Learning the Dependence Graph of Time Series with Latent Factors

Ali Jalali and Sujay Sanghavi ALIJ & SANGHAVI@MAIL.UTEXAS.EDU
University of Texas at Austin, 1 University Station Code:C0806, Austin, TX 78712 USA

Abstract

This paper considers the problem of learning, from samples, the dependency structure of a system of linear stochastic differential equations, when some of the variables are *latent*. We observe the time evolution of some variables, and never observe other variables; from this, we would like to find the dependency structure of the observed variables – *separating out* the spurious interactions caused by the latent variables' time series. We develop a new convex optimization based method to do so in the case when the number of latent variables is smaller than the number of observed ones. For the case when the dependency structure between the observed variables is sparse, we theoretically establish a high-dimensional scaling result for structure recovery. We verify our theoretical result with both synthetic and real data (from the stock market).

1. Introduction

Linear stochastic dynamical systems are classic processes that are widely used due to their simplicity and effectiveness in practice to model time series data in a huge number of fields: financial data (Cochrane, 2005), biological networks of species (Lawrence et al., 2010) or genes (Bar-Joseph, 2004), chemical reactions (Gillespie, 2007; Higham, 2008), control systems with noise (Young, 1984), etc. An important task in several of these domains is learning the model from data which is often the first step in both data interpretation, prediction of future values or perturbation analysis. Often one is interested in learning the *dependency structure*; i.e., identifying, for each variable, which set of other variables it directly interacts with.

This paper considers the problem of structure learning in linear stochastic dynamical systems, in a setting where only a subset of the time series are observed, and others are unobserved/latent. In particular, we consider a system with state vectors $x(t) \in \mathbb{R}^p$ and $u(t) \in \mathbb{R}^r$, for $t \in \mathbb{R}^+$ and dynamics described by

$$\frac{d}{dt}\begin{bmatrix} x(t) \\ u(t) \end{bmatrix} = \underbrace{\begin{bmatrix} A^* & B^* \\ C^* & D^* \end{bmatrix}}_{\mathcal{A}^*}\begin{bmatrix} x(t) \\ u(t) \end{bmatrix} + \frac{d}{dt}w(t), \qquad (1)$$

where, $w(t) \in \mathbb{R}^{p+r}$ is an independent standard Brownian motion vector and A^*, B^*, C^*, D^* are system parameters. We observe the process $x(t)$ for some time horizon $0 \leq t \leq T$, but not the process $u(\cdot)$. We are interested in learning the matrix A^* (both for the continuous and discrete time systems), which captures the interactions between the observed variables. However, the presence of latent time series $u(\cdot)$, if not properly accounted for by the model learning procedure, will result in the appearance of spurious interactions between observed variables especially for classic max-likelihood estimators even over infinite horizon.

Suppose, for illustration, that we are interested in learning the dependency structure between the prices of a set of stocks $x(\cdot)$ via model (1). Clearly, stock prices depend not only on each other, but are also jointly influenced by several variables $u(\cdot)$ that may not be observed, for example, currency markets, commodity prices, etc. The presence of $u(\cdot)$ means that a naive learning algorithm (say LASSO) will report several spurious interactions; say, e.g. between all stocks that fluctuate with the price of oil.

Clearly there are several issues with regards to fundamental identifiability, and sample and computational complexity, that need to be defined and resolved. We do so below in the specific context of our model setting and provide both theoretical guarantees on the problem, as well as numerical illustrations for both synthetic and real data extracted from stock market.

2. Related Work

We organize the most directly related work as follows (recognizing of course that these descriptions overlap).

Appearing in *Proceedings of the 29^{th} International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012. Copyright 2012 by the author(s)/owner(s).

Sparse Recovery and Gaussian Graphical Model Selection: It is now well recognized (Tibshirani, 1996; Wainwright, 2009; Meinshausen & Buhlmann, 2006) that a sparse vector can be tractably recovered from a small number of linear measurements; and also that these techniques can be applied to do model selection (i.e. inferring the Markov graph structure and parameters) in Gaussian graphical models (Meinshausen & Buhlmann, 2006; Ravikumar et al., 2008; d'Aspremont et al., 2007; Friedman et al., 2007; Yuan & Lin, 2007). Two differences between our setting and these papers are that they do not have any latent factors, and theoretical guarantees typically require independent (over time) samples. In particular, latent factors imply that these techniques will in effect attempt to find models that are dense, and hence not be able to have a high-dimensional scaling. Correlation among samples means we cannot directly use standard concentration results, and also brings in the interesting issue of the effect of sampling frequency; in our setting, one can get more samples by finer sampling, but increased correlation means these do not result in better consistency.

Sparse plus Low-Rank Matrix Decomposition: Our results are based on the possibility of separating a low-rank matrix from a sparse one, given their sum (either the entire matrix, or randomly sub-sampled elements thereof) – see (Chandrasekaran et al., 2011; Candes et al., 2009; Chen et al., 2011; Zhou et al., 2010; Candes & Plan, 2010) for some recent results, as well as its applications in graph clustering (Jalali et al., 2011; Jalali & Srebro, 2012), collaborative filtering (Srebro & Jaakkola, 2003), image coding (Hazan et al., 2005), etc. Our setting is different because we observe correlated linear functions of the sum matrix, and furthermore these linear functions are generated by the stochastic linear dynamical system described by the matrix itself. Another difference is that several of these papers focus on recovery of the low-rank component, while we focus on the sparse one. These two objectives have a very different high-dimensional behavior.

Inference with Latent Factors: In real applications of data driven inference, it is always a concern that whether or not there exist influential factors that have never been observed (Loehlin, 1984; West, 2003). Several approaches to this problem are based on Expectation Maximization (EM) (Dempster et al., 1977; Redner & Walker, 1984); while this provides a natural and potentially general method, it suffers from the fact that it can get stuck in local optima (and hence is sensitive to initialization), and that it comes with weak theoretical guarantees. The paper (Chandrasekaran et al., 2010) takes an alternative, convex optimization approach to the latent factor problem in Gaussian graphical models, and is of direct relevance to our paper. In (Chandrasekaran et al., 2010), the objective is to find the number of latent factors in a Gaussian graphical model, given iid samples from the distribution of observed variables; they also use sparse and low-rank matrix decomposition. Differences between our paper and theirs is that we focus on recovering the support of the "sparse part", i.e. the interactions between the observed variables exactly, while they focus on recovery the rank of the low-rank part (i.e. the number of latent variables). Our objective requires $O(\log p)$ samples, theirs requires $\Omega(p)$. Another major difference is that our observations are correlated, and hence sample complexity itself needs a different definition (viz. it is no more the number of samples, but rather the overall time horizon over which the linear system is observed).

System Identification: Linear dynamical system identification is a central problem in Control Theory (Ljung, 1999). There is a long line of work on this problem in that field including expectation maximization (EM) methods (Martens, 2010), Subspace Identification (4SID) methods (Van Overschee & De Moor, 1993), Prediction Error Method (PEM) (Ljung, 2002; Peeters et al.; Fazel et al., 2011). Our problem can be considered as a special case of system identification $\dot{\mathcal{X}} = \mathcal{A}\mathcal{X} + \mathcal{B}\mathcal{U} + \mathcal{W}$ with output $\mathcal{Y} = \mathcal{C}\mathcal{X} + \mathcal{D}\mathcal{U}$, when $\mathcal{X} = [x; u]$, $\mathcal{U} = 0$ and $\mathcal{C}$ is a matrix with identity matrix of size $p \times p$ on its diagonal and zero elsewhere. However, the results in the literature do not provide high-dimensional guarantees for system identification and perhaps our paper is an initial step in that direction. Recently, (Bento et al., 2010) considered a problem similar to ours, *without* any latent variables, i.e., the matrix $\mathcal{C}$ is identity. They implement the LASSO; the main contribution is characterizing sample complexity in the presence of sample dependence. In our setting, with latent variables, their method returns several spurious graph edges caused by marginalization of latent variables.

Time-series Forecasting: Motivated by finance applications, time-series forecasting has got a lot of attention during the past three decades (Chatfield, 2000). In the model based approaches, it is assumed that the time-series evolves according to some statistical model such as linear regression model (Bowerman & O'Connell, 1993), transfer function model (Box et al., 1990), vector autoregressive model (Wei, 1994), etc. In each case, researchers have developed different methods to learn the parameters of the model for the purpose of forecasting. In this paper, we focus on linear stochastic dynamical systems that are an instance of vector autoregressive models. Previous work toward estimating this model parameters include adhoc use of neural network (Azoff, 1994) or support vector machine method (Kim, 2003), all without providing theoretical guarantees on the performance of the algorithm. Our work is different from these results because although our method provides better prediction, our main focus is sparse model selection not prediction. Perhaps, once a sparse model is selected, one can study the prediction as a separate subject.

3. Problem Setting and Main Idea

Other than the continuous time model (1), we are interested in a similar objective for an analogous *discrete time* system with parameter $0 < \eta < \frac{2}{\sigma_{\max}(\mathcal{A}^*)}$ for $\sigma_{\max}(\cdot)$ being the maximum singular value:

$$\begin{bmatrix} x(n+1) \\ u(n+1) \end{bmatrix} - \begin{bmatrix} x(n) \\ u(n) \end{bmatrix} = \eta \begin{bmatrix} A^* & B^* \\ C^* & D^* \end{bmatrix} \begin{bmatrix} x(n) \\ u(n) \end{bmatrix} + w(n) \tag{2}$$

for all $n \in \mathbb{N}_0$. Here, $w(n)$ is a zero-mean Gaussian noise vector with covariance matrix $\eta I_{(p+r)\times(p+r)}$. The parameter η can be thought of as the sampling step; in particular notice that as $\eta \to 0$, we recover model (1) from model (2). The upper bound on η ensures the stability of the discrete time system as required by our theorem. Intuitively, $\sigma_{\max}(\mathcal{A}^*)$ corresponds to the fastest convergence rate (Nyquist sampling rate).

(A1) Stable Overall System: We only consider stable systems. In fact, we impose an assumption slightly stronger than the stability on the overall system. For the continuous system (1), we require $D := -\Lambda_{\max}(\frac{\mathcal{A}^* + \mathcal{A}^{*T}}{2}) > 0$, where $\Lambda_{\max}(\cdot)$ is the maximum eigenvalue. With slightly abuse of notation in the discrete system (2), we require $D := \frac{1-\Sigma_{\max}^2}{\eta} > 0$, where, $\Sigma_{\max} := \sigma_{\max}(I + \eta\mathcal{A}^*)$. ■

As a consequence of this assumption, by Lyapunov theory, the continuous system (1) has a unique stationary measure which is a zero-mean Gaussian distribution with positive definite (otherwise, it is not unique) covariance matrix $\mathcal{Q}^* \in \mathbb{R}^{(p+r)\times(p+r)}$ given by the solution of $\mathcal{A}^*\mathcal{Q}^* + \mathcal{Q}^*\mathcal{A}^{*T} + I = 0$. Similarly, for the discrete time system (2), we have $\mathcal{A}^*\mathcal{Q}^* + \mathcal{Q}^*\mathcal{A}^{*T} + \eta\mathcal{A}^*\mathcal{Q}^*\mathcal{A}^{*T} + I = 0$. This matrix $\mathcal{Q}^*$ has the form $\mathcal{Q}^* = [Q^*\, R^{*T}\, ;\, R^*\, P^*]$, where, Q^* and P^* are the steady-state covariance matrices of the observed and latent variables, respectively, and R^* is the steady-state cross-covariance between observed and latent variables. By stability, we have $\mathcal{C}_{\min} := \Lambda_{\min}(\mathcal{Q}^*) > 0$ and $\mathcal{D}_{\max} := \Lambda_{\max}(\mathcal{Q}^*) < \infty$, where, $\Lambda_{\min}(\cdot)$ is the minimum eigenvalue.

Identifiability: Clearly, the above objective of identifying A^* is in general impossible without some additional assumptions on the model; in particular, several different choices of the overall model (including different choices of A^*) can result in the same *effective* model for the $x(\cdot)$ process. $x(\cdot)$ would then be statistically identical under both models, and correct identification would not be possible even over an infinite time horizon. Additionally, it would in general be impossible to achieve identification if the number of latent variables is comparable to or exceeds the number of observed variables. Thus, to make the problem well-defined, we need to restrict (via appropriate assumptions) the set of models of interest.

3.1. Main Idea

Consider the discrete-time system (2) in steady state and suppose, for a moment, that we ignored the fact that there may be latent time series; in this case, we would be back in the classical setting, for which the (population version of) the likelihood is

$$\mathcal{L}(A) = \frac{1}{2\eta^2}\mathbb{E}\left[\|x(i+1) - x(i) - \eta A x(i)\|_2^2\right].$$

Lemma 1. *For $x(\cdot)$ generated by (2), the the optimum $\bar{A} := \max_A \mathcal{L}(A)$ is given by $\bar{A} = A^* + B^*R^*(Q^*)^{-1}$.*

Thus, the optimal $\bar{A}$ is a sum of the original A^* (which we want to recover) and the matrix $B^*R^*(Q^*)^{-1}$ that captures the spurious interactions obtained due to the latent time series. Notice that the matrix $B^*R^*(Q^*)^{-1}$ has the rank at most equal to number r of latent time series. We will assume that the number of latent time series is smaller than the number of observed ones – i.e. $r < p$ – and hence $B^*R^*(Q^*)^{-1}$ is a *low-rank matrix.*

3.2. Identifiability

Besides identifying the effect of the latent time series, we would need the true model to be such that A^* is uniquely identifiable from $B^*R^*(Q^*)^{-1}$. We choose to study models that have a *local-global structure* where *(a)* each of the observed time series $x_i(t)$ interacts with only a few other observed series, while *(b)* each of the latent series interacts with a (relatively) large number of observed series. In the stock market example, this would model the case where the latent series corresponds to macro-economic factors, like currencies or the price of oil, that affect a lot of stock prices.

In particular, let s be the maximum number of nonzero entries in any row or column of A^* ; it is the maximum number of other observed variables any given observed variable directly interacts with. Note that this means A^* is a *sparse* matrix. Let $L^* := B^*R^*(Q^*)^{-1}$ and assume it has SVD $L^* = U^*\Sigma^*V^{*T}$, and recall that its rank is r. Then, following (Chen et al., 2011), L^* is said to be *μ-incoherent* if $\mu > 0$ is the smallest real number satisfying

$$\max_{i,j}(\|U^{*T}\mathbf{e}_i\|, \|V^{*T}\mathbf{e}_j\|) \leq \sqrt{\frac{\mu r}{p}} \;\;,\;\; \|U^*V^{*T}\|_\infty \leq \sqrt{\frac{r\mu}{p^2}},$$

where, $\mathbf{e}_i$'s are standard basis vectors and $\|\cdot\|$ is vector 2-norm. Smaller values of μ mean the row/column spaces make larger angles with the standard bases, and hence the resulting matrix is more dense.

(A2) Identifiability: We require that the s of the sparse matrix A^* and the μ of the low-rank L^*, which has rank r, satisfy $\alpha := 3\sqrt{\frac{\mu r s}{p}} < 1$. ■

3.3. Algorithm

Recall that our task is to recover the matrix A^* given observations of the $x(\cdot)$ process. We saw that the max-likelihood estimate (in the population case) was the sum of A^* and a low-rank matrix; we subsequently assumed that A^* is sparse. It is natural to use the max-likelihood as the loss function for the *sum* of a sparse and low-rank matrix, and separate appropriate regularizers for each of the components. Thus, for the continuous-time system observed up to time T, we propose solving

$$\begin{aligned}(\widehat{A},\widehat{L}) = & \arg\min_{A,L} \frac{1}{2T}\int_{t=0}^{T}\|(A+L)x(t)\|_2^2\, dt \\ & -\frac{1}{T}\int_{t=0}^{T} x(t)^T(A+L)^T dx(t) + \lambda_A\|A\|_1 + \lambda_L\|L\|_*,\end{aligned} \tag{3}$$

and for the discrete-time system given n samples, we propose solving

$$\begin{aligned}(\widehat{A},\widehat{L}) = & \operatorname{argmin}_{A,L} \frac{1}{2\eta^2 n}\sum_{i=0}^{n-1}\|x(i+1)-x(i)-\eta(A+L)x(i)\|_2^2 \\ & + \lambda_A\|A\|_1 + \lambda_L\|L\|_*.\end{aligned} \tag{4}$$

Here $\|\cdot\|_1$ is the ℓ_1 norm (a convex surrogate for sparsity), and $\|\cdot\|_*$ is the nuclear norm (i.e. sum of singular values, a convex surrogate for low-rank). The optimum $\widehat{A}$ of (4) or (3) is our estimate of A^*, and our main result provides conditions under which we recover the support of A^*, as well as a bound on the error in values $\|\widehat{A}-A^*\|_\infty$ (maximum absolute value). We provide a bound on the error $\|\widehat{L}-L^*\|_2$ (spectral norm) for the low-rank part.

3.4. High-dimensional setting

We are interested in recovering A^* with a number of samples n that is potentially much smaller than p (for small s). In the special case when we are in steady state and $L = 0$ (i.e. λ_L large) the recovery of each row of A^* is akin to a LASSO (Tibshirani, 1996) problem with Q^* being the covariance of the design matrix. We thus require Q^* to satisfy incoherence conditions that are akin to those in LASSO (see e.g. (Wainwright, 2009) for the necessity of such conditions).

(A3) Incoherence: To control the effect of the *irrelevant* (not latent) variables on the set of *relevant* variables, we require $\theta := 1 - \max_k \|Q^*_{\mathcal{S}_k^c\mathcal{S}_k}\left(Q^*_{\mathcal{S}_k\mathcal{S}_k}\right)^{-1}\|_{\infty,1} > 0$, where, $\mathcal{S}_k$ is the support of the k^{th} row of A^* and $\mathcal{S}_k^c$ is the complement of that. The norm $\|\cdot\|_{\infty,1}$ is the maximum of the ℓ_1-norm of the rows. ■

4. Main Results

In this section, we present our main result for both Continuous and Discrete time systems. We start by imposing some assumptions on the regularizers and the sample complexity.

(A4) Regularizers: Let m be the maximum of $\frac{80}{\sqrt{D}}\|B^*\|_{\infty,1}$ and $\sqrt{\|x(0)\|_2^2 + \|u(0)\|_2^2 + (\sqrt{\eta}+1)^2}$ capturing the effect of initial condition and latent variables through matrix B^*. We impose the following assumptions on the regularizers:

(A4-1) $\lambda_A = \frac{16m(4-\theta)}{\theta\sqrt{D}}\sqrt{\frac{\log\left(\frac{4((s+2r)p+r^2)}{\delta}\right)}{n\eta}}$.

(A4-2) $\frac{\lambda_L}{\lambda_A\sqrt{p}} = \frac{1}{1-\alpha}\left(\left(\frac{3\alpha\sqrt{s}}{4} + \frac{(8-\theta)s}{\theta(4-\theta)}\right)\left(\frac{\theta\sqrt{p}}{9s\sqrt{s}}+1\right)+\frac{1}{2}\right)$.

(A5) Sample Complexity: In our setting, the smaller the η is, the more dependent two subsequent samples are. Sample complexity is thus governed by the total time horizon $\eta n = T$ over which we observe the system, and not simply n; indeed finer sampling (i.e. smaller η) requires a larger number of samples. For a probability of failure δ, we require

$$T = n\eta \geq \frac{K\ s^3}{D^2\theta^2\mathcal{C}_{\min}^2}\log\left(\frac{4((s+2r)p+r^2)}{\delta}\right).$$

Here K is a constant independent of any other system parameter; for example, $K \geq 3 \times 10^6$ suffices.

Define parameters $\nu = \frac{\alpha\theta}{2\mathcal{D}_{\max}} + \frac{(8-\theta)\sqrt{s}}{\mathcal{C}_{\min}(4-\theta)}$ and $\rho := \min\left(\frac{\alpha}{4}, \frac{\theta\alpha\lambda_A}{5\theta\alpha\lambda_A + 16\mathcal{D}_{\max}\|L^*\|_2}\right)$. The following (unified) theorem states our main result for both discrete and continuous time systems.

Theorem 1. *If assumptions (A1)-(A5) are satisfied, then with probability $1-\delta$, our algorithm outputs a pair $(\widehat{A},\widehat{L})$ satisfying*

(a) Sub Support Recovery: $Supp(\widehat{A}) \subset Supp(A^*)$.

(b) Error Bounds:

$$\|\widehat{A}-A^*\|_\infty \leq \nu\lambda_A \quad \textit{and} \quad \|\widehat{L}-L^*\|_2 \leq \frac{\rho}{1-5\rho}\|L^*\|_2.$$

(c) Exact Signed Support Recovery: *If addition-*

ally the smallest magnitude A_{min} of a non-zero element of A^* satisfies $A_{min} > \nu\lambda_A$, then we obtain full signed-support recovery Sign$(\widehat{A})$ = Sign(A^*).

Note: Note that λ_A, as defined in **(A4-1)**, depends on the sample complexity T, and goes to 0 as T becomes large. Thus it is possible to get exact signed support recovery by making T large.

Remark 1: Our result shows that, in sparse and low-rank decomposition for latent variable modeling, recovery of only the sparse component seems to be possible with much fewer samples – $O(s^3 \log p)$ – as compared to, for example, the recovery of the exact rank of the low-rank part; the latter was show to require $\Theta(p)$ samples in (Chandrasekaran et al., 2010).

Remark 2: The above theorem shows that, even in the presence of latent variables, our algorithm requires a similar number of samples (i.e. upto universal constants) as previous work (Bento et al., 2010) required in the absence of hidden variables. Of course, this is true as long as identifiability **(A2)** holds. Note that the absence of such identifiability conditions makes even simple sparse and low-rank matrix decomposition ill-posed (Chandrasekaran et al., 2011).

Remark 3: Although our theoretical result shows a scaling of s^3 for the sample complexity, the empirical result suggests that the correct scaling factor is s^2. We suspect our result as well as Bento et al. (2010) can be tightened.

Illustrative Example: Consider a simple idealized example that helps give intuition about the above theorem. Suppose that we are in the continuous time setting, where each latent variable j depends only on its own past, updating according to $\frac{dx_j}{dt} = -x_j(t) + \frac{dw_j}{dt}$ and for each observed variable i depends only on its own past and a *unique* latent variable $j(i)$, i.e., $\frac{dx_i}{dt} = -x_i(t) + x_{j(i)}(t) + \frac{dw_i}{dt}$. There are r latent variables, and assume that each latent variable affects exactly $\frac{p}{r}$ observed variables in this way.

For this idealized setting, we can exactly evaluate all the quantities we need. It is not hard to show that the steady-state covariance matrices are $Q^* = 0.5(I + B^*B^{*T})$ and $R^* = B^{*T}$ resulting in $L^* = (r/(p+r))B^*B^{*T}$, which gives $U^* = V^* = \sqrt{r/p}B^*$ and $\mu = r$. Hence, we need $r < \sqrt{p}/3$ by assumption **(A2)**. Moreover, we can show that $\theta = \frac{1}{2}$ for this example and hence the assumption **(A3)** is also satisfied. Finally by evaluating other parameters in the theorem, we get the error bounds $\|A^* - \hat{A}\|_\infty \leq (3r/(4\sqrt{p}) + 25\sqrt{s}/7)\lambda_A$ and $\|L^* - \hat{L}\|_2 \leq 3r\lambda_A/(32\sqrt{p})$. The details of this calculations can be found in the appendix available online.

5. Proof Outline

In this section, we first introduce some notations and definitions and then, provide a three step proof technique to prove the main theorem for the discrete time system. The proof of the continuous time system is done via a coupling argument in the appendix.

There are two key novel ingredients in the proof enabling us to get the low sample complexity result in our theorem. The first ingredient comes from our new set of optimality conditions inspired by (Candes et al., 2009). This optimality conditions enable us to certify an approximation of L^* while certifying the exact sign support of A^*. The second ingredient comes from the bounds on the Schur complement of the perturbation of positive semi-definite matrices (Stewart, 1995). This result enables us to get a bound on the Schur complement of a perturbation of a positive semi-definite matrix of size p with only $\log(p)$ samples.

Given a matrix A^*, let Ω be the subspace of matrices whose their support is a subset of the matrix A^*. The orthogonal projection of a matrix M to Ω is denoted by $\mathcal{P}_\Omega(M)$. Denote the orthogonal complement space with Ω^c with orthogonal projection $\mathcal{P}_{\Omega^c}(M)$.

For any matrix $L \in \mathbb{R}^{p\times p}$, if the SVD is $L = U\Sigma V^T$, then let $\mathcal{T}(L) := \{M = UX^T + YV^T \text{for some } X, Y\}$ denote the subspace spanned by all matrices that have the same column space or row space as L. The orthogonal projection of a matrix N to $\mathcal{T}$ is denoted by $\mathcal{P}_\mathcal{T}(N)$. Denote the orthogonal complement space with $\mathcal{T}^c$ with orthogonal projection $\mathcal{P}_{\mathcal{T}^c}$. We define a metric to measure the *closeness* of two subspaces $\mathcal{T}_1$ and $\mathcal{T}_2$ as $\rho(\mathcal{T}_1, \mathcal{T}_2) = \max_{N\in\mathbb{R}^{p\times p}} \frac{\|\mathcal{P}_{\mathcal{T}_1}(N) - \mathcal{P}_{\mathcal{T}_2}(N)\|_2}{\|N\|_2}$.Finally, let $\mathcal{T} = \mathcal{T}(L^*)$ to shorten the notation and $L^* = U^*\Sigma^*V^*$ be a singular value decomposition.

We outline the proof in three steps as follows:

STEP 1: Constructing a candidate primal optimal solution $(\widetilde{A}, \widetilde{L})$ with the desired sparsity pattern using the restricted support optimization problem, called *oracle problem*:

$$(\widetilde{A}, \widetilde{L}) = \arg \min_{\substack{L:\rho(\mathcal{T}(L),\mathcal{T})\leq\rho \\ A:\mathcal{P}_{\Omega^c}(A)=0}} \lambda_A\|A\|_1 + \lambda_L\|L\|_* + \frac{1}{2\eta^2 n}\sum_{i=0}^{n-1} \|x(i+1) - x(i) - \eta(A+L)x(i)\|_2^2 \,. \tag{5}$$

This oracle is similar to the one used in (Chandrasekaran et al., 2010). It ensures that the right sparsity pattern is chosen for $\widetilde{A}$ and the tangent spaces $\widetilde{L}$ and L^* come from are *close* with parameter ρ.

STEP 2: Writing down a set of sufficient (stationary) optimality conditions for $(\widetilde{A},\widetilde{L})$ to be the unique solution of the (unrestricted) optimization problem (4):

Lemma 2. *If $\Omega \cap \mathcal{T} = \{0\}$, then $(\widetilde{A},\widetilde{L})$, the solution to the oracle problem (5), is the unique solution of the problem (4) if there exists a matrix $\widetilde{Z} \in \mathbb{R}^{p\times p}$ s.t.*

(C1) $\mathcal{P}_{\Omega}(\widetilde{Z}) = \lambda_A \mathit{Sign}\left(\widetilde{A}\right).$ **(C2)** $\left\|\mathcal{P}_{\Omega^c}(\widetilde{Z})\right\|_{\infty} < \lambda_A.$

(C3) $\left\|\mathcal{P}_{\mathcal{T}}(\widetilde{Z}) - \lambda_L U^* V^{*T}\right\|_2 \leq 4\rho\lambda_L.$

(C4) $\left\|\mathcal{P}_{\mathcal{T}^c}(\widetilde{Z})\right\|_2 < (1-\alpha)\lambda_L.$

(C5) $-\frac{1}{\eta n}\sum_{i=1}^{n}\left(x(i+1) - x(i) - \eta(\widetilde{A}+\widetilde{L})x(i)\right)x(i)^T + \widetilde{Z} = 0.$

STEP 3: Constructing a dual variable $\widetilde{Z}$ that satisfies the sufficient optimality conditions stated in Lemma 2. For matrices $M \in \Omega$ and $N \in \mathcal{T}$, let

$$\begin{aligned}\mathcal{H}_M &= M - \mathcal{P}_{\mathcal{T}}(M) + \mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}}(M) - \mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}}(M) + \dots\\ \mathcal{G}_N &= N - \mathcal{P}_{\Omega}(N) + \mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}(N) - \mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}(N) + \dots.\end{aligned}$$

It has been shown in (Chen et al., 2011) that if $\alpha < 1$ then both infinite sums converge. Suppose we have the SVD decomposition $\widetilde{L} = \widetilde{U}\widetilde{\Sigma}\widetilde{V}^T$. Let

$$\widetilde{Z} = \mathcal{H}_{\lambda_A \mathrm{Sign}(\widetilde{A})} + \mathcal{G}_{\mathcal{P}_{\mathcal{T}}(\lambda_L \widetilde{U}\widetilde{V}^T)} + \Delta,$$

where, Δ is a matrix such that (C5) is satisfied. As a result of this construction, we have $\mathcal{P}_{\Omega}(\Delta) = \mathcal{P}_{\mathcal{T}}(\Delta) = 0$. Now, we can establish $P_{\Omega}(\widetilde{Z}) = \lambda_A \mathrm{Sign}(\widetilde{A})$ and $P_{\mathcal{T}}(\widetilde{Z}) = \mathcal{P}_{\mathcal{T}}(\lambda_L \widetilde{U}\widetilde{V}^T)$ and consequently the conditions (C1) and (C3) in Lemma 2 are satisfied. It suffices to show that (C2) and (C4) are satisfied with high probability. This has been shown in Lemma 6.

6. Experimental Results

6.1. Synthetic Data

Motivated by the illustrative example discussed in section 4, we simulate a similar (but different) dynamic system for the purpose of our experiments. Consider the system where each latent variable only evolves by itself, i.e., $C^* = 0$ and D^* is a diagonal matrix. Moreover, assume that each latent variable affects $2p/r$ observed variable and each observed variable is affected by exactly two latent variable. We randomly select a support of size s per row for A^* and draw all the values of A^* and B^* i.i.d. standard Gaussian. To make the matrix $\mathcal{A}^*$ stable, by Geršgorin disk theorem (Geršgorin, 1931), we put a large-enough negative value on the diagonals of A^* and D^*.

We generate the data according to the continuous time model sub-sampled at points $t_i = \eta i$ for $i = 1, 2, \dots, n$, that is

$$\begin{bmatrix} x(i) \\ u(i) \end{bmatrix} = e^{\eta\mathcal{A}} \begin{bmatrix} x(i-1) \\ u(i-1) \end{bmatrix} + \int_{\eta(i-1)}^{\eta i} e^{\mathcal{A}(\eta i - \tau)} dw(\tau)$$

The stochastic integral can be estimated by binning the interval and assuming the Brownian motion is constant over the bin and hence, can be estimated by a standard Gaussian. See Chapter 4 in Shreve (2004).

Using this data, we solve (4) using accelerated proximal gradient method (Lin et al., 2009). Motivated by our Theorem, we plot our result with respect to the control parameter $\Theta = \frac{\eta n}{s^3 \log((s+2r)p+r^2)}$.

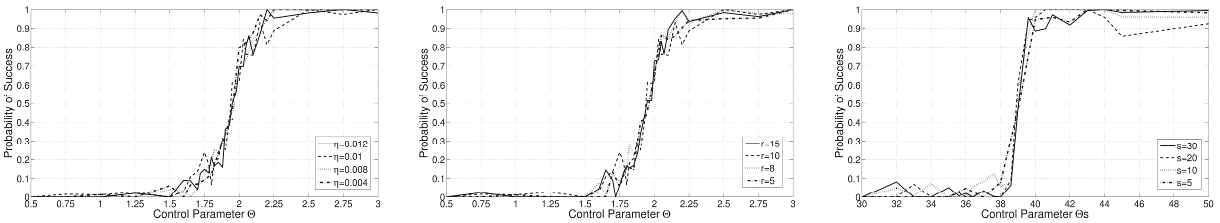
Figure 5 shows the phase transition of the probability of success in recovering the exact sign support of the matrix A^*. We ran three different experiments, each investigating the effect of one of the three key parameters of the system η (sampling frequency), r (number of latent variables) and s (sparsity of the model). These three figures show that the probability of success curves line up if they are plotted versus the correct control parameter. The first two curves for η and r line up versus Θ, indicating that our theorem suggests the correct scaling law for the sample complexity. However, from this experiment, it seems that the phase transition probability lines up with respect to Θs suggesting the scaling of s^2 instead of s^3.

6.2. Stock Market Data

We take the end-of-the-day closing stock prices for 50 different companies in the period of May 17, 2010 - May 13, 2011 (255 business days). These companies (among them, Amazon, eBay, Pepsi, etc) are consumer goods companies traded either at NASDAQ or NYSE in USD. The data is collected from Google Finance website. Applying our method and pure LASSO (Bento et al., 2010) to the data, we recover the structure of the dependencies among stocks. We present the result as a graph in Fig 6.2; where each company is a node in this graph and there is an edge between company i and j if $\hat{A}_{ij} \neq 0$. This result shows that the recovered dependency structure by our algorithm is order of magnitude sparser than the one recovered by pure LASSO.

To show the usefulness of our algorithm for prediction purposes, we apply our algorithm to this data and try to learn the model using the data for random n (consecutive) days. Then, we compute the mean squared error in the prediction of the following month (25 business days). The ratio $\frac{n}{25}$ is the training/testing ratio in our experiment.

Figure 3(b) shows the prediction error for both our and pure LASSO (Bento et al., 2010) methods as

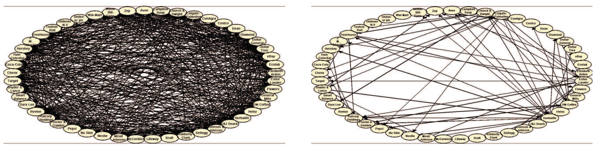


(a) Effect of η

(b) Effect of r

(c) Effect of s

Figure 1. Probability of success in recovering the true signed support of A^* versus the control parameter Θ (rescaled ηn) with $p = 200$, $r = 10$ and $s = 20$ for different values of η (left), and, with $p = 200$, $s = 20$ and $\eta = 0.01$ for different number of latent time series r (middle), and, with $p = 200$, $r = 10$ and fixed $\eta = 0.01$ for different sparsity sizes s (right). Notice that (c) is plotted versus $\Theta \times s$ which means $n\eta$ scales with s^2 not s^3.

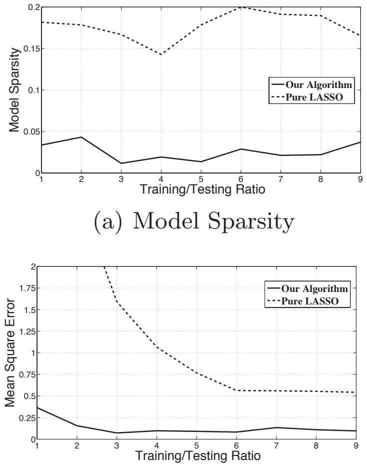


(a) Pure LASSO

(b) Our Algorithm

Figure 2. Comparison of the stock dependencies recovered by Pure LASSO (Bento et al., 2010) and our algorithm. This shows that there are latent factors affecting large number of stocks.

the train/test ratio increases. It can be seen that our method not only have better prediction, but also is more robust. Our algorithm requires only three months of the past data to give a robust estimation of the next month; in contrast with almost 6 months requirement of LASSO while the error of our algorithm is much smaller (by a factor of 6) than LASSO even in the steady state. Figure 3(a) illustrates that our estimated $\widehat{A}$ is order of magnitude sparser than the one estimated by LASSO. The number of latent variables our model finds varies from $8 - 12$ for different train/test ratios.



(a) Model Sparsity

(b) Prediction Error

Figure 3. Prediction error and model sparsity versus the ratio of the training/testing sample sizes for prediction of the stock price. Prediction error is measured using mean squared error and the model sparsity is the number of non-zero entries divided by the size of $\widehat{A}$.

References

Azoff, E.M. *Neural Network Time Series Forecasting of Financial Markets.* John Wiley & Sons, Inc., 1994.

Bar-Joseph, Z. Analyzing time series gene expression data. *Bioinformatics, Oxford University Press*, 20:2493–2503, 2004.

Bento, J., Ibrahimi, M., and Montanari, A. Learning networks of stochastic equations. In *NIPS*, 2010.

Bowerman, B.L. and O'Connell, R.T. *Forecasting and time series: An applied approach.* Duxbury Press, 1993.

Box, G.E.P., Jenkins, G.M., and Reinsel, G.C. *Time-series Analysis: Forecasting and Control.* John Wiley & Sons, Inc., 1990.

Candes, E. J. and Plan, Y. Matrix completion with noise. In *IEEE Proceedings*, volume 98, pp. 925 – 936, 2010.

Candes, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? In *Available at arXiv:0912.3599*, 2009.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. Latent variable graphical model selection via convex optimization. In *Available at arXiv:1008.1290*, 2010.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 2011.

Chatfield, C. *Time-series Forecasting.* Chapman & Hall, 2000.

Chen, Y., Jalali, A., Sanghavi, S., and Caramanis, C. Low-rank matrix recovery from errors and erasures. In *ISIT*, 2011.

Cochrane, J. H. *Time Series for Macroeconomics and Finance*. University of Chicago, 2005.

d'Aspremont, A., Bannerjee, O., and Ghaoui, L. El. First order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications*, 2007. To appear.

Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum-likelihood from incomplete datavia the em algorithm. *Journal of Royal Statistics Society, Series B.*, 39, 1977.

Fazel, M., Pong, T.K., Sun, D., and Tseng, P. Hankel matrix rank minimization with applications in system identification and realization. In *Available at http://faculty.washington.edu/mfazel/Hankelrm9.pdf*, 2011.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *BioStatistics*, 9:432–441, 2007.

Geršgorin, S. Uber die abgrenzung der eigenwerte einer matrix. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na*, 7:749–754, 1931.

Gillespie, D.T. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58:35–55, 2007.

Hazan, T., Polak, S., and Shashua, A. Sparse image coding using a 3d non-negative tensor factorization. In *ICCV*, 2005.

Higham, D. Modeling and simulating chemical reactions. *SIAM Review*, 50:347–368, 2008.

Horn, R. A. and Johnson, C. R. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

Jalali, A. and Srebro, N. Clustering using max-norm constrained optimization. In *Available at arXiv:1202.5598*, 2012.

Jalali, A., Chen, Y., Sanghavi, S., and Xu, H. Clustering partially observed graphs via convex optimization. In *ICML*, 2011.

Kim, K. Financial time series forecasting using support vector machines. *Elsevier Neurocomputing*, 55:307–319, 2003.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28:1303–1338, 1998.

Lawrence, N. D., Girolami, M., Rattray, M., and Sanguinetti, G. *Learning and Inference in Computational Systems Biology*. MIT Press, 2010.

Lin, Z., Ganesh, A., Wright, J., Wu, L., Chen, M., and Ma, Y. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. In *UIUC Technical Report UILU-ENG-09-2214*, 2009.

Ljung, L. *System identification: Theory for the user*. Prentice Hall, 1999.

Ljung, L. Prediction error estimation methods. *Circuits, systems, and signal processing*, 21(1):11–21, 2002.

Loehlin, J.C. *Latent Variable Models: An introduction tofactor, path, and structural analysis*. L. Erlbaum Associates Inc. Hillsdale, NJ, USA, 1984.

Marchal, P. Constructing a sequence of random walks strongly converging to brownian motion. In *Discrete Mathematics and Theoretical Computer Science Proceedings*, pp. 181–190, 2003.

Martens, J. Learning the linear dynamical system with asos. In *ICML*, 2010.

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

Peeters, RLM, Tossings, ITJ, and Zeemering, S. Sparse system identification by mixed l2/l1-minimization.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Technical Report 767, UC Berkeley, Department of Statistics*, 2008.

Redner, R. and Walker, H. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26, 1984.

Shreve, S. E. *Stochastic Calculus for Finance II: Continuous-Time Models*. Springer, 2004.

Srebro, N. and Jaakkola, T. Weighted low rank approximation. In *ICML*, 2003.

Stewart, G. W. On the perturbation of schur complements in positive semidefinite matrices. *Technical Report, University of Maryland, College Park*, 1995.

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

Van Overschee, P. and De Moor, B. Subspace algorithms for the stochastic identification problem. *Automatica*, 29 (3):649–660, 1993.

Wainwright, M. J. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Trans. on Information Theory*, 55:2183–2202, 2009.

Wei, W.W.S. *Time Series Analysis: Univariate and Multivariate Methods*. Addison Wesley, 1994.

West, Mike. Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics*, pp. 723–732. Oxford University Press, 2003.

Young, P. *Recursive estimation and time-series analysis*. Springer - Verlag, 1984.

Yuan, M. and Lin, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Zhou, Z., Li, X., Wright, J., Candes, E., and Ma, Y. Stable principal component pursuit. In *ISIT*, 2010.