4

Temporal Causal Modeling

Prabhanjan Kambadur[1], Aurélie C. Lozano[2] and Ronny Luss[2]
[1]*Bloomberg LP, USA*
[2]*IBM T.J. Watson Research Center, USA*

4.1 Introduction

Discovering causal relationships in multivariate time series data has many important applications in finance. Consider portfolio management, where one of the key tasks is to quantify the risk associated with different portfolios of assets. Traditionally, correlations amongst assets have been used to manage risk in portfolios. Knowledge of causal structures amongst assets can help improve portfolio management as knowing causality – rather than just correlation – can allow portfolio managers to mitigate risks directly. For example, suppose that an index fund "A" is found to be one of the causal drivers of another index fund "B." Then, the variance of B can be reduced by offsetting the variation due to the causal effects of A. In contrast simply knowing that "A" is correlated with "B" provides no guidance on how to *act* on index "B," as this does not mean that the two indexes are connected by a cause-and-effect relationship; hedging solely based on correlation does not protect against the possibility that correlation is driven by an unknown effect. Moreover, causal structures may be more stable across market regimes as they have more chance to capture effective economic relationships.

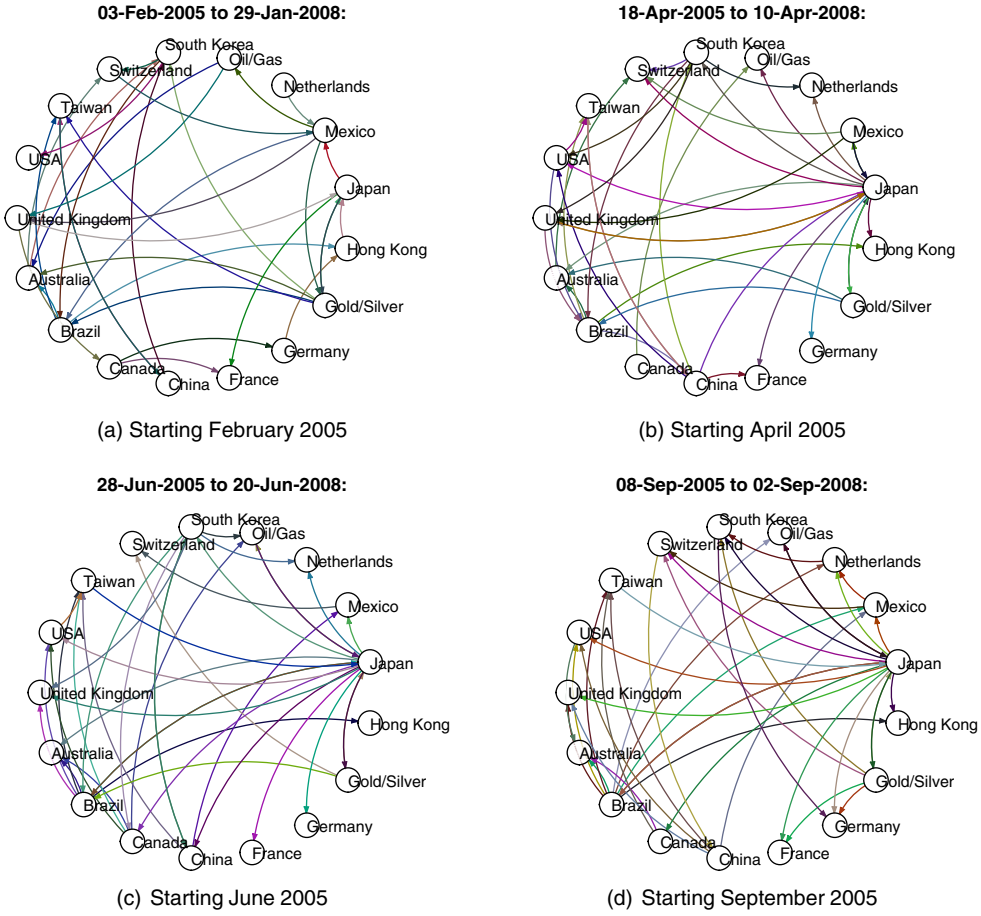
In order to mitigate risks effectively, we need several enhancements to mere causality detection. First, we need to be able to reason about the "strength" of the causal relationship between two assets using statistical measures such as p-values. Attaching well-founded strengths to causal relationships allows us to focus on the important relationships and serves as a guard against false discovery of causal relationships. Second, we need to be able to infer causality in the presence of heteroscedasticity. Typically, causal relationships are modeled by regressing to the conditional mean; this does not always give us a complete understanding of the conditional distributions of the responses based on their causalities. Finally, as the causality amongst assets might be seasonal, we need to be able to automatically identify

Financial Signal Processing and Machine Learning, First Edition.
Edited by Ali N. Akansu, Sanjeev R. Kulkarni and Dmitry Malioutov.

regime changes. If we can successfully discover a temporally accurate causal structure that encodes causal strengths, we would be able to enhance the accuracy of tasks such as explaining the effect of political or financial events on different markets and understanding the microstructure of a financial network.

In this chapter, we discuss *temporal causal modeling* (or TCM) (Lozano *et al*., 2009a,b), an approach that generalizes the notion of Granger causality to multivariate time series by linking the causality inference process to the estimation of sparse vector autoregressive (VAR) models (Section 4.2). Granger causality (Granger, 1980) is an operational definition of causality well known in econometrics, where a source time series is said to "cause" a target time series if it contains additional information for predicting the future values of the target series, beyond the information contained in the past values of the target time series. In essence, TCM combines Granger causality with sparse multivariate regression algorithms, and performs graphical modeling over the lagged temporal variables. We define and use a notion of causal strength modeling (or CSM) for TCM to investigate these potential implications (Section 4.3). We describe how TCM can be extended to the quantile loss function (or Q-TCM) to better model heteroscedactic data (Section 4.4). Finally, we extend TCM to identify regime changes by combining it with a Markov switching modeling framework (Chib, 1998); specifically, we describe a Bayesian Markov switching model for estimating sparse dynamic Bayesian networks (or MS-SDBN) (Section 4.5).

As a concrete case study that highlights the benefits of TCM, consider a financial network of various exchange-traded funds (ETFs) that represent indices tracking a mix of stocks traded on the largest exchanges of various countries. For example, the ticker symbol EWJ represents an ETF that tracks the MSCI Japan Index. We consider a dataset from a family of ETFs called iShares that contains ETF time series of 15 countries as well as an index tracking oil and gas prices and an index tracking the spot price of gold; iShares are managed by Blackrock and the data are publicly available from `finance.yahoo.com`. The causal CSM graphs formed during four different 750-day periods that cover 2005–2008 are shown in Figure 4.1; for interpretability, we use a lag spanning the previous 5 days for vector autoregression. Each feature is a monthly return computed over the previous 22 business days and the lag of 5 days is the monthly return ending on each of those 5 days. Each graph moves the window of data over 50 business days in order to view how time affects the causal networks. Each arc appearing in the CSM causal graphs represents a causal relationship with causal strength greater than a predefined threshold in [0, 1]. Causal strength, as we define it, measures the likelihood that the causal relationship between two nodes is statistically significant. For example, in Figure 4.1a, the causal strength of the relationship directed from South Korea to the United States measures the likelihood that including the South Korea data increases the performance of a United States model. This likelihood is further defined in Section 4.3 (note that this definition of causal strength differs from heuristic notions of causal strength, such as measuring coefficient magnitudes in a linear model). In particular, we are interested in the dependencies of the United States – represented by an ETF that tracks the S&P 500 – during the financial crisis of 2007 and 2008. The panels in Figure 4.1 show an interesting dependence of US-listed equities on Asian-listed equities, which focuses mostly on Japan. To analyze this further, we perform TCM to discover the United States' dependencies over several time periods beginning in 2005 and running through 2015. Table 4.1, which shows the results of our analysis, depicts the causal strength values for the three strongest relationships for each time period.



(a) Starting February 2005

(b) Starting April 2005

(c) Starting June 2005

(d) Starting September 2005

Figure 4.1 Causal CSM graphs of ETFs from iShares formed during four different 750-day periods in 2007–2008. Each graph moves the window of data over 50 business days in order to discover the effect of time on the causal networks. The lag used for VAR spans the 5 days (i.e., uses five features) preceding the target day. Each feature is a monthly return computed over the previous 22 business days.

From Table 4.1, we see that causal relationships change very quickly with time (periods are shifted by 50 business days, resulting in rapid changes to the causal networks). For example, we see that Asian dependencies almost always play a major role for the United States, but the specific Asian countries of interest change. In the early ongoing period of the financial crisis, Japan, South Korea, and China all played major roles, but after the brunt of the crisis occurred (2009 and onwards), South Korea and Germany had the biggest influences. In 2012, Japan and China superseded South Korea as the major dependencies, and 2 years later, Hong Kong played the bigger role. We also note from this analysis that, in 2012, Switzerland overtook Germany as the main European factor for US-listed equities.

Table 4.1 Results of TCM modeling on an ETF that tracks the S&P 500 between 2005 and 2015 that depicts the causal strength values for the three strongest relationships are given for each time period.

Time period	Factor 1	Factor 2	Factor 3	Factor 4
03-Feb-2005 to 29-Jan-2008	South Korea (0.9994)	Mexico (0.9264)	Australia (0.9216)	Gold & Silver (0.0136)
18-Apr-2005 to 10-Apr-2008	Japan (1)	South Korea (1)	China (0.968)	Gold & Silver (0.0162)
28-Jun-2005 to 20-Jun-2008	Brazil (1)	Japan (0.9998)	Canada (0.9994)	Gold & Silver (0.0248)
08-Sep-2005 to 02-Sep-2008	Brazil (1)	Japan (0.9998)	China (0.982)	Gold & Silver (0.0486)
17-Nov-2005 to 11-Nov-2008	China (0.9996)	Switzerland (0.993)	South Korea (0.987)	Gold & Silver (0.3646)
01-Feb-2006 to 26-Jan-2009	South Korea (0.987)	Netherlands (0.973)	Oil & Gas (0.8532)	Gold & Silver (0.5068)
13-Apr-2006 to 07-Apr-2009	Netherlands (0.9998)	South Korea (0.9958)	Oil & Gas (0.8332)	Gold & Silver (0.5622)
26-Jun-2006 to 18-Jun-2009	South Korea (0.998)	Netherlands (0.9878)	Oil & Gas (0.8874)	Gold & Silver (0.6324)
06-Sep-2006 to 28-Aug-2009	South Korea (0.979)	Germany (0.9704)	Oil & Gas (0.8208)	Gold & Silver (0.6126)
15-Nov-2006 to 09-Nov-2009	South Korea (0.981)	Germany (0.9734)	Oil & Gas (0.8078)	Gold & Silver (0.676)
31-Jan-2007 to 22-Jan-2010	South Korea (0.986)	Germany (0.9774)	Oil & Gas (0.7702)	Gold & Silver (0.6364)
13-Apr-2007 to 06-Apr-2010	South Korea (0.9962)	Germany (0.9792)	Gold & Silver (0.7118)	Oil & Gas (0.625)
25-Jun-2007 to 16-Jun-2010	South Korea (0.9922)	Germany (0.934)	Oil & Gas (0.762)	Gold & Silver (0.666)
05-Sep-2007 to 26-Aug-2010	South Korea (0.9828)	Germany (0.9524)	Oil & Gas (0.8022)	Gold & Silver (0.7132)
14-Nov-2007 to 05-Nov-2010	South Korea (0.9736)	Germany (0.9424)	Oil & Gas (0.7744)	Gold & Silver (0.745)
29-Jan-2008 to 19-Jan-2011	Germany (0.9868)	South Korea (0.9718)	Oil & Gas (0.898)	Gold & Silver (0.779)
10-Apr-2008 to 31-Mar-2011	Germany (0.953)	South Korea (0.9308)	Oil & Gas (0.9082)	Gold & Silver (0.7752)
20-Jun-2008 to 13-Jun-2011	Germany (0.9674)	Oil & Gas (0.9208)	South Korea (0.909)	Gold & Silver (0.8036)

02-Sep-2008 to 23-Aug-2011	Germany (0.9924)	Oil & Gas (0.9758)	South Korea (0.9178)	Gold & Silver (0.6388)
11-Nov-2008 to 02-Nov-2011	Brazil (0.9382)	Switzerland (0.8766)	South Korea (0.7974)	Gold & Silver (0.4654)
26-Jan-2009 to 17-Jan-2012	Japan (0.9996)	Switzerland (0.797)	South Korea (0.6104)	Gold & Silver (0.3424)
07-Apr-2009 to 28-Mar-2012	Japan (1)	China (0.99)	Switzerland (0.6638)	Gold & Silver (0.3712)
18-Jun-2009 to 08-Jun-2012	Japan (1)	China (0.9902)	Switzerland (0.749)	Gold & Silver (0.7332)
28-Aug-2009 to 20-Aug-2012	Japan (1)	China (0.9662)	Switzerland (0.8934)	Gold & Silver (0.6304)
09-Nov-2009 to 01-Nov-2012	Japan (1)	Switzerland (0.9404)	China (0.9386)	Gold & Silver (0.5754)
22-Jan-2010 to 15-Jan-2013	Japan (1)	Switzerland (0.9576)	China (0.8048)	Gold & Silver (0.4698)
06-Apr-2010 to 28-Mar-2013	Japan (0.9998)	Switzerland (0.9552)	China (0.8902)	Gold & Silver (0.244)
16-Jun-2010 to 10-Jun-2013	Switzerland (0.9996)	Japan (0.9996)	China (0.7656)	Gold & Silver (0.3372)
26-Aug-2010 to 20-Aug-2013	Switzerland (0.9966)	Brazil (0.9872)	China (0.621)	Gold & Silver (0.4168)
05-Nov-2010 to 30-Oct-2013	Switzerland (1)	Brazil (0.9964)	Hong Kong (0.9558)	Gold & Silver (0.185)
19-Jan-2011 to 13-Jan-2014	Switzerland (1)	Hong Kong (0.976)	Brazil (0.9572)	Gold & Silver (0.1658)
31-Mar-2011 to 26-Mar-2014	Switzerland (1)	Brazil (0.996)	Hong Kong (0.946)	Gold & Silver (0.2392)
13-Jun-2011 to 06-Jun-2014	Switzerland (1)	Hong Kong (0.9946)	Brazil (0.9668)	Gold & Silver (0.185)
23-Aug-2011 to 18-Aug-2014	Hong Kong (0.9604)	Taiwan (0.9598)	Switzerland (0.9564)	Gold & Silver (0.3786)
02-Nov-2011 to 28-Oct-2014	Switzerland (1)	Brazil (0.907)	Gold & Silver (0.787)	Taiwan (0.7262)

4.2 TCM

In this section, we exposit the basic methodology of TCM. In Section 4.1, we introduce Granger causality, which forms the underpinning for TCM. In Section 4.2, we expand the notion of Granger causality to grouping, which allows us to determine causality of an entire time series on other time series. Finally, in Section 4.3, we present experiments on synthetic datasets with known ground truths for the basic TCM methods.

4.2.1 Granger Causality and Temporal Causal Modeling

Granger causality (Granger, 1980), which was introduced by the Nobel prize–winning economist Clive Granger, has proven useful as an operational notion of causality in time-series analysis in the area of econometrics. Granger causality is based on the simple intuition that a cause should precede its effect; in particular, if a "source" time-series causally affects another "target" time-series, then the past values of the source should be helpful in predicting the future values of the target, beyond what can be predicted based only on the target's own past values. That is, a time-series **x** "Granger causes" another time series **y**, if the accuracy of regressing for **y** in terms of past values of **y** and **x** is (statistically) significantly better than that of regressing just with past values of **y**. Let $\{x_t\}_{t=1}^{T}$ denote the time-series variables for x and $\{y_t\}_{t=1}^{T}$ the same for y; then, to determine Granger causality, we first perform the following two regressions:

$$y_t \approx \sum_{l=1}^{L} a_l \cdot y_{t-l} + \sum_{l=1}^{L} b_l \cdot x_{t-l} \tag{4.1}$$

$$y_t \approx \sum_{l=1}^{L} a_l \cdot y_{t-l} \tag{4.2}$$

where L is the maximum "lag" allowed in past observations. To determine whether or not (4.1) is more accurate than (4.2) with a statistically significantly advantage, we perform an F-test or another suitable statistical test. We shall use the term *feature* to mean a time-series (e.g., **x**) and use temporal variables or lagged variables to refer to the individual values (e.g., x_t).

The notion of Granger causality, as introduced above, was defined for a pair of time-series; however, we typically want to determine causal relationships amongst several time-series. Naturally, we use graphical modeling over time-series data to determine conditional dependencies between the temporal variables, and obtain insight and constraints on the causal relationship between the time-series. One technique for graphical modeling is to use regression algorithms with variable selection to determine the causal relationships of each variable; for example, lasso (Tibshirani, 1996), which minimizes the sum of squared errors loss plus a sparsity-inducing ℓ_1 norm penalty on the regression coefficients. That is, we can consider the variable selection process in regression for y_t in terms of $y_{t-1}, x^1_{t-1}, x^2_{t-1}$ and so on, as an application of the Granger test on time-series y against the time-series $x^1, x^2, \ldots, x^p$.[1] When a pairwise Granger test is extended to facilitate multiple causal time-series, we can say that x^1 Granger causes y, if x^1_{t-l} is selected for any time lag $l = \{1, 2, \ldots, L\}$ in the above variable selection. If

[1] Superscripts represent features; for example, x^p is the p^{th} feature or the p^{th} time-series.

such regression-based variable selection coincides with the conditional dependence between the variables, the above operational definition can be interpreted as the key building block of the temporal causal model.

4.2.2 Grouped Temporal Causal Modeling Method

In TCM, we are typically interested in knowing whether an entire time-series $x_{t-1}, x_{t-2}, \ldots, x_{t-L}$ provides information to help predict another time-series y_t; it is of little or no consequence if an individual lag l, x_{t-l} provides additional information for predicting y_t. From a modeling perspective, the relevant variable selection question is not whether an individual lagged variable is to be included in regression, but whether the lagged variables for a given time-series as a *group* are to be included. This would allow us to make statements of the form "**x** Granger causes **y**." Therefore, a more faithful implementation of TCM methods should take into account the group structure imposed by the time-series into the modeling approach and fitting criteria that are used in the variable selection process. This is the motivation for us to turn to the recently developed methodology, group lasso (Yuan and Lin, 2006), which performs variable selection with respect to model-fitting criteria that penalize intragroup and intergroup variable inclusion differently. This argument leads to the generic procedure of the grouped graphical Granger modeling method that is shown in Figure 4.2. We now describe both regularized and greedy regression methods that can serve as **REG** in Figure 4.2. Under regularized methods, we describe with both nongrouped and grouped variable selection techniques: lasso, adaptive lasso, and group lasso. Of these three, we prefer group lasso for the subprocedure **REG** in Figure 4.2 as it performs regression with group variable selection. Lasso and adaptive lasso are not grouped methods and will be used for comparison purposes in the simulations of Section 4.2.3.

1. **Input**
   - Time-series data $\{x_t\}_{t=1,\ldots,T}$ where each $\mathbf{x}_t$ is a p-dimensional vector.
   - A regression method with group variable selection, **REG**.
2. **Initialization**
   Initialize the adjacency matrix for the p features, that is, $G = \langle V, E \rangle$, where V is the set of p features (e.g., by all 0's).
3. **Selection**
   For each feature $\mathbf{y} \in V$, run **REG** on regressing for y_t in terms of the past lagged variables, $x_{t-L}, \ldots, x_{t-1}$, for all the features $\mathbf{x} \in V$ (including **y**). That is, regress $(y_T, y_{T-1}, \ldots, y_{1+L})^T$ in terms of

$$\begin{pmatrix} x^1_{T-1} & \cdots & x^1_{T-L} & \cdots & x^p_{T-1} & \cdots & x^p_{T-L} \\ x^1_{T-2} & \cdots & x^1_{T-1-L} & \cdots & x^p_{T-2} & \cdots & x^p_{T-1-L} \\ \vdots & \vdots & \vdots & \vdots & \vdots\vdots & \vdots & \vdots \\ x^1_L & \cdots & x^1_1 & \cdots & x^p_L & \cdots & x^p_1 \end{pmatrix}$$

   where $V = \{\mathbf{x}^j, j = 1, \ldots, p\}$. For each feature $\mathbf{x}^j \in V$, place an edge $\mathbf{x}^j \to \mathbf{y}$ into E, if and only if $\mathbf{x}^j$ was selected as a group by the grouped variable selection method **REG**.

Figure 4.2 Generic TCM algorithm.

Regularized Least-Squares Methods

Let $\mathbf{y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ be a response vector and let $\mathbf{X} = [\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^p] \in \mathbb{R}^{n \times p}$ be the predictor matrix, where $\mathbf{x}^j = (x_1^j, \dots, x_n^j)^T$, $j = 1, \dots p$, are the covariates. Typically the pairs $(\mathbf{X}_i, \mathbf{y}_i)$ are assumed to be independently identically distributed (i.i.d.) but most results can be generalized to stationary processes given a reasonable decay rate of dependencies such as conditions on the mixing rates. As we are interested in selecting the most important predictors in a high-dimensional setting, the ordinary-least-squares (OLS) estimate is not satisfactory; instead, procedures performing coefficient shrinkage and variable selection are desirable. A popular method for variable selection is the lasso (Tibshirani, 1996), which is defined as:

$$\hat{\boldsymbol{\theta}}_{\text{lasso}}(\lambda) = \arg \min_{\boldsymbol{\theta}} (\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|_1),$$

where λ is a penalty parameter. Here the ℓ_1 norm penalty $\|\boldsymbol{\theta}\|_1$ automatically introduces variable selection, that is $\hat{\boldsymbol{\theta}}_j(\lambda) = 0$ for some $j's$, leading to improved accuracy and interpretability. The lasso procedure – with lag $L = 1$ – has been used for causality analysis in Fujita *et al.* (2007). Unfortunately, lasso tends to overselect the variables and to address this issue, Zou (2006) proposed the adaptive lasso, a two-stage procedure solving:

$$\hat{\boldsymbol{\theta}}_{\text{adapt}}(\lambda) = \arg \min_{\boldsymbol{\theta}} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda \sum_{j=1}^{p} \frac{|\theta_j|}{|\hat{\theta}_{\text{init},j}|} \right),$$

where $\hat{\theta}_{\text{init}}$ is an initial root-n consistent estimator such as that obtained by OLS or Ridge Regression. Notice that if $\hat{\theta}_{\text{init},j} = 0$ then $\forall \lambda > 0$, $\hat{\boldsymbol{\theta}}_{\text{adapt}}(\lambda) = 0$. In addition if the penalization parameter λ is chosen appropriately, adaptive lasso is consistent for variable selection, and enjoys the "Oracle Property", which (broadly) signifies that the procedure performs as well as if the true subset of relevant variables were known. Our final regression method – group lasso (Yuan and Lin, 2006; Zhao *et al.*, 2006) – shines in situations where natural groupings exist between variables, and variables belonging to the same group should be either selected or eliminated as a whole. Given J groups of variables that partition the set of predictors, the group lasso estimate of Yuan and Lin (2006) solves:

$$\hat{\boldsymbol{\theta}}_{\text{group}}(\lambda) = \arg \min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2 + \lambda \sum_{j=1}^{J} \|\boldsymbol{\theta}_{\mathcal{G}_j}\|_2,$$

where $\boldsymbol{\theta}_{\mathcal{G}_j} = \{\theta_k; k \in \mathcal{G}_j\}$ and $\mathcal{G}_j$ denotes the set of group indices. Notice that the penalty term $\lambda \sum_{j=1}^{J} \|\boldsymbol{\theta}_{\mathcal{G}_j}\|_2$ in the above equation corresponds to the sparsity-inducing ℓ_1 norm applied to the J groups of variables, where the ℓ_2 norm is used as the intragroup penalty. In TCM, groups are of equal length as they correspond to the maximum lag that we wish to consider, so the objective does not need to account for unequal group size. By electing to use the ℓ_2 norm as the intragroup penalty, group lasso encourages the coefficients for variables within a given group to be similar in amplitude (as opposed to using the ℓ_1 norm, for example). Note that Granger causality always includes previous values of $\mathbf{y}$ in the model for $\mathbf{y}$; we omit this in the above equation as we assume that the effect of the previous values of $\mathbf{y}$ has "removed" from $\mathbf{y}$. This is done as there is no means to force $\mathbf{y}$ into the model for $\mathbf{y}$ using group-lasso.

Greedy Methods
In lieu of regularized least-squares methods, we could use greedy methods such as the orthogonal patching pursuit algorithm (Lozano *et al*., 2009d) (or OMP) and its variant for group variable selection, group OMP (Lozano *et al*., 2009d). These procedures are iterative and pick the best feature (or feature group) in each iteration, with respect to reduction of the residual error, and then re-estimate the coefficients, $\boldsymbol{\theta}^{(k)}$, via OLS on the restricted sets of selected features (or feature groups). The group OMP procedure is described in Figure 4.3; the classical OMP version can be recovered from Figure 4.3 by considering groups of individual features. Note that – to strictly satisfy the definition of Granger causality – we can forcibly select $\mathbf{y}$ as one of the selected predictors of $\mathbf{y}$ by initializing $\boldsymbol{\theta}^{(0)} = \mathbf{X}_{G_y}$ in Figure 4.3.

4.2.3 Synthetic Experiments

We conducted systematic experimentation using synthetic data in order to test the performance of group lasso and group OMP against that of the nongroup variants (lasso and adaptive lasso) for TCM. We present our findings in this section.

Data Synthesis
As models for data generation, we employed the vector autoregression (VAR) models (Enders, 2003). Specifically, let $\mathbf{x}_t$ denote the vector of all feature values at time t, then a VAR model is defined as $\mathbf{x}_t = \Theta_{t-1} \cdot \mathbf{x}_{t-1} + \ldots + \Theta_{t-T} \cdot \mathbf{x}_{t-T}$, where $\boldsymbol{\Theta}$s are coefficient matrices over the features. We randomly generate an adjacency matrix over the features that determines the structure

1. **Input**
   - The data matrix $\mathbf{X} = [\mathbf{f}_1, \ldots, \mathbf{f}_p] \in \mathbb{R}^{n \times p}$,
   - Group structure $G_1, \ldots, G_J$,
   - The response $\mathbf{y} \in \mathbb{R}^n$.
   - Precision $\epsilon > 0$ for the stopping criterion.
2. **Output**
   - The selected groups $\mathcal{G}^{(k)}$.
   - The regression coefficients $\boldsymbol{\theta}^{(k)}$.
3. **Initialization**
   - $\mathcal{G}^{(0)} = \emptyset$, $\boldsymbol{\theta}^{(0)} = 0$.
4. **Selection**
   For $k = 1, 2, \ldots$
   1. Let $r^{(k-1)} = \mathbf{X}\boldsymbol{\theta}^{(k-1)} - \mathbf{y}$.
   2. Let $j^{(k)} = \arg \min_j \|\mathbf{r}^{(k-1)} - \mathbf{X}_{G_j} \mathbf{X}^+_{G_j} \mathbf{r}^{(k-1)}\|_2$. That is, $j^{(k)}$ is the group that minimizes the residual for the target $\mathbf{r}^{(k-1)}$. $\mathbf{X}^+_{G_j} = (\mathbf{X}^\top_{G_j} \mathbf{X}_{G_j})^{-1} \mathbf{X}_{G_j}$.
   3. **If** $(\|\mathbf{r}^{(k-1)} - \mathbf{X}_{G_j} \mathbf{X}^+_{G_j} \mathbf{r}^{(k-1)}\|_2 \leq \epsilon)$ **break**,
   4. Set $\mathcal{G}^{(k)} = \mathcal{G}^{(k-1)} \cup G_{j^{(k)}}$; $\boldsymbol{\theta}^{(k)} = \mathbf{X}^+_{\mathcal{G}^{(k)}} \mathbf{y}$.

   End

Figure 4.3 Method *group OMP*.

of the true VAR model, and then randomly assign the coefficients – $\mathbf{\Theta}$ – to each edge in the graph. We use the model thus generated on a random initial vector $\mathbf{x}_1$ to generate time-series data $\mathbf{X} = \{\mathbf{x}_t\}_{t=1,\ldots,T}$ of a specified length T. Following Arnold *et al*. (2007), during data generation, we made use of the following parameters: (1) *affinity*, which is the probability that each edge is included in the graph, was set at 0.2; and (2) *sample size per feature per lag*, which is the total data size per feature per maximum lag allowed, was set at 10. We sampled the coefficients of the VAR model according to a normal distribution with mean 0 and standard deviation 0.25. The noise standard deviation was set at 0.1, and so was the standard deviation of the initial distribution.

Evaluation

For all the variable selection subprocedures, the penalty parameter λ is tuned so as to minimize the *BIC* criterion (as recommended in Zou *et al*. (2006)), with degrees of freedom estimated as in Zou *et al*. (2006) for lasso and adaptive lasso, and as in Yuan and Lin (2006) for group lasso. Following Arnold *et al*. (2007), we evaluate the performance of all methods using the F_1 measure, viewing the causal modeling problem as that of predicting the inclusion of the edges in the true graph, or the corresponding adjacency matrix. Briefly, given precision P and recall R, the F_1-measure is defined as $F_1 = \frac{2PR}{(P+R)}$, and hence strikes a balance in the trade-off between the two measures.

Results

Table 4.2 summarizes the results of our experiments, which reports the average F_1 values over 18 runs along with the standard error. These results clearly indicate that there is a significant gap in performance between group lasso and the nongroup counterparts (lasso and adaptive lasso). Figure 4.4 shows some typical output graphs along with the true graph. In this particular instance, it is rather striking how the nongroup methods tend to overselect, whereas the grouped method manages to obtain a perfect graph.

Our experiments demonstrate the advantage of using the proposed TCM method (using group lasso and group OMP) over the standard (nongrouped) methods (based on lasso or adaptive lasso). Note that the nongrouped method based on lasso can be considered as an extension of the algorithm proposed in Fujita *et al*. (2007) to lags greater than one time unit.

Remark 4.1 *In the remainder of this chapter, we present various extensions of TCM, where we alternatively employ group OMP, group lasso, and Bayesian variants to serve as* ***REG*** *in Figure 4.2. We do so in order to expose the reader to the variety of group variable selection approaches. In general, the TCM extensions presented here can be extended to use any of the group OMP, group lasso, or Bayesian variants interchangeably.*

Table 4.2 The accuracy (F_1) and standard error in identifying the correct model of the two nongrouped TCM methods, compared to those of the grouped TCM methods on synthetic data

Method	lasso	adalasso	grplasso	grpOMP
Accuracy (F_1)	0.62 ± 0.09	0.65 ± 0.09	0.92 ± 0.19	0.92 ± 0.08

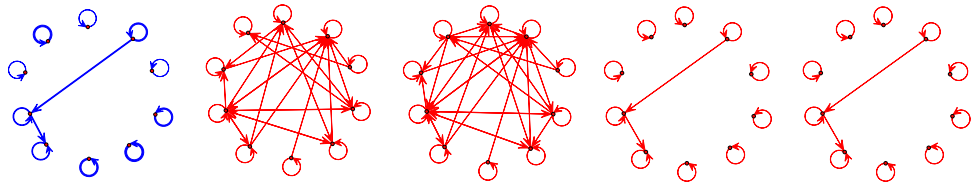


Figure 4.4 Output causal structures on one synthetic dataset by the various methods. In this example, the group-based method exactly reconstructs the correct graph, while the nongroup ones fail badly.

4.3 Causal Strength Modeling

In Section 4.2, we discussed modeling of causal relationships using Granger causality. In particular, in Figure 4.3, we discussed group OMP, a greedy mechanism to compute the causal graph. As causal relationships are defined by multiple lags of each time series, the next logical statistic to consider is the strength of a group, where a group of variables is defined by the different lags of the same time series. In the iShares example discussed in Section 4.1, the causal graphs made use of a notion called causal strength modeling (or CSM). To recap, the causal graphs in Figure 4.1 maintain only those relationships where the causal strength is greater than a given threshold. In this section, we define the notion of causal strength for each causal relationship discovered when *applying the group OMP method to TCM*.[2] The focus here is to determine the likelihood that a selected group of features is a true predictor for the model.

Method

Group OMP (see Figure 4.3) is a greedy feature selection procedure that produces, at each iteration, a new linear model that is fit using least-squares regression. Causal strength is a concept that tells information about the significance of a causal relationship. There are various ways to describe this relationship; a simple heuristic, for example, would be to measure coefficient magnitudes. In this framework, we measure causal strength as the likelihood that the coefficients for a group G in the linear model are nonzero. This likelihood is estimated by testing the probability of the null hypothesis $H_0 : \theta_G = 0$ for all coefficient indices in group G. In general, the distribution of the coefficients of a group of variables can be shown to be Gaussian (assuming Gaussian noise), and the null hypothesis can be tested using an F-test. However, as group OMP selects group G in a greedy fashion, the distribution of the coefficients of a group of variables is only *conditionally* Gaussian and the F-test offers conservative estimates for p-values, which leads to incorrectly declaring certain relationships as causal. To estimate the p-values accurately, we use Monte Carlo simulation. Loftus and Taylor (2014) analyze the conditional Gaussian distribution of the coefficients that group OMP (which they refer to as forward stepwise model selection) discovers at each iteration and discuss how to compute what they term the truncated χ test statistic. This statistic has a uniform distribution and leads to unbiased p-values for the corresponding hypothesis tests.[3]

[2] Although causal strengths can be estimated for group lasso, we have not tried it in practice and therefore omit discussion.

[3] While this procedure is significantly less computationally intensive, the Monte Carlo simulations are sufficient for our purposes. However, we have not had the opportunity to try out this method.

The proxy for testing whether the coefficients of a group added at iteration k are zero tests the statistic $\max_{j \in \bar{J}_{sel}} \|\mathbf{X}_{G_j}^* \hat{\mathbf{r}}^{(k-1)}\|_2$ as a proxy for $\|\boldsymbol{\theta}_{G_j}^{(k)}\|_2$, where $\bar{J}_{sel}$ is the set of remaining groups that can be selected and includes the group selected at iteration k, and $\hat{\mathbf{r}}^{(k-1)}$ is the normalized residual $\mathbf{X}\boldsymbol{\theta}^{(k-1)} - \mathbf{y}$. We want to test whether the normalized residual $\hat{\mathbf{r}}^{(k-1)}$ is noise or contains information. At each iteration, the probability that the null hypothesis holds is estimated by Monte Carlo simulation. First, random vectors are generated in order to create an empirical distribution of $\max_{j \in \bar{J}_{sel}} \|\mathbf{X}_{G_j}^* \mathbf{z}\|_2$ where $\mathbf{z}$ has a standard normal distribution. This represents the distribution of the above statistic that would be observed if the normalized residual were Gaussian. Then, the probability that the null hypothesis should be rejected can be computed from the quantiles of this empirical distribution. Refer to Loftus and Taylor (2014) for further discussion of this Monte Carlo estimation as well as their novel test statistic for testing the null hypothesis. Note that our values of causal strength are defined as $1 - p$ where p is the p-value corresponding to the null hypothesis.

4.4 Quantile TCM (Q-TCM)

The causal models discussed in Section 4.2 consider the causal relationship of the conditional mean of the responses given predictors; yet in many relevant applications, interest lies in the causal relationships for certain quantiles. These relationships may differ from those for the conditional mean or might be more or less pronounced. Therefore, it is critical to develop a complete understanding of the conditional distributions of the responses based on their predictors. In addition, real-world time-series often deviate from the Gaussian distribution, while quantile estimation obtained via the distribution function of the conditional mean regression models is very sensitive to these distributional assumptions. However, quantile regression does not rely on a specific distributional assumption on the data and provides more robust quantile estimation, and is thus more applicable to real-world data. In view of the above desiderata, we present in this section a quantile TCM approach, which extends the traditional VAR model to estimate quantiles, and promotes sparsity by penalizing the VAR coefficients (Aravkin *et al*., 2014). In particular, we extend the group OMP algorithm in Figure 4.3 by replacing the ordinary-least-squares solution for $\hat{\boldsymbol{\theta}}^{(k)}$ at each iteration k with the solution to quantile regression. Section 4.4.1 details the new algorithm for quantile TCM, and Section 4.4.2 uses quantile TCM to perform feature selection on the example from Section 4.1 with additional outliers.

4.4.1 Modifying Group OMP for Quantile Loss

Overview

Figure 4.3 details the group OMP algorithm for ℓ_2 norm loss function. We modify two steps of this algorithm in order to apply it to Q-TCM. First, we generalize the group selection step (4.4.1) for the quantile loss function. Second, the refitting step (4.4.3) is modified to learn the new linear model using a quantile loss function. We begin with a discussion of quantile regression. Let $\mathbf{y}$ and $\mathbf{X}$ denote the response vector and data matrix, respectively. Quantile regression assumes that the τ-th quantile is given by

$$F_{\mathbf{y}|\mathbf{X}}^{-1}(\tau) = \mathbf{X}\bar{\boldsymbol{\theta}}_\tau \tag{4.3}$$

where $\bar{\theta}_\tau \in \mathbb{R}^p$ is the coefficient vector that we want to estimate in p dimensions and $F_{\mathbf{y}|\mathbf{X}}$ is the cumulative distribution function for a multivariate random variable with the same distribution as $\mathbf{y}|\mathbf{X}$. Let $\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\theta}$ be the vector of residuals. Quantile regression is traditionally solved using the following "check-function":

$$c_\tau(\mathbf{r}) = \sum_i (-\tau + 1\{r_i \geq 0\})r_i,$$

where the operations are taken element-wise; note that setting $\tau = 0.5$ yields the least absolute deviation (LAD) loss. Denote the quantile regression solution for a linear model as:

$$\hat{\boldsymbol{\theta}}_{QR,X}(\mathcal{G}, \mathbf{y}) = \arg \min_{\boldsymbol{\theta}} c_\tau(\mathbf{y} - \mathbf{X}_\mathcal{G}\boldsymbol{\theta}).$$

Quantile regression can be solved efficiently using various methods; for example, the regression problem can be rewritten as a linear program. In the case of Q-TCM, quantile regression is used to fit a linear model with the currently selected groups at each iteration. Thus, each iteration solves a larger regression problem, but in practice the total number of groups selected in most applications tends to be small. Our implementation uses an interior point (IP) method described in Aravkin *et al*. (2013) to solve the quantile regression problem. In short, the IP method rewrites the check function in a variational form,

$$c_\tau(\mathbf{r}) = \max_{\mathbf{u}} \{\langle \mathbf{u}, \mathbf{r} \rangle \ : \ \mathbf{u} \in [-\tau, 1-\tau]^n\},$$

and applies the Newton method to solve the optimality conditions of the resulting min-max optimization problem. Structure of the problem is crucial to an efficient implementation. We now discuss how to modify Algorithm 4.3 for Q-TCM. The group selection step is generalized to select the group that maximizes the projection onto the direction of steepest descent (i.e., gradient) with respect to the loss function. In the case of Q-TCM, the selection step becomes:

$$j^{(k)} = \arg \max_j \|\mathbf{X}^*_{G_j}((1-\tau)(\mathbf{r}^{(k-1)})_+ + \tau(\mathbf{r}^{(k-1)})_-)\|_2,$$

where $\mathbf{r}^{(k-1)} = \mathbf{y} - \mathbf{X}\boldsymbol{\theta}^{(k-1)}$ and $(1-\tau)(\mathbf{r}^{(k-1)})_+ + \tau(\mathbf{r}^{(k-1)})_-$ is a subgradient of the quantile loss evaluated at $\mathbf{r}^{(k-1)}$. The second change replaces the refitting step with $\boldsymbol{\theta}^{(k)} = \hat{\boldsymbol{\theta}}_{QR,\mathbf{X}}(\mathcal{G}, y)$. The fully modified version of group OMP tailored for Q-TCM is given in Figure 4.5.

4.4.2 Experiments

We analyze the iShares dataset used in Section 4.1, after introducing a few outliers, and illustrate how using a quantile loss function can discover the same causal relationships in the presence of outliers as an l_2 loss function can discover without the outliers (but cannot discover with them). United States The focus is on learning a model for the United States for the period April 18, 2005, through April 10, 2008 (the second row of Table 4.1). In Section 4.1, we showed that a TCM model for the United States is (statistically) significantly improved by including time-series data pertaining to Japan, South Korea, and China. This is a period where Asian-traded companies have a major influence on US-traded companies. *We introduce three outliers into the original data*. Figure 4.6 displays the original time-series, during the period of interest, with the noisy dates represented by red circles. Note that each data point is a monthly

1. **Input**
   - Data $\mathbf{X} = [\mathbf{f}_1, \ldots, \mathbf{f}_p] \in \mathbb{R}^{n \times p}$
   - Group structure $G_1, \ldots, G_J$, such that $\mathbf{X}_{G_j}^* \mathbf{X}_{G_j} = \mathbf{I}_{d_j}$.
   - The response $\mathbf{y} \in \mathbb{R}^n$
   - Precision $\epsilon > 0$ for the stopping criterion.
2. **Output**
   - The selected groups $\mathcal{G}^{(k)}$.
   - The regression coefficients $\boldsymbol{\theta}^{(k)}$.
3. **Initialization**
   $\mathcal{G}^{(0)} = \emptyset$, $\boldsymbol{\theta}^{(0)} = 0$.
4. **Selection**
   For $k = 1, 2, \ldots$
   - Let $j^{(k)} = \arg \max_j \|\mathbf{X}_{G_j}^* ((1-\tau)(\mathbf{r}^{(k-1)})_+ + \tau(\mathbf{r}^{(k-1)})_-)\|_2$, where $\mathbf{r}^{(k-1)} = \mathbf{y} - \mathbf{X}\boldsymbol{\theta}^{(k-1)}$.
   - **If** $(\|\mathbf{X}_{G_{j(k)}}^* (\mathbf{X}\boldsymbol{\theta}^{(k-1)} - \mathbf{y})\|_2 \leq \epsilon)$ **break**,
   - Set $\mathcal{G}^{(k)} = \mathcal{G}^{(k-1)} \cup G_{j^{(k)}}$. Let $\boldsymbol{\theta}^{(k)} = \hat{\theta}_{QR,X}(\mathcal{G}, \mathbf{y})$.

   End

Figure 4.5 Method *Quantile group OMP*.

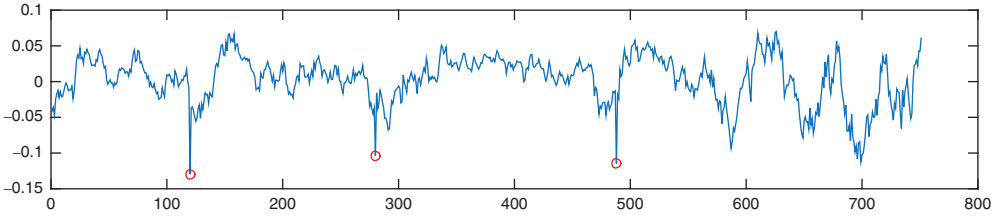


Figure 4.6 Log-returns for ticker IVV (which tracks S&P 500) from April 18, 2005, through April 10, 2008. Outliers introduced on 10/26/2005, 12/14/2007, and 01/16/2008 are represented by red circles.

return and that consecutive points are consecutive days, so that the period for computing the return overlaps on all but one day (and hence there are long periods of correlation as opposed to daily returns that would appear as white noise). While these outliers have been introduced, they are on the same magnitude with the largest magnitude dates of the original data. Outlier detection algorithms could likely detect at least two of the outliers simply because they occur suddenly. The outlier on 01/16/2008 occurs during an event with an already highly negative return.

A look at the data shows possible outliers and that Q-TCM should be considered. The grid of quantiles considered here is a default grid that can be used when no other information is known. Given extra information about the noise, a particular set of quantiles could be considered. We consider the following quantiles: $\tau \in \{0.10, 0.25, 0.35, 0.50, 0.65, 0.75, 0.90\}$. Models here are limited to selecting a total of five regressor time-series, including the target IVV (which tracks S&P 500), so only four additional time-series can be included. The selected regressors for each of the models after running Q-TCM and TCM are given in Table 4.3.

Table 4.3 Time-series selected for IVV, which tracks S&P 500, using Q-TCM and TCM on noisy data. The correct features are South Korea, Japan, and China, which are discovered by Q-TCM

Q-TCM $\tau = 0.10$	Q-TCM $\tau = 0.25$	Q-TCM $\tau = 0.35$	Q-TCM $\tau = 0.50$	Q-TCM $\tau = 0.65$	Q-TCM $\tau = 0.75$	Q-TCM $\tau = 0.90$	TCM
Gld & Slvr	Gld & Slvr	Gld & Slvr	Gld & Slvr	Gld & Slvr	Gld & Slvr	Gld & Slvr	Gld & Slvr
Japan	Oil & Gas	Japan	South Korea	South Korea	South Korea	South Korea	Taiwan
Brazil	France	Mexico	Japan	Japan	Japan	Japan	Canada
South Korea	Mexico	South Korea	China	China	China	China	China

Table 4.4 MSE on test period for Q-TCM and TCM models for IVV on noisy data

Q-TCM $\tau = 0.10$	Q-TCM $\tau = 0.25$	Q-TCM $\tau = 0.35$	Q-TCM $\tau = 0.50$	Q-TCM $\tau = 0.65$	Q-TCM $\tau = 0.75$	Q-TCM $\tau = 0.90$	TCM
0.1229e-3	0.1246e-3	0.1226e-3	0.1228e-3	0.1226e-3	0.1225e-3	0.1226e-3	0.3956e-3

In order to select the best model, we use data for the three months following the model generation period to learn which model fits best. The test period is 04/11/2008 through 07/11/2008. The Q-TCM and TCM models learned above are used to predict returns in the test period, and the mean squared error (MSE) is computed for the predictions on the test data for each model. The losses for the seven Q-TCM models and TCM are given in Table 4.4. TCM clearly has the worst loss on the test data and is thus not a good model to use. The best model, by a very small margin, is the Q-TCM model for the 0.90 quantile, so that model is selected. As we have seen here, this model selects Japan, South Korea, and China as significant factors, which is consistent with the uncorrupted data (see Section 4.1). Note that models with each of the quantiles $\tau \in \{0.50, 0.65, 0.75, 0.90\}$ select the same model as TCM selected on the uncorrupted data (which, while not a ground truth, is an estimate). In this example, the squared loss grossly penalized the outliers, but various versions of an absolute value penalty (i.e., quantile penalty at different quantiles) would have sufficed. With additional noise, Q-TCM was required here to learn a sufficient model, whereas if TCM was used here with the default squared loss, then Taiwan, Canada, and China would have been selected instead of the model selected on the uncorrupted data.

4.5 TCM with Regime Change Identification

In Section 4.2, we described TCM to accurately model causal relationships. However, we did not attempt to capture time-dependent variations in causal relationships; in Section 4.1, we saw an example of iShares, where the causal relationships vary with time. In this section, we extend TCM to incorporate such temporal information and identity regime changes (Jiang *et al.*, 2012). Formally, the main goal of this section is to describe a computationally efficient methodology to model the time-varying dependency structure underlying multivariate time-series data, with a particular focus on regime change identification. For this purpose, we

marry the TCM framework with the Markov switching modeling framework; specifically, we describe a Bayesian Markov switching model for estimating TCMs (MS-TCM).

The key idea is to introduce a latent state variable that captures the regime from which observations at each time period are drawn (Chib, 1998). We allow this latent variable to return to any previous state, which closely resembles reality and allows one to overcome sample scarcity by borrowing strength across samples that are not adjacent in time. Each regime is governed by a TCM model. For group variable selection with TCM, we extend a hierarchical Bayesian framework that adds flexibility to the original group lasso (Yuan and Lin, 2006; Zhao *et al.*, 2006). Briefly, a hierarchical prior is specified for the regression coefficients, which results in *maximum a posteriori* (MAP) estimation with sparsity-inducing regularization; this can be seen as an iteratively reweighted adaptive group lasso estimator. Here, adaptivity refers to the fact that the penalty amount may differ across groups of regression coefficients, similar to adaptive lasso (Zou, 2006). Moreover, the penalty parameter λ is iteratively updated, therefore alleviating the need for parameter tuning (as opposed to non-Bayesian approaches). An additional benefit of such a quasi-Bayesian approach is its computational efficiency, which allows for graceful accommodation of high-dimensional datasets.

By combining a Markov-switching framework with Bayesian group lasso, MS-TCM provides a natural and integrated modeling framework to both capture regime changes and estimate TCM. The rest of this section is laid out as follows. In Section 4.5.1, we present the combined Markov switching model for TCM. In Section 4.5.2, we present algorithms to efficiently solve the combined model. In Sections 4.5.3 and 4.5.4, we present experiments that demonstrate the power of TCM with regime change identification.

4.5.1 Model

Markov-switching Model for TCM

To extend the TCM model in Section 4.2, we propose a Markov switching VAR model as follows: introduce a latent state variable S_t, $S_t \in \{1, 2, \dots, K\}$ for each time point, where K is the total number of possible states and S_t stands for the state at time t. Given the state variables $\{S_t\}$, we can model the observed data $y_{j,t}$ (j^{th} time series at time t) using the VAR model,

$$y_{j,t} = \sum_{i=1}^{p} \sum_{l=1}^{L} \theta_{ijS_t,l} y_{i,t-l} + \epsilon_{j,t}; \quad \epsilon_{j,t} \sim N(0, \sigma^2_{S_t}). \tag{4.4}$$

As before, p is the number of features, L is the maximum lag, and θ_{ijS_tl} is the coefficient of the l^{th} lagged variable of the i^{th} time series for the model of the j^{th} response variable when the regime is given by S_t. Note that the state variables $S_t \in \{1, \dots, K\}$ are defined jointly on all responses; that is, they do not depend on j. This introduces a tight coupling amongst models for the different time-series, which is a departure from Sections 4.2 and 4.4 where a model for each response could be estimated independently. Note that without introducing such coupling amongst the different responses, we would not be able to define and identify regimes and associated change points that are common across all responses. The states S_t are modeled as a Markov chain using $\mathbf{P} \in \mathbb{R}^{K \times K}$ as the transition probability matrix, where P_{ij} is the transition probability from state i to j:

$$P_{ij} = \mathbb{P}(S_t = i | S_{t-1} = j), \forall i, j \in \{1, \dots, K\}.$$

We do not impose any restrictions on the structure of $\mathbf{P}$, which allows for the random process to go back to a previous state or forward to a new state (unlike in Chib (1998). From (4.4), if two time points are from the same state, they will have the same set of autoregression coefficients; that is, $\theta_{ijS_{t1}l} = \theta_{ijS_{t2}l} = \theta_{ijkl}$ iff $S_{t1} = S_{t2} = k$. For simplicity, we denote the regression coefficients at state k by θ_{ijkl} in the rest of this chapter.

Temporal Causal Modeling via the Bayesian Group Lasso

To map the VAR model coefficients into the dependency structure of the TCMs, we make use of a group lasso technique for variable group selection. For a given (i, j, k), we define $\boldsymbol{\theta}_{ijk} = [\theta_{ijk1}, \ldots, \theta_{ijkL}]$ as a coefficient group. We adapt the Bayesian hierarchical framework for group variable selection in Lee *et al*. (2010) as follows:

$$\begin{aligned} \boldsymbol{\theta}_{ijk}|\sigma^2_{ijk} &\sim N(0, \sigma^2_{ijk}), \\ \sigma^2_{ijk}|\tau_{ijk} &\sim \mathrm{G}\left(\frac{L+1}{2}, 2\tau^2_{ijk}\right), \\ \tau_{ijk}|a_{ijk}, b_{ijk} &\sim \mathrm{IG}(a_{ijk}, b_{ijk}), \end{aligned} \tag{4.5}$$

where $\mathrm{G}(a, b)$ represents a gamma distribution with density function $f(x) = x^{a-1}b^{-a}\Gamma(a)^{-1}$ exp $(-x/b)$, and $\mathrm{IG}(a, b)$ represents an inverse gamma distribution whose density function is $f(x) = \frac{b^a}{\Gamma(a)}x^{-a-1}\exp\ (-b/x)$. This hierarchical formulation implies an adaptive version of the group lasso algorithm and allows for automatic update of the smoothing parameters. As suggested in Lee *et al*. (2010), $\boldsymbol{\theta}_{ijk}$ can be estimated by the following MAP estimate:

$$\mathrm{argmax}_{\boldsymbol{\theta}_{ijk}} \log\ \mathcal{L}(\boldsymbol{\theta}_{ijk}|a_{ijk}, b_{ijk}) + \log\ \mathbb{P}(\boldsymbol{\theta}_{ijk}|a_{ijk}, b_{ijk})\ .$$

Integrating out σ^2_{ijk}, τ_{ijk} in (4.5), the marginal density for $\boldsymbol{\theta}_{ijk}$ can be written as

$$\mathbb{P}(\boldsymbol{\theta}_{ijk}|a_{ijk}, b_{ijk}) = \frac{(2b_{ijk})^{-L}\pi^{-(L-1)/2}\Gamma(L + a_{ijk})}{\Gamma((L+1)/2)\Gamma(a_{ijk})}\left(\frac{\|\boldsymbol{\theta}_{ijk}\|_2}{b_{ijk}} + 1\right)^{-a_{ijk}-L},$$

where $\|\boldsymbol{\theta}_{ijk}\|_2 = \sqrt{\sum_{l=1}^{L}\theta^2_{ijkl}}$ is the ℓ_2 norm of $\boldsymbol{\theta}_{ijk}$. We note that the marginal distribution includes the ℓ_2 norm of $\boldsymbol{\theta}_{ijk}$, which is directly related to the penalty term in the group lasso. However, the marginal likelihood resulting from the hierarchical group lasso prior is not concave, which means that search for the global mode by direct maximization is not feasible. An alternative approach proposed in Lee *et al*. (2010) is to find local modes of the posterior using the expectation–maximization (EM) algorithm (McLachlan and Krishnan, 2008) with τ_{ijk} being treated as latent variables. This leads to the following iteratively reweighted minimization algorithm,

$$\boldsymbol{\theta}^{(m+1)}_{ijk} = \mathrm{argmax}_{\boldsymbol{\theta}_{ijk}} \log\ \mathbb{P}(\mathbf{Y}|\boldsymbol{\theta}_{ijk}) - w^{(m)}_{ijk}\|\boldsymbol{\theta}_{ijk}\|_2$$

where $w^{(m)}_{ijk} = \frac{a_{ijk}+L}{\|\boldsymbol{\theta}^{(m)}_{ijk}\|_2 + b_{ijk}}$. For the parameters in the transition probability matrix, we assign the Dirichlet distribution as their prior distributions: for a given state k, the transition probabilities to all possible states $\mathbf{P}_{k\cdot} = [P_{k1}, \ldots, P_{kK}]$ take the form $\mathbf{P}_{k\cdot} \propto \prod_{k'=1}^{K} P_{kk'}^{\alpha_{k'}-1}$, where $\alpha_{k'}$ are the hyperparameters in the Dirichlet distribution. A popular choice is $\alpha_{k'} = 1$, corresponding to a

noninformative prior on the $P_{kk'}$s. The Dirichlet distribution – the conjugate prior for the multinomial distribution – is a popular prior for the probabilities of discrete random variables due to its computational efficiency. Note that the noninformative prior on the transition probability does not imply that the states are equally likely at a transition; the transition probabilities will also be updated according to the data likelihood. Finally, to complete the Bayesian hierarchical model, we assign the following noninformative prior to the variances, $q(\sigma_k^2) \propto \frac{1}{\sigma_k^2}$.[4]

Choosing the Number of States

An important parameter in the previous sections was K, the number of states in the Markov switching network. To determine this, we utilize the Bayesian information criterion (BIC) (Schwarz, 1978), which is defined as $-2\log\ \mathcal{L}(\hat{\boldsymbol{\psi}}_K) + d_K \log\ (N)$, where $\log\ \mathcal{L}(\hat{\boldsymbol{\psi}}_K)$ is the log likelihood of the observed data under $\hat{\boldsymbol{\psi}}_K$, d_K is the number of parameters, and N is the number of observations. The first term in BIC measures the goodness of fit for a Markov model with K states, while the second term is an increasing function of the number of parameters d_K, which penalizes the model complexity. Following (Yuan and Lin, 2006), the complexity of our model with an underlying group sparse structure can be written as:

$$d_K = \sum_{k=1}^{K}\sum_{i=1}^{p}\sum_{j=1}^{p} I(\|\boldsymbol{\theta}_{ijk}\| > 0) + \sum_{k=1}^{K}\sum_{i=1}^{p}\sum_{j=1}^{p} \frac{\|\boldsymbol{\theta}_{ijk}\|}{\|\boldsymbol{\theta}_{ijk}^{LS}\|}(L-1),$$

where $\boldsymbol{\theta}_{ijk}^{LS}$ are the parameters estimated by ordinary-least-squares estimates. We thus estimate $\hat{K}$, the total number of states, by the value that has minimum BIC value.

4.5.2 Algorithm

Let $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_{ijk}\}_{i,j=1,\ldots,d;k=1,\ldots,K}$, be the tensor of parameters, $\boldsymbol{\sigma}^2 = \{\sigma_k^2\}_{k=1,\ldots,K}$, and recall that $\mathbf{P}$ is the transition matrix and $\mathbf{Y}$ is the observed data. The unknown parameters in our model are $\boldsymbol{\psi} = (\boldsymbol{\Theta}, \boldsymbol{\sigma}^2, \mathbf{P})$, which we estimate using MAP estimates obtained by maximizing the posterior distribution $q(\boldsymbol{\psi}|\mathbf{Y})$. In this section, we develop an efficient algorithm to find the MAP estimates using an EM approach (McLachlan and Krishnan, 2008) where the state variables S_t are treated as missing data. When the goal is to find MAP estimates, the EM algorithm converges to the local modes of the posterior distribution by iteratively alternating between an expectation (E) step and a maximization (M) step, as follows. In the E-step, we compute the expectation of the joint posterior distribution of latent variables and unknown parameters, conditional on the observed data, denoted by $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(m)})$. Let $\mathbf{y}_t$ be the p-dimensional vector of observations at time t, and $\mathbf{Y}_{t_1:t_2}$ is the collection of the measurements from time t_1 to t_2. For simplicity, we set $\mathbf{D}_0 = \{\mathbf{y}_1, \ldots, \mathbf{y}_L\}$ to be the initial information consisting of the first L observations and then relabel $\mathbf{y}_t \Rightarrow \mathbf{y}_{t-L}$ for $t > L$. Then we have,

$$\begin{aligned} Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(m)}) &= \mathbb{E}_{\mathbf{S},\boldsymbol{\tau}|\mathbf{Y},\mathbf{D}_0,\boldsymbol{\psi}^{(m)}}[\log\ \mathbb{P}(\boldsymbol{\psi}, S, \boldsymbol{\tau}|\mathbf{Y}_{1:T}, \mathbf{D}_0)] \\ &= \sum_{t=1}^{T}\sum_{k=1}^{K} \mathcal{L}_{tk} \log f(\mathbf{y}_t|\mathbf{Y}_{t-1:t-L}, S_t = k, \boldsymbol{\theta}, \boldsymbol{\sigma}^2) \end{aligned}$$

[4] We use $q(\cdot)$ instead of the commonly used $p(\cdot)$ to denote probability distributions as we use p to refer to the number of features/time-series and P_{ij} to refer to the individual probabilities in $\mathbf{P}$, the probability transition matrix.

$$+\sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{k'=1}^{K} H_{t,k'k}\log\ \mathbb{P}(S_t=k|S_{t-1}=k',\mathbf{P})$$
$$+\sum_{k=1}^{K}\mathcal{L}_{1k}\log\ \pi_k-\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{K} w_{ijk}\|\boldsymbol{\theta}_{ijk}\|_2$$
$$-\sum_{k=1}^{K}\log\ \sigma_k^2+\sum_{k=1}^{K}\sum_{k'=1}^{K}(\alpha_k-1)\log\ P_{k'k}+\text{constant},$$

where $f(\cdot)$ denotes the probability density function, $\mathcal{L}_{tk}=\mathbb{P}(S_t=k|\mathbf{Y}_{1:T},\mathbf{D}_0,\boldsymbol{\psi}^{(m)})$ and $H_{t,k'k}=\mathbb{P}(S_{t-1}=k',S_t=k|\mathbf{Y}_{1:T},\mathbf{D}_0,\boldsymbol{\psi}^{(m)})$ is the posterior probability of all hidden state variables, and $\pi_k=\mathbb{P}(S_1=k|\mathbf{Y}_{1:T},\mathbf{D}_0)$ is the probability of the initial state being k. In the E-step, the posterior probability $\mathcal{L}_{tk}$ and $H_{t,k'k}$ can be calculated using the three-step backward-and-forward algorithm (Baum *et al*., 1970) as follows:

1. Compute the forward probability $\alpha_k^{(m+1)}(t)=\mathbb{P}(\mathbf{Y}_{1:t},S_t=k|\mathbf{D}_0,\boldsymbol{\psi}^{(m)})$ by going forward iteratively in time

$$\begin{aligned}\alpha_k^{(m+1)}(1)&=\mathbb{P}(\mathbf{y}_1,S_1=k|\mathbf{D}_0,\boldsymbol{\psi}^{(m)})\\&=\pi_k^{(m)}\mathbb{P}(\mathbf{y}_1|S_1=k,\mathbf{D}_0,\boldsymbol{\psi}^{(m)})\\\alpha_k^{(m+1)}(t)&=\mathbb{P}(\mathbf{Y}_{1:t},S_t=k|\mathbf{D}_0,\boldsymbol{\psi}^{(m)})\\&=\sum_{k'=1}^{K} f(\mathbf{y}_t|\mathbf{Y}_{t-L:t-1},S_t=k,\mathbf{D}_0,\boldsymbol{\psi}^{(m)})\times P_{k'k}^{(m)}\alpha_{k'}^{(m+1)}(t-1)\end{aligned}$$

2. Compute the backward probability $\beta_k^{(m+1)}(t)=f(\mathbf{Y}_{t+1:T}|\ \mathbf{Y}_{1:t},\mathbf{D}_0,S_t=k,\boldsymbol{\psi}^{(m)})$ by going backward iteratively in time

$$\begin{aligned}\beta_k^{(m+1)}(T)&=1\\\beta_k^{(m+1)}(t)&=f(\mathbf{Y}_{t+1:T}|\mathbf{Y}_{1:t},\mathbf{D}_0,S_t=k,\boldsymbol{\psi}^{(m)})\\&=\sum_{k'=1}^{K} f(\mathbf{y}_{t+1}|\mathbf{Y}_{t:t-L+1},\mathbf{D}_0,S_{t+1}=k',\boldsymbol{\psi}^{(m)})\times\beta_{k'}(t+1)P_{kk'}^{(m)}\end{aligned}$$

3. Compute the posterior probability

$$\begin{aligned}\mathcal{L}_{tk}^{(m+1)}&=\mathbb{P}(S_t=k|\mathbf{Y}_{1:T},\mathbf{D}_0,\boldsymbol{\psi}^{(m)})\\&=\frac{\alpha^{(m+1)}(t)\beta^{(m+1)}(t)}{f(\mathbf{Y}_{1:T}|\mathbf{D}_0,\boldsymbol{\psi}^{(m)})}\ ,\end{aligned}$$

and

$$\begin{aligned}H_{t,k'k}^{(m+1)}&=\mathbb{P}(S_{t-1}=k',S_t=k|\mathbf{Y}_{1:T},\mathbf{D}_0,\boldsymbol{\psi}^{(m)})\\&=f(\mathbf{y}_t|S_t=k,\mathbf{Y}_{t-1:t-L},\boldsymbol{\psi}^{(m)})\times\frac{\alpha_{k'}^{(m+1)}(t-1)p_{k'k}^{(m)}\beta_k^{(m+1)}(t)}{f(\mathbf{Y}|\mathbf{D}_0,\boldsymbol{\psi}^{(m)})}\end{aligned}$$

In the M-step, we update $\boldsymbol{\psi}$ by maximizing $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(m)})$, as follows.

1. The VAR coefficients $\boldsymbol{\theta}_{ijk}^{(m+1)}$ are estimated by minimizing

$$\sum_{t=1}^{T} \frac{1}{2} \mathcal{L}_{tk}^{(m+1)} \left(y_{j,t} - \sum_{i=1}^{p} \mathbf{x}_{ti} \boldsymbol{\theta}_{ijk} \right)^2 / \sigma_k^{2,(m)} + \sum_{i=1}^{p} w_{ijk}^{(m+1)} \|\boldsymbol{\theta}_{ijk}\|_2 \ ,$$

   where $\mathbf{x}_{ti} = (y_{i,t-1}, \ldots, y_{i,t-L})$ and the updated weights are calculated as $w_{ijk}^{(m+1)} = \frac{a_{ijk}+L}{\|\boldsymbol{\theta}_{ijk}^{(m)}\|_2 + b_{ijk}}$. This regularized minimization problem can be transformed into a standard group lasso formulation (Yuan and Lin, 2006) by appropriately rescaling $y_{j,t}$ and $\mathbf{x}_{ti}$. The resulting group lasso problem can be solved efficiently by using the optimization procedure proposed by Meier *et al*. (2008).
2. The variance $\sigma_{0,k}^{2,(m+1)}$ of each of the Markov states is updated as

$$\sigma_k^{2,(m+1)} = \sum_{j=1}^{p} \sum_{t=1}^{T} \frac{\mathcal{L}_{tk}^{(m+1)}}{pT_k^{(m+1)} + 2} \left(y_{j,t} - \sum_{i=1}^{p} \mathbf{x}_{ti} \boldsymbol{\theta}_{ijk}^{(m+1)} \right)^2$$

   where $T_k^{(m+1)} = \sum_{t=1}^{T} \mathcal{L}_{tk}^{(m+1)}$.
3. The transition probability $P_{k'k}^{(m+1)}$ is updated as

$$P_{k'k}^{(m+1)} = \frac{\sum_{t=1}^{T} H_{t,k'k}^{(m+1)} + \alpha_k - 1}{\sum_{t=1}^{T} \mathcal{L}_{tk'} + \sum_{k=1}^{K} (\alpha_k - 1)}$$

4. The initial probability $\pi_k^{(m+1)} = \mathcal{L}_{1k}^{(m+1)}$.

To summarize, the proposed EM algorithm is computationally efficient; the $Q(\boldsymbol{\psi}, \boldsymbol{\psi}^{(m)})$ can be derived in closed form in the E-step, the maximization in the M-step can be transformed into a standard group lasso formulation (Yuan and Lin, 2006), and the maximization can be carried out very efficiently by using the optimization procedure in Meier *et al*. (2008). The algorithm iterates between the E-step and M-step until it converges.

4.5.3 Synthetic Experiments

Data Synthesis

We investigate the performance of MS-TCM on synthetic data with respect to identifying the switching states and the resulting TCM networks. We considered $K = 2$ and $K = 5$ states while generating the synthetic data. The synthetic data are generated according to the following steps.

1. *Generate the state assignment sequence*. Instead of assuming that the true generative process for the states is a Markov switching process, we randomly sampled $T/60$ change points and randomly assign states to each block. We relax the Markov assumption to test if our model still enables us to identify the underlying true process under a more general and realistic condition.

2. *Generate a random directed graph (the true network) with a specified edge probability*. We generated a set of $p \times p$ adjacency matrices $\mathbf{A}_1, \ldots, \mathbf{A}_K$, where the entry $a_{ijk} = 1$ indicates that at Markov state k, $\mathbf{y}_i$ influences $\mathbf{y}_j$, and $a_{ijk} = 0$ otherwise. The value of each entry was chosen by sampling from a binomial distribution, where the probability that an entry equals to one was set to 0.2.
3. *Create a sparse Markov switching VAR model that corresponds to the Markov states and the networks generated in the previous two steps*. For each Markov state, the VAR coefficients $a_{ijkl}, l = 1, \ldots, L$ were sampled according to a normal distribution with mean 0 and SD 0.20 if $a_{ijkl} = 1$, and set to be 0 otherwise. The noise SD was set at 0.01 for all states.
4. *Simulate data from the above switching VAR model*. The sample size per feature per lag per state was set at 10. We considered $p = 12$ and maximum lags $L = 1$ and $L = 3$.

Evaluation

To evaluate the accuracy of the switching states estimation, we use the Rand index (Rand, 1971), where we treat the switching modeling problem as clustering T multivariate data vector $\mathbf{y}_t = (y_{1,t}, \ldots, y_{p,t})$ into K Markov states. The Rand index is often used to measure the clustering error and is defined as the fraction of all misclustered pairs of data vectors $(\mathbf{y}_t, \mathbf{y}_s)$. Letting $\mathcal{C}^*$ and $\hat{\mathcal{C}}$ denote the true and estimated clustering maps, respectively, the Rand index is defined by $\mathcal{R} = \frac{\sum_{t<s} I(\hat{\mathcal{C}}(\mathbf{y}_t, \mathbf{y}_s) \neq \mathcal{C}^*(\mathbf{y}_t, \mathbf{y}_s))}{\binom{T}{2}}$. To evaluate the accuracy of the DBN estimation, we use the F_1 score: larger values of the Rand index and F_1 score indicate higher accuracy.

We compare our method (MS-TCM) with two comparison methods: Fused-DBN and TV-DBN. Fused-DBN is a change point detection method extending the method of Kolar *et al*. (2009) to the VAR setting, which estimates change points via fused lasso penalized regression and subsequently estimates sparse static networks for each segment separately. TV-DBN (Song *et al*., 2009) is used, although the simulation setting is not favorable to it, only to illustrate that in real-world settings where networks are not smoothly and constantly varying over time, approaches like TV-DBN usually fall short.

Results

The results of our experiments are summarized in Table 4.5, where we show respectively the average F_1 score for MS-TCM, Fused-DBN, and TV-DBN, and the average Rand index for our method only (since to our knowledge our method is the only method allowing for regime identification in DBNs) over 100 runs, along with the standard errors. Overall, MS-TCM has very good accuracy under a varying number of Markov states. Rand indices do not change much from small- to large-number states, which suggests similar accuracy in change point detection. The F_1 scores, smaller under $L = 1$, suggest that the dependency structure may be more accurately recovered when the relationships involve multiple time lags, illustrating the value of our Bayesian group lasso subprocedure. This approach achieves significantly better F_1 accuracy than the other methods for all the cases considered. Computationally, our algorithm is more efficient when compared to Fused-DBN as Fused-DBN involves applying randomized lasso to a transformed data matrix of dimensions $T \times Tp$; this implies that the number of features effectively considered is T times higher than the actual number of features. This can be a significant impediment even in low-dimensional settings; for example, for 10 features and 1000 time points, one has to work with 10,000 features after the transformation. In addition, Fused-DBN is unable to leverage the grouping structure corresponding to dependencies with $L > 1$.

Table 4.5 Accuracy of comparison methods in identifying the correct Bayesian networks measured by the average Rand index and F_1 score on synthetic data with a varying number of Markov states (K) and lags (L). The numbers in the parentheses are standard errors

Type	Method	K = 2 L = 1	K = 2 L = 3	K = 5 L = 1	K = 5 L = 3
Rand index	MS-TCM	0.94 (0.02)	0.96 (0.02)	0.97 (0.005)	0.97 (0.06)
	Fused-DBN	N/A	N/A	N/A	N/A
	TV-DBN	N/A	N/A	N/A	N/A
F_1 score	MS-TCM	0.76 (0.05)	0.91 (0.12)	0.76 (0.03)	0.87 (0.17)
	Fused-DBN	0.48 (0.02)	0.50 (0.03)	0.53 (0.04)	0.52 (0.02)
	TV-DBN	0.36 (0.02)	0.38 (0.02)	0.50 (0.06)	0.40 (0.04)

Another strength of MS-TCM, compared to Fused-DBN and other change point–based methods proposed in the literature, is in the setting of regime change identification. In the existing methods, once the change points have been estimated, the coefficients are estimated individually for each interval. Hence, if some of the intervals between two subsequent change points are small (which may happen in many practical situations), the algorithms may be forced to work with extremely small sample size, thus leading to poor estimates. In contrast, our method considers all states and allows for return to previous states. It is thus able to borrow strength across a wider number of samples that may be far away in time.

We now turn our attention to the results obtained by MS-TCM, using the synthetic data with $K = 2$ and $L = 1$. According to BIC, the number of states K is estimated as 2. In the left panel of Figure 4.7, we show the Markov path estimated by our method and the true path for a particular simulation run, where the transition jumps highlighted in red in the true path (upper panel) are those missed by our method. As shown in the plot, MS-TCM is able to detect the change points with very little delay. We also observe that MS-TCM tends to miss a transition when the process remains in a single state for too short a duration. In practice, however, such transient jumps rarely happen or may be of less interest in real applications. In the right panel of Figure 4.7 we show the corresponding estimated Bayesian networks along with the true networks. In the true networks (left column), we highlighted in red the *false negatives* (edges that exist in the true graphs but are missed in the estimated graphs) and in the estimated networks (right column) we highlighted in green the *false positives* (edges that do not exist in the true graphs but are selected in the method). As we can see from the plot, the estimated Bayesian networks exhibit reasonable agreement with the true networks.

4.5.4 Application: Analyzing Stock Returns

To demonstrate MS-TCM's utility in finance, we apply MS-TCM on monthly stock return data from 60 monthly stock observations from 2004-10-01 to 2009-09-01 for 24 stocks in six

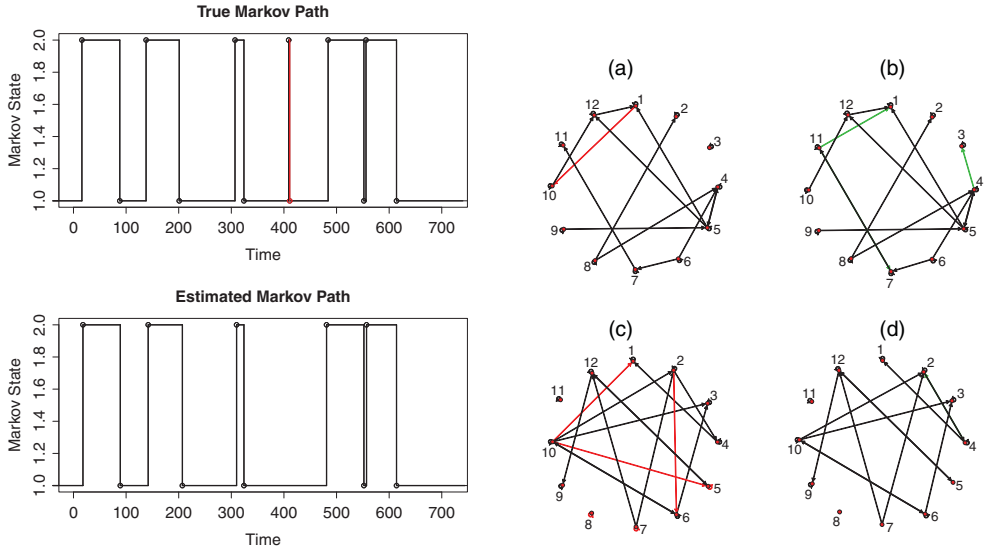


Figure 4.7 (Left) Output switching path on one synthetic dataset with two Markov states. Transition jumps missing in the estimated Markov path are highlighted in red. (Right) The corresponding output networks: (a) true network at state 1; (b) estimated network at state 1; (c) true network at state 2; and (d) estimated network at state 2. Edges coded in red are the false positives, and those in green are the false negatives.

industries. For simplicity, we consider that each state is associated to a VAR model with lag 1. MS-TCM identifies one change point at the 19th time point. The VAR model coefficients for the TCM models concerning Citigroup (C), before and after that change point, are presented in Figure 4.8 along with the model produced when assuming a single static causal model (i.e., running vanilla TCM). By examining the models, one can get some insights on the varying causal relationships. Another benefit of this approach is the potential to improve forecasting accuracy. For instance, if we treat the 60th time point as unobserved, then MS-TCM can forecast this point with relative error of 0.5%, while a static TCM approach has a higher relative error of 26%.[5]

4.6 Conclusions

In this chapter, we presented TCM, a method that builds on Granger causality and generalizes it to multivariate time-series by linking the causality inference process to the estimation of sparse VAR models. We also presented extensions to TCM that allow users to determine causal strengths of causal relationships, use quantile loss functions, and automatically identify and efficiently model regime changes with TCM. TCM and its extensions that were presented in this chapter can be an important tool in financial analysis.

[5] It is possible to leave out and predict more points than 1; we choose to predict just the 60th point for simplicity.

	Model 1	Model 2	Model all
C	-0.2627136	0.21964992	-0.1238139
KEY	0	0	0
WFC	0	0	0
JPM	0	0	0
SO	-1.7279345	-0.1946859	-0.7924643
DUK	-0.7895888	0	-0.2395446
D	0	0	0
HE	1.71362655	0	1.40037993
EIX	0.72517224	0	0.2604309
LUV	1.38827242	0	0.11521195
CAL	0	0	0
AMR	0.39029193	0.11319507	0.3189362
AMGN	-1.3494361	0.02012697	-0.1708315
GILD	0	0	-0.2630309
CELG	0	0	0
GENZ	0	0	0
BIIB	-0.8281948	0	-0.401407
CAT	0	0	0
DE	-0.0239606	0	0
HIT	-0.3431499	0	0
IMO	-0.0428805	0	0
MRO	0	0.00209084	0
HES	0	0	0
YPF	0.34016702	0	0
X.GSPC	0	0	0.41745692

Figure 4.8 Results of modeling monthly stock observations using MS-TCM. MS-TCM uncovered a regime change after the 19th time step; columns `Model 1` and `Model 2` contain the coefficients of the corresponding two TCM models. The column `Model all` gives the coefficients when plain TCM without regime identification is used. The symbols C, KEY, WFC, and JPM are money center banks; SO, DUK, D, HE, and EIX are electrical utilities companies; LUX, CAL, and AMR are major airlines; AMGN, GILD, CELG, GENZ, and BIIB are biotechnology companies; CAT, DE, and HIT are machinery manufacturers; IMO, HES, and YPF are fuel refineries; and X.GPSC is an index.

Further Reading

There are some extensions to TCM that were not covered in this chapter. For example, in order to expose the readers to a variety of approaches, we presented some extensions of TCM using group OMP, and others using group lasso or Bayesian variants. We encourage the readers to experiment with these methods to determine the "best" method for their specific problems. Other extensions that were not covered include extending TCM for modeling spatiotemporal data (Lozano *et al*., 2009c), modeling multiple related time-series datasets (Liu *et al*., 2010), and modeling nonlinear VAR models (Sindhwani *et al*., 2013).

References

Aravkin, A. Y., Burke, J. V. and Pillonetto, G. (2013). Sparse/robust estimation and Kalman smoothing with non-smooth log-concave densities: modeling, computation, and theory. *Journal of Machine Learning Research*, 14, 2689–2728.

Aravkin, A.Y., Kambadur, A., Lozano, A.C., Luss, R. (2014). Orthogonal matching pursuit for sparse quantile regression. IEEE International Conference on Data Mining (ICDM).

Arnold, A., Liu, Y. and Abe, N. (2007). Temporal causal modeling with graphical Granger methods. Proceedings of the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. http://www.cs.cmu.edu/~aarnold/cald/frp781-arnold.pdf

Baum, L., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1), 164–171.

Biard, D. and Ducommun, B. (2008). Moderate variations in CDC25B protein levels modulate the response to DNA damaging agents. *Cell Cycle*, 7(14), 2234–2240.

Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19, 1141–1164.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86, 221–241.

Dahlhaus, R. and Eichler, M. (2003). Causality and graphical models in time series analysis. In *Highly structured stochastic systems* (ed. P. Green, N. Hjort and S. Richardson). Oxford: Oxford University Press.

Davison, A.C. and Hinkley, D. (2006). *Bootstrap methods and their applications*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press.

Enders, W. (2003). *Applied econometric time series*, 2nd ed. Hoboken, NJ: John Wiley & Sons.

Fujita, A., *et al*. (2007). Modeling gene expression regulator networks with the sparse vector autoregressive model. *BMC Systems Biology*, 1, 39.

Furstenthal, L., Kaiser, B.K., Swanson, C. and Jackson, P.K.J. (2001). Cyclin E uses Cdc6 as a chromatin-associated receptor required for DNA replication. *Journal of Cell Biology*, 152(6), 1267–1278.

Granger, C. (1980). Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and Control*, 2, 329–352.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. London: Chapman and Hall.

Jiang, H., Lozano, A.C. and Liu, F. (2012). A Bayesian Markov-switching model for sparse dynamic network estimation. Proceedings of 2012 SIAM International Conference on Data Mining. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.370.3422&rep=rep1&type=pdf

Kolar, M., Song, L. and Xing, E. P. (2009). Sparsistent learning of varying-coefficient models with structural changes. *Advances in Neural Information Processing Systems*, 1006–1014.

Lee, A., Caron, F., Doucet, A. and Holmes, C. (2010). A hierarchical Bayesian framework for constructing sparsity-inducing priors. http://arxiv.org/abs/1009.1914

Li, X., *et al*. (2006). Discovery of time-delayed gene regulatory networks based on temporal gene expression profiling. *BMC Bioinformatics*, 7, 26.

Liu, Y., Fan, S.W., Ding, J.Y., Zhao, X., Shen, W.L. (2007). Growth inhibition of MG-63 cells by cyclin A2 gene-specific small interfering RNA. *Zhonghua Yi Xue Za Zhi*, 87(9), 627–33 (in Chinese).

Liu, Y., Niculescu-Mizil, A., Lozano, A.C. and Lu, Y. (2010). Learning temporal graphs for relational time-series analysis. Proceedings of the 27th International Conference on Machine Learning (ICML). http://www.niculescu-mizil.org/papers/hMRF-ICML_final.pdf

Loftus, J. R. and Taylor J. E. (2014). A significance test for forward stepwise model selection. arXiv:1405.3920

Lozano, A.C., Abe N., Liu Y. and Rosset, S. (2009a). Grouped graphical Granger modeling for gene expression regulatory network discovery. Proceedings of 17th Annual International Conference on Intelligent System for Molecular Biology (ISMB) and Bioinformatics, 25 (12). http://bioinformatics.oxfordjournals.org/content/25/12/i110.short

Lozano, A.C., Abe N., Liu Y., Rosset, S. (2009b). Grouped graphical Granger modeling methods for temporal causal modeling. In *Proceedings of 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2009*, 577–586.

Lozano, A.C., Li, H., Niculescu-Mizil, A., Liu Y., Perlich C., Hosking, J.R.M. and Abe, N. (2009c). Spatial-temporal causal modeling for climate change attribution. In *Proceedings of 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2009*, 587–596.

Lozano, A.C., Swirszcz, G.M. and Abe, N. (2009d). Group orthogonal matching pursuit for variable selection and prediction. *Proceedings of the Neural Information Processing Systems Conference (NIPS)*. http://papers.nips.cc/paper/3878-grouped-orthogonal-matching-pursuit-for-variable-selection-and-prediction.pdf

McLachlan, G.J. and Krishnan, T. (2008). *The EM algorithm and extensions*. New York: Wiley-Interscience.

Meier, L., van de Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society*, 70(1), 53–71.

Meinshausen, N. and Buhlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3), 1436–1462.

Meinshausen, N. and Yu, B. (2006). Lasso-type recovery of sparse representations for high-dimensional data. Technical Report, Statistics. Berkeley: University of California, Berkeley.

Mukhopadhyay, N.D. and Chatterjee, S. (2007). Causality and pathway search in microarray time series experiment. *Bioinformatics*, 23(4).

Ong, I.M., Glasner, J.D. and Page, D. (2002) Modelling regulatory pathways in *E. coli* from time series expression profiles. *Bioinformatics*, 18, S241–S248.

Opgen-Rhein, R. and Strimmer, K. (2007). Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC Bioinformatics*, 8(Suppl. 2), S3.

Rand, W. (1971). Objective criteria for the evaluation of clusterings methods. *Journal of the American Statistical Association*, 66(336), 846–850.

Ray, D. and Kiyokawa, H. (2007). CDC25A levels determine the balance of proliferation and checkpoint response. *Cell Cycle*, 156(24), 3039–3042.

Salon, C., *et al*. (2007). Links E2F-1, Skp2 and cyclin E oncoproteins are upregulated and directly correlated in high-grade neuroendocrine lung tumors. *Oncogene*, 26(48), 6927–6936.

Sambo, F., Di Camillo, B. and Toffolo G. (2008). CNET: an algorithm for reverse engineering of causal gene networks. In: *Bioinformatics methods for biomedical complex systems applications*, 8th Workshop on Network Tools and Applications in Biology NETTAB2008, Varenna, Italy, 134–136.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2).

Segal, E., *et al*. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34, 166–176.

Sindhwani, V., Minh, H.W., Lozano, A.C. (2013). Scalable matrix-valued kernel learning and high-dimensional nonlinear causal inference. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*. http://arxiv.org/ftp/arxiv/papers/1408/1408.2066.pdf

Song, L., Kolar, M. and Xing, E. P. (2009). Time-varying dynamic Bayesian networks. *Advances in Neural Information Processing Systems*, 1732–1740.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1), 267–288.

Whitfield, M.L., *et al*. (2002). Identification of genes periodically expressed in the human cell cycle and their expression in tumors, *Molecular Biology of the Cell*, 13(6). Dataset available at http://genome-www.stanford.edu/Human-CellCycle/Hela/

Xu, X., Wang, L. and Ding, D. (2004). Learning module networks from genome-wide location and expression data. *FEBS Letters*, 578, 297–304.

Yamaguchi, R., Yoshida, R., Imoto, S., Higuchi, T. and Miyano, S. (2007). Finding module-based gene networks in time-course gene expression data with state space models. *IEEE Signal Processing Magazine*, 24, 37–46.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B,* 68, 49–67.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Annals of Statistics*, 35(5), 2173–2192.

Zhao, P., Rocha, G. and Yu, B. (2006). Grouped and hierarchical model selection through composite absolute penalties. http://statistics.berkeley.edu/sites/default/files/tech-reports/703.pdf