SUPPLEMENTARY MATERIALS: The Convex Mixture Distribution: Granger Causality for Categorical Time Series*

Alex Tank† ‡, Xiudi Li† §, Emily B. Fox¶, and Ali Shojaie‖

SM1. Experiments.

SM1.1. mLTD Bach Analysis. For the mLTD Bach analysis, we performed a 5-fold cross validation to select the tuning parameter λ, then thresholded the final connection weights, given by the standardised L_2 norm of $\mathbf{Z}^{ij}$, at .01, as in the MTD case. First, we note that with only 5 total zero weights the final mLTD model is much less sparse than the MTD model. We display the final graph in Figure SM1, where, for interpretability, we bold edges with total weight greater than .45. In this graph there are strong connections in the counter-clockwise direction between G#, C#, F#, and B. However, the other connections on the circle of fifths are relatively weaker, and there are many more connections between notes far away on the circle of fifths. The mLTD graph also shows that the chord note both affects and is affected by many harmony notes. Furthermore, we see that the bass category is effected by most harmony notes as well. Overall, however, this graph is much less interpretable than the MTD graph and fails to find the full circle of fifths structure.

SM1.2. iEEG Segmentation. To segment the iEEG time series into a sequence of categorical states, we use a Markov switching autoregressive model. The model assumes that each channel in the d-dimensional EEG signal, $\mathbf{y}_t \in \mathbb{R}^d$, follows a Markov switching univariate autoregressive process (AR) each with the same m dynamic regimes. Specifically, let $\mathbf{a}^1, \ldots, \mathbf{a}^m$, where $\mathbf{a}^i = \left(a_1^i, \ldots, a_h^i\right)$, denote the lag h $\mathbf{AR(h)}$ parameters for each of the m dynamic regimes and let x_{jt} be the latent m-dimensional categorical state that governs the dynamics for channel j at time t. The model assumes that y_{jt} follows a locally stationary $\mathbf{AR(h)}$ model with m state dynamics:

$$y_{jt} = \sum_{l=1}^{h} a_k^{x_{jt}} y_{j(t-l)} + e_{jt}, \tag{SM1.1}$$

where the lag l AR dynamics at time t, $\mathbf{a}^{x_{jt}}$, are indexed by the latent state, x_{jt}, and e_{jt} is mean zero Gaussian noise independent across series, $E\left(e_{jt}\right) = 0$ and $E\left(e_{jt}e_{j't'}\right) = 0$ for all $(j,t) \neq (j',t')$. The transitions between dynamic regimes are assumed to evolve independently

†Denotes equal contribution.

*Submitted to the editors April 6th, 2020.

Funding: This work was partially funded by grants from the National Science Foundation (CAREER IIS-1350133, DMS-1161565, DMS-1561814) and the National Institutes of Health (R01GM114029).

‡The Voleon Group, Berkeley, CA (alextank@uw.edu)

§Department of Biostatistics, University of Washington, Seattle WA (xiudil@uw.edu)

¶Departments of Computer Science & Engineering and Statistics, University of Washington, Seattle WA (ebfox@uw.edu)

‖Department of Biostatistics, University of Washington, Seattle WA (ashojaie@uw.edu)

between series according to a hidden Markov model. See [SM11] for more details on the model. Due to the long length of the series, we use a stochastic gradient MCMC algorithm [SM7] to fit the model with $m = 5$ categorical states. We display the segmentation of a single channel using this method in Figure SM2.

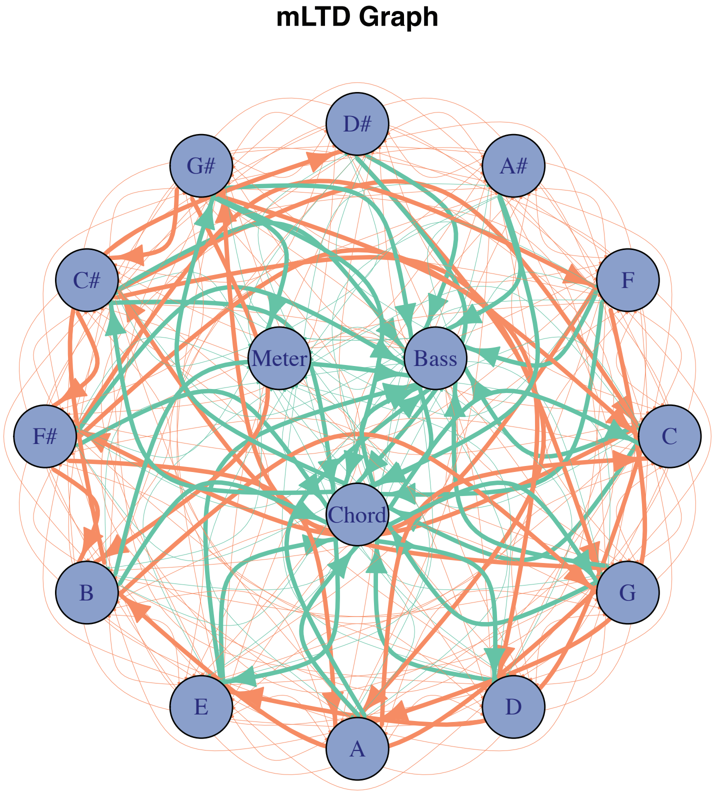

Figure SM1. *The Granger causality graph for the 'Bach Choral Harmony' data set using the mLTD method. The harmony notes are displayed around the edge in a circle corresponding to the circle of fifths. Orange links display directed interactions between the harmony notes while green links display interactions to and from the 'bass', 'chord', and 'meter' variables.*

SM1.3. Additional Simulation Results.

Figure SM3 compares the signal strengths in the mLTD and MTD models for the case where each series has $m = 4$ possible states and $d = 15$. To capture the effect of time series j on time series i, we unfold the transition probability tensor $p(x_{it}|x_{1(t-1)}, \ldots, x_{d(t-1)})$ along the mode defined by $x_{j(t-1)}$, and obtain an $m \times m^d$ matrix. We then compute the l_2 distances between any two rows of the resulting matrix. For the MTD model, this is equivalent (up to scaling) to the l_2 distance between columns of $\mathbf{Z}^{ij}$, since the effect is additive. We repeat this procedure for all (i, j) pairs and aggregate the results over 20 replications. Figure SM3 shows a histogram of nonzero signals in the MTD and mLTD models.

We observe that, in our simulation settings, the difference among transition probabilities in the mLTD model is larger than that in the MTD model, leading to stronger connections.

Next, we present median ROC curves over 20 replications for the proposed methods, under

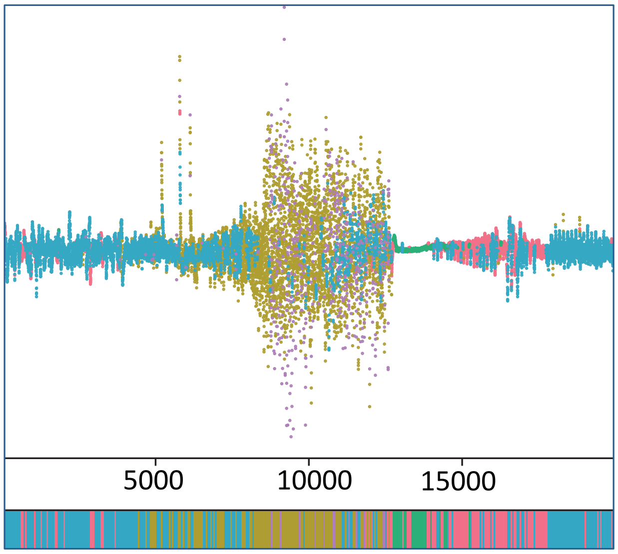

Figure SM2. *Colored segmentation with $m = 5$ states of a single iEEG channel during a seizure using the Markov switching autoregressive model.*

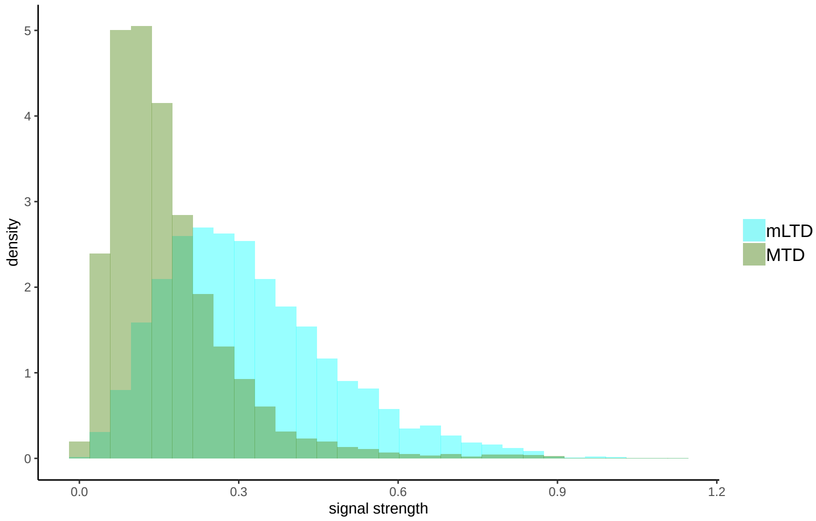

Figure SM3. *Signal strengths in the mLTD and MTD models.*

different simulation settings. The results displayed in Figures SM4-SM5, Figures SM6-SM7 and Figures SM8-SM9, correspond to data generated by MTD, mLTD and latent VAR models, respectively. We observe that for all three methods, the performance improves with increasing sample size T and worsens with increasing dimension d.

We also show the points on the ROC curves that correspond to tuning parameter values chosen by BIC and cross-validation. In general, cross-validation tends to over-select Granger causality relationships. This highlights the importance of thresholding when using cross-validation in practice. In contrast, BIC generally gives an overly sparse model when sample size is small; but it performs much better with large sample sizes.

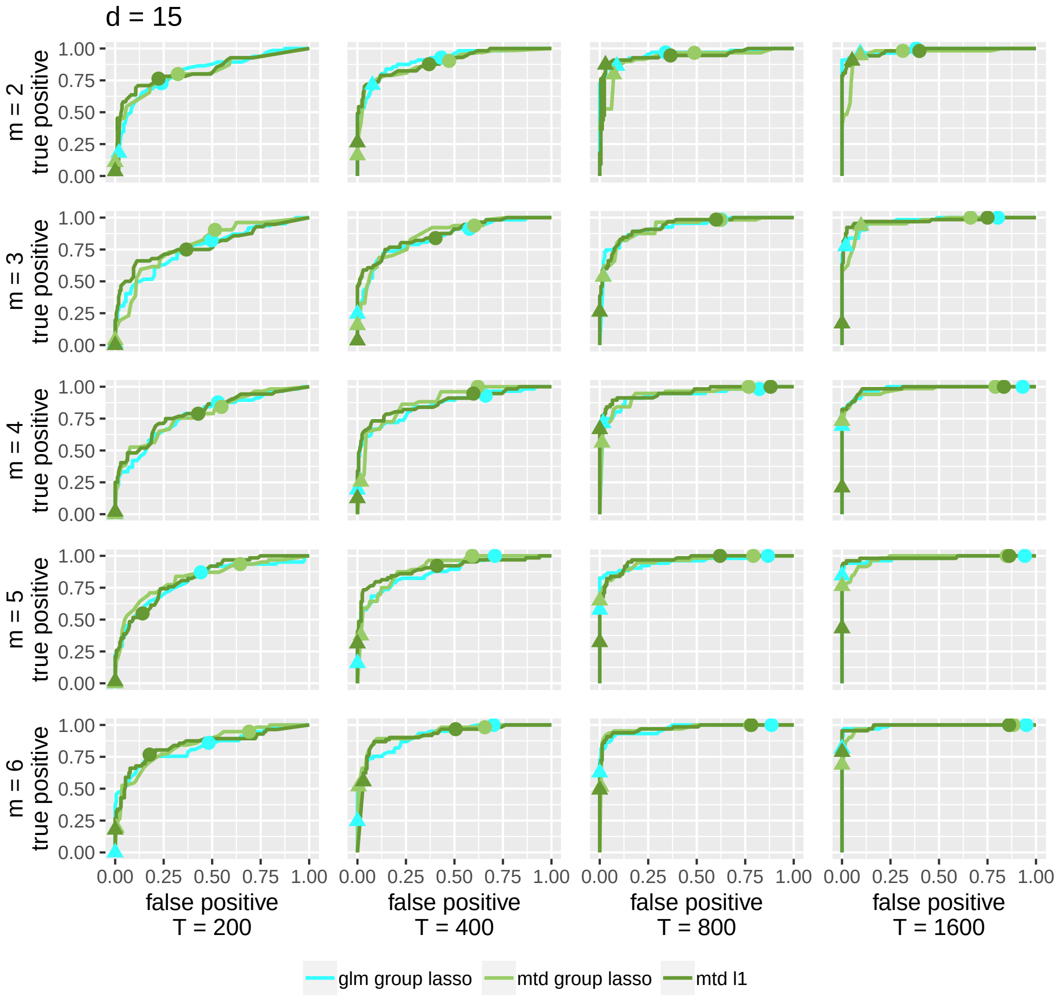

Figure SM4. *Median ROC curves over 20 simulation runs, for data generated by a sparse MTD process with $d = 15$. Triangles correspond to tuning parameter values chosen by BIC; while dots correspond to the values chosen by cross-validation.*

Finally, in Figure SM10, we show the average run time of the three proposed methods under different sample size T and number of time series d, where each time series has 4 categories. We observe that in general mLTD group lasso runs faster than MTD with either group lasso or lasso penalty. This is due to the constraints on the parameter set in the MTD model, which requires additional projection steps. For all three methods, the run time scales

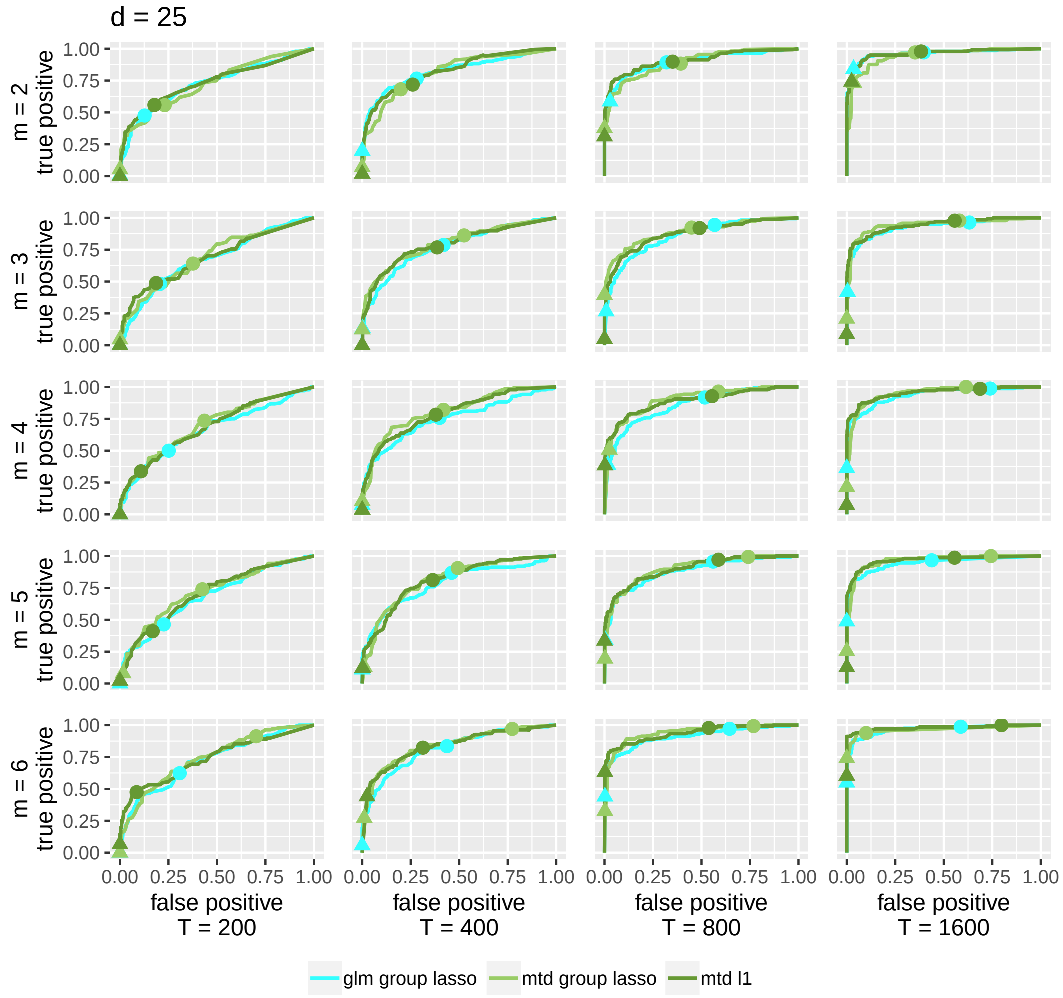

Figure SM5. *Median ROC curves over 20 simulation runs, for data generated by a sparse MTD process with $d = 25$. Triangles correspond to tuning parameter values chosen by BIC; while dots correspond to the values chosen by cross-validation.*

nearly linearly in sample size.

SM2. Proofs of Results in Section 3.

Proof of Proposition 3.3. If the columns of $\mathbf{Z}^j$ are all equal, then for all fixed values of $x_{\setminus j(t-1)}$ the conditional distribution is the same for all values of $x_{j(t-1)}$. If one column is different, then the conditional distribution for all values of $x_{\setminus j(t-1)}$ will depend on $x_{j(t-1)}$.

To prove the second claim, we let $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ be two parameterizations for the same MTD model. Suppose that they give different causality conclusions. Then, there exists some $j \in \{1, \ldots, d\}$ such that the columns of $\mathbf{Z}^j$ are all equal, while the columns of $\tilde{\mathbf{Z}}^j$ are not, or the other way around. There must thus exist a row where at least two columns differ in this row.

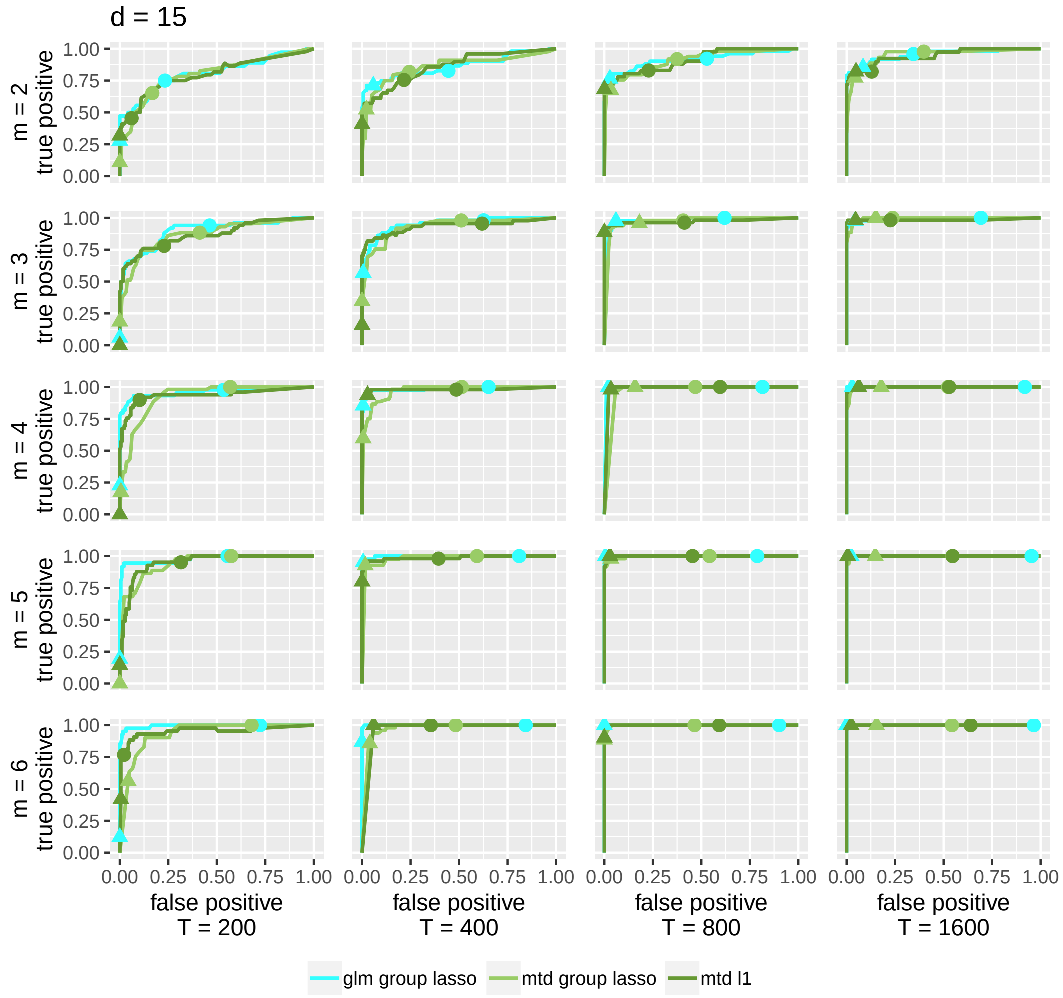

Figure SM6. *Median ROC curves over 20 simulation runs, for data generated by a sparse mLTD process with $d = 15$. Triangles correspond to tuning parameter values chosen by BIC; while dots correspond to the values chosen by cross-validation.*

Without loss of generality, we assume that $\mathbf{Z}^1_{11} \neq \mathbf{Z}^1_{12}$ but $\tilde{\mathbf{Z}}^1_{11} = \tilde{\mathbf{Z}}^1_{12}$. Then under $\mathbf{Z}$, we have that

$$P\left(x_{it} = 1 | x_{1(t-1)} = 1, x_{2(t-1)}, \ldots, x_{d(t-1)}\right) \neq P\left(x_{it} = 1 | x_{1(t-1)} = 2, x_{2(t-1)}, \ldots, x_{d(t-1)}\right).$$

However, under $\tilde{\mathbf{Z}}$ we have that

$$P\left(x_{it} = 1 | x_{1(t-1)} = 1, x_{2(t-1)}, \ldots, x_{d(t-1)}\right) = P\left(x_{it} = 1 | x_{1(t-1)} = 2, x_{2(t-1)}, \ldots, x_{d(t-1)}\right).$$

This is a clear contradiction, as $\tilde{\mathbf{Z}}$ and $\mathbf{Z}$ are different parameterizations of the same model, and hence all conditional probabilities should be the same. ∎

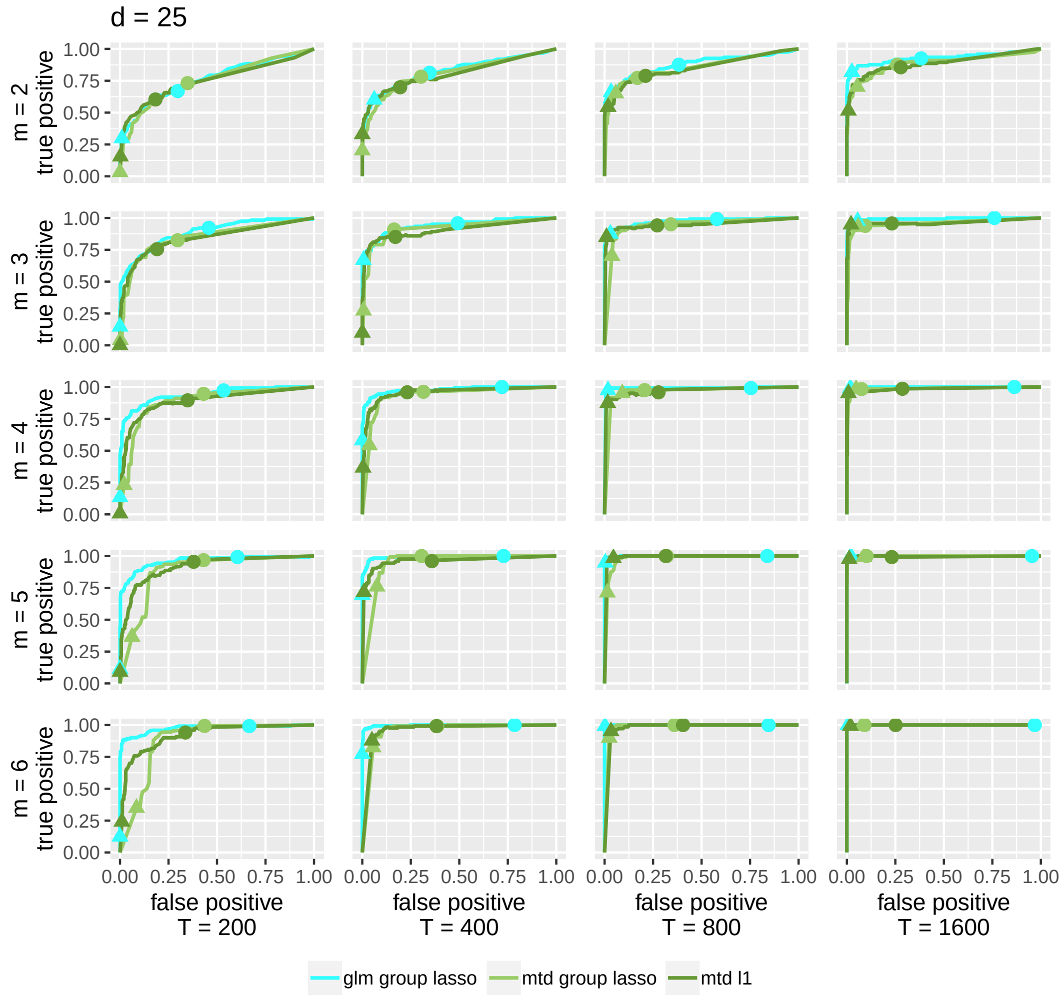

Figure SM7. *Median ROC curves over 20 simulation runs, for data generated by a sparse mLTD process with $d = 25$. Triangles correspond to tuning parameter values chosen by BIC; while dots correspond to the values chosen by cross-validation.*

Proof of Theorem 1. First we show that any parameter set $\mathbf{Z}$ can be converted to another set $\tilde{\mathbf{Z}}$ that contains at least one 0 element in each row of each matrix; and that $\tilde{\mathbf{Z}}$ satisfies the constraints of the MTD model. Let $\mathbf{Z}$ be the parameter set for an MTD model. For each $\mathbf{Z}^j$ let the vector α^j be the minimal element in each row, $\alpha^j_k = \min \mathbf{Z}^j_{k:}$. Let $\tilde{\mathbf{Z}}^j = \mathbf{Z}^j - \alpha_j$ and $\tilde{\mathbf{z}}^0 = \mathbf{z}^0 + \sum_{j=1}^{d} \alpha_j$. This $\tilde{\mathbf{Z}}$ gives the same MTD distribution as $\mathbf{Z}$. Furthermore, this $\tilde{\mathbf{Z}}$ has a zero element in each row of each $\tilde{\mathbf{Z}}^j$ by construction.

The non-negativity constraint is trivially satisfied by $\tilde{\mathbf{Z}}$ as we subtract the minimum in each row. For all j, we have that $\mathbf{1}^T\mathbf{Z}^j = \gamma_j \mathbf{1}^T$. Then $\mathbf{1}^T\tilde{\mathbf{Z}}^j = \mathbf{1}^T(\mathbf{Z}^j - \alpha^j\mathbf{1}^T) = \left(\gamma_j - \mathbf{1}^T\alpha^j\right)\mathbf{1}^T = \tilde{\gamma}_j\mathbf{1}^T$, where we define $\tilde{\gamma}_j = \gamma_j - \mathbf{1}^T\alpha^j$. We note that $\tilde{\gamma}_j \geq 0$ as we subtract the row minimum.

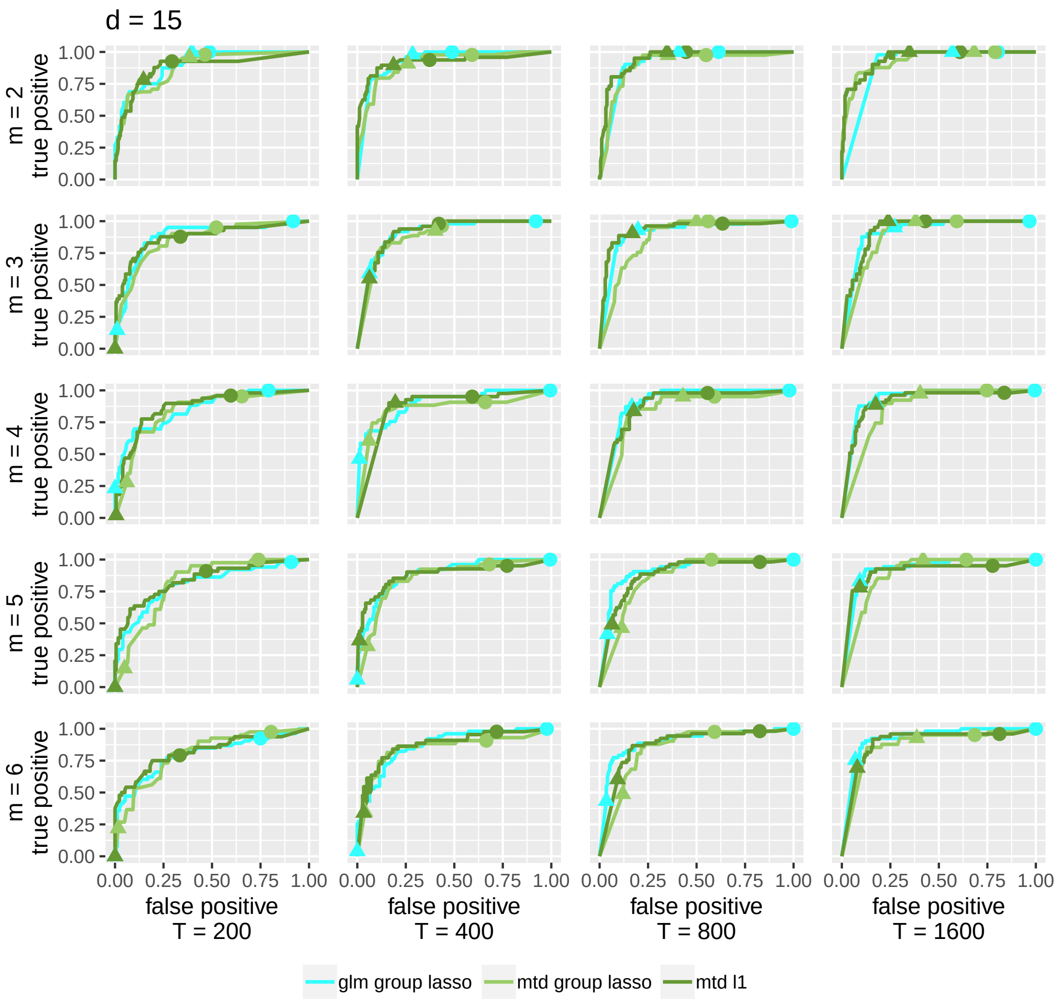

Figure SM8. *Median ROC curves over 20 simulation runs, for data generated by a sparse latent VAR process with $d = 15$. Triangles correspond to tuning parameter values chosen by BIC; while dots correspond to the values chosen by cross-validation.*

Hence within each $\tilde{\mathbf{Z}}^j$, the column sums are all equal. Finally, we have that $\tilde{\gamma}_0 = \gamma_0 + \sum_{j=1}^{d} \mathbf{1}^T \alpha^j$ and $\sum_{j=0}^{d} \gamma_j = 1$, so $\sum_{j=0}^{d} \tilde{\gamma}_j = \gamma_0 + \sum_{j=1}^{d} \mathbf{1}^T \alpha^j + \sum_{j=1}^{d} \left(\gamma_j - \mathbf{1}^T \alpha^j\right) = \sum_{j=0}^{d} \gamma_j = 1$. Hence $\tilde{\gamma}_j$'s sum up to 1.

Next, we show that this new parameter set is uniquely determined. Suppose two parameter sets $\mathbf{X}$ and $\mathbf{Y}$ provide the same MTD distribution. Let $\tilde{\mathbf{X}}$ be as above for $\mathbf{X}$ and $\tilde{\mathbf{Y}}$ of $\mathbf{Y}$.

We use a proof by contradiction. Suppose that $\tilde{\mathbf{Y}} \neq \tilde{\mathbf{X}}$. There must exist some j and some row k such that $\tilde{\mathbf{X}}^j_{k:} \neq \tilde{\mathbf{Y}}^j_{k:}$. Let l_X be the index of the zero element for $\mathbf{X}^j$, i.e., such that $\tilde{\mathbf{X}}^j_{kl} = 0$, and likewise for l_Y . If there are more than one zero elements, pick any. Furthermore, if $\tilde{\mathbf{X}}^j_{k:}$ and $\tilde{\mathbf{Y}}^j_{k:}$ share a zero in the same location (if there are one or more zero elements in

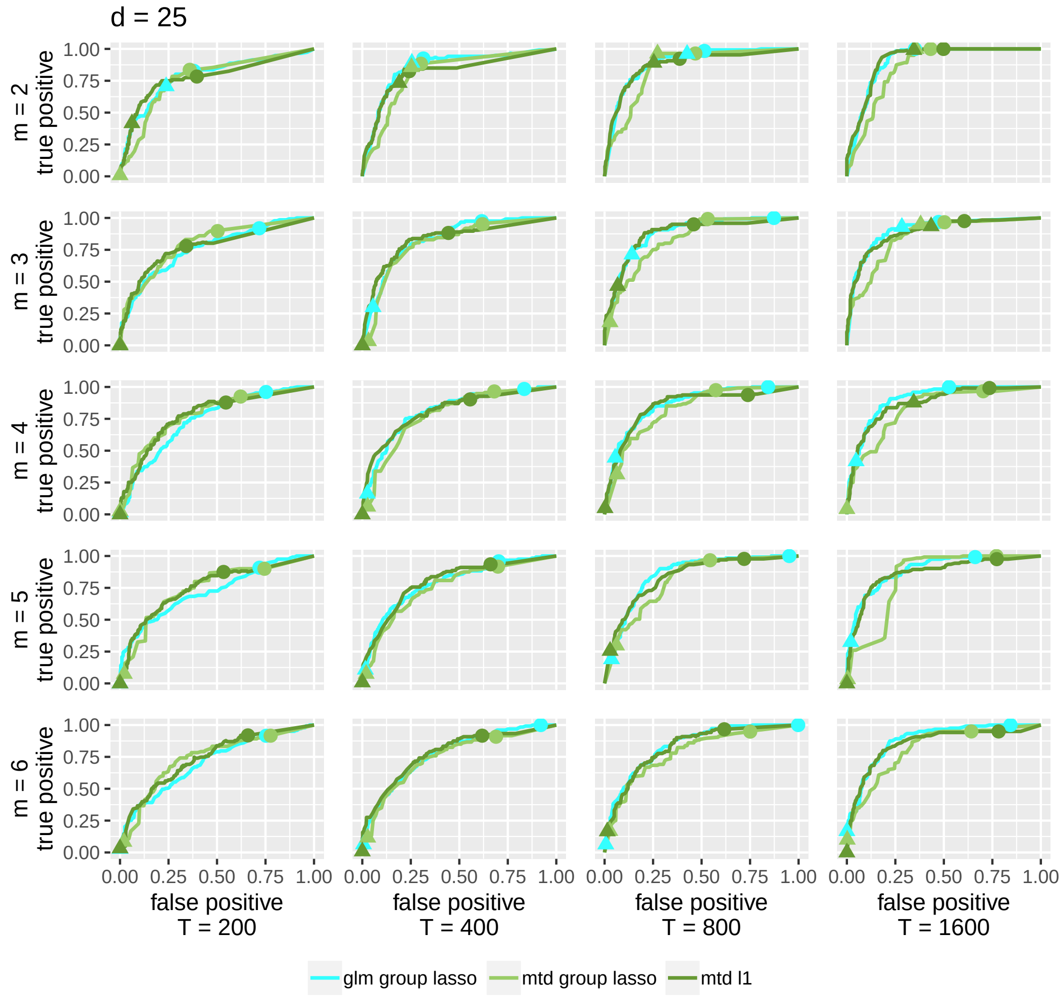

Figure SM9. *Median ROC curves over 20 simulation runs, for data generated by a sparse latent VAR process with $d = 25$. Triangles correspond to tuning parameter values chosen by BIC; while dots correspond to the values chosen by cross-validation.*

each), then let l_X and l_Y be that index so that $l_X = l_Y$.

If $l_X = l_Y$, let l' be an index such that $\tilde{\mathbf{X}}^j_{kl'} \neq \tilde{\mathbf{Y}}^j_{kl'}$. This index must exist by construction. Let the categories of other series (not for series j), $x_{\setminus j(t-1)}$, be fixed arbitrarily. The difference

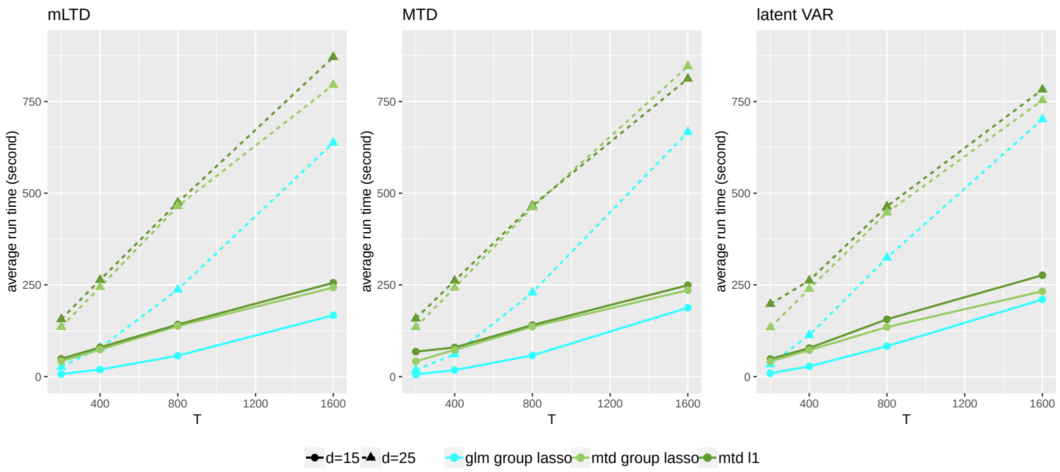

Figure SM10. *Average run time of three proposed methods over 10 replications, with $m = 4$ and $\lambda = 100$ for MTD group lasso, MTD L_1 and $\lambda = 12.5$ for mLTD group lasso.*

between the conditional distributions for $\mathbf{X}$ are

$$\begin{aligned}
\tilde{\mathbf{X}}^j_{kl'} &= \tilde{\mathbf{X}}^j_{kl'} - \tilde{\mathbf{X}}^j_{kl_X} \\
&= \left(\tilde{\mathbf{X}}^j_{kl'} + \alpha_{jk}\right) - \left(\tilde{\mathbf{X}}^j_{kl_X} + \alpha_{jk}\right) \\
&= \mathbf{X}^j_{kl'} - \mathbf{X}^j_{kl_X} \\
&= \left(\mathbf{x}^0_k + \sum_{i \in \setminus j} \mathbf{X}^i_{kx_{i(t-1)}} + \mathbf{X}^j_{kl'}\right) - \left(\mathbf{x}^0_k + \sum_{i \in \setminus j} \mathbf{X}^i_{kx_{i(t-1)}} + \mathbf{X}^j_{kl_X}\right) \\
&= p_X\left(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l'\right) - p_X\left(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l_X\right).
\end{aligned}$$

A similar calculation for $\mathbf{Y}$ shows that

$$\tilde{\mathbf{Y}}^j_{kl'} = p_Y\left(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l'\right) - p_Y\left(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l_Y\right).$$

However, $\tilde{\mathbf{Y}}^j_{kl'} \neq \tilde{\mathbf{X}}^j_{kl'}$, thus showing that

$$\begin{aligned}
p_Y\left(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l'\right) - p_Y\left(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l_Y\right) \neq \\
p_X\left(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l'\right) - p_X\left(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l_X\right).
\end{aligned}$$

This inequality contradicts our assumption that the MTD distributions parametrized by $\mathbf{X}$ and $\mathbf{Y}$ are the same since $l_X = l_Y$.

If $l_X \neq l_Y$, then

$$p_X\left(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l_Y\right) - p_X\left(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l_X\right) = \tilde{\mathbf{X}}^j_{kl_Y},$$

and

$$p_Y\left(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l_Y\right) - p_Y\left(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l_X\right) = -\tilde{\mathbf{Y}}^j_{kl_X}.$$

However, $-\tilde{\mathbf{Y}}^j_{kl_X} \neq \tilde{\mathbf{X}}^j_{kl_Y}$ since at least one of $\tilde{\mathbf{Y}}^j_{kl_X}$ and $\tilde{\mathbf{X}}^j_{kl_Y}$ are nonzero and both are nonnegative. Again, this shows that

$$p_Y\left(x_t = k|x_{\backslash j(t-1)}, x_{j(t-1)} = l_Y\right) - p_Y\left(x_t = k|x_{\backslash j(t-1)}, x_{j(t-1)} = l_X\right) \neq$$
$$p_X\left(x_t = k|x_{\backslash j(t-1)}, x_{j(t-1)} = l_Y\right) - p_X\left(x_t = k|x_{\backslash j(t-1)}, x_{j(t-1)} = l_X\right),$$

which contradicts our assumption that the MTD distributions parametrized by $\mathbf{X}$ and $\mathbf{Y}$ are the same.

The same argument shows that the reduction is unique. ■

Proof of Proposition 3.1. First we check the parameter set satisfies the constraints of MTD model. Since $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are valid MTD parameter sets, we have that $\forall j, \mathbf{1}^T\mathbf{Z}^j = \gamma_j\mathbf{1}^T, \mathbf{Z}^j \geq 0; \mathbf{1}^T\tilde{\mathbf{Z}}^j = \tilde{\gamma}_j\mathbf{1}^T, \tilde{\mathbf{Z}}^j \geq 0$, and $\mathbf{1}^T\gamma = 1, \gamma \geq 0; \mathbf{1}^T\tilde{\gamma} = 1, \tilde{\gamma} \geq 0$. Consider the new parameter set $\alpha\mathbf{Z} + (1-\alpha)\tilde{\mathbf{Z}}$; we have that for all j,

$$\begin{aligned}
&\mathbf{1}^T(\alpha\mathbf{Z}^j + (1-\alpha)\tilde{\mathbf{Z}}^j) \\
&= \alpha(\mathbf{1}^T\mathbf{Z}^j) + (1-\alpha)(\mathbf{1}^T\tilde{\mathbf{Z}}^j) \\
&= (\alpha\gamma_j + (1-\alpha)\tilde{\gamma}_j)\mathbf{1}^T \\
&= \bar{\gamma}_j\mathbf{1}^T,
\end{aligned}$$

where we define $\bar{\gamma}_j = \alpha\gamma_j + (1-\alpha)\tilde{\gamma}_j$ for all j. Then

$$\mathbf{1}^T\bar{\gamma} = \mathbf{1}^T(\alpha\gamma + (1-\alpha)\tilde{\gamma}) = \alpha + (1-\alpha) = 1. \quad \text{(SM2.1)}$$

Finally since $\mathbf{Z}^j, \tilde{\mathbf{Z}}^j, \gamma$ and $\tilde{\gamma}$ are all non-negative, we have that $\alpha\mathbf{Z}^j + (1-\alpha)\tilde{\mathbf{Z}}^j \geq 0 \; \forall j$ and $\bar{\gamma} \geq 0$.

Next we demonstrate that the probability tensor given by this new parameter set is the same as those given by $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$. For any two MTD factorizations $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ that have the same conditional distribution $p\left(x_{kt}|x_{t-1}\right)$ for all x_{kt} and x_{t-1}, then for any $0 < \alpha < 1$, the probability tensor of the MTD model for the parameter set $\alpha\mathbf{Z} + (1-\alpha)\tilde{\mathbf{Z}}$ is given by

$$\begin{aligned}
&\alpha\mathbf{z}^0_{x_{kt}} + (1-\alpha)\tilde{\mathbf{z}}^0_{x_{kt}} + \sum_{j=1}^d\left(\alpha\mathbf{Z}^j_{x_{kt}x_{j(t-1)}} + (1-\alpha)\tilde{\mathbf{Z}}^j_{x_{kt}x_{j(t-1)}}\right) \\
&= \alpha\left(\mathbf{z}^0_{x_{kt}} + \sum_{j=1}^d\mathbf{Z}^j_{x_{kt}x_{j(t-1)}}\right) + (1-\alpha)\left(\tilde{\mathbf{z}}^0_{x_{kt}} + \sum_{i=1}^d\tilde{\mathbf{Z}}^j_{x_{kt}x_{j(t-1)}}\right) \\
&= \alpha p\left(x_{kt}|x_{(t-1)}\right) + (1-\alpha)p\left(x_{kt}|x_{(t-1)}\right) \\
&= p\left(x_{kt}|x_{(t-1)}\right).
\end{aligned}$$

This shows that $\alpha\mathbf{Z} + (1-\alpha)\tilde{\mathbf{Z}}$ has the same distribution as both $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$, so that the set of parameters with the same distribution is a convex set. ■

Proof of Theorem 2. First, we note that a solution always exists since the log likelihood $L(\mathbf{Z}) = -\sum_{t=1}^T\log\left(\mathbf{z}^0_{x_{it}} + \sum_{j=1}^d\mathbf{Z}^j_{x_{it}x_{j(t-1)}}\right)$ and penalty are both bounded below by zero and the feasible set is closed and bounded. Suppose an optimal solution is $\mathbf{Z}$ for which there exists some j such that one row, call it k, of $\mathbf{Z}^j$ does not have a zero element. Let $\alpha = \min\left(\mathbf{Z}^j_{k:}\right)$ be the minimum value in row k and let $\tilde{\mathbf{Z}}^j$ be equal to $\mathbf{Z}^j$ $\forall j$ except that $\tilde{\mathbf{Z}}^j_{k:} = \mathbf{Z}^j_{k:} - \alpha$ and $\tilde{z}^0_k = z^0_k + \alpha$. Due to the nonidentifiability of the MTD model $L(\tilde{\mathbf{Z}}) = L(\mathbf{Z})$, while we have that $\Omega\left(\tilde{\mathbf{Z}}^j\right) < \Omega\left(\mathbf{Z}^j\right)$, implying for $\lambda > 0$

$$L(\tilde{\mathbf{Z}}) + \lambda\Omega(\tilde{\mathbf{Z}}) < L(\mathbf{Z}) + \lambda\Omega(\mathbf{Z}),$$

showing that $\mathbf{Z}$ cannot be an optima. ∎

SM3. Proof of Estimation Consistency.

First, we re-introduce some of our notations. Recall that we define a covariate vector $W \in \mathbb{R}^{m+dm^2}$ as follows: $W_t = (W_{t0}^T, W_{t1}^T, \ldots, W_{td}^T)^T$; $W_{t0} = \left(W^1_{t0}, \ldots, W^m_{t0}\right)^T \in \mathbb{R}^m$ where $W^l_{t0} = I\left\{x_{it} = l\right\}$; and $W_{tj} = \left((W^1_{tj})^T, \ldots, (W^m_{tj})^T\right)^T \in \mathbb{R}^{m^2}$, for $j \in \{1, \ldots, d\}$, where $W^l_{tj} = (W^{l1}_{tj}, \ldots, W^{lm}_{tj})^T$ and $W^{lk}_{tj} = I\left\{x_{it} = l, x_{j(t-1)} = k\right\}$. Let $\mathcal{A}_t$ denote the sub σ-algebra generated by $x_1, \ldots, x_t$. Then $\{W_t\}$ is adapted to $\{\mathcal{A}_t\}$. For a general MTD parameter set, we collect the parameters in a vector form $\beta \in \mathbb{R}^{m+dm^2}$ where $\beta = \left(\beta_0^T, \beta_1^T, \ldots, \beta_d^T\right)^T$, $\beta_0 = \mathbf{z}^0$ and $\beta_j = \mathrm{vec}(\mathbf{Z}^j)$ for $j \in \{1, \ldots, d\}$. The MTD model can be written as

$$p(x_{it}|x_{t-1}) = W_t^T\beta. \tag{SM3.1}$$

For a general β, we define R_n and R to be the empirical and conditional expected negative log-likelihood risks, respectively,

$$R_n(\beta) = -\frac{1}{T}\sum_{t=1}^{T}\log(W_t^T\beta); \quad R(\beta) = -\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\log(W_t^T\beta)|\mathcal{A}_{t-1}\right]. \tag{SM3.2}$$

Denote the group lasso penalty by $\Omega(\beta) = \sum_{j=1}^d \|\beta_j\|_2 = \sum_{j=1}^d \|\mathbf{Z}^j\|_F$. In the remainder of this section, we will use the superscript 0 to denote the true parameter value.

We now turn to the proofs of the estimation consistency results.

SM3.1. Proof of Lemma 6.2.

By definition, we have

$$\begin{aligned} &R_n(\beta) - R(\beta) - (R_n(\beta^0) - R(\beta^0)) \\ &= -\frac{1}{T}\sum_{t=1}^{T}\left\{(\log(W_t^T\beta) - \log(W_t^T\beta^0)) - \mathbb{E}\left[(\log(W_t^T\beta) - \log(W_t^T\beta^0))|\mathcal{A}_{t-1}\right]\right\}. \end{aligned} \tag{SM3.3}$$

For simplicity, we define $\tilde{\Omega}(\beta) = \|\beta_0\|_1 + \Omega(\beta)$. We will consider the following empirical process indexed by f,

$$M_n(f) = \frac{1}{T}\sum_{t=1}^{T}\left(f(W_t) - \mathbb{E}\left[f(W_t)|\mathcal{A}_{t-1}\right]\right), \quad f \in \mathcal{F}, \tag{SM3.4}$$

where the function class $\mathcal{F}$ is defined as

$$\mathcal{F} = \left\{ f : f(W_t) = \log(W_t^T \beta) - \log(W_t^T \beta^0), \tilde{\Omega}(\beta - \beta^0) \le M \right\}. \tag{SM3.5}$$

In the following, we will consider expectation of the supremum of this empirical process. Since $W^T\beta^0$ is the transition probability, values of W such that $W^T\beta^0 = 0$ will not contribute to the expectation as these types of transition occur with probability 0.

Take $M_{\max} = c(T,d)/2$. If $\tilde{\Omega}(\beta - \beta^0) \le M_{\max}$, $\left|W_t^T(\beta - \beta^0)\right| \le M_{\max}$. Then by Assumption 2, we can regard $\mathcal{F}$ as a class of $[\log(c(T,d)/2), -\log(c(T,d)/2)]$-valued functions for some function c that only depends on the sample size T and the number of time series d. Hence we rescale it by multiplying $c(T,d)/2$, and denote the new class by $\tilde{\mathcal{F}}$ so that $\tilde{\mathcal{F}}$ is bounded by 1 and is Lipschitz-continuous with Lipschitz constant 1.

We use the notion of sequential Rademacher complexity and covering number developed in [SM9], which generalizes the definition of Rademacher complexity and covering number to the setting of dependent samples. For a general function class $\mathcal{G}$ mapping from $\mathcal{Z}$ to $\mathbb{R}$, its sequential Rademacher complexity is defined as

$$\mathcal{R}_n = \sup_{\mathbf{z}} \mathcal{R}_n(\mathcal{G}, \mathbf{z}), \quad \text{where } \mathcal{R}_n(\mathcal{G}, \mathbf{z}) = \mathbb{E}\left[\sup_{g \in \mathcal{G}} \frac{1}{T} \sum_{t=1}^{T} \epsilon_t g(\mathbf{z}_t(\epsilon))\right], \tag{SM3.6}$$

where $(\epsilon_t)_{t=1}^T$ is a sequence of independent Rademacher random variables, i.e., Uniform $\{-1,1\}$ and $\mathbf{z}$ is a $\mathcal{Z}$-valued tree of depth T. Further, define

$$\mathcal{D}_n(\mathcal{G}) = \sup_{\mathbf{z}} \mathcal{D}_n(\mathcal{G}, \mathbf{z}), \quad \text{where } \mathcal{D}_n(\mathcal{G}, \mathbf{z}) = \inf_{\alpha} \left\{ 4\alpha + 12/\sqrt{T} \int_{\alpha}^{1} \sqrt{\log \mathcal{N}_2(\delta, \mathcal{G}, \mathbf{z})} d\delta \right\}, \tag{SM3.7}$$

and $\mathcal{N}_2(\cdot, \mathcal{G}, \mathbf{z})$ is the l_2 covering number of $\mathcal{G}$ over a tree $\mathbf{z}$ of depth T. See [SM9] for a complete introduction to sequential Rademacher complexities and covering numbers.

By Theorem 2 and Theorem 4 in [SM9] we can bound the expectation by the sequential Rademacher complexity and a Dudley-type entropy integral,

$$\mathbb{E}\left[\sup_{f \in \tilde{\mathcal{F}}} |M_n(f)|\right] = \mathbb{E}\left[\sup_{f \in \tilde{\mathcal{F}} \cup -\tilde{\mathcal{F}}} M_n(f)\right] \le 2\mathcal{R}_n(\tilde{\mathcal{F}} \cup -\tilde{\mathcal{F}}) \le 2\mathcal{D}_n(\tilde{\mathcal{F}} \cup -\tilde{\mathcal{F}}). \tag{SM3.8}$$

We note that since β^0 is fixed, the covering number of $\tilde{\mathcal{F}}$ is the same as that of $\mathcal{G} = \{g(\cdot) : g(W_t) = \log(W_t^T\beta), \tilde{\Omega}(\beta - \beta^0) \le M\}$. Using the same arguments as in Lemma 13 of [SM9], we can show that

$$\log \mathcal{N}_2(\delta, \tilde{\mathcal{F}}, \mathbf{z}) = \log \mathcal{N}_2(\delta, \mathcal{G}, \mathbf{z}) \le \log \mathcal{N}_\infty(\delta, \mathcal{H}, \mathbf{z}), \tag{SM3.9}$$

where $\mathcal{H} = \{h : h(W_t) = W_t^T\beta - W_t^T\beta^0, \tilde{\Omega}(\beta - \beta^0) \le M\}$. Hence we have that

$$\begin{aligned}
\mathcal{D}_n(\tilde{\mathcal{F}} \cup -\tilde{\mathcal{F}}) &= \sup_{\mathbf{z}} \inf_{\alpha} \left\{ 4\alpha + 12/\sqrt{T} \int_{\alpha}^{1} \sqrt{\log \mathcal{N}_2\left(\delta, \tilde{\mathcal{F}} \cup -\tilde{\mathcal{F}}, \mathbf{z}\right)} d\delta \right\} \\
&\le \sup_{\mathbf{z}} \inf_{\alpha} \left\{ 4\alpha + 12/\sqrt{T} \int_{\alpha}^{1} \sqrt{\log \mathcal{N}_\infty\left(\delta, \mathcal{H} \cup -\mathcal{H}, \mathbf{z}\right)} d\delta \right\} \\
&= \mathcal{D}_n^\infty(\mathcal{H} \cup -\mathcal{H}).
\end{aligned} \tag{SM3.10}$$

Applying Lemma 9 in [SM9], we then get

$$\mathcal{D}_n^{\infty}(\mathcal{H} \cup -\mathcal{H}) \leq 8\mathcal{R}_n(\mathcal{H} \cup -\mathcal{H}) \left(1 + 4\sqrt{2}\log^{3/2}\left(eT^2\right)\right). \tag{SM3.11}$$

Our last step is to bound the Rademacher complexity of the class $\mathcal{H} \cup -\mathcal{H}$. Note that by definition,

$$\begin{aligned}
\mathcal{R}_n(\mathcal{H} \cup -\mathcal{H}) &= \sup_{\mathbf{w}} \mathbb{E}\left[\sup_{h\in\mathcal{H}} \left|\frac{1}{T}\sum_{t=1}^{T} \epsilon_t h(\mathbf{w}_t(\epsilon))\right|\right] \\
&= \sup_{\mathbf{w}} \mathbb{E}\left[\sup_{\beta:\tilde{\Omega}(\beta-\beta^0)\leq M} \left|\frac{1}{T}\sum_{t=1}^{T} \epsilon_t \mathbf{w}_t(\epsilon)^T(\beta-\beta^0)\right|\right] \\
&\leq \sup_{\mathbf{w}} \mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^{T} \epsilon_t \mathbf{w}_t(\epsilon)\right\|_{\infty}\right] \sup_{\beta:\tilde{\Omega}(\beta-\beta^0)\leq M} \left\|\beta-\beta^0\right\|_1 \\
&\leq mM \sup_{\mathbf{w}} \mathbb{E}\left[\max_{j\in\{1,\ldots,m+dm^2\}} \left|\frac{1}{T}\sum_{t=1}^{T} \epsilon_t \mathbf{w}_{tj}(\epsilon)\right|\right] \\
&\leq mM\sqrt{\frac{2\log(2(m+dm^2))}{T}} \qquad \text{(SM3.12)}
\end{aligned}$$

where the fourth line follows from Lemma SM3.1 and the fifth line follows by applying the finite class lemma in the dependent setting [SM9] and a union bound.

Finally, combining (SM3.8), (SM3.10), (SM3.11) and (SM3.12), we have that

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |M_n(f)|\right] \leq \frac{32}{c(T,d)} mM\sqrt{\frac{2\log(2(m+dm^2))}{T}} \left(1 + 4\sqrt{2}\log^{3/2}\left(eT^2\right)\right). \tag{SM3.13}$$

Thus, by Markov inequality, we can take

$$\lambda_\epsilon = O_p\left(\frac{1}{c(T,d)}\sqrt{\frac{\log(d)\log^3(T)}{T}}\right). \tag{SM3.14}$$

Finally, we need

$$\frac{32\lambda_\epsilon(1+\delta)^2|S|}{\delta^2\phi^2(1/(1-\delta),S,\tau)} \leq \frac{1}{2}c(T,d), \tag{SM3.15}$$

which holds with probability tending to 1 by Assumption 2 and Assumption 3.

SM3.2. Useful lemmas. Before proving our main theorem, we first establish several lemmas which will be useful later in the proof.

The first lemma establishes a margin condition for the negative loglikelihood loss.

Lemma SM3.1. *(Margin condition) For all β satisfying the MTD model constraints, $R(\beta) - R(\beta^0) \geq \frac{1}{2}\tilde{\tau}^2(\beta - \beta^0)$, where $\tilde{\tau}(\beta)$ is a semi-norm defined as*

$$\tilde{\tau}(\beta) = \sqrt{\frac{1}{T}\beta^T \left(\sum_{t=1}^{T} \mathbb{E}\left[W_t W_t^T | \mathcal{A}_{t-1}\right]\right)\beta}. \tag{SM3.16}$$

Proof. As β^0 is the true parameter in the conditional distribution specified by MTD model, it maximizes $\mathbb{E}[\log(W_t^T\beta)|\mathcal{A}_{t-1}]$ for all t, and hence minimizes $R(\beta)$. (The minimizer is not unique, as in general the MTD model is not identifiable. But restricting each row to have at least one zero can make the solution unique.)

Let $H(\beta) = 0$ denote the set of equality constraints on a valid MTD parameter set. Then, consider the Lagrangian form of the MTD optimization,

$$R(\beta) + \lambda_1^T H(\beta) + \lambda_2^T(-\beta), \tag{SM3.17}$$

where λ_1 and λ_2 are the Lagrange multipliers associated with the equality and inequality constraints respectively. Then β^0 satisfies the following KKT conditions:

$$\frac{\partial R(\beta)}{\partial \beta}|_{\beta^0} + (\lambda_1^0)^T \frac{\partial H(\beta)}{\partial \beta}|_{\beta^0} - \lambda_2^0 = 0; \tag{SM3.18}$$

$$H(\beta^0) = 0; \tag{SM3.19}$$

$$(\lambda_2^0)^T \beta^0 = 0; \tag{SM3.20}$$

$$\lambda_2^0 \geq 0, \beta^0 \geq 0. \tag{SM3.21}$$

We define a new function

$$\tilde{R}(\beta) = R(\beta) + (\lambda_1^0)^T H(\beta) + (\lambda_2^0)^T(-\beta). \tag{SM3.22}$$

Note that for all β satisfying the MTD model constraints, $H(\beta) = 0$. Thus,

$$\begin{aligned}
\tilde{R}(\beta) - \tilde{R}(\beta^0) &= R(\beta) - R(\beta^0) + (\lambda_1^0)^T(H(\beta) - H(\beta^0)) + (\lambda_2^0)^T(\beta^0 - \beta) \\
&= R(\beta) - R(\beta^0) + (\lambda_2^0)^T(\beta^0 - \beta) \qquad \text{(SM3.23)}\\
&= R(\beta) - R(\beta^0) - (\lambda_2^0)^T\beta, \qquad \text{(SM3.24)}
\end{aligned}$$

where the last line follows from the KKT conditions. At the same time, using a first order Taylor expansion and noting that the derivative of $\tilde{R}(\beta)$ at β^0 is 0, we get

$$\tilde{R}(\beta) - \tilde{R}(\beta^0) = (\beta - \beta^0)^T \frac{\partial^2 \tilde{R}}{\partial \beta^2}|_{\beta^*}(\beta - \beta^0)/2, \tag{SM3.25}$$

for some β^* between β and β^0. Then, we have

$$R(\beta) - R(\beta^0) = (\lambda_2^0)^T\beta + (\beta - \beta^0)^T \frac{\partial^2 \tilde{R}}{\partial \beta^2}|_{\beta^*}(\beta - \beta^0)/2. \tag{SM3.26}$$

Since the equality and inequality constraints are both linear, $\partial^2 \tilde{R}/\partial\beta^2 = \partial^2 R/\partial\beta^2$ and we have

$$\frac{\partial^2 R}{\partial \beta^2} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{(W_t^T\beta)^2}W_tW_t^T|\mathcal{A}_{t-1}\right]. \tag{SM3.27}$$

Here, $W_t^T\beta$ models conditional probability, and is bounded between 0 and 1. Hence the above expression is lower bounded by $\sum_{t=1}^{T}\mathbb{E}[W_tW_t^T|\mathcal{A}_{t-1}]/T$. Also, we have that $(\lambda_2^0)^T\beta \geq 0$. Together, we have

$$R(\beta) - R(\beta^0) \geq \frac{1}{2}(\beta-\beta^0)^T\frac{\sum_{t=1}^{T}\mathbb{E}[W_tW_t^T|\mathcal{A}_{t-1}]}{T}(\beta-\beta^0). \tag{SM3.28}$$ ∎

Recall that S denotes the active set of β^0, i.e., $S = \{j : j > 0,\ \beta_j^0 \neq \mathbf{0}\}$ and S^c denotes its complement in $\{1, \ldots, d\}$. We define $\Omega^+(\beta) = \sum_{j\in S}\|\beta_j\|_1$ and $\Omega^-(\beta) = \sum_{j\in S^c}\|\beta_j\|_1$. The next lemma shows some basic properties of the penalty $\Omega(\cdot)$.

Lemma SM3.2. *(Properties of the penalty) The penalty $\Omega(\cdot)$ satisfies the following for any β:*

1. $\|\beta\|_1 \leq \|\beta_0\|_1 + m\Omega(\beta)$.
2. $\Omega(\beta^0) - \Omega(\beta) \leq \Omega^+(\beta-\beta^0) - \Omega^-(\beta-\beta^0)$.

Proof.

1. $\|\beta\|_1 = \sum_{j=0}^{d}\|\beta_j\|_1$. For $j \neq 0, \beta_j \in \mathbb{R}^{m^2}$. By Lyapunov inequality $\frac{1}{m^2}\|\beta_j\|_1 \leq \sqrt{\frac{1}{m^2}\|\beta_j\|_2^2}$, and hence $\|\beta_j\|_1 \leq m\|\beta_j\|_2$. Invoking the definition of $\Omega(\beta)$ completes the proof.
2. We note that $\Omega(\beta) = \Omega^+(\beta) + \Omega^-(\beta)$. By the triangle inequality, $\|\beta_j^0\|_1 \leq \|\beta_j^0 - \beta_j\|_1 + \|\beta_j\|_1$. Summing over $j \in S$ we have $\Omega^+(\beta^0) - \Omega^+(\beta) \leq \Omega^+(\beta^0 - \beta)$. By definition $\Omega^-(\beta^0) = 0$ and $\beta_j - \beta_j^0 = \beta_j$ for $j \in S^c$, which implies that $\Omega^-(\beta-\beta^0) = \Omega^-(\beta)$. Thus,

$$\begin{aligned}\Omega(\beta^0) - \Omega(\beta) &= \Omega^+(\beta^0) - \Omega^+(\beta) - \Omega^-(\beta)\\ &\leq \Omega^+(\beta-\beta^0) - \Omega^-(\beta) = \Omega^+(\beta-\beta^0) - \Omega^-(\beta-\beta^0).\end{aligned} \tag{SM3.29}$$ ∎

Recall that we have defined a semi-norm $\tilde{\tau}(\beta) = \sqrt{\beta^T\sum_{t=1}^{T}\mathbb{E}[W_tW_t^T|\mathcal{A}_{t-1}]\beta/T}$. However, this semi-norm itself is random as we condition on the past. The next lemma shows that it is close to a deterministic semi-norm $\tau(\cdot)$, and the compatibility constants defined with $\tilde{\tau}$ and τ are close. To this end, we will use concentration inequalities for Markov chains developed in [SM8].

Lemma SM3.3. *Under Assumption 1 and Assumption 4, with probability at least $1 - 1/T$,*

$$\frac{\phi^2(L,S,\tilde{\tau})}{\phi^2(L,S,\tau)} \geq 1 - (1+(1+L)m)^2C'\sqrt{\frac{\log(2(m+dm^2)^2)+\log(T)}{T\gamma_{ps}}}|S|/\phi^2(L,S,\tau). \tag{SM3.30}$$

Thus, under Assumptions 1, 3 and 4, for T sufficiently large, $\phi^2(L,S,\tilde{\tau})/\phi^2(L,S,\tau) > 1/2$ with probability at least $1 - 1/T$.

Proof. For any $j, k \in \{1, \ldots, m+dm^2\}$, $W_j W_k$ is bounded between 0 and 1. For simplicity, we will assume, for now, that $x_0 \sim \pi$, i.e., the chain starts in the stationary distribution. We will relax this assumption later. Applying Theorem 3.11 in [SM8],

(SM3.31)
$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[W_{tj}W_{tk}|\mathcal{A}_{t-1}] - \mathbb{E}_\pi[W_{1j}W_{1k}]\right| \geq t\right) \leq 2\exp\left(-\frac{T^2t^2\gamma_{ps}}{8(T+1/\gamma_{ps})+20Tt}\right).$$

And, using a union bound,

$$\mathbb{P}\left(\sup_{j,k}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[W_{tj}W_{tk}|\mathcal{A}_{t-1}] - \mathbb{E}[W_{1j}W_{1k}]\right| \geq t\right) \leq 2(m+dm^2)^2\exp\left(-\frac{T^2t^2\gamma_{ps}}{8(T+1/\gamma_{ps})+20Tt}\right). \tag{SM3.32}$$

In order to obtain a concentration bound, we will choose $t = o(1)$ and consider large T. Hence, the right-hand-side is of the same order as $2(m+dm^2)^2\exp(-CTt^2\gamma_{ps})$, provided that $1/\gamma_{ps} = o(T)$. Now setting $t = \sqrt{\log(2(m+dm^2)^2/\alpha)/CT\gamma_{ps}}$,

$$\mathbb{P}\left(\max_{j,k}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[W_{tj}W_{tk}|\mathcal{A}_{t-1}] - \mathbb{E}[W_{1j}W_{1k}]\right| \geq \sqrt{\frac{\log(2(m+dm^2)^2/\alpha)}{CT\gamma_{ps}}}\right) \leq \alpha, \tag{SM3.33}$$

for T sufficiently large.

Then, for all β

$$\begin{aligned}
\left|\tau^2(\beta) - \tilde{\tau}^2(\beta)\right| &= \left|\beta^T\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[W_tW_t^T|\mathcal{A}_{t-1}] - \mathbb{E}_\pi[W_1W_1^T]\right)\beta\right| \\
&\leq \|\beta\|_1^2\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[W_tW_t^T|\mathcal{A}_{t-1}] - \mathbb{E}_\pi[W_1W_1^T]\right\|_\infty \\
&\leq \|\beta\|_1^2 C'\sqrt{\frac{\log(2(m+dm^2)^2/\alpha)}{T\gamma_{ps}}}, \qquad \text{(SM3.34)}
\end{aligned}$$

where by (SM3.33) the last line holds with probability at least $1-\alpha$.

Recall the definition of Γ and compatibility constant ϕ,

$$\Gamma_\Omega(L,S,\tau) = \left(\min\left\{\tau(\beta) : \|\beta_0\|_1 + \Omega^+(\beta) = 1, \Omega^-(\beta) \leq L\right\}\right)^{-1} \tag{SM3.35}$$

$$\phi^2(L,S,\tau) = \Gamma_\Omega^{-2}(L,S,\tau)|S|. \tag{SM3.36}$$

Thus,

$$\begin{aligned}
\frac{\phi^2(L,S,\tilde{\tau})}{\phi^2(L,S,\tau)} = \frac{\Gamma_\Omega^2(L,S,\tau)}{\Gamma_\Omega^2(L,S,\tilde{\tau})} = \frac{\min\tilde{\tau}^2(\beta)}{\min\tau^2(\beta)} &\geq 1 + \frac{\min\tilde{\tau}^2(\beta) - \tau^2(\beta)}{\min\tau^2(\beta)} \\
&\geq 1 - (1+(1+L)m)^2C'\sqrt{\frac{\log(2(m+dm^2)^2/\alpha)}{T\gamma_{ps}}}|S|/\phi^2(L,S,\tau), \qquad \text{(SM3.37)}
\end{aligned}$$

with probability at least $1-\alpha$. Setting $\alpha = 1/T$, we see that with probability approaching 1, the ratio is greater than $\frac{1}{2}$ for sufficiently large T, provided that $|S|\sqrt{\log(d)/T\gamma_{ps}} = o(1)$ and $\phi^2(L,S,\tau)$ is bounded away from 0.

If the chain does not start in stationary distribution, a result similar to (SM3.31) can be established, provided that the distribution of x_0 is not too far away from π. In the rest of this subsection, we use $\mathbb{P}_q$ to denote the probability under the case $x_0 \sim q$. Define

$$N_q = \begin{cases} \mathbb{E}_\pi\left[\left(\frac{q(x)}{\pi(x)}\right)^2\right] & \text{if } q \text{ is absolutely continuous with respect to } \pi, \\ +\infty & \text{otherwise.} \end{cases} \tag{SM3.38}$$

Applying Proposition 3.15 in [SM8], we get

$$\begin{aligned} &\mathbb{P}_q\left(\left|\frac{1}{T}\sum_{t=1}^T \mathbb{E}[W_{tj}W_{tk}|\mathcal{A}_{t-1}] - \mathbb{E}_\pi[W_{1j}W_{1k}]\right| \geq t\right) \\ &\leq N_q^{1/2}\left[\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^T \mathbb{E}[W_{tj}W_{tk}|\mathcal{A}_{t-1}] - \mathbb{E}_\pi[W_{1j}W_{1k}]\right| \geq t\right)\right]^{1/2} \\ &\leq 2N_q^{1/2}\exp\left(-\frac{T^2t^2\gamma_{ps}}{16(T+1/\gamma_{ps}) + 40Tt}\right). \end{aligned} \tag{SM3.39}$$

This bound is essentially the same as in (SM3.31), except that we are working with different constants. The rest of the proof follows. ∎

SM3.3. Proof of Theorem 6.1. Next we prove our main theorem, which is a modification of the proof of Theorem 7.2 in [SM10]. The difference is that we handle the unpenalized intercept as in [SM2] and we have time dependence in the data. For notational convenience, define

$$M = \frac{4\lambda(1+\delta)^2|S|}{\delta\phi^2(1/(1-\delta),S,\tau)}, \text{ and } t = \frac{M}{M + \Omega(\hat{\beta}-\beta^0) + \|\hat{\beta}_0 - \beta_0^0\|_1}.$$

Define $\tilde{\beta} = t\hat{\beta} + (1-t)\beta^0$. With this construction, $\|\tilde{\beta}_0 - \beta_0^0\|_1 + \Omega(\tilde{\beta}-\beta^0) \leq M$.

We note that although in general $\tilde{\beta}$ may not have a zero in each row of the corresponding $\mathbf{Z}^j$ matrices, and hence may not be identifiable, it does satisfy the equality and inequality constraints of the MTD model. By the convexity of $R_n + \lambda\Omega$, we have that

$$\begin{aligned} R_n(\tilde{\beta}) + \lambda\Omega(\tilde{\beta}) &\leq tR_n(\hat{\beta}) + t\lambda\Omega(\hat{\beta}) + (1-t)R_n(\beta^0) + (1-t)\lambda\Omega(\beta^0) \\ &\leq R_n(\beta^0) + \lambda\Omega(\beta^0). \end{aligned} \tag{SM3.40}$$

We rewrite this and apply Lemma 6.2 and Lemma SM3.2,

$$\begin{aligned} 0 \leq R(\tilde{\beta}) - R(\beta^0) &\leq -\left[[R_n(\tilde{\beta}) - R(\tilde{\beta})] - [R_n(\beta^0) - R(\beta^0]\right] + \lambda\Omega(\beta^0) - \lambda\Omega(\tilde{\beta}) \\ &\leq \lambda_\epsilon M + \lambda\Omega(\beta^0) - \lambda\Omega(\tilde{\beta}) \\ &\leq \lambda_\epsilon M + \lambda\Omega^+(\tilde{\beta}-\beta^0) - \lambda\Omega^-(\tilde{\beta}-\beta^0). \end{aligned} \tag{SM3.41}$$

We consider two cases.

- Case 1: If $\lambda\|\tilde{\beta}_0 - \beta_0^0\|_1 + \lambda\Omega^+(\tilde{\beta} - \beta^0) \leq (1-\delta)\lambda_\epsilon M/\delta$, we have that

$$\delta\lambda\left(\|\tilde{\beta}_0 - \beta_0^0\|_1 + \Omega^+(\tilde{\beta} - \beta^0)\right) \leq \lambda_\epsilon M, \tag{SM3.42}$$

and

$$\delta\lambda\Omega^-(\tilde{\beta} - \beta^0) \leq \lambda_\epsilon M. \tag{SM3.43}$$

Hence,

$$\delta\lambda\left(\|\tilde{\beta}_0 - \beta_0^0\|_1 + \Omega(\tilde{\beta} - \beta^0)\right) \leq 2\lambda_\epsilon M. \tag{SM3.44}$$

- Case 2: If instead $\lambda\|\tilde{\beta}_0 - \beta_0^0\|_1 + \lambda\Omega^+(\tilde{\beta} - \beta^0) \geq (1-\delta)\lambda_\epsilon M/\delta$, then by (SM3.41)

$$\begin{aligned} R(\tilde{\beta}) - R(\beta^0) + \lambda\Omega^-(\tilde{\beta} - \beta^0) &\leq \lambda\Omega^+(\tilde{\beta} - \beta^0) + \frac{\delta}{(1-\delta)}\lambda\left(\Omega^+(\tilde{\beta} - \beta^0) + \|\tilde{\beta}_0 - \beta_0^0\|_1\right) \\ &\leq \lambda\left(\Omega^+(\tilde{\beta} - \beta^0) + \|\tilde{\beta}_0 - \beta_0^0\|_1\right)/(1-\delta), \end{aligned} \tag{SM3.45}$$

where the second inequality holds because $0 < \delta < 1$. Since $R(\tilde{\beta}) - R(\beta^0) \geq 0$,

$$\Omega^-(\tilde{\beta} - \beta^0) \leq \left(\Omega^+(\tilde{\beta} - \beta^0) + \|\tilde{\beta}_0 - \beta_0^0\|_1\right)/(1-\delta), \tag{SM3.46}$$

which allows us to use the compatibility condition later. Again from (SM3.41),

$$\begin{aligned} &R(\tilde{\beta}) - R(\beta) + \lambda\Omega^-(\tilde{\beta} - \beta^0) + \delta\lambda\left(\Omega^+(\tilde{\beta} - \beta^0) + \|\tilde{\beta}_0 - \beta_0^0\|_1\right) \\ &\leq \lambda_\epsilon M + (1+\delta)\lambda\left(\Omega^+(\tilde{\beta} - \beta^0) + \|\tilde{\beta}_0 - \beta_0^0\|_1\right) \\ &\leq \lambda(1+\delta)\tilde{\tau}(\tilde{\beta} - \beta^0)\Gamma_\Omega(1/(1-\delta), S, \tilde{\tau}) + \lambda_\epsilon M \\ &\leq \frac{1}{2}(\lambda^2(1+\delta)^2\Gamma_\Omega^2(1/(1-\delta), S, \tilde{\tau})) + \frac{1}{2}\tilde{\tau}^2(\tilde{\beta} - \beta^0) + \lambda_\epsilon M \\ &\leq \frac{1}{2}(\lambda^2(1+\delta)^2\Gamma_\Omega^2(1/(1-\delta), S, \tilde{\tau})) + R(\tilde{\beta}) - R(\beta) + \lambda_\epsilon M, \end{aligned} \tag{SM3.47}$$

where the second inequality follows by applying Assumption 3 with stretching factor $1/(1-\delta)$, and the fourth inequality follows from Lemma SM3.1. It follows that

$$\begin{aligned} \delta\lambda\left(\Omega(\tilde{\beta} - \beta^0) + \|\tilde{\beta}_0 - \beta_0^0\|_1\right) &\leq \frac{1}{2}(\lambda(1+\delta)\Gamma_\Omega(1/(1-\delta), S, \tilde{\tau}))^2 + \lambda_\epsilon M \\ &= \frac{1}{2}(\lambda(1+\delta))^2\frac{|S|}{\phi^2(L, S, \tilde{\tau})} + \lambda_\epsilon M \\ &\leq \frac{(\lambda(1+\delta))^2|S|}{\phi^2(L, S, \tau)} + \lambda_\epsilon M, \end{aligned} \tag{SM3.48}$$

with probability approaching 1.

Hence, in both cases we have that with probability going to 1,

$$\delta\lambda\left(\Omega(\tilde{\beta}-\beta^0)+\|\tilde{\beta}_0-\beta_0^0\|_1\right)\leq 2\lambda_\epsilon M+(\lambda(1+\delta)\Gamma_\Omega(1/(1-\delta),S,\tau))^2$$

$$= \delta\lambda M/4+2\lambda_\epsilon M\leq \delta\lambda M/2, \tag{SM3.49}$$

where the inequality follows from the fact that $\lambda \geq 8\lambda_\epsilon/\delta$ and the equality follows from the definition of M. Finally, this implies that

$$\Omega(\tilde{\beta}-\beta^0)+\|\tilde{\beta}_0-\beta_0^0\|_1\leq M/2, \tag{SM3.50}$$

which in turn, by the construction of $\tilde{\beta}$, implies that

$$\Omega(\hat{\beta}-\beta^0)+\|\hat{\beta}_0-\beta_0^0\|_1\leq M. \tag{SM3.51}$$

SM4. Optimization Algorithms.

In the main text, we presented a projected gradient algorithm for optimization. Here, we present some alternative methods for optimization of the MTD objective and discuss in what contexts they might be applicable.

SM5. Frank-Wolfe.

In very high-dimensional settings, with large state spaces, the projection step in the MTD projected gradient algorithm presented in the main text becomes increasingly more computationally intensive. Frank-Wolfe algorithms, on the other hand, are projection free algorithms for solving constrained convex optimization problems and have recently gained popularity due to their simplicity and scalability in sparse, high-dimensional regression and machine learning [SM4]. Fortunately, the Frank-Wolfe algorithm for MTD also takes a simple form that allows updating only a small number of parameters at a time. In very sparse, high dimensional problems with large state spaces, where most entries are zero, this is typically advantageous [SM4]. We develop the algorithm and provide a timing comparison to the projected gradient algorithm in the main text. We leave the development of Frank-Wolfe using various variants [SM5] for future work.

SM5.1. Frank-Wolfe MTD.

Let $\mathbf{Z}^{(0)}$ be the initial MTD model. Let $L(\mathbf{Z}) = L_{MTD}(\mathbf{Z}) + \lambda\Omega(\mathbf{Z})$. The Frank-Wolfe algorithm iterates between the following steps starting with $k = 0$:

1. Find a direction $\hat{\mathbf{D}}$ that maximizes the dot product with the gradient while staying in the constraint set:

$$\hat{\mathbf{D}} = \underset{\mathbf{D}}{\operatorname{argmin}}\left(\mathbf{z}^0\right)^T\nabla_{\mathbf{z}^0}L\left(\mathbf{Z}^{(k)}\right)+\sum_{j=1}\operatorname{trace}\left(\left(\mathbf{D}^j\right)^T\nabla_{\mathbf{Z}^j}L\left(\mathbf{Z}^{(k)}\right)\right) \tag{SM5.1}$$

$$\text{subject to } \mathbf{1}^T\mathbf{D}^j=\gamma_j\mathbf{1}^T,\ \mathbf{D}^j\geq 0\ \forall j,\quad \mathbf{1}^T\gamma=1,\gamma\geq 0.$$

2. Choose θ by line search or set $\theta = \frac{2}{2+k}$.
3. Set $\mathbf{Z}^{(k+1)} = \theta\hat{\mathbf{D}} + (1-\theta)\mathbf{Z}^{(k)}$.

Step 1 involves solving a linear programming problem. Since the solution to Step 1 stays in the constraint set, any step taken in Step 2 for $\theta \in (0, 1)$ remains in the constraint set. Fortunately, the linear program in Step 1 has a simple, closed form solution with linear complexity in the number of parameters, $O\left(m^2d + m\right)$.

Proposition SM5.1. *First let* $\mathbf{F}^j = \nabla_{\mathbf{Z}^j} L\left(\mathbf{Z}^{(k)}\right)$. *Let* q_k^j *be the row index of the minimal element in column* k *of* $\mathbf{F}^j_{:k}$ *and let* s^j *be the sum of the minimal elements in each column:* $s^j = \sum_{k=1}^m \mathbf{Z}^j_{q_k^j k}$. *Furthermore, let* j^* *be the index of the minimum* s^j : $j^* = \underset{j}{\operatorname{argmin}}\left(s^j\right)$. *Then* $\mathbf{D}^*$ *is given by*

$$\hat{\mathbf{D}}^j = 0 \ \forall j \neq j^*,$$
$$\hat{\mathbf{D}}^{j*}_{lk} = \begin{cases} 1 \ \textit{if } l = q_k^j \\ 0 \ \textit{if } l \neq q_k^j \end{cases}.$$

Intuitively, to stay in the MTD constraint set any feasible step must place equal mass on each column of a $\mathbf{Z}^j$, and that the minima is attained by only taking steps in the direction of $\mathbf{Z}^j$ with a minimal sum of columnwise minima.

Proposition SM5.1 implies that if the model is initialized with $\left(\mathbf{Z}^j\right)^{(0)} = 0$ for all j, then at step k at most only km entries in $\mathbf{Z}^{(k)}$ will be nonzero, and typically less in high-dimensional sparse settings since certain entries with strong signal will be updated repeatedly. The final Frank-Wolfe algorithm for MTD is shown in Algorithm SM5.1.

Proof of Proposition SM5.1. We study the KKT conditions. The Lagrangian is given by:

$$\sum_j \sum_l \sum_k \mathbf{D}^j_{lk}\mathbf{F}^j_{lk} + \sum_j \sum_k \lambda_k^j \left(\left(\sum_l \mathbf{D}^j_{lk}\right) - \gamma_j\right) + \nu\left(1^T\gamma - 1\right) + \sum_j \sum_k \sum_l \phi^j_{lk}\mathbf{D}^j_{lk}.$$

So that the KKT conditions for an optima are given by:

$$\mathbf{F}^j_{lk} = \lambda_k^j + \gamma^j_{lk}; \tag{SM5.2}$$

$$\sum_k^{m_j} \lambda_k^j = \nu \quad \forall j; \tag{SM5.3}$$

$$\phi^j_{lk} \geq 0 \ (\text{dual feasibility}) \ ; \tag{SM5.4}$$

$$\phi^j_{lk}\hat{\mathbf{D}}^j_{lk} = 0 \quad (\text{complimentary slackness}) \ . \tag{SM5.5}$$

We show that for the primary feasible solution given in Proposition SM5.1, there exists a set of dual variables that obey the KKT conditions, showing that the solution in Proposition SM5.1 is indeed the global optima.

For the primal solution given in Proposition SM5.1, let the dual variables for j^* be

$$\lambda_k^{j*} = \mathbf{F}^{j*}_{q_k^{j*}k} \text{ and } \phi^{j*}_{q_k^{j*}k} = 0 \ \forall k \in (1, \ldots, m_j),$$

which obeys (SM5.2) and the complimentary slackness in (SM5.5) since $\hat{\mathbf{D}}^{j*}_{q_k^{j*}} = 1$. For all other entries of $\hat{\mathbf{D}}^{j*}$, $\phi^{j*}_{lk} = F^j_{lk} - \lambda_k^j = F^j_{lk} - F^{j*}_{q_k^{j*}k}$, so that all entries in ϕ^{j*}_{lk} and λ_k^{j*} obey the KKT conditions for all l, k in (SM5.4). The complimentary slackness holds in (SM5.5) since for these l, k $\hat{\mathbf{D}}^{j*}_{lk} = 0$. Finally, set $\nu = \sum_k^{m_{j*}} \lambda_k^{j*} = \sum_k^{m_{j*}} F^{j*}_{q_k^{j*}k}$ which by construction satisfies condition (SM5.3).

For $j \neq j^*$, let $\lambda_k^j = F^j_{q_k^j k} - \frac{\tilde{\nu}^j - \nu}{m_j}$ where $\tilde{\nu}^j = \sum_k^{m_j} F^j_{q_k^j k}$. By construction, $\sum_j^{m_j} \lambda_k^j = \nu$ satisfying (SM5.3). Furthermore, letting $\phi_{lk}^j = F_{lk}^j - \lambda_k^j$, we have that $\phi_{lk}^j > 0$ since $F_{lk}^j > F^j_{q_k^j k} > F^j_{q_k^j k} - \frac{\tilde{\nu}^j - \nu}{m_j} = \lambda_k^j$ and $\tilde{\nu}^j - \nu = \sum_k^{m_j} F^j_{q_k^j k} - \sum_k^{m_{j*}} F^{j*}_{q_k^{j*} k} > 0$ satisfying (SM5.4). For all these entries the complimentary slackness condition holds since $\hat{\mathbf{D}}_{lk}^j = 0$, satisfying (SM5.5).

Taken together, we have found a set of dual feasible points that obey the KKT conditions for the solution in Proposition SM5.1, showing that the solution is the optima. ■

Algorithm SM5.1 Projection free Frank-Wolfe algorithm for MTD.

Initialize $\left(\mathbf{Z}^j\right)^{(0)} = 0 \quad \forall j, \left(\mathbf{z}^0\right)^{(0)} = \frac{1}{m}$
for $k = 0, 1, 2, \ldots$ **do**
  compute $\nabla L\left(\mathbf{Z}^{(k)}\right)$
  determine $\hat{\mathbf{D}}$ according to Proposition SM5.1
  determine θ by line search or $\theta = \frac{2}{2+k}$
  $\mathbf{Z}^{(k)} = (1-\theta)\mathbf{Z}^{(k+1)} + \theta\hat{\mathbf{D}}$
end for

SM5.2. Run time comparison between Frank-Wolfe and Projected Gradient.

We compare the Frank-Wolfe algorithm for MTD to the projected gradient algorithm in the main text. In Figure SM11 we show the value of the objective as a function of time for Frank-Wolfe, projected gradient descent, and accelerated projected gradient descent on a synthetic data set. For Frank-Wolfe, we use the step size of $\theta = \frac{2}{2+k}$. In this case, the Frank-Wolfe algorithm is slower to converge than the projected or accelerated projected gradient algorithm. We suspect that the gains of Frank-Wolfe over projected gradient will be in very high-dimensional settings with large state spaces, but we leave that exploration for future work.

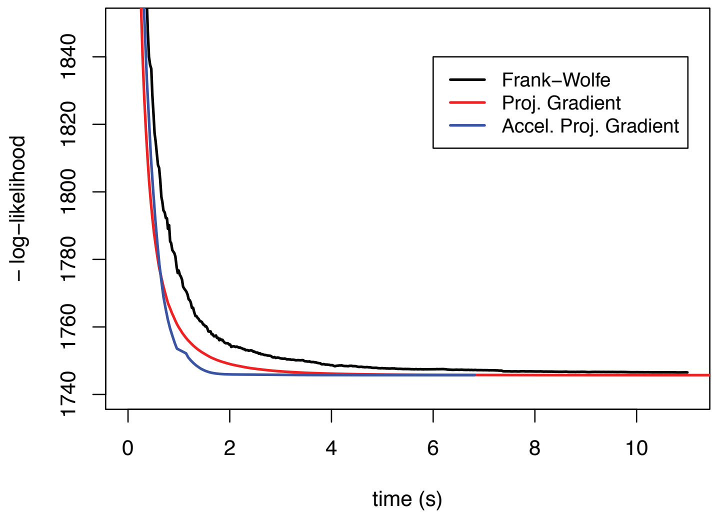

Figure SM11. *Run time comparison between Frank-Wolfe, projected gradient, and accelerated projected gradient on a $d = 25$, $T = 400$, and $m = 5$ synthetic data set.*

SM5.3. Majorization-Minimization. Here we use the convex formulation of MTD in the main text to derive a majorization-minimization (MM) algorithm [SM3]. The closed form updates are only given when there is no penalty function $\Omega(\mathbf{Z})$, so that this algorithm is not as generally applicable as the projected gradient algorithm presented in the main text. Interestingly, we find that the MM updates of the convex formulation correspond exactly to the MTD EM algorithm of [SM6] for the non-convex parameterization. This proves that the EM algorithm for MTD converges to a global optima even though the log-likelihood is non-convex.

We derive the MM algorithm for the convex MTD formulation with no penalty term (and no intercept):

$$\begin{aligned} &\underset{\mathbf{Z},\gamma}{\text{minimize}}\ L_{\text{MTD}}(\mathbf{Z}) \\ &\text{subject to } \mathbf{1}^T\mathbf{Z}^j = \gamma_j \mathbf{1}^T, \mathbf{Z}^j \geq 0\ \forall j, \quad \mathbf{1}^T\gamma = 1, \gamma \geq 0. \end{aligned} \tag{SM5.6}$$

To derive the MM algorithm, we first form the surrogate function

$$Q\left(\mathbf{Z}, \mathbf{Z}^{(n)}\right) = \sum_{t=1}^{T}\sum_{j=1}^{d} p_{jt} \log \frac{Z^j_{x_{it}x_{j(t-1)}}}{p_{jt}},$$

where $p_{jt} = \frac{Z^{j(n)}_{x_{it}x_{j(t-1)}}}{\sum_{l=1}^{d} Z^{l(n)}_{x_{it}x_{l(t-1)}}}$. Now, $Q\left(\mathbf{Z}, \mathbf{Z}^{(n)}\right)$ satisfies the MM algorithm conditions that $Q\left(\mathbf{Z}, \mathbf{Z}^{(n)}\right) \geq L_{\text{MTD}}(\mathbf{Z})$ and $Q(\mathbf{Z}, \mathbf{Z}) = L_{\text{MTD}}(\mathbf{Z})$. This implies we may iteratively minimize $Q\left(\mathbf{Z}, \mathbf{Z}^{(n)}\right)$:

$$\mathbf{Z}^{(n+1)} = \underset{\mathbf{Z},\gamma}{\text{argmin}} Q\left(\mathbf{Z}, \mathbf{Z}^{(n)}\right),$$

and that this sequence of $\mathbf{Z}^{(n+1)}$ converges to a global optima since Problem (SM5.6) is convex.

Proposition SM5.2. *The solution to Problem (SM5.6) under the MTD constraints is given in closed form:*

$$\mathbf{Z}^{j(n+1)}_{lk} = \left(\frac{\tilde{p}^j_{lk}}{\sum_l \tilde{p}^j_{lk}}\right)\left(\frac{\sum_{lk}\tilde{p}^j_{lk}}{\sum_j \sum_{lk} \tilde{p}^j_{lk}}\right), \tag{SM5.7}$$

where $\tilde{p}^j_{lk} = \sum_{t=1} p_{jt} 1_{\left(x_{it}=l, x_{j(t-1)}=k\right)}$.

Corollary SM5.3. *The EM algorithm for the unpenalized MTD model in the original* $(\gamma, \mathbf{P})$ *parameterization converges to a global optima of the non-convex log-likelihood.*

Proof of Proposition SM5.2 and Corollary SM5.3. The optimization problem for the MM update in Problem (SM5.6) is given by

$$\underset{\mathbf{Z},\gamma}{\text{minimize}} - \sum_{t=1}^{T}\sum_{j=1}^{d} p_{jt} \log \frac{Z^j_{x_{it}x_{j(t-1)}}}{p_{jt}} \tag{SM5.8}$$

$$\text{subject to } \mathbf{1}^T\mathbf{Z}^j = \gamma_j \mathbf{1}^T \ \forall j, \quad \mathbf{1}^T\gamma = 1,$$

where we have removed the non-negativity constraints because these are automatically enforced in the log terms of the $Q\left(Z, Z^{(n)}\right)$ objective. We may first rewrite the objective in (SM5.8) equivalently as

$$\text{(SM5.9)} \qquad \underset{\mathbf{Z},\gamma}{\text{minimize}} - \sum_{j=1}^{d}\sum_{l=1}^{m}\sum_{k=1}^{m} \tilde{p}^j_{lk} \log Z^j_{lk}$$

$$\text{subject to } \mathbf{1}^T\mathbf{Z}^j = \gamma_j \mathbf{1}^T \ \forall j, \mathbf{1}^T\gamma = 1,$$

where $\tilde{p}^j_{lk} = \sum_{t=1} p_{jt} 1_{\left(x_{it}=l, x_{j(t-1)}=k\right)}$. We derive the solution by solving the KKT conditions. The Lagrangian of (SM5.9) is given by

$$\sum_{j=1}^{d}\sum_{l=1}^{m}\sum_{k=1}^{m} \tilde{p}^j_{lk} \log Z^j_{lk} + \sum_j \sum_k \lambda^j_k \left(\left(\sum_l Z^j_{lk}\right) - \gamma_j\right) + \nu\left(1^T\gamma - 1\right),$$

where λ^k_j and ν are Lagrange multipliers. The solution must satisfy the KKT conditions: [SM1]

$$\text{(SM5.10)} \qquad Z^j_{lk} = \frac{\tilde{p}^j_{lk}}{\lambda^j_k} \ \forall j, l, k,$$

$$\text{(SM5.11)} \qquad \nu = \sum_k \lambda^j_k \ \forall j,$$

$$\text{(SM5.12)} \qquad \mathbf{1}^T\mathbf{Z}^j = \gamma_j \mathbf{1}^T \ \forall j, \quad \mathbf{1}^T\gamma = 1.$$

Summing over Equation (SM5.10) for all rows l gives

$$\gamma_j = \frac{\sum_l \tilde{p}^j_{lk}}{\lambda^j_k}.$$

Re-arranging and summing over k gives

$$\frac{\sum_{lk} \tilde{p}^j_{lk}}{\gamma_j} = \sum_k \lambda^j_k = \nu,$$

and finally re-arranging once more and summing over j gives

$$\frac{\sum_j \sum_{lk} \tilde{p}^j_{lk}}{\nu} = \sum_j \gamma_j = 1.$$

Plugging these results back into those above implies that $\nu = \sum_j \sum_{lk} \tilde{p}^j_{lk}$, $\gamma_j = \frac{\sum_{lk} \tilde{p}^j_{lk}}{\sum_j \sum_{lk} \tilde{p}^j_{lk}}$, $\lambda^j_k = \frac{(\sum_l \tilde{p}^j_{lk})(\sum_j \sum_{lk} \tilde{p}^j_{lk})}{\sum_{lk} \tilde{p}^j_{lk}}$. Plugging into Equation (SM5.10) gives the final update for $\mathbf{Z}^{(n+1)}$ as

$$Z^{j(n+1)}_{lk} = \left(\frac{\tilde{p}^j_{lk}}{\sum_l \tilde{p}^j_{lk}}\right)\left(\frac{\sum_{lk} \tilde{p}^j_{lk}}{\sum_j \sum_{lk} \tilde{p}^j_{lk}}\right) \tag{SM5.13}$$

$$= P^{j(n+1)}_{lk} \gamma^{(n+1)}_j, \tag{SM5.14}$$

where $P^{j(n+1)}_{lk} = \left(\frac{\sum_{lk} \tilde{p}^j_{lk}}{\sum_j \sum_{lk} \tilde{p}^j_{lk}}\right)$ and $\gamma^{(n+1)}_j = \left(\frac{\tilde{p}^j_{lk}}{\sum_l \tilde{p}^j_{lk}}\right)$.

This update for $P^{j(n+1)}_{lk}$ and $\gamma^{(n+1)}_j$ is identical to the updates for the EM algorithm in the original $(\mathbf{P}, \gamma)$ parameterization [SM6]. Since the MM algorithm on a convex problem converges to a global optima, it follows that the EM algorithm for the original non-convex MTD parameterization also converges to a global optima. ■

REFERENCES

[1] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.
[2] A. Haris, N. Simon, and A. Shojaie, *Generalized sparse additive models*, arXiv preprint arXiv:1903.04641, (2019).
[3] D. R. Hunter and K. Lange, *A tutorial on mm algorithms*, The American Statistician, 58 (2004), pp. 30–37, https://doi.org/10.1198/0003130042836, https://doi.org/10.1198/0003130042836, https://arxiv.org/abs/https://doi.org/10.1198/0003130042836.
[4] M. Jaggi, *Revisiting frank-wolfe: Projection-free sparse convex optimization.*, in ICML (1), 2013, pp. 427–435.
[5] S. Lacoste-Julien and M. Jaggi, *On the global linear convergence of frank-wolfe optimization variants*, in Advances in Neural Information Processing Systems 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds., Curran Associates, Inc., 2015, pp. 496–504, http://papers.nips.cc/paper/5925-on-the-global-linear-convergence-of-frank-wolfe-optimization-variants.pdf.
[6] S. Lèbre and P.-Y. Bourguignon, *An em algorithm for estimation in the mixture transition distribution model*, Journal of Statistical Computation and Simulation, 78 (2008), pp. 713–729, https://doi.org/10.1080/00949650701266666, https://doi.org/10.1080/00949650701266666, https://arxiv.org/abs/https://doi.org/10.1080/00949650701266666.
[7] Y.-A. Ma, N. J. Foti, and E. B. Fox, *Stochastic gradient mcmc methods for hidden markov models*, in Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 2265–2274.
[8] D. Paulin et al., *Concentration inequalities for markov chains by marton couplings and spectral methods*, Electronic Journal of Probability, 20 (2015).
[9] A. Rakhlin, K. Sridharan, and A. Tewari, *Sequential complexities and uniform martingale laws of large numbers*, Probability Theory and Related Fields, 161 (2015), pp. 111–153.
[10] S. Van de Geer, *Estimation and testing under sparsity*, Lecture notes in mathematics, 2159 (2016).
[11] D. Wulsin, E. Fox, and B. Litt, *Parsing epileptic events using a markov switching process model for correlated time series*, in International Conference on Machine Learning, 2013, pp. 356–364.