SPARSE AUTOREGRESSIVE MODEL ESTIMATION FOR LEARNING GRANGER CAUSALITY IN TIME SERIES

Jitkomut Songsiri

Department of Electrical Engineering
Faculty of Engineering, Chulalongkorn University
254 Phayathai Road, Pathumwan, Bangkok, Thailand 10330
Email: jitkomut.s@chula.ac.th, Tel: +662-2186487

ABSTRACT

This paper considers a problem of estimating multivariate autoregressive (AR) models with sparse coefficient matrices. A joint zero pattern of AR coefficients reveals a Granger causality structure of the variables, which is typically depicted as a graphical model. The problem of estimating the graph topology is then formulated as a least-squares problem with an ℓ_1-type regularization to promote a sparsity in the AR coefficients. We obtain a convex framework of the estimation problem which can become challenging to solve in a large scale setting due to the nondifferentiability of the cost function. We apply a recent powerful algorithm, namely, the alternating direction method of multipliers (ADMM) for solving topology selection problems in Granger graphical models of AR processes. We illustrate the idea and verify the performance of the ADMM algorithm on randomly generated data sets. This approach is finally applied on Google Flu Trends data learn a causal structure of flu activities in the 51 states of the USA.

Index Terms— Sparse autoregressive models, Granger causality, Topology selection.

1. INTRODUCTION

Recent studies have been focused on exploring relationship or causal structures in multivariate time series. Such relationship can be represented as a graphical model where the directional edges specify the *Granger causality* structure of variables [1], which states that time series y_i is Granger-caused by time series y_j if knowing the past values of y_j helps improve the prediction of y_i. There is a nice characterization of Granger causality for autoregressive processes which is widely used to model multivariate time series in many applications. An n-dimensional autoregressive process of order p is given by

$$y(t) = A_1y(t-1)+A_2y(t-2)+\cdots+A_py(t-p)+u(t) \quad (1)$$

where $y(\cdot) \in \mathbf{R}^n$, $A_k \in \mathbf{R}^{n\times n}$, $k = 1, 2, \ldots, p$ and $u(\cdot)$ is input noise. The absence of a directed edge from note j to i illustrates that y_i is not *Granger-caused* by y_j and this can be characterized in terms of AR coefficients as [1]

$$(A_k)_{ij} = 0, \quad k = 1, 2, \ldots, p \quad (2)$$

(where $(A_k)_{ij}$ denotes the (i, j) entry of A_k.) The concept of Granger causality has been extensively used for learning graphical models for various systems, for example, gene network, electroencephalogram (EEG), or fMRI (function magnetic resonance imaging); see [2, 3, 4, 5, 6] and many references therein. A common goal of these works is to fit an autoregressive model to a time series of interest and evaluate the zero entries in the estimated AR coefficients which finally reveal the interaction structure of the variables.

This paper discusses about an estimation problem in learning topology of Granger graphical models. Section 2 shows a problem formulation that promotes a joint sparsity in A_k's. The problem is a least-squares estimation with an ℓ_1-type regularization which can also be found in [4, 5, 7]. The problem falls into a convex framework which can be solved efficiently by many existing solvers. However, the nondifferentiability of the objective function makes it challenging to solve in large scale. Section 3 presents the contribution of this paper. We apply the alternating direction method of multipliers (ADMM) to solve the estimation problem. The algorithm has been successfully shown to be efficient and suitable for solving many large-scale statistical learning problems [8]. We confirm this result in section 4 where the performance of ADMM is shown in practice. We also provide numerical examples by synthetic and real data sets to demonstrate the approach presented in this paper.

2. PROBLEM FORMULATION

The least-squares (LS) method is a common approach used for fitting an AR model (1) to the measurements $y(1), \ldots, y(N)$.

Research supported by Grants for Development of New Faculty Staff, Ratchadaphiseksomphot Endowment Fund, Chulalongkorn University.

The model parameters A_k's are chosen such that the quadratic loss $\sum_{t=p+1}^{N} \|y(t) - \sum_{k=1}^{p} A_k y(t-k)\|_2^2$ is minimized. If we define $A = \begin{bmatrix} A_1 & \cdots & A_p \end{bmatrix} \in \mathbf{R}^{n \times np}$ then the quadratic loss can be rewritten more compactly as $\|Y - AH\|_2^2$ where the matrices Y and H contain the past measurements of $y(t)$.

If a Granger causality structure is given, formulating the problem of estimating AR model subject to the zero pattern of A_k's as in (2) is straightforward. However, in most applications, the goal is to learn a causal inference from the data, so the graph topology is commonly *unknown*. The topology can be induced from a zero pattern of matrices A_k's. Therefore, the idea is to propose a formulation that favors a *group* sparsity in A_k's. This can be done by introducing a sum of ℓ_2-norm term in the cost objective as follows.

$$\text{min} \quad \tfrac{1}{2}\|Y - AH\|_2^2 + \lambda \sum_{i \neq j} \|[(A_1)_{ij} \; (A_2)_{ij} \; \cdots \; (A_p)_{ij}]\|_2 \tag{3}$$

with variables $A_k \in \mathbf{R}^{n \times n}$ for $k = 1, 2, \ldots, p$. The scalar $\lambda > 0$ is called the regularization parameter. The summation over (i, j) with $i \neq j$ plays a role of ℓ_1-type norm, which causes some (i, j) entries of A_k to zero for a sufficiently large λ. Furthermore, using the ℓ_2 norm of p-tuple of $(A_k)_{ij}$ will force all p matrices A_k's to have the same sparsity pattern. This is known as a *Group Lasso* problem introduced in [9]. The sum of ℓ_2 norms is also called *composite absolute penalties* [10] or a *sum-of-norms regularization*. The formulation (3) is also considered in [4, 5, 7]. In these studies, they have shown the advantage of using Group Lasso formulation over the standard Lasso (the matrices A_k's may have different sparsity patterns). A nice property of (3) is that it is casted in the framework of convex optimization. We will show the benefits of solving it using the alternating direction method of multipliers (ADMM) in section 3.

3. ALTERNATING DIRECTION METHOD OF MULTIPLIERS

The problem (3) can be expressed in a general form as

$$\text{minimize} \;\; f(x) := h(x) + g(x)$$

where x is the variable (representing AR coefficients), h refers to the quadratic loss, and g refers to the sum-of-norms regularization term added to promote sparsity in x. Due to nondifferentiability in g, we explore several fast gradient-based methods [11, 12] that have been shown to be efficient for related nonsmooth problems such as Lasso or sparse covariance selection problems. At iteration k, the error in the cost objective decreases as fast as $1/k^2$ which is a significant improvement from existing gradient or subgradient methods applied on nonsmooth problems. Recently, the alternating direction method of multipliers (ADMM) has been proposed to similar ℓ_1 minimization problems. The method is a combination of the dual decomposition and augmented Lagrangian methods; see the detail in [8]. The method also offers a performance that is comparable to recent competitive algorithms. As an example, the performance of ADMM, fast iterative shrinkage-thresholding algorithm (FISTA) [12], and other methods applied on the related ℓ_1-regularized LS problem were discussed in [13, 14], where it was shown that ADMM and FISTA are efficient to estimate sparse models in several settings. Compared to the problem in [13], ours is more involved since the variables are matrices and to apply the ADMM, we need to arrange (3) in the ADMM format as

$$\begin{array}{ll} \text{min} & \tfrac{1}{2}\|Y - AH\|_2^2 + \lambda \sum_{i \neq j} \|[(Z_1)_{ij} \; \cdots \; (Z_p)_{ij}]\|_2 \\ \text{s.t.} & A - Z = 0. \end{array} \tag{4}$$

We define an auxillary variable $Z = \begin{bmatrix} Z_1 & Z_2 & \cdots & Z_p \end{bmatrix}$ and $Z_k \in \mathbf{R}^{n \times n}$, $k = 1, \ldots, p$. The ADMM algorithm for the problem (4) is as follows.

ADMM for sparse AR estimation. Initialize $A^{(0)}, Z^{(0)}$ and $U^{(0)}$ and set an ADMM parameter $\rho > 0$. For $k = 0, 1, \ldots$, repeat the following steps

$$A^{(k+1)} = \underset{A}{\text{argmin}} \;\; \frac{1}{2}\|Y - AH\|_2^2 + \frac{\rho}{2}\|A - Z^{(k)} + U^{(k)}\|_F^2$$

$$Z^{(k+1)} = \underset{Z}{\text{argmin}} \;\; \Big\{ (\rho/2)\|A^{(k+1)} + U^{(k)} - Z\|_F^2 + \lambda \sum_{i \neq j} \Big\| \begin{bmatrix} (Z_1)_{ij} & (Z_2)_{ij} & \cdots & (Z_p)_{ij} \end{bmatrix} \Big\|_2 \Big\}$$

$$U^{(k+1)} = U^{(k)} + A^{(k+1)} - Z^{(k+1)}$$

until a termination criterion is satisfied.

The A-update step can be analytically calculated by

$$A^{(k+1)} = \left[\rho(Z^{(k)} - U^{(k)}) + YH^T \right] (HH^T + \rho I)^{-1}.$$

Since $\rho > 0$, the A-update takes the form of a *ridge regression*. We perform the Cholesky factorization of $(HH^T + \rho I)$ once and use the factor in the next A-update.

For each (i, j), the Z-update step takes the form

$$\min_{z} \;\; (1/2)\|z - a\|_2^2 + \nu\|z\|_2, \qquad \nu > 0$$

where $z, a \in \mathbf{R}^p$. The solution of this problem is unique and given by $z = 0$ if $\|a\|_2 \leq \nu$ and $z = \dfrac{a}{\|a\|_2} \cdot (\|a\|_2 - \nu)$ otherwise. This expression is widely known as *a soft thresholding* applied on $\|a\|_2$ which can be cheaply computed in an elementwise manner for the Z-update.

In the context of ADMM, $A^{(k)}$ and $Z^{(k)}$ are called the *primal* and *dual* variables, respectively. The updates are terminated when the primal and dual residuals are small.

4. NUMERICAL EXAMPLES

In this section, we demonstrate the performance of the ADMM method for solving the topology selection on random sparse AR processes and on Google flu data set.

4.1. Convergence of ADMM in practice

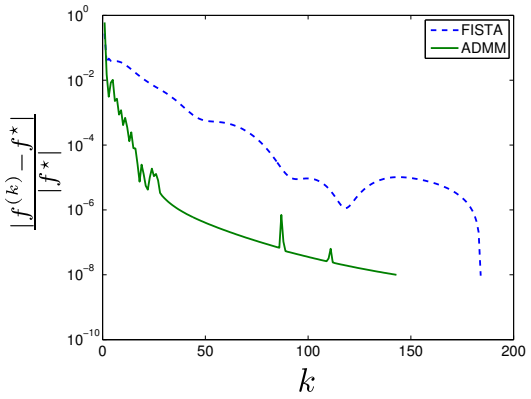


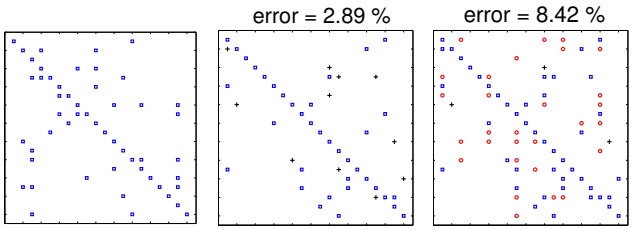
Fig. 1. Comparison of convergence in the relative error of the objective between ADMM and FISTA methods.

The ADMM algorithm is computationally cheap in each step so it is suitable for large-scale problems. In this experiment, we compare the ADMM method with the fast iterative shrinkage-thresholding algorithm (FISTA) [12]. The latter is another recent fast gradient-based method that is applicable to nonsmooth problems [14]. We use $n = 100, p = 3$ which gives $30,000$ variables in total. The density of nonzero entries in A_k's is set to 0.01 (high sparsity setting). We solve (3) with regularization parameter $\lambda = 0.1\lambda_{\max}$ where $\lambda_{\max}$ is the critical value such that using any $\lambda \geq \lambda_{\max}$ yields the solution A_k's of (3) as diagonal matrices (sparsest solution). The calculation of $\lambda_{\max}$ will be shown in section 4.3. We observed that selecting ρ in ADMM in the range of 10λ to 50λ gives a desirable performance. The algorithm FISTA is implemented with a backtracking line search to find a step size. We plot the relative errors $\frac{|f^{(k)}-f^\star|}{|f^\star|}$ in Figure 1 where we compute $f^\star$ by solving (3) using SeDuMi solver in `CVX` [15]. We can see that the number of iterations required to reach a desired accuracy for ADMM is much less than FISTA. Given that we have $30,000$ variables in total, it requires only a few hundred iterations to converge and finishes the task within 15-30 seconds.

4.2. Model selection on synthetic data set

In this section, we investigated the effect of the regularization parameter λ on the averaged error in the estimated topologies, chosen from two approaches; cross validation and a model selection criterion. The latter is the idea from [16] which incorporates a Bayes information criterion (BIC) score for ranking a small set of candidate topologies obtained by solving (3).

We generate a sparse AR model with dimension $n = 20, p = 3$ and $N = 1000$. Solving (3) by using nine values of λ in the range of $(0.01\lambda_{\max}, \lambda_{\max})$ results in nine estimated topologies, ranging from densest to sparsest graphs. The chosen λ corresponds to the model that minimizes the BIC score. In this example, the best model according to BIC yields the error of 2.89% in the estimated topology. We also tune the



(a) True topology (b) BIC (c) Cross validation

Fig. 2. Comparison of the true and estimated sparsity patterns. The blue squares are the correctly identified nonzero entries. The red circles are misclassified entries as nonzero. The black crosses are misclassified entries as zeros.

regularization parameter λ by using K-fold cross validation; see the method in [17, §7]. We choose 5 folds and 10 values of λ in the range $(0.01\lambda_{\max}, \lambda_{\max})$. From Figure 2, the cross validation technique tends to favor a denser graph while BIC prefers a sparse graph because the penalty term in BIC score suggests that BIC tends to reject complex models. The technique of tracing trade-off curves in [16] and using BIC score is a less computational burden compared to the cross-validation technique and also provides a smaller error in the estimated topology. Therefore, this approach is recommended when a simple model is favored.

4.3. Critical value of the regularization parameter

We can derive $\lambda_{\max}$ such that for any $\lambda \geq \lambda_{\max}$, the solution $A_1, \ldots, A_p$ of (3) are diagonal. The optimality (KKT) condition of (3) is obtained by setting the subgradient of the objective to zero. Define $a_{ij} = \begin{bmatrix}(A_1)_{ij} & (A_2)_{ij} & \cdots & (A_p)_{ij}\end{bmatrix}^T$. When $A_1, \ldots, A_p$ are *diagonal*, *i.e.*, $a_{ij} = 0$, a subgradient of the cost objective f in (3) with respect to a_{ij} is

$$\begin{bmatrix}\frac{\partial f}{\partial (A_1)_{ij}} \\ \vdots \\ \frac{\partial f}{\partial (A_p)_{ij}}\end{bmatrix} = -b_{ij} + \lambda g_{ij}, \quad \text{where} \quad b_{ij} = \begin{bmatrix}((Y - AH)H_1^T)_{ij} \\ \vdots \\ ((Y - AH)H_p^T)_{ij}\end{bmatrix},$$

H_k is the k^{th} block row of H, and g_{ij} is a subgradient of $\|a_{ij}\|_2$ which is any vector in $\mathbf{R}^p$ such that $\|g_{ij}\|_2 \leq 1$. The KKT condition is expressed as

$$\|b_{ij}\|_2 = \lambda \|g_{ij}\|_2 \quad \leq \quad \lambda, \quad \text{for } i \neq j. \tag{5}$$

To derive $\lambda_{\max}$, we need to compute the diagonals of A_k's which can be obtained from the zero gradient condition of the cost objective with respect to $(A_k)_{ii}$: $[(Y - AH)H_k^T]_{ii} = 0,\ k = 1, \ldots, p$. Since A_k's are diagonal, these equations are

$$\begin{bmatrix} (YH_1^T)_{ii} \\ \vdots \\ (YH_p^T)_{ii} \end{bmatrix} = \begin{bmatrix} (H_1H_1^T)_{ii} & (H_2H_1^T)_{ii} & \cdots & (H_pH_1^T)_{ii} \\ \vdots & & \vdots & \\ (H_1H_p^T)_{ii} & (H_2H_p^T)_{ii} & \cdots & (H_pH_p^T)_{ii} \end{bmatrix} \begin{bmatrix} (A_1)_{ii} \\ \vdots \\ (A_p)_{ii} \end{bmatrix}.$$

By solving for $(A_k)_{ii}$ and substituting them in (5), we define

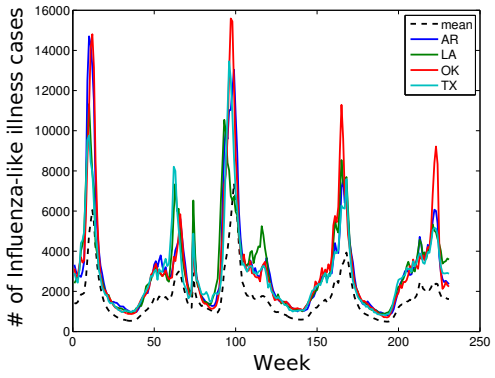
$$\lambda_{\max} = \max_{i \neq j} \|b_{ij}\|_2 \tag{6}$$

and conclude that if $\lambda \geq \lambda_{\max}$ then the solution of (3) are diagonal matrices (the sparsest solution.)

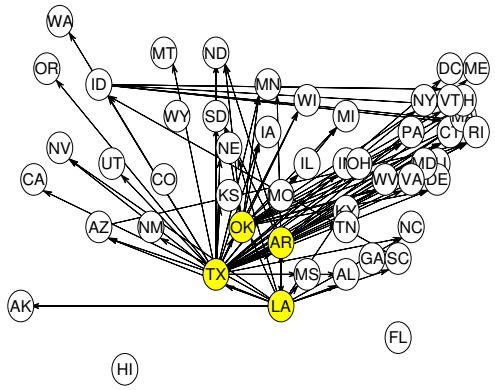
4.4. Real data set

Google Flu Trends is a project of analyzing Google search queries related to influenza-like illness (ILI) and using these information to estimate the number of actual flu activities occurred across the world; see `http://www.google.org/flutrends/`. The result of [18] shows that the estimated number of patients with ILI based on queries such as "flu" or "influenza" are close to the traditional flu activities.

In this experiment[1], we wish to investigate a causal structure of flu trends occurred in 51 states across the USA via Google Flu Trends data collected during Dec 2007 - Apr 2012 in a weekly basis. Figure 3(a) shows time series examples of Arkansas, Texas, Oklahoma and Louisiana. The y-axis represents the number of influenza-like illness (ILI) cases per 100,000 population (estimated by Google). Figure 3(b) shows the estimated topology of sparse AR models using the model selection approach with BIC score. It shows that Texas, Oklahoma, Louisiana and Arkansas have significant influences on many states. They are among the states that have higher numbers of ILI cases than the mean value and this result agrees with the fact that they are geographically neighbors.



(a) Time series of flu trend in AR, TX, OK, and LA states



(b) Granger graphical model for Google flu data

Fig. 3. Learning a graphical model for Google flu data.

5. RELATION TO PRIOR WORK

Our formulation (3) is similar to the ones considered in [4, 5, 7]. The paper [7] focused on the behaviour and consistency of the solution, while we focus on an efficient algorithm to solve the estimation problem. In [4], the regularization term is posed as a second-order cone (SOCP) constraint instead, and they applied the active-set algorithm with the SOCP solver. However, the problem dimension considered in this paper is small ($n = 7$), while we are more concerned with an algorithm that can handle larger problem dimensions. The paper [5] applied the Group Lasso procedure from [9] in combination with their methodology to learn a gene network system. While their method is shown to be promising, we opt to alternatively solve the problem by a convex optimization framework. The method in [5] requires tuning the parameter λ to enforce different levels of sparsity. Our explicit formula of the critical value of λ in (6) can be useful for selecting a range of λ for tuning purpose as well. We applied the ADMM method given in [8], which requires formulating (3) as the form of (4) to obtain a computationally efficient algorithm. The problem of estimating sparse AR model we considered here can be served as another evidence for the advantage of ADMM in the optimization or machine learning community.

6. CONCLUSION

We have presented a convex framework for learning a topology in Granger graphical models, which is equivalent to estimating autoregressive models and promoting a joint sparsity in the AR coefficients simultaneously. The formulation is a least-squares problem with an ℓ_1-type regularization. We have investigated the ADMM algorithm which is very simple to implement and has a desirable rate of convergence in practice. Experiment with randomly generated data sets and time series of Google flu trends were included to confirm the effectiveness of our approach.

[1]Thanks to Pancheewa Arayacheeppreecha for the experimental results.

7. REFERENCES

[1] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*, Springer, 2005.

[2] M. Eichler, "A graphical approach for evaluating effective connectivity in neural systems," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1457, pp. 953–967, 2005.

[3] P.A. Valdés-Sosa, J.M. Bornot-Sánchez, M. Vega-Hernández, L. Melie-García, A. Lage-Castellanos, and E. Canales-Rodríguez, "Granger causality on spatial manifolds: Applications to neuroimaging," in *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*, B. Schelter, M. Winterhalder, and J. Timmer, Eds. Wiley, 2006.

[4] S. Haufe, G. Nolte, and N. Kräemer, "Sparse causal discovery in multivariate time series," *Proceedings of JMLR Workshop and Conference*, vol. 6, pp. 97–106, 2008.

[5] A.C. Lozano, N. Abe, Y. Liu, and S. Rosset, "Grouped graphical granger modeling for gene expression regulatory networks discovery," *Bioinformatics*, vol. 25, pp. 110–118, 2009.

[6] A. Shojaie and G. Michailidis, "Discovering graphical granger causality using the truncating lasso penalty," *Bioinformatics*, vol. 26, no. 18, pp. 517–523, 2010.

[7] A. A. Bolstad, B. Van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2628–2641, 2011.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[9] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B Statistical Methodology*, vol. 68, no. 1, pp. 49–67, 2006.

[10] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *The Annals of Statistics*, vol. 37, no. 6A, pp. 3468–3497, 2009.

[11] Yu. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming Series A*, vol. 103, pp. 127–152, 2005.

[12] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[13] A.Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Y. Ma, "Fast ℓ_1-minimization algorithms and an application in robust face recognition: A review," Tech. Rep., Electrical Engineering and Computer Sciences, University of California, Berkeley, 2010.

[14] J. Liu and J. Ye., "Efficient ℓ_1/ℓ_q norm regularization," 2010, Preprint available at `arXiv.org` (1009.4766).

[15] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming (web page and software)," `http://stanford.edu/~boyd/cvx`, August 2008.

[16] J. Songsiri and L. Vandenberghe, "Topology selection in graphical models of autoregressive processes," *Journal of Machine Learning Research*, vol. 11, pp. 2671–2705, 2010.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2nd edition, 2009.

[18] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2008.