Journal of Multivariate Analysis 152 (2016) 1–14

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Group-wise semiparametric modeling: A SCSE approach

Song Song [a], Lixing Zhu [b,c,*]

[a] University of Texas, Austin, United States
[b] Suzhou University of Science and Technology, China
[c] Hong Kong Baptist University, Hong Kong

ARTICLE INFO

Article history:
Received 16 August 2015
Available online 6 August 2016

AMS subject classifications:
62J10
62G08
62H30

Keywords:
Covariance estimation
Regularization
Sparsity
Thresholding
Semiparametrics
Variable clustering

ABSTRACT

This paper is motivated by the modeling of a high-dimensional dataset via group-wise information on explanatory variables. A three-step algorithm is suggested for group-wise semiparametric modeling: (i) screening to reduce dimensionality; (ii) clustering according to grouped explanatory variables; (iii) sign-constraints-based estimation for coefficients to produce meaningful interpretations. As a justification, under the setup of m-dependent and β-mixing processes, the interplay between the estimator's convergence rate and the temporal dependence level is quantified and a cross-validation result about the resampling scheme for threshold selection is also proved. This method is evaluated in finite-sample cases through a Monte Carlo experiment, and illustrated with an analysis of the US consumer price index.

1. Introduction

The research described herein is motivated by the analysis of a dataset from [18], which contains 131 monthly macro indicators covering a broad range of categories such as income, industrial production, capacity, employment and unemployment, consumer prices, producer prices, and so forth. The time span is from January 1959 to December 2003. To standardize the observations, logarithm transformation to most of the series is used except those already expressed in rates. The series are transformed to attain stationarity by taking the (1st or 2nd order) differences of the raw data series (or the logarithm of the raw series). Then all of the observations are standardized. As the underlying model is unknown, flexible semiparametric modeling is considered.

For this kind of dataset, people usually group variables according to some rule of thumb. For example, to model the consumer price index (CPI: all items), people might subjectively organize the variables into groups, say as follows:

Group 1: CPI: apparel & upkeep, transportation, medical care, commodities, durables, services';
Group 2: CPI: all items less food, all items less shelter, all items less medical care;
Group 3: "Producer Price Index (PPI)—finished goods, finished consumer goods, intermed mat. supplies & components, crude materials";
Group 4: "Implicit Price Deflator of Personal Consumption Expenditures (PCE)—all items, durables, nondurables, services";

* Corresponding author at: Hong Kong Baptist University, Hong Kong.
E-mail addresses: ssoonngg123@gmail.com (S. Song), lzhu@hkbu.edu.hk (L. Zhu).

http://dx.doi.org/10.1016/j.jmva.2016.07.006
0047-259X/

and all other variables in Group 5. However, it is uncertain whether such grouping would fit the actual data structure. See Section 5 for discussion.

In general, properly grouping the explanatory variables is of importance for modeling, which can make the model more accurate and efficient. In effect, in regression analysis with high-dimensional data, explanatory variable grouping is a common feature.

In the independent and identically distributed (i.i.d.) scenario with fixed number of explanatory variables, Li et al. [12] considered a semiparametric modeling approach called the group-wise minimum average variance estimation (GMAVE). This is a generic method capable of handling many semiparametric models. Note that the GMAVE requires both available information of variable grouping and use of high-dimensional kernel functions. This is a typical model fitting approach. However, the requirements are often infeasible when group information is not available beforehand and certain limitations could also experienced when the number of explanatory variables is large.

As is well known, there are many variables that do not significantly affect the response. Based on an initial investigation on the dataset we described above, this is the case. Thus, to make regression modeling more efficient, we may remove those variables in the modeling process rather than including all of the variables in a model first and then removing those "insignificant variables" later. The GMAVE cannot directly do this variable selection/screening.

In the literature, there are some relevant methods. Wang et al. [20] considered an additive multi-index model and discussed the variable selection when GMAVE is applied. This is a "screening first; fitting later" approach for modeling high-dimensional data because the grouping has been given beforehand. Fan and Lv [8] is another example that does not concern grouping information. Alternatively, Bickel et al. [3] and Meinshausen Bühlmann [15] considered the "fitting first; screening later" approach. But when the spatial structure is complex and the modeling is semiparametric, the "fitting first; screening later" approach might have limitations. Furthermore, for parameter estimation, no semiparametric method handles the sign-constrained issue that is important for practical interpretation, but less attention has been paid to it in the literature.

As such, to facilitate the semiparametric modeling procedure combined with grouping information and sparse structure, we propose a novel modeling method. Our algorithm comprises three steps: (i) screening for selection; (ii) clustering for variable grouping; and (iii) estimating sign-constrained parameters. Therefore, this is an *integrated algorithm* of screening/clustering/sign-constrained estimation (SCSE). Unlike existing approaches, our algorithm could be considered a "screening first; grouping second; fitting third" approach. As the algorithm is related to large-dimensional covariance matrices, for justification, we will give an estimation and its theoretical properties for high-dimensional time series together with a special handling of the cross-validation procedure.

The article is organized as follows. In Section 2, the details of the SCSE procedure are described. Section 3 presents the investigation on covariance matrix estimation when the process is either m-dependent or β-mixing. For more general processes, the theoretical investigation deserves further research. The spatial structure of large panels of macroeconomic and financial times series is studied in Section 5 such that proper semiparametric structure for estimating several key economic measures can be found. Section 4 is devoted to numerical evaluation for the performance of the SCSE algorithm. Section 6 contains a brief discussion of further research topics and all technical proofs are sketched in the Appendix.

2. Screening–clustering–sign-constrained estimation

2.1. A visualized description of the algorithm

To describe the basic idea of the algorithm, we first study the differences among various semiparametric models from a graphical perspective. If we use a vertex in the graph to represent a relevant variable, a crossed vertex to represent an "unrelated" variable, a solid edge in a "block" to represent the linear relationships among variables inside and a banded edge (connecting a "block" with the dependent variable Y) to represent a nonparametric link function, then we can visualize different semiparametric models through corresponding graphs.

For example, we can create Fig. 1 (left panel) for the single index model, Fig. 1 (middle panel) for the additive model and Fig. 1 (right panel) for the more general multiple index model, among many others. The underlying difference among the various semiparametric models is the allocation of the nonparametric link function and linearity through variable clustering.

Consequently, assuming that all of the variables have been included, if we can find the corresponding graph types, we can construct the right class of semiparametric models. Note that sparse correlation matrices are very useful in graphs because zero partial correlations help establish independence and conditional independent relations in the context of graphs and thus imply a graphical structure.

For example, if we have a sparse covariance matrix for $Y, X_1, \ldots, X_9$ as displayed in Fig. 2, we know that $X_1, \ldots, X_6$ are "relevant" to Y, and due to the "block" structure with respect to X_1, X_2, X_3 and X_4, X_5, we can construct the following class of semiparametric models:

$$\mathrm{E}(Y) = g_1(X_1\beta_1 + X_2\beta_2 + X_3\beta_3) + g_2(X_4\beta_4 + X_5\beta_5) + g_3(X_6\beta_6). \tag{1}$$

More generally, the modeling is as follows. For given variables $X_1, \ldots, X_{J-1}$ and Y (or X_J), we try to find the index sets $\mathcal{A}_1, \ldots, \mathcal{A}_S$ (possibly with overlapping elements) such that Y could be well approximated by:

$$\sum_{s=1}^{S} g_s\Big(\sum_{\ell=1}^{|\mathcal{A}_s|} \beta_{s\ell} X_{\ell\in\mathcal{A}_s}\Big) \stackrel{\mathrm{def}}{=} \sum_{s=1}^{S} g_s\Big(\beta_s^\top X_{\mathcal{A}_s}\Big), \tag{2}$$