Granger Causality Networks for Categorical Time Series

Alex Tank
Department of Statistics
University of Washington
alextank@uw.edu

Emily Fox
Department of Statistics
University of Washington
ebfox@uw.edu

Ali Shojaie
Department of Biostatistics
University of Washington
ashojaie@uw.edu

ABSTRACT

We present two model-based methods for learning Granger causality networks for multivariate categorical time series. Our first proposal is based on the mixture transition distribution (MTD) model. Traditionally, MTD is plagued by a nonconvex objective, non-identifiability, and presence of many local optima. To circumvent these problems, we recast inference in the MTD as a convex problem. The new formulation facilitates the application of MTD to high-dimensional multivariate time series. Our second proposal is based on a multi-output logistic autoregressive model, which while a straightforward extension, has not been previously applied to the analysis of multivariate categorial time series. We investigate identifiability conditions of both methods, devise novel optimization algorithms for the MTD, and compare the MTD and mLTD in simulated experiments. Our approach simultaneously provides a comparison of methods for network inference in categorical time series and opens the door to modern, regularized inference in MTD model.

1. INTRODUCTION

Granger causality [1] is a popular framework for assessing the relationships between time series, and has been widely applied in econometrics, neuroscience, and genomics, amongst other fields. Given two time series x and y, the idea is to use the temporal structure of the data to assess whether the past values of one, say x, are predictive of future values of the other, y, beyond what the past of y can predict alone; if so, x is said to *Granger cause* y. Recently, the focus has shifted to inferring Granger causality networks from multivariate time series data, with the goal of uncovering a sparse set of Granger causal relationships amongst the individual univariate time series. Building on the typical autoregressive framework for assessing Granger causality, a majority of approaches for inferring Granger causal networks have focused on real-valued Gaussian time series using the vector autoregressive model (VAR) with sparsity inducing penalties [2, 3]. More recently, this approach has been extended to non-Gaussian data such as multivariate point processes using sparse Hawkes processes [4], count data using autoregressive Poisson generalized linear models [5], or even time series with heavy tails using VAR models with elliptical errors [6]. In contrast, inferring networks for multivariate *categorical* time series has not been studied under this paradigm.

Multivariate categorical time series arise naturally in many domains. For example, we might have health states from various indicators for a patient over time, voting records for a set of politicians, action labels for players on a team, social behaviors for kids in a school, or musical notes in an orchestrated piece. There are also many datasets that can be viewed as binary multivariate time series based on the presence or absence of an action for some set of entities. Likewise, in some applications, collections of continuous-valued time series are each quantized into a set of discrete values, like the weather data from multiple stations analyzed in [7], wind data in [8], stock returns in [9], or sales volume for a collection of products in [10].

Most literature on multivariate categorical time series is based on the *mixture transition distribution* (MTD) model [11, 8, 10]. The MTD model—which is more generally applicable to modeling high-dimensional probability tables—simplifies the transition probability tensor for multivariate Markov chain as a convex sum of pairwise probability tables. The MTD model was originally developed for modeling higher order Markov chains [8, 12], but has since been adopted for multivariate Markov chains [10, 13, 9]. The resulting structure is one that provides a nice analog to VAR processes: the probability of each component of the multivariate series at time t given past values decomposes into a sum over weightings on terms based on individual components at lagged times. While alluring due to its elegant construction and intuitive interpretation, widespread use of the MTD model has been simultaneously plagued by a non-convex objective with many local optima and serious identifiability issues [9, 13, 14]. For this reason, most applications of the MTD model to multivariate time series have looked at a maximum of three or four time series. Another recent line of work has proposed an autoregressive probit transition model to capture the transition dynamics in multivariate Markov chains [9]. While developed to side step the computational thorns of the MTD, the probit model is still highly nonconvex, both in terms of the objective function, as well as the constraints on parameters.

We present a scalable framework for inferring Granger causality networks of categorical time series using the multivariate MTD model. Through a re-parameterization, we

DOI: 10.1145/1235

simultaneously provide the first convex formulation and rigorous identifiability conditions for the MTD model. The convex MTD objective immediately invites both regularized maximum likelihood inference for model selection and the modern suite of convex optimization algorithms with attractive computational properties.

In addition to the MTD framework, we also consider an alternative approach based on generalized linear models (GLMs), in particular a method we call *multinomial logistic transition distribution* (mLTD). GLMs have been used in a suite of structure learning problems, including learning time-varying Ising models [15] and sparse autoregressive networks of multivariate binary [16] and count [17] time series. Multinomial logistic autoregressions have also been developed for univariate categorical time series [18]. Although the proposed mLTD is a straightforward extension that we simply view as a comparison point to the MTD framework, we have not seen such a model applied to multivariate categorical time series. Importantly, our re-parameterization of the MTD model allows us to more easily compare and contrast these alternative procedures. Historically, the general MTD and GLM frameworks appeared nearly concurrently in the 1980's; however, the GLM framework won out because of its superior computational properties. Based on the computational advances presented herein, it is possible to now consider the MTD model as a potential competitor to the GLM framework. Studying the potential theoretical or practical benefits (e.g., in the small sample size regime) of one framework over the other is left as future work.

Our paper is structured as follows. After discussing identifiability conditions for both MTD and mLTD models, we introduce the convex re-parametrization of MTD along with a set of regularization approaches for model selection using both MTD and mLTD. We then develop accelerated proximal gradient algorithms [19] for both the MTD and mLTD models. For the MTD model, this computational approach provides enormous gains over past methods, enabling this model to be applied to large, modern datasets for the first time. Importantly, the computational insights provided in this paper carry over to the suite of other applications of MTD models, beyond the categorical time series which are the focus herein.

2. CATEGORICAL TIME SERIES AND GRANGER CAUSALITY

Let $x_t = (x_{1t}, \dots x_{pt}) \in \mathcal{X}$ denote a p dimensional categorical random variable indexed by time where $\mathcal{X} = (\mathcal{X}_1 \times \mathcal{X}_2 \dots \times \mathcal{X}_p)$, and $\mathcal{X}_i$ denotes the set of possible values of x_{it}. Let $m_i = |\mathcal{X}_i|$ be the cardinality of set $\mathcal{X}_i$, the number of categories series i may take. A length T multivariate categorical time series is the sequence $X = \{x_1, \dots, x_t, \dots, x_T\}$. An order k multivariate Markov chain models the transition probability between the categories at lagged times $t-1, \dots, t-k$ and those at time t using a transition probability tensor:

$$p(x_t|x_{t-1}, \dots) = p(x_t|x_{t-1}, \dots, x_{t-k}). \qquad (1)$$

Due to the complexity of fully parameterizing this transition distribution, it is common to simplify the model and assume that the categories at time t are conditionally independent of one another given the past realizations:

$$p(x_t|x_{t-1}, \dots, x_{t-k}) = \prod_{i=1}^{p} p(x_{it}|x_{t-1}, \dots, x_{t-k}). \qquad (2)$$

For simplicity, we assume $k = 1$, but stress that all models and results equally apply to higher order k. Based on the decomposition in Eq. (2), we define Granger non-causality for two categorical time series x_i and x_j as follows:

Definition 1. *Time series x_j is not Granger causal for time series x_i iff*

$$p(x_{it}|x_{1(t-1)}, \dots, x_{j(t-1)}, \dots x_{p(t-1)}) = p(x_{it}|x_{1(t-1)}, \dots, x_{(j-1)(t-1)}, x_{(j+1)(t-1)}, \dots, x_{p(t-1)})$$

That is, the probability that time series x_i is in a given state at time t is conditionally independent of the value of x_j at time $t-1$ given the values of all other series x_k, $k \neq i, j$, at lag $t-1$. In Sections 2.1 and 2.2, we present modeling frameworks in which we will devise methods for identifying such Granger non-causality statements, using the results and methods in Sections 3 and 4, respectively.

In specifying our models, and throughout the remainder of the paper, we focus in on a single conditional of x_{it} given x_{t-1}. We do this for notational simplicity; otherwise, we would add additional i indices to all model parameters. Recall that based on the decomposition of Eq. (2), the problem of inference and estimation decomposes into independent subproblems over i.

2.1 MTD model

The MTD model [8] provides an elegant and intuitive parameterization of the multivariate Markov transition distribution as a convex combination of pairwise transition probabilities. Specifically, the MTD model is given by:

$$p(x_{it}|x_{1(t-1)}, \dots, x_{p(t-1)}) = \gamma_0 p_0(x_{it}) + \sum_{j=1}^{p} \gamma_j p_j(x_{it}|x_{j(t-1)}), \qquad (3)$$

where p_0 is a probability vector, $p_j(.|.)$ is a pairwise transition probability table between $x_{j(t-1)}$ and x_{it} and $\gamma = (\gamma_0, \gamma_1, \dots, \gamma_p)$ is a $p+1$ dimensional probability distribution such that $\mathbf{1}^T\gamma = 1$ with $\gamma_j \geq 0$, $j = 0, \dots, p$. We let the matrix $\mathbf{P}^j \in \mathbb{R}^{m_i \times m_j}$ with $\mathbf{1}^T\mathbf{P}^j = \mathbf{1}^T$, $\mathbf{P}^j_{lk} \geq 0$, $l = 1, \dots, m_i$, $k = 1, \dots, m_j$, denote the pairwise transitions and $\mathbf{p}^0 \in \mathbb{R}^{m_i}$ the intercept. While past formulations of the MTD model neglect the p_0 intercept term, we show below that it is crucial for model identifiability and consequently, Granger causality inference. Finally, we note that the MTD model may be extended by adding in interaction terms for pairwise effects [11], such as $p_{jk}(x_{it}|x_{j(t-1)}, x_{k(t-1)})$, though we focus our presentation on the simple case above.

2.2 mLTD model

The multinomial logistic transition distribution (mLTD) model is given by:

$$p(x_{it}|x_{1(t-1)}, \dots, x_{p(t-1)}) = \frac{\exp(\mathbf{z}^0_{x_{it}} + \sum_{j=1}^{p} \mathbf{Z}^j_{x_{it}, x_{j(t-1)}})}{\sum_{x' \in \mathcal{X}_i} \exp(\mathbf{z}^0_{x'} + \sum_{j=1}^{p} \mathbf{Z}^j_{x', x_{j(t-1)}})} \qquad (4)$$

where $\mathbf{Z}^j \in \mathbb{R}^{m_i \times m_j}$, and $\mathbf{z}^0 \in \mathbb{R}^{m_i}$. While not used before to model multivariate categorical time series, its close cousin, the probit model, has been utilized for this purpose [9]. The model in [9] is not a natural fit for inferring Granger causality networks both due to the non-convexity of the probit model and the non-convex constraints imposed on the $\mathbf{Z}^j$ matrices, as explained in more detail in the Supplement. Note that, like the MTD model, the mLTD model naturally allows adding interaction terms, though we focus our presentation on the simple case above.

3. IDENTIFIABILITY AND GRANGER CAUSALITY

In this section, we examine conditions under which our model parameterizations are equivalent to statements of Granger non-causality. We also provide conditions on the model parameterizations to ensure identifiability of the model parameters. The proofs of all results are in the Supplement.

To analyze the identifiability of the MTD model, as well as simplify the inference procedure (see Section 4), we introduce the parameterization $\mathbf{Z}^j = \gamma_j \mathbf{P}^j$ and $\mathbf{z}^0 = \gamma_0 \mathbf{p}^0$:

$$p(x_{it}|x_{1(t-1)}, \dots, x_{p(t-1)}) = \mathbf{z}^0_{x_{it}} + \sum_{j=1}^{p} \mathbf{Z}^j_{x_{it}, x_{j(t-1)}}, \qquad (5)$$

where the constraints now become $\mathbf{1}^T \mathbf{Z}^j = \gamma_j \mathbf{1}^T$, $\mathbf{Z}^j \geq 0$ for all j, and $\mathbf{1}^T \gamma = 1$, $\gamma \geq 0$. Under this MTD parameterization and the mLTD specification of Eq. (4), we have the following simple result:

Proposition 2. *In both the MTD model of Eq.* (5) *and the mLTD model of Eq.* (4)*, time series x_j is Granger non-causal for time series x_i iff the columns of $\mathbf{Z}^j$ are all equal.*

Intuitively, if all columns of $\mathbf{Z}^j$ are equal, the transition distribution for x_{it} does not depend on $x_{j(t-1)}$. Based on this simple observation, we might select for Granger non-causality by penalizing the columns of $\mathbf{Z}^j$ to be the same. While this approach is potentially interesting, a more direct, stable method takes into account the conditions required for identifiability of the $\mathbf{Z}^j$ under both models.

Identifiability for the MTD model.

It is well known that the MTD model is non-identifiable [20]. However, the re-parameterization of the MTD model in terms of $\mathbf{Z}^j$ instead of $\gamma_j, \mathbf{P}^j$, combined with the introduction of an intercept term, allows us to explicitly characterize identifiability conditions for this model.

Theorem 3. *Every MTD distribution has a unique parameterization such that the minimal element in each row of $\mathbf{P}^j$ ($\mathbf{Z}^j$) is zero for all j.*

Under these identifiability conditions we may provide an interpretation of the parameters in the MTD model. Specifically, the element Z^j_{mn} denotes the additive increase in probability that x_i is in state m given that x_j is in state n.

Identifiability for the mLTD model.

The non-identifiability of multinomial logistic models is also well known, as is the non-identifiability of generalized linear models with categorical covariates. Combining the standard identifiability restrictions for both settings [21] gives:

Proposition 4. *Every mLTD has a unique parameterization such that first column and last row $\mathbf{Z}^j$ are zero for all j and the last element of $\mathbf{z}^0$ is zero.*

Under the identifiability constraints, at least one element in each row of $\mathbf{Z}^j$ is fixed to zero. This implies that under the required identifiability restrictions for both MTD and mLTD models, x_j is Granger non-causal for x_i iff $\mathbf{Z}^j = 0$ (a special case of all columns being equal). Taken together, if we enforce the identifiability constraints, we may uniquely select for Granger non-causality by enforcing that $\mathbf{Z}^j$ is equal to zero. The identifiability constraints for the mLTD model are handled with linear constraints, while the constraints for the MTD model are non-convex and are instead enforced indirectly, as explained in Section 4.

4. ESTIMATION AND OPTIMIZATION

We now turn to procedures for inferring Granger non-causality statements from observed multivariate categorical time series. In Section 3, we derived that if $\mathbf{Z}^j = 0$, then x_j is Granger non-causal for x_i. To perform model selection, we take a penalized likelihood approach and present a set of penalty terms that encourage $\mathbf{Z}^j = 0$ while maintaining convexity of the overall objective. At first glance, this seems like an ill-suited approach to the MTD model due to the non-convex maximum likelihood problem in the $(\gamma, \mathbf{P})$ parameterization of Eq. (3). A key insight we make in this section is that the same re-parameterization considered in Section 3, which provided a framework for identifiability conditions to be stated, also allows for a convex objective. This change-of-variable trick opens up an array of possibilities for the MTD framework beyond our multivariate categorical time series focus, eliminating the primary barrier to adoption of this method.

4.1 Convex MTD

Maximum likelihood inference for the MTD model under the $(\gamma, \mathbf{P})$ parameterization is given by the non-convex optimization problem:

$$\begin{aligned} &\underset{\mathbf{P},\gamma}{\text{minimize}} - \sum_{t=1}^{T} \log \left(\gamma_0 \mathbf{p}^0_{x_{it}} + \sum_{j=1}^{p} \gamma_j \mathbf{P}^j_{x_{it}\ x_{j(t-1)}} \right) \\ &\text{subject to} \quad \mathbf{1}^T \mathbf{P}^j = \mathbf{1}^T, \ \mathbf{P}^j \geq 0, \ \forall j \qquad \mathbf{1}^T \gamma = 1 \ , \gamma \geq 0 \end{aligned}$$

The non-convexity follows from the multiplication of the γ_j and $\mathbf{P}^j$ terms in the log. The surface is highly non-convex with many local optima, made even worse by the general non-identifiability. Indeed, the set of equivalent models forms a non-convex region in the $(\gamma, \mathbf{P})$ parameterization (i.e., the convex combination of equivalent models is not necessarily another equivalent model), leading to many non-convex shaped ridges and sets of equal probability. Fortunately, optimization may be recast into a convex program using the re-parameterization $\mathbf{Z}^j = \gamma_j \mathbf{P}^j$ and $\mathbf{z}^0 = \gamma_0 \mathbf{p}^0$:

$$\begin{aligned} &\underset{\mathbf{Z},\gamma}{\text{minimize}} - \sum_{t=1}^{T} \log \left(\mathbf{z}^0_{x_{it}} + \sum_{j=1}^{p} \mathbf{Z}^j_{x_{it}\ x_{j(t-1)}} \right) \\ &\text{subject to} \quad \mathbf{1}^T \mathbf{Z}^j = \gamma_j \mathbf{1}^T, \ \mathbf{Z}^j \geq 0, \ \forall j \qquad \mathbf{1}^T \gamma = 1 \ , \gamma \geq 0 \end{aligned}$$

which is convex since it is a linear function composed with log and linear equality and inequality constraints. Furthermore, the sets of equivalent MTD parameterizations have the appealing property:

Proposition 5. *The set of MTD parameters,* $\mathbf{Z}$*, that yield the same factorized conditional distribution* $p(x_{it}|x_{(t-1)})$ *forms a convex set.*

Due to non-identifiability, the maximum likelihood solution is not unique, potentially leading to difficulties assessing convergence. Rather than enforce the non-convex constraints given in Theorem 3 for model identifiability, we instead add a penalty term $\Omega(\mathbf{Z})$, or prior, that biases the solution towards the uniqueness constraints. Letting $L_{\text{MTD}}(\mathbf{Z}) = -\sum_{t=1}^{T} \log\left(\mathbf{z}^0_{x_{it}} + \sum_{j=1}^{p} \mathbf{Z}^j_{x_{it}\ x_{j(t-1)}}\right)$ the regularized estimation problem is given by

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

Theorem 6. *For any* $\lambda > 0$ *and* $\Omega(\mathbf{Z})$ *that does not depend on* $\mathbf{z}^0$ *and is increasing with respect to the absolute value of entries in* $\mathbf{Z}^j$*, the solution to the problem in Eq.* (6) *is contained in the set of identifiable MTD models described in Theorem 3.*

Intuitively, by penalizing the size of the $\mathbf{Z}^j$ matrices, but not the intercept term, the excess probability mass on the $\mathbf{Z}^j$ matrices is shifted over to the intercept. Thus, by introducing a very small penalty or prior, we constrain the solution space to the set of identifiable models. As we explain in Section 4.2, a convenient choice for $\Omega(\mathbf{Z})$ coincides with a regularizer for selecting for Granger causality.

4.2 Model selection in MTD

Recall from Section 3 that Granger non-causality occurs when the columns of $\mathbf{Z}^j$ are identical, and for identification we restricted this to the case of $\mathbf{Z}^j = 0$. Combining with the discussion in Section 4.1, we may thus select for Granger causality by performing penalized maximum likelihood estimation under penalties that encourage the $\mathbf{Z}^j$ matrices to be zero. Ideally, we would solve the problem:

$$\begin{aligned} &\underset{\mathbf{Z},\gamma}{\text{minimize}}\ \ L_{\text{MTD}}(\mathbf{Z}) + \lambda||\gamma_{1:p}||_0 \\ &\text{subject to}\ \ \mathbf{1}^T\mathbf{Z}^j = \gamma_j\mathbf{1}^T,\ \ \mathbf{Z}^j \geq 0\ \forall j,\ \ \mathbf{1}^T\gamma = 1\ , \gamma \geq 0 \end{aligned} \tag{7}$$

where $\lambda \geq 0$ is a regularization parameter, $||\gamma_{1:p}||_0$ is the L_0 norm over the γ weights and we do not regularize the intercept weight γ_0. The L_0 penalty simply counts the number of non-zero γ_j, which is equivalent to the number of non-zero $\mathbf{Z}^j$. This results in a non-convex objective. Instead, we introduce and compare two convex relaxations of this problem. One is the standard L_1 relaxation, as in lasso regression, which simply sums the absolute values of γ_j. One can show that this penalty leads to *soft-thresholding*, where some estimated γ_j are set exactly to zero while others are shrunk relative to the estimates from the unpenalized objective. Note that if γ_0 were included in the L_0 regularization, the L_1 relaxation would not be a suitable regularizer since $1^T\gamma = 1,\ \ \gamma \geq 0$ so the L_1 norm would always be equal to one [22]. Fortunately, the intercept is not included so we may solve:

$$\begin{aligned} &\underset{\mathbf{Z},\gamma}{\text{minimize}}\ \ L_{\text{MTD}}(\mathbf{Z}) + \lambda\sum_{i=1}^{p}\gamma_i \\ &\text{subject to}\ \ \mathbf{1}^T\mathbf{Z}^j = \gamma_j\mathbf{1}^T,\ \ \mathbf{Z}^j \geq 0\ \forall j,\ \ \mathbf{1}^T\gamma = 1\ , \gamma \geq 0, \end{aligned} \tag{8}$$

where the absolute value of the L_1 norm is dropped due the $\gamma \geq 0$ constraint.

Another convex relation of the objective in Eq. (7), as we show in the Supplement, is given by a group lasso penalty on each $\mathbf{Z}^j$:

$$\begin{aligned} &\underset{\mathbf{Z},\gamma}{\text{minimize}}\ \ L_{\text{MTD}}(\mathbf{Z}) + \lambda\sum_{i=1}^{p}||\mathbf{Z}^j||_F \\ &\text{subject to}\ \ \mathbf{1}^T\mathbf{Z}^j = \gamma_j\mathbf{1}^T,\ \ \mathbf{Z}^j \geq 0\ \forall j,\ \ \mathbf{1}^T\gamma = 1\ , \gamma \geq 0 \end{aligned} \tag{9}$$

where $||.||_F$ is the Frobenius norm. Here, we are penalizing $\mathbf{Z}^j$ directly, rather than indirectly via γ_j. The group lasso penalty drives all elements of $\mathbf{Z}^j$ to zero together, such that the optimal solution automatically selects some $\mathbf{Z}^j$ to be all zero and others not. This effect naturally coincides with our conditions of Granger non-causality that *all* elements of $\mathbf{Z}^j = 0$.

Both Eq. (8) and (9) may be rewritten solely in terms of the $\mathbf{Z}^j$ terms by noting that $\gamma_j = \frac{1}{m_j}1^T\mathbf{Z}^j 1$. Defining $\tilde{z}^T = (\text{vec}(\mathbf{Z}_1)^T, \ldots, \text{vec}(\mathbf{Z}_p)^T)$, we can rewrite the constraints as:

$$(I_p \otimes A)\tilde{z} = 0,\ \ \mathbf{1}^T\tilde{z} = m,\ \ \tilde{z} \geq 0\ \ \text{where} \tag{10}$$

$$A = \begin{pmatrix} \mathbf{1}^T_m & -\mathbf{1}^T_m & 0 & 0 & \cdots \\ 0 & \mathbf{1}^T_m & -\mathbf{1}^T_m & 0 & \cdots \\ \cdots & \cdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathbf{1}^T_m & -\mathbf{1}^T_m \end{pmatrix} \tag{11}$$

and I_p is a p dimensional identity matrix and we have assumed the same number of categories across time series, $|\mathcal{X}_i| = m\ \forall i$, for simplicity of presentation. As both Eq. (8) and (9) are continuously differentiable in the interior of the constraint set, we use a projected gradient algorithm for optimization. At each gradient step we solve a quadratic program to project onto the constraint set (see the Appendix for more details).

Finally, we note that the L_1 and group regularizers in Eq. (8) and (9) satisfy the conditions for Theorem 6, implying that their solutions automatically lay in the identifiability restricted set, without having to explicitly impose the constraints.

4.3 Model selection in mLTD

To select for Granger causality in the mLTD model, we add a group lasso penalty to each of the $\mathbf{Z}^j$ matrices, analogously to Eq. (9), leading to the following optimization problem:

$$\begin{aligned} &\underset{\mathbf{Z}}{\text{minimize}}\ \ \sum_{t=1}^{T}\mathbf{z}^0_{x_{it}} + \sum_{j=1}^{p}\mathbf{Z}^j_{x_{it},x_{j(t-1)}} + \\ &\log\left(\sum_{x'\in\mathcal{X}_i}\exp\left(\mathbf{z}^0_{x'} + \sum_{j=1}^{p}\mathbf{Z}^j_{x',x_{j(t-1)}}\right)\right) + \lambda\sum_{j=1}^{p}||\mathbf{Z}^j||_F \\ &\text{subject to}\ \ \mathbf{Z}^j_{:1} = 0, \mathbf{Z}^j_{m_i:} = 0\ \ \forall j. \end{aligned} \tag{12}$$

For two categories, $m_i = 2\ \ \forall i$, this problem reduces to sparse logistic regression for binary time series, which was recently studied theoretically [5]. For optimization, we utilize an accelerated proximal gradient algorithm [19] as described in the Supplement.

4.4 Comparing model selection in MTD and mLTD

Approaches to model selection in MTD and mLTD models are conceptually similar; both add regularizing penalties to enforce elements in $\mathbf{Z}^j$ to zero. However, in practice these two approaches differ.

Inference in the MTD model is still more computationally demanding due to the large number of linear equality and inequality constraints. The constraint projections that we solve by a quadratic program solver become increasingly costly as the number of time series and categories increases. On the other hand, the mLTD model has no constraints and scales more gracefully to higher dimensions. An ongoing line of work is developing approximate projection methods for the MTD that will scale to larger number of time series.

MTD and mLTD are also quite different models; the resulting probability tensor is an additive combination in MTD while in mLTD the parameter combination passes through a nonlinearity. Our experiments in Section 5 explore the difference in modeling power between MTD and mLTD in the context of inferring sparse networks. We see that in some cases the MTD formulation can outperform mLTD.

5. EXPERIMENTS

We perform a set of simulated experiments to compare the MTD and mLTD model selection methods. Specifically, we compare the MTD group lasso, MTD L_1, and mLTD group lasso methods on simulated categorical time series generated from 1) a sparse MTD model 2) a sparse mLTD model and 3) a sparse latent VAR model with quantized outputs. For all experiments we consider time series of length $T = 200$, dimension $p = 15$, and number of categories $m = 3$. Future work will explore a wider range of settings.

Sparse MTD.

For the MTD model, we randomly generate parameters by $\gamma_{ij} \sim \frac{z_{ij}\phi_{ij}}{\sum_{l=1}^{p} z_{il}\phi_{il}}$ where $\phi_i \sim \text{Dirichlet}(\alpha)$ and $z_{ij} \sim \text{Bin}(p)$. We let $p = .15, \alpha = 5$. Columns of $\mathbf{Z}^{ij}$ are generated according to $\mathbf{Z}^{ij}_{:l} \sim \text{Dirichlet}(\gamma)$ with $\gamma = .7$. (Note that here we have added a superscript i to $\mathbf{Z}$ to specifically indicate the j to i interaction, whereas previously we dropped the i index for notational simplicity by assuming we were just looking at the series i term.) To ensure that the columns are not nearly identical in $\mathbf{Z}^{ij}$ (and thus Granger non-causal), $\mathbf{Z}^{ij}$ is sampled until the average total variation norm between the columns is greater than some tolerance ρ. For our simulations, we set $\rho = .3$. A lower value of ρ makes it more difficult to learn the Granger causality graph since some true interactions might be extremely weak.

Sparse mLTD.

For the mTLD model, the $\mathbf{Z}^{ji}$ parameters are generated by $\mathbf{Z}^{ji}_{lk} \sim z_{ji}N(0, \sigma_Z^2)$ where $z_{ji} \sim \text{Bin}(p)$ with $p = .15$.

Sparse Latent VAR.

To examine data generated from neither of the models considered, we simulate from a continuous time series $y_t \in \mathbb{R}^p$ according to a sparse VAR(1):

$$y_t = Ay_{t-1} + \epsilon_t$$

where $\epsilon_t \sim N(0, \sigma^2 I_p)$. We then quantize each dimension, y_{ti}, into m categories to create a categorical time series x_{ti}. For example, when $m = 3$, $x_{ti} = 1$ if y_{ti} is in the $(0, .33)$ quantile of $\{y_{1i}, \dots y_{Ti}\}$, and so forth. The sparse matrix A is generated by first sampling entries $B_{ij} \sim N(0, \sigma_A^2)$ and then setting $A_{ij} = B_{ij}z_{ij}$, where $z_{ij} \sim \text{Bin}(p)$ with $p = .15$.

	mLTD group	MTD group	MTD L_1
mLTD	**0.930**	0.915	0.903
MTD	0.833	**0.850**	0.837
latent VAR	0.667	**0.770**	0.616

Table 1: Average AUC for each data generating / method pair. Results are averages over 5 different simulation seeds.

For all methods, MTD L_1, MTD group lasso, and mLTD group lasso, we use 5-fold cross validation to select the λ tuning parameter over a λ range from $[0, 100]$. To compare the different methods, we calculate the AUC, or area under the ROC curve between the true Granger causality graph and the estimated graph at various thresholds for the λ chosen by cross validation.

The results are displayed in Table 1. We note that the mLTD model performs best when the data are generated from a mLTD, and likelwise for the MTD. Furthermore, it seems the MTD with group lasso consistently outperforms the MTD with L_1 across all conditions. Finally, the MTD with group lasso penalty performs signficantly better than the mLTD model when both methods are misspecified on the latent VAR example. Taken together, these results display both the utility of the MTD model for inference of categorical Granger causality networks and highlight that the group penalty performs best for MTD.

6. DISCUSSION

We have presented and compared two model-based methods for inferring Granger causality networks from multivariate categorical time series. The penalized MTD method leverages both a novel regularized convex objective that simultaneously promotes sparsity and constrains the solution to an identifiable space. The mLTD model, while thoroughly explored in i.i.d. settings, is also introduced for multivariate categorical time series. For optimization, we have developed a novel projected gradient algorithm for the MTD model that harnesses the new convex formulation. Our experiments demonstrate the utility of the MTD model for inference of Granger categorical networks, even under model misspecification. They also consistently suggest that the group lasso MTD method is better than the L_1 method at this task.

There are a number of potential directions for future work. First, we are currently exploring more scalable methods for optimizing the convex sparse MTD objective based on approximate projections and/or active set methods. Active set methods have the potential to greatly reduce computation since the dimensionality of the constraint set becomes significantly smaller and more tractable when many parameters are fixed at zero. We also intend to develop complimentary theory for network recovery rates for both MTD and mLTD models in the high dimensional setting. Finally, it would also be interesting to explore other regularized MTD objectives, such as the nuclear norm on $\mathbf{Z}^j$ when the number of categories per time series is large.

Acknowledgements.
This work was supported in part by ONR Grant N00014-15-1-2380 and NSF CAREER Award IIS-1350133. AT was partially funded by an IGERT fellowship.

7. APPENDIX

7.1 Proofs

Proof of Proposition 2.
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 $x_{j(t-1)}$.

Proof of Theorem 3.
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. Let $\tilde{\mathbf{Z}}^j = \mathbf{Z}^j - \alpha_j$ and $\tilde{z} = z + \sum_{j=1}^p \alpha_j$. This $\tilde{\mathbf{Z}}$ gives the same MTD distribution as $\mathbf{Z}$.

Suppose two parameter sets $\mathbf{X}$ and $\mathbf{Y}$ provide the same MTD distribution. Let $\tilde{\mathbf{X}}$ be the unique reduction of $\mathbf{X}$ and $\tilde{\mathbf{Y}}$ of $\mathbf{Y}$. Suppose $\tilde{\mathbf{Y}} \neq \tilde{\mathbf{X}}$. 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 such that $\tilde{\mathbf{X}}^j_{kl} = 0$ and likewise for l_Y.

If $l_X = l_Y$, let l' be an index such that $\tilde{\mathbf{X}}^j_{kl'} \neq \tilde{\mathbf{Y}}^j_{kl'}$. Let $x_{\setminus j(t-1)}$ be fixed arbitrarily. The value of

$$\begin{aligned}
&p_X(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l') \\
&-p_X(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l_X) = \tilde{\mathbf{X}}^j_{kl'} \\
&\qquad\qquad\neq \tilde{\mathbf{Y}}^j_{kl'} \\
&p_Y(x_t = k | x_{\setminus j(t-1)}, x_{(t-1)j} = l') \\
&-p_Y(x_t = k | x_{\setminus j(t-1)}, x_{(t-1)j} = l_Y) =
\end{aligned}$$

showing the MTD distributions parametrized by $\mathbf{X}$ and $\mathbf{Y}$ are not the same.

If $l_X \neq l_Y$, then

$$\begin{aligned}
&p_X(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l_Y) \\
&-p_X(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l_X) = \tilde{\mathbf{X}}^j_{kl_Y} \\
&\qquad\qquad\neq -\tilde{\mathbf{Y}}^j_{kl_X} \\
&p_Y(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l_Y) \\
&-p_Y(x_t = k | x_{\setminus j(t-1)}, x_{j(t-1)} = l_X) =
\end{aligned}$$

showing the MTD distributions parametrized by $\mathbf{X}$ and $\mathbf{Y}$ are not the same, leading to a contradiction so that $\tilde{\mathbf{X}} = \tilde{\mathbf{Y}}$. The same argument shows that the reduction is unique.

Proof of Proposition 5.
For any two MTD factorizations $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ and any x_{kt} and $x_{(t-1)}$

$$\begin{aligned}
&\sum_{j=1}^p \left(\alpha \mathbf{Z}^j_{x_{kt} x_{j(t-1)}} + (1-\alpha) \tilde{\mathbf{Z}}^j{}_{x_{kt} x_{j(t-1)}} \right) \\
&= \alpha \sum_{j=1}^p \mathbf{Z}^j_{x_{kt} x_{j(t-1)}} + (1-\alpha) \sum_{i=1}^p \tilde{\mathbf{Z}}^j{}_{x_{kt} x_{j(t-1)}} \\
&= \alpha p(x_{kt} | x_{(t-1)}) + (1-\alpha) p(x_{kt} | x_{(t-1)}) \\
&= p(x_{kt} | x_{(t-1)}).
\end{aligned} \tag{13}$$

Proof of Theorem 6.
First, we note that a solution always exists since the log likelihood $L(\mathbf{Z}) = -\sum_{t=1}^T \log \left(z_{x_{jt}} + \sum_{i=1}^p \mathbf{Z}^j_{x_{jt}\ x_{i(t-1)}} \right)$ and penalty are both bounded below by zero and the feasible set is closed and bounded. Suppose an optimal solution is $\mathbf{Z}$ such that there exists some i such that one row, call it k, of $\mathbf{Z}^j$ does not have a zero element. Let $\alpha = \min(\mathbf{Z}^j_{k:})$ be the minimum value in row k and let $\tilde{\mathbf{Z}}^j$ be equal to $\mathbf{Z}^j$ $\forall i$ except that $\tilde{\mathbf{Z}}^j{}_{k:} = \mathbf{Z}^j_{k:} - \alpha$ and $\tilde{z}^j_k = z^j_k + \alpha$. Due to the nonidentifiability of the MTD model $L(\tilde{\mathbf{Z}}) = L(\mathbf{Z})$, while we have that $||\tilde{\mathbf{Z}}^j||_2 < ||\mathbf{Z}^j||_2$, implying for $\lambda > 0$

$$L(\tilde{\mathbf{Z}}) + \lambda \Omega(\tilde{\mathbf{Z}}) < L(\mathbf{Z}) + \lambda \Omega(\mathbf{Z}), \tag{14}$$

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

7.2 Optimization

7.2.1 *Group lasso MTD*

We show that a group lasso over entries in $\mathbf{Z}^j$ is a convex relaxation to the L_0 norm over $\gamma_{1:p}$. For simplicity assume $m_j = m$ $\forall j$. Due to the equality and greater than zero constraints

$$\begin{aligned}
||\gamma_{1:p}||_0 &= || \left(\mathbf{1}^T \mathrm{vec}(\mathbf{Z}^1), \ldots, \mathbf{1}^T \mathrm{vec}(\mathbf{Z}^p) \right) ||_0 && (15) \\
&= \mathrm{rank}(H_1^T H_1) && (16) \\
&= \mathrm{rank}(H_1) && (17)
\end{aligned}$$

where

$$H_1 = \begin{pmatrix} \mathrm{vec}(\mathbf{Z}^1) & 0 & \ldots & 0 \\ 0 & \mathrm{vec}(\mathbf{Z}^2) & \ldots & 0 \\ 0 & \ldots & \ddots & \vdots \\ 0 & \ldots & \ldots & \mathrm{vec}(\mathbf{Z}^p) \end{pmatrix} \tag{18}$$

Thus we can use the nuclear norm on H_1 as a convex relaxation,

$$||H_1||_* = \sum_{i=1}^p ||\mathbf{Z}^j||_F. \tag{19}$$

7.2.2 *Projected gradient MTD*

The gradient of the MTD model over the feasible set is given by:

$$\begin{aligned}
\frac{dL}{dZ^j_{x',x''}} &= \sum_{t=1}^T 1_{x_{it}=x', x_{j(t-1)}=x''} \frac{1}{\mathbf{z}^0_{x_{it}} + \sum_{j=1}^p \mathbf{Z}^j_{x_{it}, x_{j(t-1)}}} \\
&\quad + \lambda \frac{d\Omega}{dZ^j_{x',x''}}.
\end{aligned} \tag{20}$$

Note that for the L_1 and L_2 norms $\Omega(Z)$ is not differentiable when elements are equal to zero. However, note that due to our problem constraints we have that $\mathbf{Z}^j \geq 0$. Since the point of non-differentiability occurs when elements are identically zero, we modify the problem constraints so that $\mathbf{Z}^j \geq \epsilon$ for some small ϵ, so we may ignore the non-differentiability. Following the notation from the main text, let the set $C = \{\tilde{z} | \tilde{z} \geq \epsilon, (I_p \otimes A)\tilde{z} = 0, 1^T \tilde{z} = m\}$. We perform projected gradient descent:

$$\tilde{z}^{k+1} = P_C \left(\tilde{z}^k - \gamma_k \frac{dL}{d\tilde{z}} \right) \tag{21}$$

where γ_k is the step size, which we chose by line search, and $P_C(x)$ is the projection of x onto the set C:

$$\begin{aligned} &\underset{z}{\text{minimize}} \quad ||z - x||_2^2 \\ &\text{subject to} \quad z \geq \epsilon, \quad (I_p \otimes A)z = 0, \quad 1^T z = m. \end{aligned}$$

This is a quadratic program which we solve using the dual method of Goldfarb and Idnani (1982, 1983) as implemented in the R quadratic programming package *quadprog*. Note that simply projecting onto the simplex may be done efficiently in $(\mathcal{O})n \log n$ time [23]. Perhaps there is a similar type algorithm for fast projection onto C. We also utilize a simple acceleration method which we find vastly improves convergence in practice.

8. REFERENCES

[1] Clive WJ Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980.

[2] Fang Han, Huanran Lu, and Han Liu. A direct estimation of high dimensional stationary vector autoregressions. *arXiv preprint arXiv:1307.0293*, 2013.

[3] Ali Shojaie and George Michailidis. Discovering graphical granger causality using the truncating lasso penalty. *Bioinformatics*, 26(18):i517–i523, 2010.

[4] Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional hawkes processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 641–649, 2013.

[5] E. C. Hall, G. Raskutti, and R. Willett. Inference of High-dimensional Autoregressive Generalized Linear Models. *ArXiv e-prints*, May 2016.

[6] Huitong Qiu, Sheng Xu, Fang Han, Han Liu, and Brian Caffo. Robust estimation of transition matrices in high dimensional heavy-tailed vector autoregressive processes. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1843–1851, 2015.

[7] Finale Doshi, David Wingate, Josh Tenenbaum, and Nicholas Roy. Infinite dynamic bayesian networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 913–920, 2011.

[8] Adrian E Raftery. A model for high-order markov chains. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 528–539, 1985.

[9] João Nicolau. A new model for multivariate markov chains. *Scandinavian Journal of Statistics*, 41(4):1124–1135, 2014.

[10] Wai-Ki Ching, Eric S Fung, and Michael K Ng. A multivariate markov chain model for categorical data sequences and its applications in demand predictions. *IMA Journal of Management Mathematics*, 13(3):187–199, 2002.

[11] André Berchtold and Adrian E Raftery. The mixture transition distribution model for high-order markov chains and non-gaussian time series. *Statistical Science*, pages 328–356, 2002.

[12] Adrian Raftery and Simon Tavaré. Estimation and modelling repeated patterns in high order markov chains with the mixture transition distribution model. *Applied Statistics*, pages 179–199, 1994.

[13] Dong-Mei Zhu and Wai-Ki Ching. A new estimation method for multivariate markov chain model with application in demand predictions. In *Business Intelligence and Financial Engineering (BIFE), 2010 Third International Conference on*, pages 126–130. IEEE, 2010.

[14] Andre Berchtold. Estimation in the mixture transition distribution model. *Journal of Time Series Analysis*, 22(4):379–397, 2001.

[15] Amr Ahmed and Eric P Xing. Recovering time-varying networks of dependencies in social and biological studies. *Proceedings of the National Academy of Sciences*, 106(29):11878–11883, 2009.

[16] Eric C Hall, Garvesh Raskutti, and Rebecca Willett. Inference of high-dimensional autoregressive generalized linear models. *arXiv preprint arXiv:1605.02693*, 2016.

[17] Mohammad Taha Bahadori, Yan Liu, and Eric P Xing. Fast structure learning in generalized stochastic processes with latent factors. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 284–292. ACM, 2013.

[18] Benjamin Kedem and Konstantinos Fokianos. Regression models for categorical time series. *Regression Models for Time Series Analysis*, pages 89–137, 2005.

[19] Neal Parikh and Stephen P Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[20] Sophie Lèbre and Pierre-Yves Bourguignon. An em algorithm for estimation in the mixture transition distribution model. *Journal of Statistical Computation and Simulation*, 78(8):713–729, 2008.

[21] Alan Agresti and Maria Kateri. *Categorical data analysis*. Springer, 2011.

[22] Mert Pilanci, Laurent E Ghaoui, and Venkat Chandrasekaran. Recovery of sparse probability measures via convex programming. In *Advances in Neural Information Processing Systems*, pages 2420–2428, 2012.

[23] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l 1-ball for learning in high dimensions. In *Proceedings of the 25th international conference on Machine learning*, pages 272–279. ACM, 2008.