Learning Low-Complexity Autoregressive Models with Limited Time Sequence Data

Fu Lin and Jie Chen

***Abstract*— We consider the estimation of the state transition matrix in vector autoregressive models when the time sequence data is limited but nonsequence steady-state data is abundant. To leverage both sources of data, we formulate the problem as the least-squares minimization regularized by a Lyapunov penalty. Explicit cardinality or rank constraints are imposed to reduce the complexity of the model. The resulting nonconvex, nonsmooth problem is solved by using the proximal alternating linearization method (PALM). An advantage of PALM is that for the problem under investigation, it is globally convergent to a critical point and the objective value is monotonically decreasing. Furthermore, the proximal operators therein can be efficiently computed by using explicit formulas. Our numerical experiments demonstrate the effectiveness of the developed approach.**

***Index Terms*— Autoregressive models, low-rank matrix, Lyapunov penalty, proximal alternating linearized minimization, sparse matrix, steady-state data**

I. Introduction

Vector autoregressive (VAR) models are widely used in the analysis of linear interdependence in time series. A key step in building the autoregressive model is the identification of the state transition matrix. When time sequence data is readily available, the standard approach is to solve a least-squares problem. In several modern applications, however, the dimension of the model is significantly larger than the number of time sequence measurements, which makes the model unidentifiable through the simple least-squares. Such scenarios include, for example, tracking the progression of brain neurological diseases, because the number of comprehensive brain scans is limited due to the cost or medical concerns [1]. In inferring the gene expression networks, the number of genes is typically much larger than the number of measurements because of the intrusive nature of the measuring techniques [2], [3]. In such situations, regularization is a typical rescue. Regularization techniques may introduce structures to the transition matrix. In particular, sparsity and low-rank structures are extensively studied, in part because of the convenience of convex optimization [1], [3]–[5].

F. Lin is with Systems Department, United Technologies Research Center, 411 Silver Ln, East Hartford, CT 06118, USA. E-mail: linf@utrc.utc.com.

J. Chen is with IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598, USA. E-mail: chenjie@us.ibm.com.

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under contract number DE-AC02-06CH11357. J. Chen is supported in part by XDATA program of the Defense Advanced Research Projects Agency (DARPA), administered through Air Force Research Laboratory contract FA8750-12-C-0323.

In this work, we consider a regularization that takes advantage of additional data sources. When the VAR model is stable and the steady-state data are abundant, the steady-state data can be used to improve the estimation accuracy [1], [3], [6]. In [1], the Lyapunov penalty is proposed to make use of the steady-state nonsequence data in conjunction with the standard least-squares estimator. In [3], the perturbed steady-state data is used to infer the sparse, stable gene expression models. In [6], both steady-state and temporal data are integrated for estimating gene regulatory networks.

Hence, we move one step further, leveraging the availability of limited time sequences and the abundant steady-state data, meanwhile imposing additional structural constraints to reduce the complexity of the model. The resulting estimator is a least-squares estimator, regularized by the Lyapunov penalty, and constrained by an explicit sparsity or low-rank requirement. That is, the state transition matrix is constrained to have a specific number of nonzeros or a specific rank. The identification problem then becomes nonconvex (due to the Lyapunov penalty and the low-complexity constraint) and nonsmooth (due to the low-complexity constraint). We propose solving the problem by the proximal alternating linearization method (PALM). We show that PALM is globally convergent to a critical point and the objective value monotonically decreases. Moreover, the proximal operators therein admit closed-form expressions and therefore the implementation is particularly simple. It is also straightforward for PALM to handle stability constraints and other convex low-complexity constraints (e.g., ℓ_1 constraint or nuclear norm constraint).

Our contributions can be summarized as follows. First, we propose the use of explicit cardinality or rank constraints for the identification of low-complexity VAR models. By doing so, one can directly control the number of nonzero elements or the rank of the transition matrix. Second, we reformulate the objective function that lends itself to a powerful PALM algorithm. In particular, we show that the proximal operators therein can be computed via closed-form expressions. Third, we prove that the problem formulation satisfies Lipschitz conditions and the Kurdyka-Lojasiewicz (KL) property. As a result, we establish the global convergence of our algorithm to a critical point.

The presentation is organized as follows. In Section II, we formulate the optimization problem for a low-complexity VAR model. In Section III, we present the PALM algorithm and derive explicit formulas for the proximal operators. In Section IV, we verify the convergence of PALM by using numerical experiments. We summarize our contributions in Section V.

II. VAR Model Identification via Lyapunov Penalty

Consider a p-dimensional vector autoregressive model:

$$\phi(t+1) \;=\; A\phi(t) + \epsilon(t), \tag{1}$$

where $\phi(t) \in \mathbb{R}^p$ is the state vector at time t, $A \in \mathbb{R}^{p\times p}$ is the state transition matrix, and $\epsilon(t) \in \mathbb{R}^p$ is a zero-mean white stochastic process. We assume that the autoregressive model (1) is stable; that is, all eigenvalues of A have modulus less than one. Then, the state vector $\phi(t)$ has a steady-state distribution whose covariance matrix P is determined by the discrete-time Lyapunov equation

$$APA^T \,+\, Q \;=\; P, \tag{2}$$

where $Q \in \mathbb{R}^{p\times p}$ is the covariance matrix of $\epsilon(t)$ and $(\cdot)^T$ denotes the matrix transpose operation. Linear systems theory says that P is positive definite if and only if A is stable.

Our objective is to identify the state transition matrix A. For the convenience of developing optimization details, we use X to replace the unknown A in what follows. Given a set of n time sequence measurements of $\phi(t)$, the standard least-squares estimation amounts to

$$\underset{X\in\mathbb{R}^{p\times p}}{\text{minimize}} \;\; \frac{1}{2}\|XC \,-\, D\|_F^2, \tag{3}$$

where $C := [\,\phi(1), \cdots, \phi(n-1)\,] \in \mathbb{R}^{p\times(n-1)}$, $D := [\,\phi(2), \cdots, \phi(n)\,] \in \mathbb{R}^{p\times(n-1)}$, and $\|\cdot\|_F$ denotes the Frobenius norm. When the number of time sequence data is less than the dimension of the states (i.e., $p > n-1$), infinitely many solutions exist for (3) and hence the state transition matrix is unidentifiable.

We are interested in the scenario when the time sequence data is scarce but the steady-state nonsequence data is readily available. This situation arises in several applications [1], [5], [7] including identification of cancer expression networks [5] and gene regulatory networks [7]. In this case, Huang and Schneider [1] propose to use the Lyapunov penalty

$$\|XPX^T \,+\, Q \,-\, P\|_F^2$$

as a regularization term. They show that the Lyapunov penalty in conjunction with least-squares may improve the accuracy of the estimation. Since the covariance matrix P is unknown in practice, we approximate it by using the sample covariance

$$S \;:=\; \frac{1}{N-1}\sum_{i=1}^{N}(z^i-\bar{z})(z^i-\bar{z})^T \text{ with } \bar{z} \;:=\; \frac{1}{N}\sum_{i=1}^{N} z^i,$$

where $\{z^i\}_{i=1}^N$ is the steady-state nonsequence data. Then, the identification problem becomes

$$\underset{X\in\mathbb{R}^{p\times p}}{\text{minimize}} \;\; \frac{1}{2}\|XC \,-\, D\|_F^2 \,+\, \frac{\rho}{2}\|XSX^T \,+\, Q - S\|_F^2, \tag{4}$$

where ρ is a positive coefficient that balances the estimation error between time sequence and nonsequence data.

A. Stability Constraint

Since stability is a necessary condition for the use of Lyapunov penalty, it is natural to constrain the formulation (4) by stability. Let $\tau(X)$ denote the spectral radius of X, that is, $\tau(X) := \max\{|\lambda_i|\}_{i=1}^p$. Since $\tau(X) \leq \|X\|_2$ and since $\|X\|_2 \leq \|X\|_F$, we have

$$\tau^2(X) \;\leq\; \|X\|_2^2 \;=\; \|XX^T\|_2 \;\leq\; \|XX^T\|_F.$$

Thus, a stable VAR model can be obtained by solving the following problem:

$$\begin{array}{ll} \underset{X\in\mathbb{R}^{p\times p}}{\text{minimize}} & \frac{1}{2}\|XC - D\|_F^2 \,+\, \frac{\rho}{2}\|XSX^T + Q - S\|_F^2 \\ \text{subject to} & \|XX^T\|_F^2 \;<\; 1. \end{array} \tag{5}$$

Alternatively, one can incorporate the stability constraint in the cost function with a sufficiently large coefficient $\mu \geq 0$

$$\begin{array}{ll} \underset{X\in\mathbb{R}^{p\times p}}{\text{minimize}} & \frac{1}{2}\|XC - D\|_F^2 \,+\, \frac{\rho}{2}\|XSX^T + Q - S\|_F^2 \\ & +\, \frac{\mu}{2}\|XX^T\|_F^2. \end{array} \tag{6}$$

Note that both the Lyapunov penalty and the stability penalty are quadratic functions of X in the Frobenius norm (e.g., they coincide when $S = Q = I$). Thus, the additional term $\|XX^T\|_F^2$ is inconsequential in the design of solution algorithms. For this reason and for the ease of presentation, in what follows we omit the stability constraint in the general discussion, but comment on the modifications of algorithm when necessary.

B. Low-Complexity Models

In several applications, it is desired to impose sparsity or low-rank structures on the state transition matrix [1], [3]–[5], [8]. In identification of gene expression networks, for example, the nonzero elements of A determine the interaction graph of the expression network [3], [4]. In this context, a sparse A is desired because one can construct a sparse network to explain data.

A common approach to promote sparsity is to impose an ℓ_1 constraint:

$$\|X\|_{\ell_1} := \sum_{i,j=1}^{p} |X_{ij}| \;\leq\; l, \tag{7}$$

where l is a prescribed positive number. Since the ℓ_1 norm promotes sparsity *implicitly*, the actual number of nonzeros in the solution is indirectly controlled by the threshold l. However, given a desired level of sparsity, the correct choice of l is typically not known a priori. Practical choice of l typically requires grid search or cross validation.

Alternatively, an *explicit* way to guarantee sparsity is to directly control the number of nonzeros by the cardinality constraint:

$$\mathbf{card}(X) \;:=\; \text{number of nonzero entries of } X \;\leq\; s, \tag{8}$$

where s is a positive integer. Note that the cardinality constraint is harder to deal with than the ℓ_1 constraint, because cardinality is nonconvex and nonsmooth. An essential ingredient of this work is the solution method that handle this difficulty.

Another approach to obtain low-complexity VAR models is to impose a low-rank constraint. A low-rank state transition matrix is useful because it implies that the data can be explained by a simpler VAR model. It is common practice to employ model reduction techniques to obtain a low-rank model as a post-processing step [9], [10].

An implicit way to promote low-rank solutions in system identification is to use a nuclear norm constraint [11]–[13]

$$\|X\|_* := \sum_{i=1}^{p} \sigma_i(X) \leq \nu, \tag{9}$$

where ν is a positive number and the σ_i's denote the singular values of X. Similar to the sparsity case, the threshold ν may be challenging to prescribe.

Alternatively, we may impose a low-rank constraint by explicitly controlling the rank of the state transition matrix:

$$\mathbf{rank}(X) := \text{number of nonzero singular values of } X \leq r, \tag{10}$$

where r is a positive integer. This integer may be determined from resource limitations in practice.

To summarize, we consider the estimation problem of the low-complexity VAR model:

$$\widehat{A} = \underset{X \in \mathbb{R}^{p \times p}}{\text{argmin}} \ \frac{1}{2}\|XC - D\|_F^2 + \frac{\rho}{2}\|XSX^T + Q - S\|_F^2 \quad \text{subject to constraint (7) or (8) or (9) or (10).} \tag{11}$$

For convex constraints (i.e., the ℓ_1 and the nuclear norm), one may employ gradient projection methods, namely, taking a descent direction of the objective function and projecting onto the convex constraint sets. A gradient projection method is proposed in [1] to solve (11) with the ℓ_1 constraint (7). For nonconvex constraints (i.e., the cardinality and the rank constraints), we develop PALM in the subsequent section. Note that PALM can also be applied to convex constraints.

III. Proximal Alternating Linearized Method (PALM)

In this section, we reformulate (11) into a form well suited to PALM. One advantage of PALM is its global convergence to a critical point for both convex and nonconvex constraints. We begin with replacing one of the two X's in the quadratic Lyapunov term by a new variable Y and rewrite (11) as

$$\begin{array}{ll} \underset{X,Y}{\text{minimize}} & \frac{1}{2}\|XC - D\|_F^2 + \frac{\rho}{2}\|YSX^T + Q - S\|_F^2 \\ \text{subject to} & Y - X = 0, \\ & \text{(7) or (8) or (9) or (10).} \end{array}$$

This change of variable will allow us to establish the Lipschitz conditions required for convergence analysis.

Let f denote the least-squares term

$$f(X) = \frac{1}{2}\|XC - D\|_F^2, \tag{12}$$

and let H denote the coupling term

$$H(X,Y) = \frac{\rho_1}{2}\|YSX^T + Q - S\|_F^2 + \frac{\rho_2}{2}\|X - Y\|_F^2, \tag{13}$$

where the penalty parameter $\rho_1 > 0$ resumes the role of ρ in (11) and $\rho_2 > 0$ is sufficiently large to penalize the discrepancy between X and Y. Then we obtain the following formulation

$$\underset{X,Y}{\text{minimize}} \ \Phi(X,Y) := f(X) + g(Y) + H(X,Y), \tag{14}$$

where g is the indicator function of the constraints (7)-(10). For example,

$$g(Y) = \begin{cases} 0, & \mathbf{card}(Y) \leq s \\ \infty, & \text{otherwise,} \end{cases} \tag{15}$$

for the cardinality constraint (8), or

$$g(Y) = \begin{cases} 0, & \mathbf{rank}(Y) \leq r \\ \infty, & \text{otherwise,} \end{cases} \tag{16}$$

for the rank constraint (10).

A. Generic PALM Method

PALM alternates between computing the proximal operators of the uncoupled functions f and g around the linearization of the coupling function H at the previous iterate, hence the name [14], [15]. To put PALM in the context of other alternating methods, suppose for the moment that $\Phi(X,Y)$ is a *strictly* convex function. One approach for minimizing Φ is the Gauss-Seidel iteration (also known as coordinate descent):

$$\begin{aligned} X^{k+1} &\in \underset{X}{\text{argmin}} \ \Phi(X, Y^k) \\ Y^{k+1} &\in \underset{Y}{\text{argmin}} \ \Phi(X^{k+1}, Y). \end{aligned}$$

For convergence, a unique solution in each minimization step is required. Otherwise, Gauss-Seidel may cycle indefinitely. When Φ is convex but *not strictly* convex, uniqueness can be achieved by including a quadratic proximal term

$$X^{k+1} \in \underset{X}{\text{argmin}} \left\{ \Phi(X, Y^k) + \frac{c_k}{2}\|X - X^k\|^2 \right\} \tag{18a}$$

$$Y^{k+1} \in \underset{Y}{\text{argmin}} \left\{ \Phi(X^{k+1}, Y) + \frac{d_k}{2}\|Y - Y^k\|^2 \right\}, \tag{18b}$$

where c_k and d_k are positive coefficients. This class of proximal methods is well studied; see [14] for a recent survey. It is worth noting that the alternating direction method of multipliers (ADMM) can be viewed as a class of proximal methods for convex problems [14].

When Φ is nonconvex, as in our case (14), we need to modify the proximal terms to ensure convergence. As opposed to taking the proximal term around X^k as in (18a), we take the term around X^k modified with a scaled partial gradient of H:

$$X^{k+1} \in \underset{X}{\operatorname{argmin}} \left\{ f(X) + \frac{c_k}{2}\|X - U^k\|_F^2 \right\}, \qquad (19)$$

where

$$U^k = X^k - \frac{1}{c_k}\nabla_X H(X^k, Y^k).$$

The parameter c_k is chosen to be greater than the Lipschitz constant of $\nabla_X H$. That is, if L_1 satisfies

$$\|\nabla_X H(X_1, Y^k) - \nabla_X H(X_2, Y^k)\| \le L_1(Y^k)\|X_1 - X_2\|,$$

for all X_1 and X_2, then we let

$$c_k = \gamma_1 L_1(Y^k)$$

for some $\gamma_1 > 1$.

Similarly, we take the proximal term around Y^k modified with a scaled partial gradient of H,

$$Y^{k+1} \in \underset{Y}{\operatorname{argmin}} \left\{ g(Y) + \frac{d_k}{2}\|Y - V^k\|_F^2 \right\}, \qquad (20)$$

where

$$V^k = Y^k - \frac{1}{d_k}\nabla_Y H(X^{k+1}, Y^k).$$

The parameter d_k is given by $d_k = \gamma_2 L_2(X^{k+1})$ for some $\gamma_2 > 1$ where the Lipschitz constant L_2 satisfies

$$\|\nabla_Y H(X^{k+1}, Y_1) - \nabla_Y H(X^{k+1}, Y_2)\| \le L_2(X^{k+1})\|Y_1 - Y_2\|,$$

for all Y_1 and Y_2. PALM updates (X, Y) by using the iterations (19)-(20).

B. Explicit Formulas

To implement the PALM iterations (19)-(20), one needs the Lipschitz constants L_1 and L_2 in order to determine the coefficients c_k and d_k, respectively. To this end, we take the partial gradients of H and obtain

$$\nabla_X H = \rho_1(XS^TY^TYS + (Q - S)^TYS) + \rho_2(X - Y)$$
$$\nabla_Y H = \rho_1(YSX^TXS^T + (Q - S)XS^T) + \rho_2(Y - X).$$

Since $\nabla_X H$ is linear in X and $\nabla_Y H$ is linear in Y, it follows that the Lipschitz constants admit explicit formulas

$$\begin{aligned} L_1(Y) &= \|\rho_1 S^TY^TYS + \rho_2 I\|_F \\ L_2(X) &= \|\rho_1 SX^TXS^T + \rho_2 I\|_F. \end{aligned} \qquad (21)$$

We next show that the proximal operators can be computed efficiently. The proximal operator (19) can be expressed as

$$X^{k+1} \in \underset{X}{\operatorname{argmin}} \left\{ \frac{1}{2}\|XC - D\|_F^2 + \frac{c_k}{2}\|X - U^k\|_F^2 \right\}.$$

Solving this least-squares problem yields

$$X^{k+1} = (DC^T + c_kU^k)(CC^T + c_kI)^{-1}, \qquad (22)$$

where I denotes the identity matrix. When $p > n-1$, one can gain some computation efficiency by inverting $C^TC + c_kI$ instead of $CC^T + c_kI$, since the Woodbury formula gives equivalently,

$$X^{k+1} = (c_k^{-1}DC^T + U^k)(I - C(c_kI + C^TC)^{-1}C^T). \qquad (23)$$

On the other hand, the proximal operator (20) is expressed as

$$\begin{array}{ll} \underset{Y}{\text{minimize}} & \frac{d_k}{2}\|Y - V^k\|_F^2 \\ \text{subject to} & (7) \text{ or } (8) \text{ or } (9) \text{ or } (10). \end{array}$$

For the cardinality constraint (8), the optimal Y is obtained by keeping the s largest elements of V^k in magnitude and zero out the rest of the elements in V^k. This is because the Frobenius norm squared is the squared sum of the elements of $Y - V^k$. For the rank constraint (10), by the Eckart–Young theorem, the optimal Y is the best rank-r approximation of V^k obtained by the truncated SVD.

For the ℓ_1 constraint (7), the projection onto the ℓ_1-ball can be computed by an algorithm developed in [16]. For the nuclear norm constraint (9), the optimal solution Y can be computed by performing the singular value decomposition of V^k and then projecting the singular values of V^k onto the ℓ_1-ball.

With these details, we summarize the computational steps in Algorithm 1. We show that Algorithm 1 globally converges to a critical point of the nonconvex, nonsmooth problem (14). Furthermore, the objective value is monotonically decreasing. Our analysis builds upon the seminal work in [15] for the global convergence of the generic PALM algorithm. Our contributions are the establishments of Lipschitz conditions and the KL property of (14). In particular, the Lipschitz conditions ensure sufficient decrease of the objective value and the KL property is the key step for the convergence to a critical point. Due to space limitation, the detailed proofs are reported in [17].

Remark 1 (Stability): As discussed in Section II-A, we can incorporate the stability constraint by penalizing $\|XX^T\|_F^2$ in the cost function. In this case, the coupling term becomes

$$H(X,Y) = \frac{\rho_1}{2}\|YSX^T + Q - S\|_F^2 + \frac{\rho_2}{2}\|X - Y\|_F^2 + \frac{\mu}{2}\|YX^T\|_F^2.$$

Its partial gradients are given by

$$\begin{aligned} \nabla_X H &= \rho_1(XS^TY^TYS + (Q - S)^TYS) + \rho_2(X - Y) \\ &\quad + \mu XY^TY \\ \nabla_Y H &= \rho_1(YSX^TXS^T + (Q - S)XS^T) + \rho_2(Y - X) \\ &\quad + \mu YX^TX, \end{aligned}$$

and the Lipschitz constants are

$$\begin{aligned} L_1(Y) &= \|\rho_1 S^TY^TYS + \mu Y^TY + \rho_2 I\|_F \\ L_2(X) &= \|\rho_1 SX^TXS^T + \mu X^TX + \rho_2 I\|_F. \end{aligned}$$

With these modifications, Algorithm 1 is directly applied to the problem (6) subject to low-complexity constraints.

IV. Numerical Experiments

In this section, we evaluate the performance of PALM on synthetic data. Additional numerical experiments and

Algorithm 1 PALM for problem (14) with cardinality constraint (15) or rank constraint (16).

Initialization: Start with any (X^0, Y^0).
for $k = 0, 1, 2, \ldots$ until convergence **do**
  // Update X^{k+1}
  Compute the Lipschitz constant
    $L_1(Y^k) = \|\rho_1 S^T Y^{kT} Y^k S + \rho_2 I\|_F$.
  Compute $c_k = \gamma_1 L_1(Y^k)$ for some $\gamma_1 > 1$.
  Compute partial gradient
    $\nabla_X H(X^k, Y^k) = \rho_1(X^k S^T Y^{kT} Y^k S + (Q-S)^T Y^k S) + \rho_2(X^k - Y^k)$.
  Update the proximal point
    $U^k = X^k - \frac{1}{c_k}\nabla_X H(X^k, Y^k)$.
  if $p \le n-1$ **then**
    $X^{k+1} = (DC^T + c_k U^k)(CC^T + c_k I)^{-1}$
  else
    $X^{k+1} = (c_k^{-1} DC^T + U^k)(I - C(c_k I + C^T C)^{-1} C^T)$.
  end if
  // Update Y^{k+1}
  Compute the Lipschitz constant
    $L_2(X^{k+1}) = \|\rho_1 S X^{(k+1)T} X^{k+1} S^T + \rho_2 I\|_F$.
  Compute $d_k = \gamma_2 L_2(X^{k+1})$ for some $\gamma_2 > 1$.
  Compute partial gradient
    $\nabla_Y H(X^{k+1}, Y^k) = \rho_1(Y^k S X^{(k+1)T} X^{k+1} S^T + (Q-S) X^{k+1} S^T) + \rho_2(Y^k - X^{k+1})$.
  Update the proximal point
    $V^k = Y^k - \frac{1}{d_k}\nabla_Y H(X^{k+1}, Y^k)$.
  if g is the cardinality constraint (8) **then**
    $Y^{k+1} = \mathcal{I}_s \circ V^k$, where $(\mathcal{I}_s)_{ij} = 1$ if $(|V^k|)_{ij}$ is greater than or equal to the s-th largest element of $|V^k|$, and $(\mathcal{I}_s)_{ij} = 0$ otherwise.
  else if g is the rank constraint (10) **then**
    Y^{k+1} is the rank-r truncated SVD of V^k.
  end if
end for

comparison with convex methods can be found in [17]. We demonstrate that the PALM solution converges to a matrix with the prescribed level of sparsity or rank. Furthermore, the objective value decreases monotonically as predicted by the convergence analysis. In our experiments, we assume $Q = \sigma^2 I$ for the covariance matrix of $\epsilon(t)$ in (1). We set $\gamma_1 = \gamma_2 = 2$ in Algorithm 1. The hyper-parameters ρ_1 and σ are determined through cross validation.

We test the performance of PALM on a sparse example and a low-rank example with the matrix size 200×200. In both examples, we use time series of length $n = 50$ for training and $m = 800$ for testing. For the length of steady-state data, $N = 1600$. The performance of the identified VAR model is evaluated by using the normalized error and the cosine score proposed in [1]

$$\text{Normalized error: } \frac{1}{m-1}\sum_{t=1}^{m-1}\frac{\|x(t+1) - \widehat{A}x(t)\|}{\|x(t+1) - Ax(t)\|}$$

Cosine score:

$$\frac{1}{m-1}\sum_{t=1}^{m-1}\frac{|(x(t+1) - Ax(t))^T(x(t) - \widehat{A}x(t))|}{\|x(t+1) - Ax(t)\|\|x(t) - \widehat{A}x(t)\|}.$$

A smaller normalized error (lower bounded by 0) and a higher cosine score (upper bounded by 1) imply better performance.

A. Sparse Example

The sparse matrix is generated by using the rule $A = (0.95M)/\max_k(|\lambda_k(M)|)$, where M has 5000 normally distributed nonzeros and $\lambda_k(M)$ denotes the eigenvalues of M. We set $s = 5000$ in the cardinality constraint (8).

Figure 1 shows the convergence results. The objective value monotonically decreases. The errors in two consecutive PALM steps, namely,

$$e_X^k = \|X^{k+1} - X^k\|_F, \qquad e_Y^k = \|Y^{k+1} - Y^k\|_F,$$
$$e_{XY}^k = \|X^k - Y^k\|_F,$$

all decrease quickly. It takes PALM fewer than 300 iterations to reach $e_X, e_Y \le 10^{-5}$ and $e_{XY} \le 2 \times 10^{-3}$. Note that the solution has exactly 5000 nonzeros as required by the cardinality constraint. For the estimated matrix $\widehat{A}$, the normalized error is 0.3268 and the cosine score is 0.9447.

B. Low-Rank Example

The low-rank matrix is generated by using the rule $A = \mathcal{U}\Sigma\mathcal{V}$, where $\Sigma \in \mathbb{R}^{25\times 25}$ is a diagonal matrix with random diagonal entries uniformly distributed in $[0, 1)$, and $\mathcal{U} \in \mathbb{R}^{200\times 25}$ and $\mathcal{V} \in \mathbb{R}^{25\times 200}$ are random orthonormal matrices. By construction $A \in \mathbb{R}^{200\times 200}$ is stable with $\mathbf{rank}(A) = 25$. Thus we set $r = 25$ in the rank constraint (10).

Figure 2 shows the convergence results. Similar to those for the sparse example in Figure 1, we observe that the objective value Φ monotonically decreases and the errors in two consecutive PALM steps decrease quickly. It takes PALM fewer than 300 iterations to reach $e_X, e_Y \le 2 \times 10^{-6}$ and $e_{XY} \le 2 \times 10^{-4}$. The solution has a numerical rank 25, as required by the rank constraint. For the estimated matrix $\widehat{A}$, the normalized error is 0.6977 and the cosine score is 0.7164.

V. Conclusions

In this paper, we formulate a constrained optimization problem for estimating the state transition matrix of a vector autoregressive model, with limited time sequence data but abundant nonsequence steady-state data. To reduce the complexity of the model, we propose imposing a cardinality or a rank constraint on the transition matrix. We develop a PALM algorithm to solve the resulting nonconvex, nonsmooth problem and establish its global convergence. Our numerical experiments verify the convergence theory and demonstrate the effectiveness of the developed method.

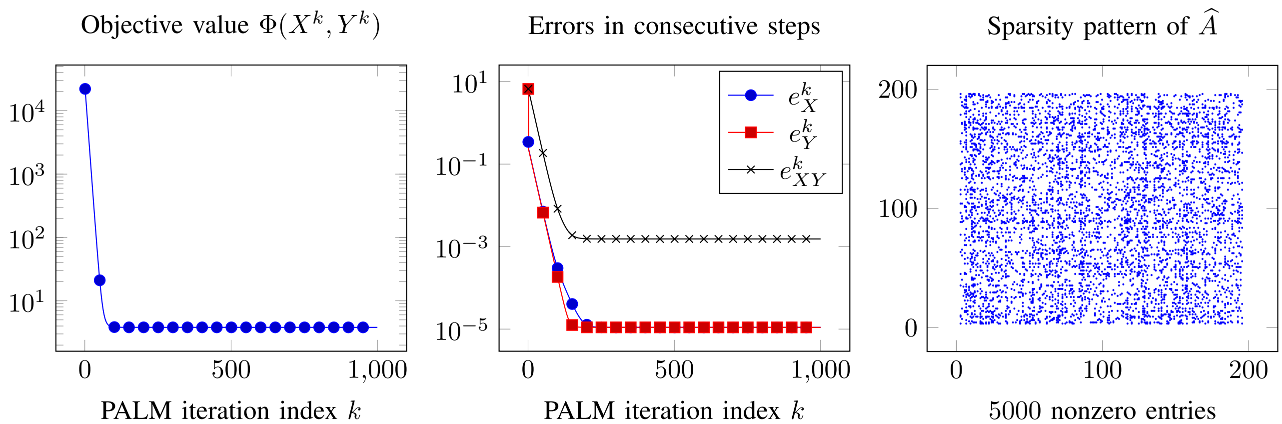

Fig. 1: Convergence results of PALM for the sparse example: the objective value (left), the errors in consecutive steps (middle), and the sparse solution with 5000 nonzero entries (right).

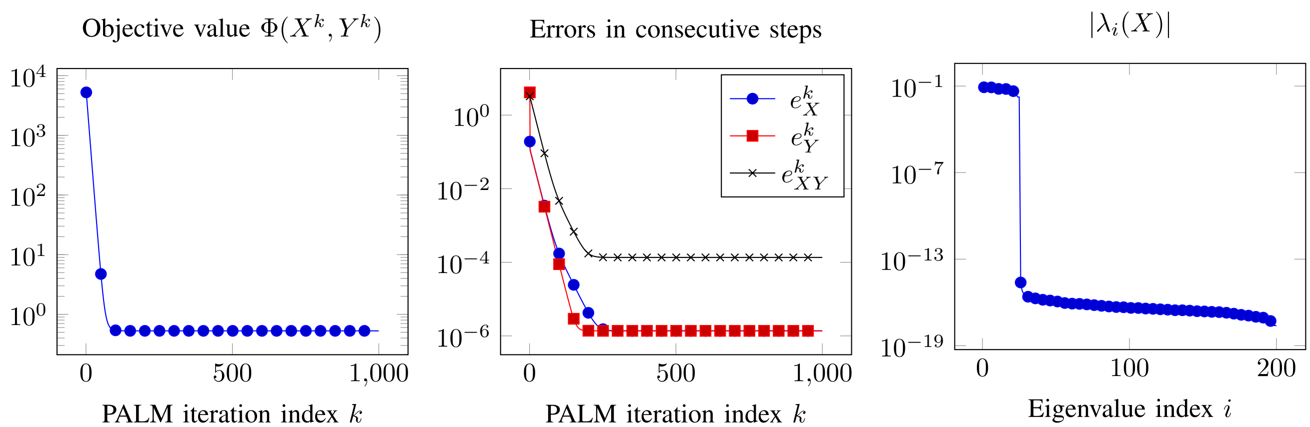

Fig. 2: Convergence results of PALM for the low-rank example: the objective value (left), the errors in consecutive steps (middle), and the low-rank solution with 25 nonzero eigenvalues (right).

References

[1] T.-K. Huang and J. G. Schneider, "Learning auto-regressive models from sequence and non-sequence data," in *Advances in Neural Information Processing Systems*, 2011, pp. 1548–1556.

[2] R. Yoshida, S. Imoto, and T. Higuchi, "Estimating time-dependent gene networks from time series microarray data by dynamic linear models with Markov switching," in *Proceedings of the 2005 Computational Systems Bioinformatics Conference*, 2005, pp. 289–298.

[3] M. M. Zavlanos, A. A. Julius, S. P. Boyd, and G. J. Pappas, "Inferring stable genetic networks from steady-state data," *Automatica*, vol. 47, no. 6, pp. 1113–1122, 2011.

[4] A. Fujita, J. R. Sato, H. M. Garay-Malpartida, R. Yamaguchi, S. Miyano, M. C. Sogayar, and C. E. Ferreira, "Modeling gene expression regulatory networks with the sparse vector autoregressive model," *BMC Systems Biology*, vol. 1, no. 1, p. 39, 2007.

[5] A. Gupta and Z. Bar-Joseph, "Extracting dynamics from static cancer expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, pp. 172–182, 2008.

[6] Y. K. Wang, D. G. Hurley, S. Schnell, E. J. Crampin *et al.*, "Integration of steady-state and temporal gene expression data for the inference of gene regulatory networks," *PloS one*, vol. 8, no. 8, p. e72103, 2013.

[7] A. Lozano, N. Abe, Y. Liu, and S. Rosset, "Grouped graphical granger modeling for gene expression regulatory networks discovery," *Bioinformatics*, vol. 25, no. 12, pp. i110–i118, 2009.

[8] H. Wang, G. Li, and C.-L. Tsai, "Regression coefficient and autoregressive order shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 1, pp. 63–78, 2007.

[9] M. Verhaegen, "Identification of the deterministic part of MIMO state space models given in innovations form from input-output data," *Automatica*, vol. 30, no. 1, pp. 61–74, 1994.

[10] L. Ljung, "System identification," in *Signal Analysis and Prediction*. Springer, 1998, pp. 163–173.

[11] M. Fazel, H. Hindi, and S. P. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the 2001 American Control Conference*, 2001, pp. 4734–4739.

[12] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 3, pp. 1235–1256, 2009.

[13] M. Fazel, T. K. Pong, D. Sun, and P. Tseng, "Hankel matrix rank minimization with applications to system identification and realization," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 3, pp. 946–977, 2013.

[14] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2013.

[15] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.

[16] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the ℓ_1-ball for learning in high dimensions," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 272–279.

[17] F. Lin and J. Chen, "Learning low-complexity autoregressive models via proximal alternating minimization," *arXiv:1609.05341*, 2016.