal{N}_{ijt}^0} \widetilde{\eta}_{0jj'}^{(q)} S_i^{*(q-1)}(u_{j'}, t-1) + \sum_{u_{j'} \in \mathcal{N}_{ijt}^1} \widetilde{\eta}_{1jj'}^{(q)} S_i^{*(q-1)}(u_{j'}, t-1)\big\}$
        where $\tilde{\boldsymbol{\theta}}^{(q)} = (\tilde{\boldsymbol{\beta}}^{(q)\mathrm{T}}, \tilde{\boldsymbol{\eta}}_0^{(q)\mathrm{T}}, \tilde{\boldsymbol{\eta}}_1^{(q)\mathrm{T}})^{\mathrm{T}}$ are bias-corrected as in (5) for $\widehat{\boldsymbol{\theta}}^{(q)} \longleftarrow \arg_{\boldsymbol{\theta}} \max F_\lambda(\boldsymbol{\theta} \mid \boldsymbol{X}_i, S_i^{(q-1)}(u_j, t)$ for $\forall i, j, t)$
    Update hidden states for each $i$ and $j$ as in Section 2.3,
        $\boldsymbol{S}_{ij}^{(q)} = \{S_i^{(q)}(u_j, 0), \ldots, S_i^{(q)}(u_j, n_i)\} \longleftarrow \arg\max_{\delta \in \{0,1\}} \nu_\delta(n_i \mid \widehat{\pi}_{1ij}^{(q)}, \widehat{\boldsymbol{E}}^{(q)}, \widetilde{\boldsymbol{A}}_{ijt}^{(q)})$
        where $\widehat{\pi}_{1ij}^{(q)} = \widehat{P}(S_i^{(q-1)}(u_j, 0) = 1 \mid \boldsymbol{X}_i)$ is the predicted marginal probability for each $i$ and $j$
until $|\widehat{e}^{(q)} - \widehat{e}^{(q-1)}| < \epsilon_e$ and $|\tilde{\boldsymbol{\theta}}^{(q)} - \tilde{\boldsymbol{\theta}}^{(q-1)}| < \epsilon_\theta$ for small enough $\epsilon_e$ and $\epsilon_\theta$, respectively.
```

the same package is used to select the tuning parameter λ using K-fold GCV method with the default value $K = 10$. Also Appendix C2 provides an R function that computes the bias-corrected estimates (5) and their standard errors.

2.3 | Hidden states by Viterbi algorithm

Given the observed sequences $\boldsymbol{Y}_{ij} = \{Y_i(u_j, 1), \ldots, Y_i(u_j, n_i)\}^{\mathrm{T}}$, we obtain the probability distribution of hidden states, $\boldsymbol{S}_{ij} = \{S_i(u_j, 1), \ldots, S_i(u_j, n_i)\}^{\mathrm{T}}$, by using the Viterbi algorithm.[10] Unlike other techniques such as the forward algorithm and the forward-backward algorithm, the Viterbi algorithm finds the entire sequence of hidden states at once, rather than a single hidden state at a time.[5] The detail of this algorithm is given below.

For the sequence of observed states up to t recorded as $\{Y_i(u_j, 1), \ldots, Y_i(u_j, t)\}^{\mathrm{T}} = (\gamma_1, \ldots, \gamma_t)^{\mathrm{T}}$ for $\gamma_1, \ldots, \gamma_t \in \{0, 1\}$, we define the joint probability of the most probable sequence of hidden states up to t with the ending state $S_i(u_j, t) = \delta$ for $\delta \in \{0, 1\}$ as

$$v_\delta(t) = \max_{S_i(u_j,1), \ldots, S_i(u_j,t-1)} P\left(S_i(u_j, 1), \ldots, S_i(u_j, t-1),\ S_i(u_j, t) = \delta,\ Y_i(u_j, 1) = \gamma_1, \ldots, Y_i(u_j, t) = \gamma_t\right), \tag{6}$$

for each subject i and location j. This probability (6) can be recursively calculated as

$$v_\delta(t) = \max_{\delta' \in \{0,1\}} v_{\delta'}(t-1)\, a_{\delta'\delta}\, e_\delta(\gamma_t) \tag{7}$$

with $v_\delta(1) = \pi_\delta e_\delta(\gamma_1)$ at $t = 1$ where $\pi_\delta = P\left(S_i(u_j, 0) = \delta\right)$ denotes the marginal probability of initial hidden state being $\delta \in \{0, 1\}$ (a.k.a. the prior probability or starting probability); $a_{\delta'\delta} = P\left(S_i(u_j, t) = \delta | S_i(u_j, t-1) = \delta'\right)$ denotes the transition probability of hidden states changing from δ' to δ as in Section 2.2; and $e_\delta(\gamma_t) = P\left(Y_i(u_j, t) = \gamma_t | S_i(u_j, t) = \delta\right)$ denotes the emission probability of the observed state $Y_i(u_j, t) = \gamma_t$ given the hidden state, $S_i(u_j, t) = \delta$, as in Section 2.1.

We then estimate the hidden states by finding the most likely sequence of hidden states such that $\widehat{\boldsymbol{S}}_{ij} = \arg\max_{\delta \in \{0,1\}} v_\delta(n_i)$ for each i and j, given the estimated probabilities, $\widehat{\pi}$, $\widehat{a}$, and $\widehat{e}$. The function `viterbi()` in the R package `HMM` can be used to implement the above algorithm.

Finally, we repeat the three-step procedure in Sections 2.1 to 2.3 until all parameter estimates for the probabilities are converged. The iterative algorithm with a reasonable convergence criterion produces solutions very close to the theoretical roots.[19] The initial set of hidden states can be defined as the same as the observed states, except for a correction to account for the absorbing feature of the impaired states. Specifically, the healthy states following the impaired states are forced to be always the impaired states, that is, $S_i(u_j, 0) = Y_i(u_j, 0)$ and $S_i(u_j, t) = \max\{Y_i(u_j, t), S_i(u_j, t-1)\}$ for $t > 0$ for all i and j at the first iteration. Algorithm 1 shows the initial set up and the detailed following steps for the entire procedure including the estimation of the model parameters described in the previous sections.

3 | DATA APPLICATION

3.1 | ALS patients' data

We analyzed data from the EMPOWER Study, a double-blind and placebo-controlled phase III clinical trial on dexpramipexole in ALS patients.[11] As there is no treatment effect, data from the two arms of the EMPOWER study are pooled together in this research. There are a total of $N = 926$ patients, 18 to 80 years old with first symptom onset 24 months or less before study entry. Their muscle strengths at $M = 16$ body locations were measured at study entry and every two months thereafter for up to 12 months ($t = 0, 2, 4, 6, 8, 10, 12$). The measured muscles were the right and left side of wrist extensors (WRSTEXT), elbow extensors (ELBEXT), elbow flexors (ELBFLEX), shoulder flexors (SHDFLEX), ankle dorsiflexors (ANKLDOR), knee flexors (KNEFLEX), knee extensors (KNEEXT) and hip flexors (HIPFLEX), as illustrated in Figure 2. Each patient's demographical characteristics, such as age, sex, and weight, and clinical records, such as symptom onset site and symptom duration, were also recorded.

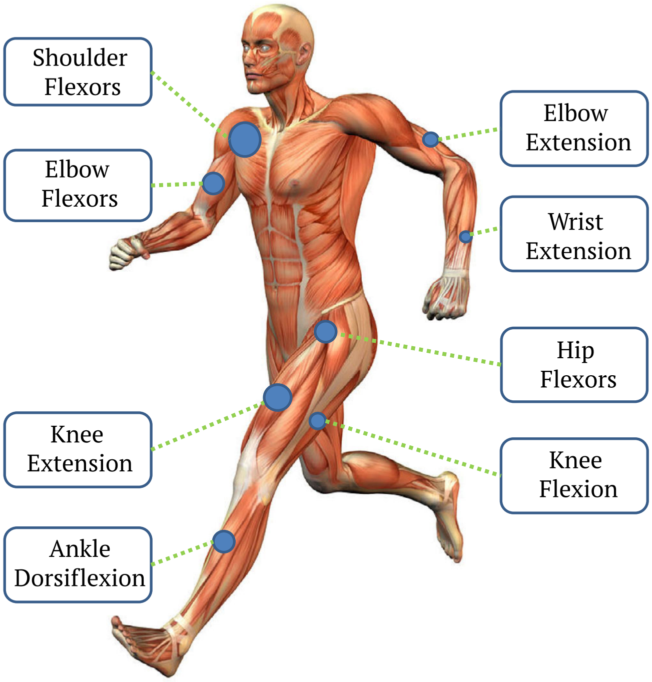


FIGURE 2 Muscles measured for the ALS patients in EMPOWER Study; the right and left sides of eight pairs of muscle groups (16 muscles in total) are examined [Colour figure can be viewed at wileyonlinelibrary.com]

The muscle strength data are dichotomized into states of "healthy" or "impaired" by comparing a patient's muscle strengths to the strengths expected when he or she were healthy. Here, the expected strengths were computed based on their gender, age, height, and weight following the results of previous studies on muscle strengths of healthy people.[20,21] We then assessed each muscle's disease state as impaired (=1) if its strength is 40% less than the computed expected strength and healthy (=0) otherwise. Figure 3 shows the trajectories of observed muscle strengths and their binary states for a selected patient, who could not control the left ankle muscle at all at the initial visit. Although the ALS disease is incurable, some muscles seemed to gain some strength over time (eg, the elbow flexors at $t = 8$ and the right knee extension at $t = 8$ in Figure 3). This is probably due to measurement errors. In this analysis, we considered such dichotomized muscle strength data as the observed states, which may not be absorbing at the impaired state, and assumed that these observed states depend on the hidden disease states, which are absorbing at the impaired state.

3.2 | Setup

We included the visit time (t), the symptom onset site (binary variable whether it is on bulbar (=1) or others (=0)), and the symptom duration when patients entered the study for the independent variables, $\boldsymbol{X}$. Note that patients' biological information such as age, sex, height, and weight were not included in $\boldsymbol{X}$ because they were used to compute their expected muscle strengths, which were the reference for the data dichotomization as described in Section 3.1.

We assumed the complete network among muscles, $\mathcal{N}_j = \{j' | u_{j'} \neq u_j\}$ for any fixed $j = 1, \ldots, 16$, so that any group of muscles (any set of autocovariates) can affect any other muscles (outcome), regardless of their physical distances. As a result, each vector of $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ had the length $16 \times 15 = 240$, and the selected subset of this complete network would be estimated by the bias-corrected LASSO.

We estimated the marginal probability of initial hidden states, π_δ, for each muscle u_j at every iteration using the standard logistic regression (`glm()` in R), where the outcomes are $\{S_1(u_j, 0), \ldots, S_N(u_j, 0)\}^{\mathrm{T}}$ and the independent variables include the onset site and the symptom duration that are recorded at the initial visit ($t = 0$).

At the first iteration of Algorithm 1, we obtained the optimal $\widehat{\lambda}$ as the largest value of pre-specified sequence of λ's such that error is within 1-standard error of the λ with the minimum of deviances computed by 10-fold cross-validation (ie, `lambda.1se` of `cv.glmnet` object), and we fixed $\widehat{\lambda}$ at this optimal value for the subsequent iterations such that $\widehat{\lambda}^{(q)} = \widehat{\lambda}$ for $\forall q > 1$. The iterations were stopped when all estimates had only little change as much as less than 5% relative difference.

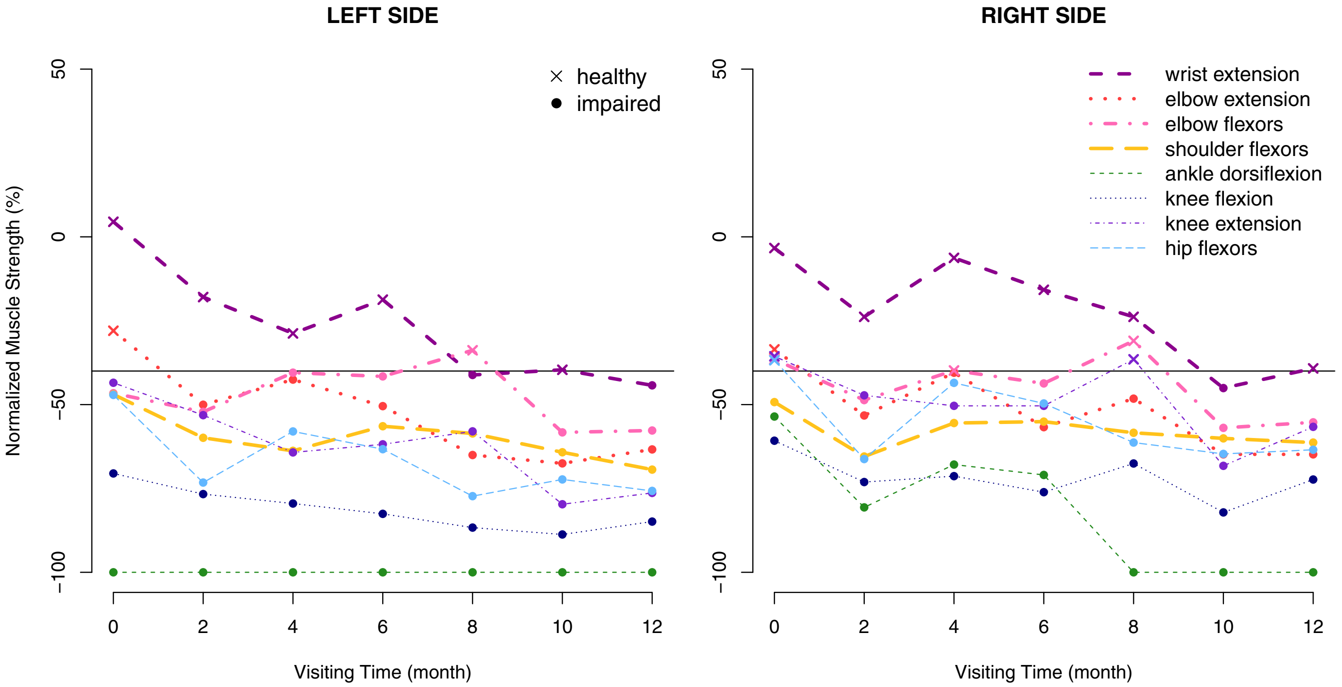


FIGURE 3 Normalized muscle strength measurements, (observed strength − expected strength)/expected strength ×100%, for a selected patient; binary states are obtained as "impaired (= 1)" if strength is 40% less than healthy people's expected strength, drawn by horizontal solid lines, and "healthy (= 0)" otherwise [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Estimated β-coefficients of the autologistic model for the ALS patients' data from EMPOWER Study

X	$\hat{\beta}$	95% Confidence Interval	P-value
Intercept	−0.841	(−1.000, −0.682)	<0.001
Visiting Time (t)	−0.023	(−0.037, −0.010)	<0.001
Onset Site (bulbar or others)	0.083	(−0.017, 0.184)	0.103
Symptom Duration Before Entry	−0.009	(−0.017, −0.001)	0.037

3.3 | Result summary

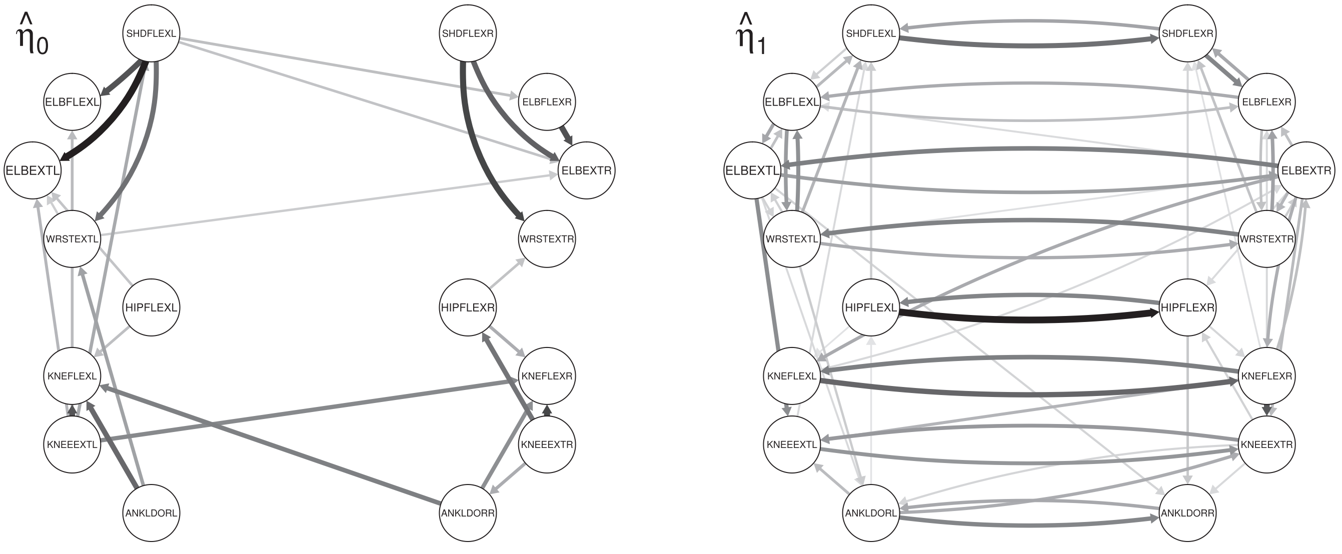
3.3.1 | Emission probability

The probabilities of misclassification between the observed and hidden states, which are the elements of the emission probability matrix $\boldsymbol{E}$, were estimated as $\hat{e}_0(1) = 863/23493 = 0.0367$ (false positive rate) with 95% confidence interval, $(0.0344, 0.0392)$, and $\hat{e}_1(0) = 1783/50539 = 0.0353$ (false negative rate) with 95% confidence interval, $(0.0337, 0.0369)$. Note that these results were based on 40% cut-off for dichotomized muscle strength data as described in Figure 3.

3.3.2 | Transition probability

The transition probability of disease states are modeled as in (3), where the β-estimates are summarized in Table 1 and the η-estimates are illustrated in Figure 4 (see also Tables B1 and B2 for the numerical results with 95% confidence intervals calculated using the bias-corrected LASSO inference).

The estimates of $\boldsymbol{\beta}$ can be better interpreted with the predicted probability values computed as $\text{logit}^{-1}(\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{\beta}})$. For example, the estimated overall probability of disease progression with no contributions from independent variables or autocovariates was $\text{logit}^{-1}(-0.841) \approx 0.30$, which is reasonably low, and this would decrease over time because $\hat{\beta} = -0.023$ for the visiting time with $p < 0.001$. This implies that the individual muscles would stay healthy if there were no



(A) $\boldsymbol{\eta}_0$: effects of previously healthy muscles

(B) $\boldsymbol{\eta}_1$: effects of previously impaired muscles

FIGURE 4 Estimated η-coefficients of the autologistic model for the ALS patients' data from the EMPOWER Study; the circle nodes with abbreviated muscle names followed by "R"(right) or "L"(left) are drawn in their relative positions on a human body; the arrows show the direction of effects from autocovariate muscles to outcome muscles ($u_{j'} \rightarrow u_j$); the arrows with darker and wider edges indicate stronger effects

inter-muscle lag-1 spatial dependency and other risk factors. However, a patient who entered the study shortly after the first symptom tends to have a higher probability of disease progression as $\hat{\beta} = -0.009$ for the symptom duration variable with $p = 0.037$.

The estimates of $\boldsymbol{\eta}_0$ in Figure 4A indicate the autocovariate effects of the muscles when they were previously healthy. Specifically, $\hat{\eta}_{0jj'} > 0$ indicates that muscle u_j is likely to stay healthy when muscle $u_{j'}$ was previously healthy (here, positive estimates indicate the negative impacts on the probability of transition from healthy to impaired because autocovariate terms, $S_i(u_j, t-1) = 0$, are centered to the negative terms, $S_i^*(u_j, t-1) = -0.5$). For example, if the left shoulder muscle was healthy at $t-1$, the elbow and wrist muscles on the same side would likely be healthy at t ($\hat{\eta}_{0jj'} = 1.91$ and 1.32 for $u_j =$ ELBEXTL and WRSTEXTL, respectively, and $u_{j'} =$ SHDFLEXL). Similarly, the knee flexor muscles tend to stay healthy if the hip and ankle muscles were previously healthy ($\hat{\eta}_{0jj'} = 1.39$ and 0.55 for $u_j =$ KNEFLEXL and $u_{j'} =$ ANKLDORL and HIPFLEXL, respectively).

The estimates of $\boldsymbol{\eta}_1$ in Figure 4B indicate the autocovariate effects of the muscles when they were previously impaired. Specifically, $\hat{\eta}_{1jj'} > 0$ indicates that muscle u_j is likely to get impaired when muscle $u_{j'}$ was previously impaired. It was remarkable that the every muscle had strong effect to its opposite side (right to left, or left to right) despite being physically far apart. For example, if the left hip flexor was at the impaired state at $t-1$, the right hip flexor would be the most likely infected at t, and vice versa ($\hat{\eta}_{1jj'} = 4.01$ and $\hat{\eta}_{1j'j} = 2.34$ for $u_j =$ HIPFLEXR and $u_{j'} =$ HIPFLEXL).

Comparing $\hat{\boldsymbol{\eta}}_0$ and $\hat{\boldsymbol{\eta}}_1$, the muscles had different effects depending on their previous states; for example, the shoulder muscles had stronger effects on the elbow and wrist muscles when they were healthy than when they were impaired (eg, $\hat{\eta}_{0jj'} = 1.56 > \hat{\eta}_{1jj'} = 0.83$ for $u_j =$ SHDFLEXL and $u_{j'} =$ ELBFLEXL). The previously healthy muscles had directional effects ($\hat{\eta}_{0jj'} \neq \hat{\eta}_{0j'j}$ for $j \neq j'$) while the previously impaired muscles mostly had bidirectional effects ($\hat{\eta}_{1jj'} \approx \hat{\eta}_{1j'j}$ for $j \neq j'$). Also the impaired muscles had a denser network of autoregressive effects than that of the healthy muscles ($\sum_{j \neq j'} I(\eta_{0jj'} > 0) = 26 < \sum_{j \neq j'} I(\eta_{1jj'} > 0) = 61$).

3.3.3 | Hidden states

Figure 5 shows the time sequence of the hidden states, estimated by the Viterbi algorithm as in Section 2.3, for the observed states of a selected subject and muscle; $\hat{\boldsymbol{S}}_{ij}$ and $\boldsymbol{Y}_{ij}$ for a selected i and j. As expected, the estimated sequence refined

FIGURE 5 The time sequence of the observed and hidden states for a selected subject and muscle

visiting time (t)	0		2		4		6		8		10		12
hidden states	**0**	→	**0**	→	**0**	→	**1**	→	**1**	→	**1**	→	**1**
↓	↓		↓		↓		↓		↓		↓		↓
observed states	**1**		**0**		**0**		**1**		**1**		**0**		**1**

the observed sequence to follow the absorbing features of the disease; $Y_i(u_j, 0) = 1$ turned out to be a false positive with $S_i(u_j, 0) = 0$, and $Y_i(u_j, 10) = 0$, followed by two 1's, turned out to be a false negative with $S_i(u_j, 10) = 1$.

4 | DISCUSSION

We developed a HMM-based approach to delineate the spreading patterns of muscle weakness in patients with ALS, where the probability of one-step transition of hidden spatio-temporal binary processes is modeled by the logistic autoregression in which neighboring locations' previous states are included as autocovariates. The proposed model is flexible in that it can describe disease progression by allowing different effects conditioning on previous states, asymmetric and directional effects between locations, and a data-driven approach to determining actual neighborhood structures through a complete network structure of muscles. Along with the estimation of the emission probability for examining the accuracy of ALS disease diagnosis, the estimated most probable hidden states can help clinicians better understand ALS patients' underlying disease conditions.

Our research indicates that muscle weakness in ALS mostly spreads from previously impaired muscles to the contralateral sides, followed by within the upper left/right muscle groups and within the lower left/right groups. Our finding also suggests that healthy muscles, particularly for shoulder muscles, can slow down the spreading, although their influence is not as strong as that of impaired muscles. The estimated mis-classification error rates were less than 4%, indicating that it is reasonable to diagnose a muscle's disease status by comparing to 40% of healthy people's expected strength.

The proposed methodology can be, with some straightforward modifications, applicable to a range of spatiotemporal binary processes to study spreading patterns. ALS disease serves as a motivating example to illustrate the usefulness of our approach. Basically, our model requires three general components in the model specification: (i) an observed binary spatiotemporal process of interest; (ii) the existence of an unknown complex network (maybe more complex than a purely spatial distance-based network) that dictates the disease/event spreading; and (iii) the inclusion of absorbing states in the hidden process. Examples include crop virus spreading in farm land, and virus attacks of computer networks. Depending on specific real-world problems, our model may also be further extended (i) to allow estimate the emission probability through a regression model with some covariates, instead of simple empirical frequencies considered in this article, and (ii) to allow multi-states for observed or hidden states in which the dimension of emission or transition probability matrix will be large.

Other penalty methods can be useful to achieve sparsity. Examples include adaptive LASSO,[22] smoothly clipped absolute deviation (SCAD),[23] minimax concave penalty (MCP),[24] and reciprocal L_1-regularization (rlasso).[25] These are all undergone the L_1-norm penalization, and thus produce shrinkage in estimation with biases (possibly smaller biases) similar to the LASSO estimator. In this type regularized estimation, bias correction is a necessary step to obtain an asymptotically distributed inferential quantity. Other methods of high-dimensional inferences such as sample splitting[26] and bootstrap[27] are potential alternatives.

We previously proposed an autologistic network model (ANM) for spatiotemporal binary data of disease progression in ALS.[18] The HMM-based model proposed in this article differs from the previous ANM mainly in two aspects. First, ANM ignores potential mis-classification and treats the observed states as the hidden states in the modeling. Second, ANM focuses on the modeling of autoregression in the space domain for current observed states and only uses limited information from previous states for classifying neighbors, while the HMM model proposed in this article captures the lag-1 spatial dependence of current and previous hidden states under the Markovian assumption and is hence more suitable for predicting muscle's future states. In like manner, our approach methodologically differs from other HMM-based spatiotemporal data analyses[28-30] in that we estimated autoregressive spatial dependency of hidden processes with absorbing states by maximizing L_1-penalized likelihood.

The proposed Algorithm 1 is along the line of the expectation-maximization (EM) algorithm, where we iteratively compute the posterior expected values of hidden state-sequences given observed data via the Viterbi algorithm (E-step), while maximizing the likelihood with each hidden state substituted by the corresponding expected value (M-step). Different from the EM algorithm proposed by Bartolucci et al,[31] our algorithm allows to estimate transition probabilities with the constraint of absorbing states and to learn the network with sparsity regularization via the bias-corrected LASSO method.

A future study of interest would be to explore the continuous measurements of muscle strengths rather than the dichotomized data to retain more information. One can assume that the loss of muscle strength is monotone, as assumed in this study, and that the impaired state of hidden sequence is absorbing. Another further analysis of the data would be to consider a autoregressive model with time varying coefficients. Such a model can explore the time-course dynamic change in the network structures as the disease progresses.

ACKNOWLEDGMENTS

We gratefully acknowledge that this research is financially supported by Biogen. Dr. Shin's work was done while she was a PhD student at the Department of Statistics in Texas A&M University as a part of the dissertation. Huiyan Sang's research was partly supported by NSF DMS–1737885 and NSF DMS–1854655. Peter Song's research was partly supported by NSF DMS-1811734.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Yei Eun Shin https://orcid.org/0000-0002-0739-1281

REFERENCES

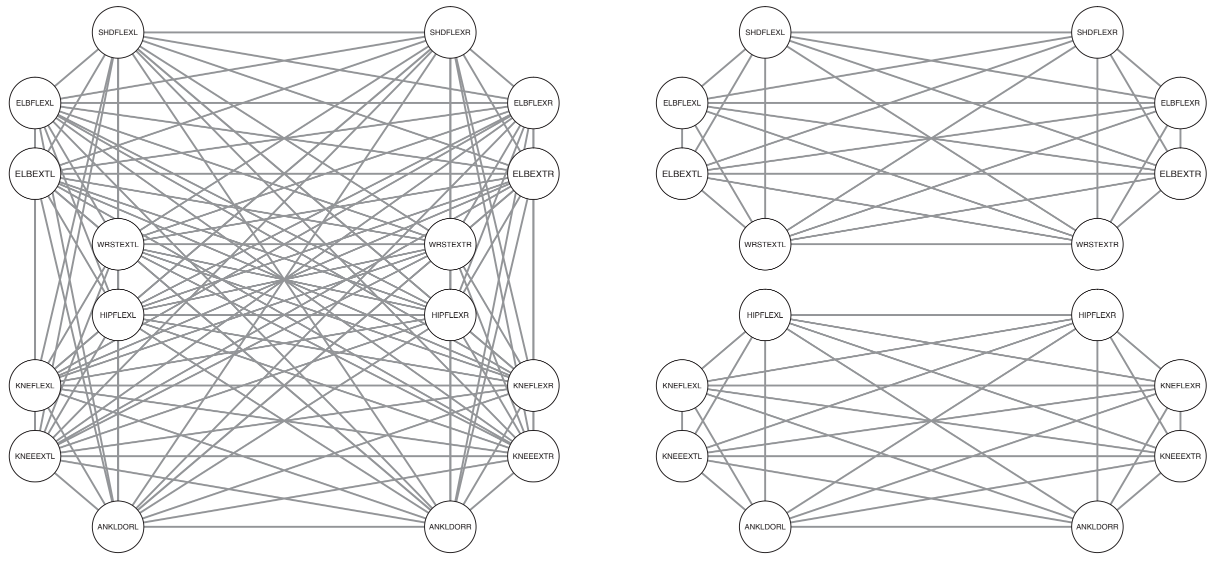
1. Besag J. Spatial interaction and the statistical analysis of lattice systems. *J Royal Stat Soc Ser B (Methodol)*. 1974;36(2):192-236.
2. Besag J. Statistical analysis of non-lattice data. *Statistician*. 1975;24(3):179-195.
3. Zhu J, Huang HC, Wu J. Modeling spatial-temporal binary data using Markov random fields. *J Agric Biol Environ Stat*. 2005;10(2):212-225.
4. Hughes J, Haran M, Caragea PC. Autologistic models for binary data on a lattice. *Environmetrics*. 2011;22(7):857-871.
5. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE*. 1989;77(2):257-286.
6. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc Ser B (Methodol)*. 1996;58(1):267-288.
7. Geer S, Bühlmann P, Ritov Y'a, Dezeure R. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann Stat*. 2014;42(3):1166-1202.
8. Tang L, Zhou L, Song PX-K. Distributed simultaneous inference in generalized linear models via confidence distribution. *J Multivar Anal*. 2020;176:104567.
9. Wang Z. *Analysis of Binary Data via Spatial-Temporal Autologistic Regression Models*. Theses and Dissertations-Statistics. Lexington, KY: University of Kentucky; 2012;3.
10. Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans Inf Theory*. 1967;13(2):260-269.
11. Cudkowicz ME, Berg LH, Shefner JM, et al. Dexpramipexole versus placebo for patients with amyotrophic lateral sclerosis (EMPOWER): a randomised, double-blind, phase 3 trial. *Lancet Neurol*. 2013;12(11):1059-1067.
12. Caragea PC, Kaiser MS. Autologistic models with interpretable parameters. *J Agric Biol Environ Stat*. 2009;14(3):281-300.
13. Hughes J, Haran M. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *J Royal Stat Soc Ser B (Stat Methodol)*. 2013;75(1):139-159.
14. Agaskar A, Lu YM. ALARM: a logistic auto-regressive model for binary processes on networks. Paper presented at: Proceedings of the 2013 IEEE Global Conference on Signal and Information Processing. Austin, TX; 2013:305-308. https://doi.org/10.1109/GlobalSIP.2013.6736876.
15. Golub GH, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*. 1979;21(2):215-223.
16. Schwarz G. Estimating the dimension of a model. *Ann Stat*. 1978;6(2):461-464.
17. Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*. 2008;95(3):759-771.
18. Shin YE, Sang H, Liu D, Ferguson TA, Song PXK. Autologistic network model on binary data for disease progression study. *Biometrics*. 2019;75(4):1310-1320.
19. Nocedal J, Wright S. *Numerical Optimization*. Berlin, Germany: Springer Science & Business Media; 2006.
20. NIMS Database of the National Isometric Muscle Strength. Muscular weakness assessment: use of normal isometric strength data. *Arch Phys Med Rehabil*. 1996;77(12):1251-1255.
21. Bohannon RW. Reference values for extremity muscle strength obtained by hand-held dynamometry from adults aged 20 to 79 years. *Arch Phys Med Rehabil*. 1997;78(1):26-32.
22. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101(476):1418-1429.
23. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96(456):1348-1360.
24. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat*. 2010;38(2):894-942.

25. Song Q, Liang F. High-dimensional variable selection with reciprocal l1-regularization. *J Am Stat Assoc*. 2015;110(512):1607-1620.
26. Wasserman L, Roeder K. High dimensional variable selection. *Ann Stat*. 2009;37(5A):2178.
27. Zhang X, Cheng G. Simultaneous inference for high-dimensional linear models. *J Am Stat Assoc*. 2017;112(518):757-768.
28. Bertarelli G, Ranalli G, Bartolucci F, D'Alò M, Solari F. Small area estimation for unemployment using latent Markov models. *Survey Methodol*. 2018;44(2):167-192.
29. Wei Z, Li H. A hidden spatial-temporal Markov random field model for network-based analysis of time course gene expression data. *Ann Appl Stat*. 2008;2(1):408-429.
30. Chen S, Langley J, Chen X, Hu X. Spatiotemporal modeling of brain dynamics using resting-state functional magnetic resonance imaging with Gaussian hidden Markov model. *Brain Connect*. 2016;6(4):326-334.
31. Bartolucci F, Pandolfi S, Pennoni F. LMest: an R package for latent Markov models for longitudinal categorical data. *J Stat Softw*. 2017;81(1):1-38.

How to cite this article: Shin YE, Liu D, Sang H, Ferguson TA, Song PXK. A binary hidden Markov model on spatial network for amyotrophic lateral sclerosis disease spreading pattern analysis. *Statistics in Medicine*. 2021;40:3035–3052. https://doi.org/10.1002/sim.8956

APPENDIX A. NETWORK SPECIFICATION

Figure A1A shows the complete network, $\mathcal{N}_j = \{j' | u_{j'} \sim u_j\}$, where every pair of muscles is connected to each other. Figure A1B shows the partial network where all is connected except between arm and leg muscles. We used the complete network for the ALS data analysis in Section 3.



(A) Complete network

(B) Partial network

FIGURE A1 Example networks

APPENDIX B. NUMERICAL RESULTS OF η-ESTIMATES

Tables B1 and B2 show the numerical results of the estimates of $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$, respectively, which were illustrated in Figure 4. The 95% confidence intervals for non-zero estimates are also included.

TABLE B1 The estimates of $\eta_{0jj'}$, the effect of $u_{j'}$ on u_j when $u_{j'}$ is previously healthy, for the ALS patients data from EMPOWER Study; the 95% confidence intervals for the non-zero estimates are in the parentheses; NA denotes not available; "." denotes zero

u_j	WRSTEXT		ELBEXT		ELBFLEX		SHDFLEX		ANKLDOR		KNEEEX		KNEFLEX		HIPFLEX	
$u_{j'}$	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R
WRSTEXTL	NA	.	0.42	0.43	.	.	.	.	.	.	.	.	.	.	.	.
			(0.03,0.81)	(0.03,0.84)												
R	.	NA	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ELBEXTL	.	.	NA	.	.	.	.	.	.	.	.	.	.	.	.	.
R	.	.	.	NA	.	.	.	.	.	.	.	.	.	.	.	.
ELBFLEXTL	.	.	.	.	NA	.	.	.	.	.	.	.	.	.	.	.
R	.	.	.	1.58	.	NA	.	.	.	.	.	.	.	.	.	.
			(1.19,1.97)													
SHDFLEXL	1.32	.	1.91	0.52	1.56	0.57	NA	.	.	.	.	.	.	.	.	.
	(0.91,1.73)		(1.47,2.34)	(0.07,0.97)	(1.18,1.95)	(0.12,1.01)										
R	.	1.67	.	1.46	.	.	.	NA	.	.	.	.	.	.	.	.
		(1.26,2.08)		(1.02,1.90)												
ANKLDORL	0.83	.	.	.	.	.	.	.	NA	.	.	.	1.39	.	.	.
	(0.36,1.31)												(0.94,1.83)			
R	.	.	.	.	.	.	.	.	.	NA	.	.	1.16	1.01	.	.
													(0.73,1.59)	(0.57,1.46)		
KNEEXTL	.	.	0.64	.	0.59	.	0.75	.	.	.	NA	.	1.56	1.08	.	.
			(0.05,1.23)		(0.02,1.17)		(0.20,1.30)						(1.03,2.08)	(0.55,1.61)		
R	.	.	.	.	.	.	.	.	.	0.73	.	NA	.	1.66	.	1.28
										(0.18,1.28)				(1.15,2.17)		(0.69,1.87)

(Continues)

TABLE B1 (Continued)

u_j	WRSTEXT		ELBEXT		ELBFLEX		SHDFLEX		ANKLDOR		KNEEEX		KNEFLEX		HIPFLEX	
$u_{j'}$	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R
KNEFLEXL	.	.	.	.	.	.	.	.	.	.	.	.	NA	.	.	.
R	.	.	.	.	.	.	.	.	.	.	.	.	.	NA	.	.
HIPFLEXL	.	.	0.51 (0.04,0.98)	.	.	.	.	.	.	.	.	.	0.55 (0.11,0.99)	.	NA	.
R	.	0.53 (0.05,1.00)	.	.	.	.	.	.	.	.	.	.	.	0.73 (0.25,1.21)	.	NA

Abbreviations: WRSTEXT, wrist extension; ELBEXT, elbow extension; ELBFLEX, elbow flexors; SHDFLEX, shoulder flexors; ANKLDOR, ankle dorsiflexion; KNEEXT, knee extension; KNEFLEX, knee flexion; HIPFLEX, hip flexors; L, left; R, right.

TABLE B2 The estimates of $\eta_{1jj'}$, the effect of $u_{j'}$ on u_j when $u_{j'}$ is previously impaired, for the ALS patients data from EMPOWER Study; the 95% confidence intervals for the non-zero estimates are in the parentheses; NA denotes not available; "." denotes zero

u_j	WRSTEXT		ELBEXT		ELBFLEX		SHDFLEX		ANKLDOR		KNEEEX		KNEFLEX		HIPFLEX	
$u_{j'}$	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R
WRSTEXTL	NA	1.57	0.84	0.55	2.09	.	1.37	.	0.91	.	.	.	.	.	.	.
		(1.16,1.97)	(0.45,1.23)	(0.14,0.95)	(1.66,2.53)		(0.83,1.91)		(0.32,1.5)							
R	2.41	NA	.	0.90	.	1.76	.	1.09	.	.	.	.	.	.	.	.
	(1.99,2.83)			(0.52,1.28)		(1.24,2.29)		(0.51,1.68)								
ELBEXTL	0.80	.	NA	1.81	1.14	.	.	.	0.66	0.77	2.18	.	.	.	.	.
	(0.29,1.30)			(1.39,2.22)	(0.55,1.73)				(0.13,1.20)	(0.22,1.32)	(1.30,3.05)					
R	.	1.34	2.47	NA	0.63	1.14	.	.	.	.	.	1.25	1.57	1.49	.	0.65
		(0.84,1.84)	(2.02,2.93)		(0.09,1.18)	(0.39,1.9)						(0.38,2.13)	(1.11,2.03)	(0.99,1.98)		(0.10,1.19)
ELBFLEXTL	1.89	.	1.66	.	NA	1.16	1.28	.	.	.	.	.	.	.	.	.
	(1.43,2.34)		(1.26,2.06)			(0.60,1.72)	(0.65,1.92)									
R	.	1.12	.	.	1.56	NA	.	1.61	.	.	.	.	.	.	.	.
		(0.72,1.52)			(1.14,1.99)			(1.08,2.14)								
SHDFLEXL	.	.	.	.	0.83	.	NA	2.78	.	.	.	.	.	.	.	.
					(0.44,1.22)			(2.28,3.27)								
R	.	.	.	.	.	2.31	1.71	NA	.	.	.	.	.	.	.	.
						(1.92,2.70)	(1.27,2.14)									
ANKLDORL	.	.	.	.	.	.	.	.	NA	2.32	1.23	1.54	.	.	0.47	.
										(1.89,2.76)	(0.63,1.84)	(0.94,2.14)			(0.04,0.9)	

(Continues)

TABLE B2 (Continued)

u_j	WRSTEXT		ELBEXT		ELBFLEX		SHDFLEX		ANKLDOR		KNEEEX		KNEFLEX		HIPFLEX	
$u_{j'}$	L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R
R	.	.	.	.	.	.	.	.	1.62	NA	.	.	.	.	.	.
									(1.17,2.07)							
KNEEXTL	.	.	.	.	.	.	0.66	.	.	.	NA	1.98	.	.	.	.
							(0.13,1.19)					(1.41,2.56)				
R	.	.	.	.	.	.	.	.	0.75	0.68	1.94	NA	.	.	.	0.80
									(0.13,1.36)	(0.14,1.22)	(1.33,2.56)					(0.22,1.39)
KNEFLEXL	.	.	.	0.69	.	.	.	.	.	.	.	.	NA	2.96	.	.
				(0.27,1.11)										(2.47,3.45)		
R	.	.	.	1.18	.	0.69	.	0.53	.	.	1.39	3.22	2.47	NA	.	.
				(0.76,1.61)		(0.20,1.18)		(0.02,1.04)			(0.24,2.54)	(1.49,4.95)	(2.10,2.84)			
HIPFLEXL	.	.	.	.	.	.	0.90	.	.	.	.	.	0.49	.	NA	4.01
							(0.41,1.4)						(0.05,0.93)			(3.54,4.48)
R	.	.	.	.	.	.	.	0.74	.	0.99	.	.	.	0.72	2.34	NA
								(0.22,1.27)		(0.45,1.54)				(0.24,1.20)	(1.88,2.81)	

Abbreviations: WRSTEXT, wrist extension; ELBEXT, elbow extension; ELBFLEX, elbow flexors; SHDFLEX, shoulder flexors; ANKLDOR, ankle dorsiflexion; KNEEXT, knee extension; KNEFLEX, knee flexion; HIPFLEX, hip flexors; L, left; R, right.

APPENDIX C. R CODES

C.1. Data transformation

We provide R function `designXy()` that transforms independent variables and two-dimensional spatiotemporal binary data for a subject into the vector of outcomes and the design matrix of covariates. The argument `DAT` is input data of $n_i \times (p+M)$ matrix, and `names_independent` and `names_location` are the names of p columns of independent variables and M columns of locations, respectively. The returning outputs include a column of the outcome vector, with length of $n_i M_i^0$, and columns of the design matrix of covariates, with dimension of $n_i M_i^0 \times (p + 2M(M-1))$ where M_i^0 is the number of previously healthy locations for subject i. The stacked outputs for all subjects, $i = 1, \ldots, N$, can then be used for the function `glmnet()` in the R package `glmnet`.

```
designXy = function(DAT, names_independent, names_location){
 y = X = NULL
 tempY = as.matrix(DAT[, names_location]); colnames(tempY) = NULL
 nloc = length(names_location)
 N = nrow(DAT)
 if (N < 2) { result = NULL }
 if (N >= 2){
    for (ntime in 2:N){
        if(sum(DAT[ntime, names_location] == DAT[ntime-1, names_location]) ==
          nloc)next
        Xb = do.call("rbind", replicate(nloc, DAT[ntime, names_independent],
          simplify = F))
        I0 = matrix(0, nloc, nloc*(nloc-1));
        I1 = matrix(0, nloc, nloc*(nloc-1));
        S = matrix(0, nloc, nloc*(nloc-1));
        for(i in 1:nloc){
           I0[i, (1+(i-1)*(nloc-1)):(i*(nloc-1))] = (tempY[ntime-1,]==0)[-i]
           I1[i, (1+(i-1)*(nloc-1)):(i*(nloc-1))] = (tempY[ntime-1,]==1)[-i]
           S[i, (1+(i-1)*(nloc-1)):(i*(nloc-1))] = (tempY[ntime-1,])[-i]
        }
        activeid = (tempY[ntime-1, ]==0);
        if (sum(activeid)==0) next
        X = rbind(X, (as.matrix(cbind(Xb, (S-.5)*I0, (S-.5)*I1)))[activeid,,drop =
          F])
        y = c(y, tempY[ntime,,drop = F][activeid])
    }
    result = cbind(Y = y, X = X)
    if(!is.null(result) & !is.null(DAT$SUBJID)) {
    rownames(result) = rep(DAT$SUBJID[1], nrow(result))
    }
 }
 return(result)
}
```

C.2. Bias correction

We provide R function `correct_bias()` for correcting the bias of the classical LASSO estimates, for example, the estimates obtained from `glmnet()`. The argument `est` is the vector of LASSO estimates, and `X` and `y` are the design matrix and the outcome vector respectively that were used to obtain `est`. The resulting matrix includes bias-corrected estimates with standard errors.

```
correct_bias = function(est, X, y){

    invlink = function(x){exp(x)/(1 + exp(x))};
    invlinkdiv = function(x){exp(x)/(1 + exp(x))^2};

    evec = function(beta, X, y){ y - sapply(drop(X%*%beta), invlink)
    pmat = function(beta, X, y){ out = sapply(drop(X%*%beta), invlinkdiv);
     diag(out)

    n = length(y); p = length(est)
    P = pmat(est, X, y)
    e = evec(est, X, y)

    Sigmahatinv = crossprod(X, P%*%X); Sigmahat = ginv(Sigmahatinv)

    A = tcrossprod(Sigmahat, X)
    est.corrected = drop(est + AA%*%e)
    sehat = sqrt(diag(Sigmahat))

    return(cbind(est = est.corrected, se = sehat))
}
```