Dimension Reduction in Time Series

by

Jin-Hong Park

(Under the direction of T. N. Sriram and Xiangrong Yin)

Abstract

We develop a new theory of dimension reduction in time series, which provides an initial phase when an adequate parsimoniously parameterized time series model is not yet available. In this thesis, we define a notion of Time Series Central Subspace and Time Series Central Mean Subspace, and estimate them using newly developed methods, when the lag of the series and minimum dimension are known. The estimators are shown to be strongly consistent. In addition, we also discuss estimation of the minimum dimension and the lag. The theory of dimension reduction in time series poses many challenges, but a variety of encouraging results presented through extensive simulations and real data analysis seem to suggest that our method has a great potential for providing a viable and meaningful alternative to traditional time series analysis. In fact, superior performance of our nonlinear or linear time series models for several real data sets serve as a testament that our methods are very useful in time series analysis. We believe that the ideas and methods presented here are of interest to time series analyst in fields such as Economics, Business, Climatology, among others. We hope that this work will stimulate a new way of analyzing time series data.

Index words: Time series central subspace, Kullback-Leibler distance, Density estimator, Nonlinear time series, Threshold, Time series central mean subspace, Kernel estimation method, BIC, RIC

Dimension Reduction in Time Series

by

Jin-Hong Park

M.S., University of Georgia, 2000

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2007

Dimension Reduction in Time Series

by

Jin-Hong Park

Approved:

Major Professors:	T. N. Sriram
	Xiangrong Yin
Committee:	Gauri Datta
	William P. McCormick
	Jaxk Reeves
	Lynne Seymour

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2007

Dedication

To the respecting memory of my father whose endless inspiration and encouragement helped me get to where I am today. I owe more than what I gave him when he was alive. He was truly a great inspiration and a dear friend.

Acknowledgments

First of all, my sincere thanks go to my advisors, Dr. T.N. Sriram and Dr. Xiangrong Yin. Let me add here that my enthusiasm alone would not have made me choose a career in academics; my advisors are overwhelmingly responsible for that. They were extremely supportive throughout my entire Ph.D. program.

I would also like to thank my committee members Dr. Gauri Datta, Dr. William P. McCormick, Dr. Jaxk Reeves and Dr. Lynn Seymour for agreeing to read this dissertation and provide wonderful advice. Special thanks, to Dr. Robert Lund at the Department of Mathematics, Clemson University, for his encouragement and his recommendations while applying for the Ph.D. program in Statistics.

I am grateful to Dr. Paul Na at Lehman Brothers and Dr. Jianwu Wang at Citigroup for providing invaluable support in the transition and adjustment to the new American system. I acknowledge with sincere thanks for the time well spent with Dr. Desale Habtzghi at the Department of Mathematics, Georgia College and State University, and Dr. Dipankar Bandyopadhyay at the Department of Biostatistics, Bioinformatics and Epidemiology, at the Medical University of South Carolina, for their advises given to this senior doctoral student and for being such good friends. Special thanks, to Chris Franklin for providing me with many good suggestions for my future career.

Finally and most importantly, I recall the support of my family in South Korea, their sacrifices, and their unconditional love. I am really lucky to learn from them the essence of education is an honorable goal of life.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

Dimension reduction in regression has been a very active area of intense research for over two decades. While many new techniques and estimation methods have been proposed for dimension reduction in regression, only a handful of recent work (Xia, Tong, Li and Zhu, 2002) deals with dimension reduction issues for time series. What is more interesting to note is that, while many ideas and techniques in time series analysis have their origin in classical regression theory, the concept of dimension reduction in regression has not been formally extended to the time series context. Such an extension is the main goal of this thesis. Our objective here is to develop the formal dimension reduction theory for time series when an adequate parsimoniously parameterized time series model is not yet available. As far as we know, the materials presented here is the first formal development of dimension reduction theory for time series. Before we delve into the main goal of the thesis, we present a brief overview of time series methodologies from the literature and describe two notions of dimension reduction in regression, known as Central Subspace (CS) and Central Mean Subspace (CMS). In Section 1.2, we describe prior work on dimension reduction in time series and briefly describe contents of the rest of thesis.

1.1 An Overview of Time Series Analysis and Dimension Reduction in Regression

Statistical analysis of data observed at adjacent time points, commonly known as time series analysis, has been an active area of research for several decades. The intrinsic nature of time series is that its observations are correlated. This severely restricts the direct applicability

of many conventional statistical methodologies, which are primarily suited for analyzing independent and identically distributed data. The unique challenges posed by the nature of time series data sets have given rise to two broad approaches, categorized as the *time domain approach* and the *frequency domain approach*. Over the years, the statistics community has witnessed development of many useful parametric and nonparametric methods for analyzing time series data. Nevertheless, there is a never-ending quest to build new and modern methodologies to analyze time series data, which occur in a variety of fields such as economics, meteorology, engineering, geophysics, social and environmental science; the list is practically endless.

There is a long tradition of using either parametric or nonparametric methods in analyzing time series data. Model-based parametric methods overcome some practical complexities associated with time series analysis. Autoregressive Moving Average (ARMA) models are customarily used in linear time series model; see, for example, Wei (2006) and Brockwell & Davis (1996), mathematical development of such model and their usefulness in analyzing real time series data. Threshold Autoregression (TAR) models, on the other hand, are special types of nonlinear models. Tong (1978, 1983, 1990) and Tong & Lim (1980) proposed piecewise linear models, where the linear relationship changes according to values of the process. There are also models incorporating a nonconstant error variance, known as heteroscedasticity. These include models such as Autoregressive Conditional Heteroscedasticity (ARCH) by Engle (1982) and Generalized Autoregressive Conditional Heteroscedasticity (GARCH) by Bollerslev (1986); and, these are also special type nonlinear models, often used to characterize volatility. Nonparametric methods make no assumption about the structure of a time series, but impose existence of density functions and associated smoothness conditions. Although an attractive alternative to parametric methods, practical implementation of nonparametric methods require specification of density estimators and selection associated bandwidth and other smoothing parameters. In addition, nonparametric methods do not necessarily perform well in high dimensions. Nevertheless, these methods provide a guid-

ance for selecting appropriate lower-dimensional parametric models and for deciding between competing models.

It is well-known that estimation approaches from classical regression theory have been useful in building linear/nonlinear models for time series data $\{x_1, \cdots, x_t; t \geq 1\}$, where there is an obvious dependence of x_t on the past values $\{x_{t-1}, \cdots, x_1\}$. In addition, the emphasis in time series is usually on forecasting future values, which is easily treated as a regression problem. For an in-depth exposition of widely studied time series models and forecasts based on these models, see e.g., Brockwell and Davis (1996), Shummway and Stoffer (2000), Fan and Yao (2003), Tsay (2005) and Wei (2006).

In this thesis, we develop a new theory for analyzing time series data which provides an initial phase when an adequate parsimoniously parameterized time series model is not yet available. Before we progress toward our goal, we briefly review the concept of dimension reduction in regression.

Let Y be a scalar response variable and $\mathbf{X}$ be a $p \times 1$ random covariate vector. The goal is to make inference about how the conditional distribution $Y|\mathbf{X}$ varies with the values of $\mathbf{X}$. Dimension reduction in regression is to find q linear combinations say, $\boldsymbol{\beta}_1^T\mathbf{X}, \cdots, \boldsymbol{\beta}_q^T\mathbf{X}$ with $q \leq p$ such that the conditional distribution of $Y|\mathbf{X}$ is same as the conditional distribution of $Y|(\boldsymbol{\beta}_1^T\mathbf{X}, \cdots, \boldsymbol{\beta}_q^T\mathbf{X})$. In other words, there would be no loss of information if $\mathbf{X}$ were replaced by the $q(\leq p)$ linear combinations. Suppose $\mathbf{B} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_q)$ denotes the $p \times q$ matrix. Then, the goal is equivalent to finding a $\mathbf{B}$ matrix such that $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^T\mathbf{X}$, that is, Y is (conditionally) independent of $\mathbf{X}$ given $\mathbf{B}^T\mathbf{X}$. Here, $\perp\!\!\!\perp$ indicates independence. As introduced in Cook (1994, 1998), the dimension reduction methods approach this problem through a Central Subspace (CS), which is a minimum dimension reduction subspace (DRS), say $\mathcal{S}(\mathbf{B_0})$, with $\dim(\mathcal{S}(\mathbf{B_0})) \leq \dim(\mathcal{S}(\mathbf{B}))$ for all DRSs $\mathcal{S}(\mathbf{B})$ for which Y is independent of $\mathbf{X}$ given $\mathbf{B}^T\mathbf{X}$. Here, "dim" denotes dimension and the columns of matrix $\mathbf{B}$ form a basis for the DRS. This approach is appealing because it does not specify a parametric model and there is no loss of information about the conditional distribution of Y given $\mathbf{X}$. For

deeper issues related to the topic of dimension reduction and estimation approaches, see Cook (1998).

Indeed, regression is understood by some to imply a study of the mean function $E(Y|\mathbf{X})$. Recently, Cook and Li (2002) introduced a concept called Central Mean Subspace (CMS), which is similar in spirit to that of a CS, but dimension reduction is aimed at reducing the mean function alone, leaving the rest of $Y|\mathbf{X}$ as the nuisance parameter. Here, CMS is a minimum mean DRS, say $\mathcal{S}_{E(Y|\mathbf{X})}(\mathbf{B_0})$, with $\dim(\mathcal{S}_{E(Y|\mathbf{X})}(\mathbf{B_0})) \leq \dim(\mathcal{S}_{E(Y|\mathbf{X})}(\mathbf{B}))$ for all DRSs $\mathcal{S}_{E(Y|\mathbf{X})}(\mathbf{B})$ for which Y is independent of $E(Y|\mathbf{X})$ given $\mathbf{B}^T\mathbf{X}$.

1.2 Dimension Reduction in Time Series

To address the issue of dimension reduction in time series, Xia and Li (1999), and Xia, Tong and Li (1999, 2002) considered a single-index model, which avoids the curse of dimensionality. Recently, Xia, Tong, Li and Zhu (2002) proposed dimension reduction methods in regression, which is also applicable to time series with known lag, but their focus is only on the estimation of dimensions in the mean function. For an ensemble of time series, Li and Shedden (2002) present a dimension reduction method, which identifies a small number of independent time series components such that each time series in the ensemble is a different linear combination of the components. Their notion of dimension reduction, however, differs from ours. Becker and Fried (2003) use a dynamic version of Sliced Inverse Regression (SIR, Li 1991) as an exploratory tool for analyzing multivariate time series where the lag is chosen using preliminary information. Hall and Yao (2005) discuss an estimation method, which approximates the conditional distribution function of x_t given the past using a single linear combination of the past. To the best of our knowledge, there is no formal sufficient dimension reduction theory in time series which overcomes curse of dimensionality without making specific model assumptions or using specific number of dimensions and lag. Development of such a formal theory for time series is the main goal of this thesis.

The primary goal of time series analysis is forecasting, which requires inference about the conditional distribution of $x_t|\mathbf{X}_{t-1}$, for some suitable lag $p \geq 1$, where $\mathbf{X}_{t-1} = (x_{t-1}, ..., x_{t-p})^T$. Typically, the lag p is not known. However, there are diagnostic ways and estimation methods for determining a value of p before proceeding with the inference (Ng and Perron, 2005). It is also important to note that with known p, we may only need a few linear combinations of $\mathbf{X}_{t-1}$ in the final model (Xia and Li 1999; Xia, Tong and Li 1999, 2002), determination of which is one of our main focus.

In Chapter 2, we propose a notion of time series central subspace, definition of a minimum dimension reduction subspace, and estimate it using a method based on Kullback-Leibler distance, when the lag and minimum dimension are known. The estimator is shown to be strongly consistent. In addition, we also discuss estimation of minimum dimension and lag. Furthermore, we show that the proposed estimator of minimum dimension (when it exists) is strongly consistent. We proposed a graphical approach for the determination of true lag of series.

In Chapter 3, we propose a notion of time series central mean subspace, definition of a minimum mean dimension reduction subspace, and estimate it using a method based on residual sum of squares, when the lag and minimum dimension are known. In order to estimate the correct dimension, we introduce two information criteria. We examine the performance of the estimators of the minimum dimension extensively through simulation and real data analysis. As for estimation of the unknown lag, we continue to use the graphical approach introduced in Chapter 2.

In Chapter 4, we give a brief discussion and conclusion summarizing the results of TSCS and the TSCMS approaches.

1.3 References

[1] Becker, C. and Fried, R. (2003), "Sliced inverse regression for high-dimensional time series," *Exploratory Data Analysis in Empirical Research: Proceedings of the 25th Annual Conference of the Gesellschaft für Klassifickation*, University of Munich, 3-11.

[2] Biren, H. J. (1994), *Topics in Advanced Econometrics: Estimation, testing, and specification of cross-section and time series models*, Cambridge: Cambridge University Press.

[3] Bollerslev, T. (1986), "Generalized autoregressive conditional heteroscedasticity," *Journal of Econometrics*, 31, 307–327.

[4] Brockwell, P. J. and Davis, R. A. (1996), *Introduction to Time Series and Forecasting*, New York: Springer-Verlag.

[5] Cook, R. D. (1998), *Regression Graphics: Ideas for studying regressions through graphics*, New York: Wiley.

[6] Cook, R. D. and Li, B. (2002), "Dimension reduction for conditional mean in regression," *The Annals of Statistics*, 30, 455-474.

[7] Engle, R. F. (1982), "Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation," *Econometrica*, 50, 987–1008.

[8] Fan, J. and Yao, Q. (2003), *Nonlinear Time Series: Nonparametric and Parametric Methods*, Springer, New York.

[9] Hall, P. and Yao, Q. (2005), "Approximating conditional distribution functions using dimension reduction." *Annals of Statistics*, 33, 1404–1421.

[10] Hart, J. D. (1996), "Some automated methods of smoothing time-dependent data," *Nonparametric Statistics*, 6, 115–142.

[11] Härdle, W., Lütkepohl, H., and Chen, R. (1997), "A review of nonparametric time series analysis," *International Statistical Review*, 65, 49–72.

[12] Härdle, W. and View, P. (1992), "Kernel regression smoothing of time series." *Journal of Time Series Analysis*, 13, 209–232.

[13] Li, K. C. (1991), "Sliced inverse regression for dimension reduction (with discussion)," *Journal of the American Statistical Association*, 86, 316–342.

[14] Li, K. and Shedden, K. (2002), "Identification of shared components in large ensembles of time series using dimension reduction," *Journal of the American Statistical Association*, 97, 759–765.

[15] Ng, S. and Perron, P. (2005). "A note on selection of time series models," *Oxford Bulletin of Economics and Statistics*, 67, 115–134.

[16] Robinson, P. M. (1983), "Nonparametric estimation for time series models," *Journal of Time Series Analysis*, 4, 185–208.

[17] Shummway, R. H. and Stoffer, D. S. (2000), *Time Series Analysis and Its Applications*, New York: Springer-Verlag.

[18] Tong, H. (1978), *On a Threshold model, in Pattern Recognition and Signal Processing (Ed. C. H. Chen)*, Sijthoff and Noordhoff, Amsterdam.

[19] Tong, H. (1983), *Threshold Models in Non-linear Time Series Analysis*, New York: Springer-Verlag.

[20] Tong, H. (1990), *Non-liner Time Series: A Dynamic System Approach*, New York: Oxford University Press.

[21] Tong, H., Lim, K. S. (1980), "Threshold autoregression, limit cycles, and cyclical data (with discussion)," *Journal of the Royal Statistical Society*, Ser. B, 42, 245–292.

[22] Tsay, R. S. (2005), *Analysis of Financial Time Series*, New York: John Wiley & Sons.

[23] Wei, W. W. S. (2006), *Time Series Analysis: Univariate and Multivariate Methods*, Boston: Pearson & Addison Wesley.

[24] Xia, Y. and Li, W. K. (1999), "On the estimation and testing of functional coefficient linear models," *Statistca Sinica*, 9, 735–757.

[25] Xia, Y., Tong, H., and Li, W. K. (1999), "On extended partially linear single-index models," *Biometrika*, 86, 831–842.

[26] Xia, Y., Tong, H., and Li, W. K. (2002), "Single-index volatility models and estimation," *Statistica Sinica*, 12, 785–799.

[27] Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002), "An adaptive estimation of dimension reduction," *Journal of the Royal Statistical Society*, Ser. B, 64, 363–410.

Chapter 2

Time Series Central Subspace [†]

[†]Park, J. H., Sriram, T. N., and Yin, X. Submitted to *The Journal of American Statistical Association,* 08/06.

2.1 Introduction

Traditionally, time series analysis involves building an appropriate model and using either parametric or nonparametric methods to make inference about the model parameters. Motivated by recent developments in dimension reduction theory in regression, we develop a similar theory for time series which does not require specification of a model but seeks to find a $p \times d$ matrix $\mathbf{\Phi}_d$, with smallest possible number $d(\leq p)$, such that the conditional distribution of $x_t|\mathbf{X}_{t-1}$ is the same as that of $x_t|\mathbf{\Phi}_d^T\mathbf{X}_{t-1}$, where $\mathbf{X}_{t-1} = (x_{t-1}, ..., x_{t-p})^T$, resulting in no loss of information about the conditional distribution of the series given its past p values. To this end, we define a notion of minimum dimension reduction subspace, called time series central subspace, and estimate it using a recent method based on Kullback-Leibler distance, when p and d are known. The estimator is shown to be strongly consistent. In addition, we also discuss estimation of dimension d and lag p. We illustrate our method via simulations for a variety of linear and nonlinear time series models and through analysis of real data on Wolf yearly sunspot numbers, U.S. GNP, and U.S. beer production data (Wei, 2006). We believe that the methods presented here offer a new approach to analyzing time series data.

In Section 2.2, we formally develop the theory of dimension reduction in time series by introducing the notion of time series central subspace (TSCS). In Section 2.3, we discuss the estimation of TSCS when its dimension d and lag p of the series are known, and discuss other related issues. More precisely, in Section 2.3.1, we introduce an objective function and study its properties. In Section 2.3.2, we give a detailed computational algorithm to compute the estimator. In Section 2.3.3, we suggest a data-dependent way of determining lag p and d linear combinations, which provides full information about the conditional distribution of $x_t|\mathbf{X}_{t-1}$. Estimating this *minimal* set of linear combinations and replacing $\mathbf{X}_{t-1}$ by the estimated linear combinations is what we call *dimension reduction in time series*. The consistency results for our estimators are stated in Section 2.3.4. In Section 2.4, we carry out several Monte Carlo simulations followed by analysis of the well-known Wolf yearly sunspot data, U.S. GNP, and

U.S. beer production data. In Section 2.5, we present a discussion of the results obtained here. All the necessary proofs are given in the Appendix. Note that our approach provides a new way of analyzing time series data, without specification of a parametric model and without loss of information about the conditional distribution of x_t given $\mathbf{X}_{t-1}$. Above all, our approach provides an initial phase when an adequate parsimoniously parameterized time series model is not yet available.

2.2 Central Subspace in Time Series

A time series data $x_1, \cdots, x_t$ naturally evolves over time and hence there is an obvious dependence of the current value x_t on the past values $x_{t-1}, \cdots, x_1$. Therefore, it would be useful to make inference about the conditional distribution of x_t given the past. However, in many real data sets it is possible to determine (using the plot of autocorrelation and partial autocorrelation functions of the series) a value of $p \geq 1$, perhaps large, such that it suffices to make inference about the conditional distribution of $x_t|\mathbf{X}_{t-1}$, for some $p \geq 1$, where $\mathbf{X}_{t-1}$ is as defined in the introduction. We will begin by assuming that such a lag value p exists and is known. Later in this chapter, we also consider the case of unknown p.

Our goal is to find finitely many linear combinations, $\Phi_1^T\mathbf{X}_{t-1}, \cdots, \Phi_q^T\mathbf{X}_{t-1}$, with $q \leq p$ such that the conditional distribution of $x_t|\mathbf{X}_{t-1}$ is same as the conditional distribution of $x_t|(\Phi_1^T\mathbf{X}_{t-1}, \cdots, \Phi_q^T\mathbf{X}_{t-1})$. As mentioned earlier, this is equivalent to finding a $p \times q$ matrix $\mathbf{\Phi} = (\Phi_1, ..., \Phi_q)$ such that

$$x_t \perp\!\!\!\perp \mathbf{X}_{t-1}|\mathbf{\Phi}^T\mathbf{X}_{t-1}, \tag{2.1}$$

that is to say, x_t is independent of $\mathbf{X}_{t-1}$ given $\mathbf{\Phi}^T\mathbf{X}_{t-1}$. Therefore, the $p \times 1$ vector $\mathbf{X}_{t-1}$ can be replaced by the $q \times 1$ vector $\mathbf{\Phi}^T\mathbf{X}_{t-1}$ without loss of information. This represents a potentially useful reduction in the dimension of $\mathbf{X}_{t-1}$, where all the information in $\mathbf{X}_{t-1}$ about x_t is contained in the q-linear combinations.

As in the introduction, we define a DRS for x_t on $\mathbf{X}_{t-1}$ as any subspace $\mathcal{S}(\mathbf{\Phi})$ of $\mathbb{R}^q$ for which (2.1) holds. Note that (2.1) holds trivially for $\mathbf{\Phi} = I_{p\times p}$, which implies that a

dimension reduction subspace always exists. Since the interest is in reducing the dimension, we want to find a minimum DRS for x_t on $\mathbf{X}_{t-1}$. To this end, we define the intersection of all DRSs as a TSCS, if the intersection is itself a DRS. We denote the TSCS by $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\mathbf{\Phi}_d)$, where $\dim(\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\mathbf{\Phi}_d)) = d$. Clearly, this TSCS is the minimum DRS. The notion of TSCS is parallel to that of the central subspace in regression (Cook, 1994, 1998a), and this provides an initial phase when an adequate parsimoniously parameterized time series model is not yet available. Note that in the definition of TSCS, we do not restrict the time series to be linear or nonlinear or impose a stationarity assumption, as done in traditional time series modeling approaches.

Whereas $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\mathbf{\Phi}_d)$ is always a subspace, it is not necessarily a DRS. However, in regression analysis, it can be shown under certain conditions (Cook 1994, 1996, 1998a) that the intersection of DRS is a DRS. For the time series setup, we have the following result that guarantees the existence of TSCS. The proof follows arguments similar to those in Cook (1998a, p. 108) and is given in the Appendix (see Section 2.6).

Proposition 1 *Let $\mathcal{S}(\boldsymbol{\eta})$ and $\mathcal{S}(\boldsymbol{\gamma})$ be DRSs for x_t on $\mathbf{X}_{t-1}$. If $\mathbf{X}_{t-1}$ has a density $\mathbf{f}(\mathbf{x}_{t-1}) > 0$ for $\mathbf{x}_{t-1} \in \Omega_{\mathbf{X}_{t-1}} \subset \mathbb{R}^p$, where $\Omega_{\mathbf{X}_{t-1}}$ is the support of $\mathbf{X}_{t-1}$, and $\mathbf{f}(\mathbf{x}_{t-1}) = 0$ otherwise, then $\mathcal{S}(\boldsymbol{\eta}) \cap \mathcal{S}(\boldsymbol{\gamma})$ is a DRS.*

We conclude this section with an example which shows that TSCS exists. If $p = 3$ and $\mathbf{X}_{t-1} = (x_{t-1}, x_{t-2}, x_{t-3})^T$, set $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_3 x_{t-3} + \varepsilon_t$, where ε_t are independent normal random variable. Then vector $(\phi_1, \phi_2, \phi_3)^T$ forms a basis of TSCS. On the other hand, if there is a degenerate case such as $x_t = \phi_1 x_{t-1}$, then $\mathcal{S}((1,0,0)^T)$, $\mathcal{S}((0,1,0)^T)$, and $\mathcal{S}((0,0,1)^T)$ are all minimum DRSs, but there does not exist a central subspace because $\cap \mathcal{S}_{DRS}$ is equal to the origin.

For the rest of this chapter, we assume the existence of TSCS. Our definition of TSCS is generally enough to cover many linear and nonlinear autoregressive (AR) models including

the AR(p), threshold autoregression, and conditionally heteroscedastic autoregression. However, our approach does not include moving average (MA) models or generalized autoregressive conditionally heteroscedastic models, which may be viewed as infinite order autoregression.

2.3 Estimation of TSCS

Our goal now is to develop a method for estimating the TSCS, $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\mathbf{\Phi}_d)$, which does not require a pre-specified model for $x_t|\mathbf{X}_{t-1}$. Our estimation method is similar to the one proposed in Yin and Cook (2005). We begin by assuming that dimension $d(\leq p)$ of TSCS and lag p in $\mathbf{X}_{t-1}$ are known. Therefore, we need to estimate only the set of vectors $(\Phi_1, ..., \Phi_d)(= \mathbf{\Phi}_d)$. One may immediately think that well-known inverse dimension reduction methods such as the Sliced Inverse Regression (Li, 1991) or the Sliced Average Variance Estimation (Cook and Weisberg, 1991) may be useful for this purpose. While these methods can be applied in some cases, as pointed out in Xia, Tong, Li and Zhu (2002; see pages 364-365), they are typically not relevant for time series data.

The assumption that the minimal dimension d of TSCS is known may be restrictive. In practice, it will be useful to develop methods to determine a value of d using the data. Also, unlike in regression, where generally p is known, one needs to determine only d of TSCS, here one also has to determine the lag p in $\mathbf{X}_{t-1}$. In Section 2.3.4, we use our estimation method to develop an iterative approach to determine both d and p.

2.3.1 Expected Log-likelihood

Let $p(\cdot,\cdot)$, $p(\cdot|\cdot)$, and $p(\cdot)$ denote joint, conditional, and marginal densities, respectively. For $p \times q$ matrices $\mathbf{h}$ with $q \leq p$, we consider an objective function $\Psi(\mathbf{h})$ defined by

$$\Psi(\mathbf{h}) = \mathrm{E}\left\{\log \frac{p(\mathbf{h}^T\mathbf{X}_{t-1}, x_t)}{p(x_t)p(\mathbf{h}^T\mathbf{X}_{t-1})}\right\} = \mathrm{E}\left\{\log \frac{p(x_t|\mathbf{h}^T\mathbf{X}_{t-1})}{p(x_t)}\right\}, \tag{2.2}$$

which is the mutual information (Cover and Thomas, 1991) between $\mathbf{h}^T\mathbf{X}_{t-1}$ and x_t. Under the assumption that $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\mathbf{\Phi}_d)$ exists, we want to maximize this objective function over

all $p \times d$ matrices $\mathbf{h}$ with $\|\mathbf{h}\| = I$. Since the marginal distribution $p(x_t)$ does not involve $\mathbf{h}$, for subspace $\mathcal{S}(\mathbf{h})$, maximizing $\Psi(\mathbf{h})$ is the same as maximizing the expected log-likelihood. This mutual information can also be thought of as the Kullback-Leibler divergence between the joint density, $p(\mathbf{h}^T\mathbf{X}_{t-1}, x_t)$, and the product of the marginal densities, $p(x_t)p(\mathbf{h}^T\mathbf{X}_{t-1})$, quantifying the dependence of x_t on $\mathbf{h}^T\mathbf{X}_{t-1}$. The following proposition shows that this is a reasonable method for identifying the TSCS.

Proposition 2 *Let $\mathcal{S}(\boldsymbol{h})$ be a DRS for x_t on $\mathbf{X}_{t-1}$ defined in Section 2.2, $\Psi(\boldsymbol{h})$ be as defined in (2.2), and let $\boldsymbol{h}_1$, $\boldsymbol{h}_2$ and $\boldsymbol{h}_d$ be $p \times q_1$, $p \times q_2$ and $p \times d$ matrices, respectively, where $q_1, q_2, d \leq p$. Also assume that $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\boldsymbol{\Phi}_d)$ exists.*

(i) *If $\mathcal{S}(\boldsymbol{h}_1) = \mathcal{S}(\boldsymbol{h}_2)$, then $\Psi(\boldsymbol{h}_1) = \Psi(\boldsymbol{h}_2)$.*

(ii) *$\Psi(I) \geq \Psi(\boldsymbol{h}_1)$ with equality if and only if $x_t \perp\!\!\!\perp \mathbf{X}_{t-1}|\boldsymbol{h}_1^T\mathbf{X}_{t-1}$. Consequently, $\Psi(I) = \Psi(\boldsymbol{\Phi}_d) \geq \Psi(\boldsymbol{h}_d)$ with equality if and only if $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\boldsymbol{\Phi}_d) = \mathcal{S}(\boldsymbol{h}_d)$.*

(iii) *$\Psi(\boldsymbol{h}_1) \geq 0$. Moreover, if $d > q_2 > q_1 \geq 1$, then $\Psi(I) = \Psi(\boldsymbol{\Phi}_d) = \max_{\boldsymbol{h}_d} \Psi(\boldsymbol{h}_d) > \max_{\boldsymbol{h}_2} \Psi(\boldsymbol{h}_2) > \max_{\boldsymbol{h}_1} \Psi(\boldsymbol{h}_1)$.*

Part (i) of the above proposition says that only the $\mathcal{S}(\mathbf{h})$ matters when maximizing $\Psi(\mathbf{h})$ and not the particular basis of the subspace. Therefore, the constraint $\mathbf{h}^T\mathbf{h} = I$, which is used for identifiability, is not at all restrictive. Part (ii) helps us confirm whether $\mathcal{S}(\mathbf{h})$ is a DRS or not (for x_t on $\mathbf{X}_{t-1}$) by comparing $\Psi(\mathbf{h})$ with $\Psi(I)$, if $\Psi(I)$ is known. More importantly, Part (ii) says that $\arg\max_{\mathbf{h}_d} \Psi(\mathbf{h}_d)$ is always a basis for the TSCS. Part (iii) provides a very useful sequential search for the TSCS by showing that information content increases with the dimension until dimension d is achieved. This result will also be useful in Section 2.3.3 below. The proof of Proposition 2 is given in the Appendix.

2.3.2 Computational Algorithm

If all the densities were known, then we could use (2.2) as the basis for a sample version $\Psi_n(\mathbf{h})$ of $\Psi(\mathbf{h})$ and define

$$\Psi_n(\mathbf{h}) = \frac{1}{n}\sum_{t=1}^{n} \log \frac{p(\mathbf{h}^T\mathbf{X}_{t-1}, x_t)}{p(x_t)p(\mathbf{h}^T\mathbf{X}_{t-1})}.$$

Then, we can maximize this sample version over all $p \times d$ matrices $\mathbf{h}$. In practice, however, the densities in $\Psi_n(\mathbf{h})$ are not known. Therefore, we have to estimate them nonparametrically. For this, we need a one-dimensional density estimate for $p(x_t)$ and, for fixed $\mathbf{h}$, we need multi-dimensional density estimates for $p(\mathbf{h}^T\mathbf{X}_{t-1}, x_t)$ and $p(\mathbf{h}^T\mathbf{X}_{t-1})$, respectively. General guidelines for choice of kernels and selection of bandwidths can be found in Silverman (1986) and Scott (1992).

In our computations, we use a density estimate based on a Gaussian kernel for the one-dimensional density, and we use a density estimate based on product Gaussian kernels for each of the two multi–dimensional densities. As observed in Yin and Cook (2005), our experience also confirms that Gaussian kernels work well in this context. More specifically, let G denote the univariate Gaussian kernel, and $\mathbf{u} = (u_1, \cdots, u_k)^T$ be the $k \times 1$ random vector, for $k \geq 1$. Denote the ith observation by $\mathbf{u}_i = (u_{1i}, \cdots, u_{ki})^T$, then the k-dimensional density estimate has the following form:

$$p_n(u_1, \cdots, u_k) = (n\prod_{j=1}^{k} a_{nj})^{-1} \sum_{i=1}^{n}\prod_{j=1}^{k} G\left(\frac{u_j - u_{ji}}{a_{nj}}\right), \tag{2.3}$$

where $a_{nj} = c_k s_j n^{-1/(4+k)}$ for $j = 1, \cdots, k$. s_j is the corresponding sample standard deviation of u_j, which must be updated during iteration. Including s_j in the bandwidth term is not necessary, however, doing so usually improves the estimation. The constant c_k comes from Silverman (1986, p. 87) or Scott (1992, p. 152).

Now, we replace the densities in $\Psi_n(\mathbf{h})$ by their corresponding estimates defined in (2.3) and maximize the function

$$\hat{\mathbf{\Psi}}_n(\mathbf{h}) = \frac{1}{n}\sum_{t=1}^{n} \log \frac{p_n(\mathbf{h}^T\mathbf{X}_{t-1}, x_t)}{p_n(x_t)p_n(\mathbf{h}^T\mathbf{X}_{t-1})}$$

over all $p \times d$ matrices $\mathbf{h}$ satisfying the constraint $\mathbf{h}^T\mathbf{h} = I$. A method that naturally incorporates this constraint is the Sequential Quadratic Programming (SQP) procedure (Gill, Murray, Wright, 1981, Ch.6). The code for our algorithm is available in *MATLAB*, where we use the function '*fmincon*' for maximization, which accommodates the SQP procedure. It should be noted that our estimation method does not require the time series to be stationary. Note that there are also other methods available in the literature for the estimation of density and conditional density. Also, it may be possible to improve our algorithm using a different density estimation method with refined bandwidths, as suggested in Fan, Heckman, and Wand (1995) and Fan, Yao, and Tong (1996).

2.3.3 Estimation of Dimension d and Lag p

Our development of consistent estimation of TSCS assumes that the minimal dimension d is known. In practice, however, prior information on d may not be available. Therefore, it will be useful to develop a data-dependent way to determine d prior to using our consistent estimation method to estimate the TSCS. Also, unlike in regression, here there may not be any prior information available on the number of lags p and hence it will be useful to develop data-dependent methods to determine p. It must be mentioned, however, that in traditional time series analysis one usually uses autocorrelation and partial autocorrelation plots or estimation approaches to determine the lag p.

For dimension reduction in regression, methods have been proposed for the determination of the minimal dimension d of CS. See, for example, Li (1992), Schott (1994), Cook (1998b) and Xia, Tong, Li and Zhu (2002). Here, we use our estimating function $\hat{\mathbf{\Psi}}_n(\mathbf{h})$ defined in Section 2.3.2 and a sequential approach to determine the *best* value of d and p. A graphical method for determining p, which differs from traditional approaches in time series analysis (Ng and Perron, 2005). More specifically, for a given data set, we propose the following iterative process to determine d and p:

Step 1: Note that if $p = 1$, then $d = 1$, and therefore there is no need for dimension reduction. Thus, the procedure starts by fixing a value of lag $p(\geq 2)$ and determines

$$\hat{d}_p = \min\{d(\leq (p-1)) : \hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,(d+1)}) - \hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,d}) \leq \tau_{p,n}\}, \tag{2.4}$$

where $\hat{\mathbf{h}}_{p,k} = \arg\max_{\mathbf{h}_k} \hat{\mathbf{\Psi}}_n(\mathbf{h}_k)$ and the maximization is over all $p \times k$ matrices $\mathbf{h}_k$, and $\{\tau_{p,n}; n \geq 1\}$ is a sequence of non-negative threshold values chosen in such a way that it converges to zero as $n \to \infty$. In our simulations and data analysis discussed in Section 2.4, we set the threshold value $\tau_{p,n} = 0$ and $\chi^2_p(\alpha)/(2n)$, where $\chi^2_p(\alpha)$ is the $100(1-\alpha)$ percentile of Chi-square distribution with p degrees of freedom.

The procedure in (2.4) is clearly iterative, which, at each stage, compares (successive) differences $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,(d+1)}) - \hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,d})$ (> 0 because of Proposition 2(iii)) with the threshold value 0 or $\chi^2_p(\alpha)/(2n)$, for a pre-specified value $\alpha > 0$, and stops at the first value of d for which the difference is below the threshold. For each $p \geq 2$, this yields an estimate $\hat{d}_p$ of d, which in turn yields an estimate $\hat{\mathbf{h}}_{p,\hat{d}_p}$ of TSCS with the maximum value $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,\hat{d}_p})$. Obviously, if the difference $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,(d+1)}) - \hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,d})$ never falls below the threshold, then $\hat{d}_p = p$.

Step 2: Repeat the process in Step 1 for each $p = 2, 3, \cdots$. This process will yield a finite sequence of estimates $\{\hat{d}_p\}$ and corresponding sequence of maximum values $\{\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,\hat{d}_p})\}$. Now plot $\{\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,\hat{d}_p})\}$ versus p, which we call the *Shoulder Plot*. In such a plot, we will look for the value of $\hat{p}$ at which $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{\hat{p},\hat{d}_{\hat{p}}})$ is essentially the largest; that is, the subsequent values of $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,\hat{d}_p})$ are about the same or less than $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{\hat{p},\hat{d}_{\hat{p}}})$. This creates a shoulder-like situation at $p = \hat{p}$, hence the name *Shoulder Plot*. This determines an estimate $\hat{p}$ of lag p.

Note that Steps 1 and 2 yield estimates $\hat{p}$ and $\hat{d}_{\hat{p}}$. The idea behind a *Shoulder Plot* is similar to an *Elbow Plot*, which plots (decreasing) eigenvalues against the serial numbers in Principal Component Analysis (PCA) in order to determine the least number of PCs to use. In Section 2.4, we use Steps 1 and 2 to determine $\hat{d}$ and $\hat{p}$ for our simulation data sets, the Wolf yearly sunspot numbers and U.S. beer production data. In fact, we show the usefulness of *Shoulder Plot* in determining $\hat{p}$ for our simulation Example 1 and for Wolf yearly sunspot

numbers and U.S. beer production data. In Section 2.4, we state a Theorem establishing the consistency of $\hat{d}_p$ defined in (2.4).

The choice of threshold value $\tau_{p,n}$ in (2.4) is critical for the computation of $\hat{d}_p$ in numerical studies. It can be easily seen from (2.4) that threshold values directly impact $\hat{d}_p$ values, which increase when $\tau_{p,n}$ values decrease. In our numerical studies (see Section 2.4), we found the threshold values $\tau_{p,n} = 0$ and $\chi^2_p(\alpha)/(2n)$ to be adequate.

2.3.4 Consistency Theorem

In this section, we state two theorems establishing the consistency of the estimate of TSCS and d, respectively. Unlike in Section 2.3.2, here we do not restrict to Gaussian kernels. Suppose $\mathbf{M}_i$ is a sequence of k-dimensional random variables with Lebesgue density p and distribution function F. We define the following kernel density estimator of p

$$f_n(\mathbf{M}) = \frac{1}{na_n^k} \sum_{i=1}^{n} K\left(\frac{\mathbf{M} - \mathbf{M}_i}{a_n}\right)$$

for $\mathbf{M} \in \mathbb{R}^k$, where $K : \mathbb{R}^k \to \mathbb{R}_+$ is a probability density, $\lim K(\mathbf{M}) = 0$ uniformly for $||\mathbf{M}|| \to \infty$, $a_n > 0$, and $\lim_{n\to\infty} a_n = 0$. Let $\kappa_\iota = \{t : p(x_t) > \iota, p(\mathbf{h}^T\mathbf{X}_{t-1}) > \iota, p(\mathbf{h}^T\mathbf{X}_{t-1}, x_t) > \iota\}$ for any fixed $p \times d$ matrix $\mathbf{h}$ such that $\mathbf{h}^T\mathbf{h} = I$ and for some ι to be chosen in the following way: Let $\epsilon \to 0$, and $\iota \to 0$, but $\frac{\epsilon}{\iota} \to 0$ as $n \to \infty$ for $\epsilon > 0$ and $\iota > 0$. Let n_ι be the number of observations whose indices are not in κ_ι. Theorems 1 and 2 stated below are proved in the Appendix (see Section 2.6).

Theorem 1 *Assume the conditions of Lemma 1 stated in the Appendix and that* $\frac{n_\iota}{n} \to 0$ *in probability as* $n \to \infty$. *Let* $\hat{\mathbf{\Phi}}_n = \arg\max_{\boldsymbol{h}} \hat{\mathbf{\Psi}}^\iota_n(\boldsymbol{h})$ *and* $\mathbf{\Phi}_d = \arg\max_{\boldsymbol{h}} \Psi(\boldsymbol{h})$, *where*

$$\hat{\mathbf{\Psi}}^\iota_n(\boldsymbol{h}) = \frac{1}{n} \sum_{t=1}^{n} J(t \in \kappa_\iota) \log \frac{f_n(\boldsymbol{h}^T\mathbf{X}_{t-1}, x_t)}{f_n(x_t) f_n(\boldsymbol{h}^T\mathbf{X}_{t-1})},$$

$J(t \in \kappa_\iota)$ *denotes the indicator function for* κ_ι, $\Psi(\boldsymbol{h})$ *is as defined in (2.2) and the maximization is over all* $p \times d$ *matrices* $\boldsymbol{h}$ *such that* $\boldsymbol{h}^T\boldsymbol{h} = I$. *Then* $\hat{\mathbf{\Phi}}_n$ *converges to* $\mathbf{\Phi}_d$ *with probability one, as* $n \to \infty$.

Theorem 2 *Assume the conditions of Lemma 1 and Lemma 2 in the Appendix, and Theorem 1. Let* $\hat{d}_p^{\iota} = \min\{k(\leq (p-1)) : \hat{c}_k^{\iota} \leq \tau_{p,n}\}$, *where* $\hat{c}_k^{\iota}$ *is same as* $\hat{c}_k = \hat{\mathbf{\Psi}}_n(\hat{\boldsymbol{h}}_{p,(d+1)}) - \hat{\mathbf{\Psi}}_n(\hat{\boldsymbol{h}}_{p,d})$ *defined (2.4) with* $\hat{\mathbf{\Psi}}_n$ *replaced by* $\hat{\mathbf{\Psi}}_n^{\iota}$ *defined in Theorem 1. If for each fixed* p, $\tau_{p,n} \to 0$ *as* $n \to \infty$, *then* $\hat{d}_p^{\iota}$ *converges to* d *with probability one, as* $n \to \infty$, *where* d *is the dimension of TSCS.*

2.4 Simulations and Data Analysis

In this section, we investigate the performance of our estimation methods proposed in Sections 2.3.2 and 2.3.3 for various simulated data sets and well-known real data on Wolf yearly sunspot numbers, U.S. GNP, and U.S. beer production data (Wei, 2006).

2.4.1 Simulations

In order to assess the accuracy of our estimates of TSCS in our simulations, we use the measures proposed by Ye and Weiss (2003) and Xia, Tong, Li and Zhu (2002). Both of these methods assess the accuracy of estimates by measuring the distance between the estimated TSCS, $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\hat{\mathbf{\Phi}}_d)$, and the TSCS, $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\mathbf{\Phi}_d)$. More precisely, Ye and Weiss (2003) measure the distance using the so called *vector correlation coefficient* (Hotelling, 1936) defined by $\rho = \sqrt{|\hat{\mathbf{\Phi}}_d^T \mathbf{\Phi}_d \mathbf{\Phi}_d^T \hat{\mathbf{\Phi}}_d|}$, where $|\mathbf{A}|$ denotes the determinant of a matrix $\mathbf{A}$. Note that $0 \leq \rho \leq 1$, and when $\rho = 1$, $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\hat{\mathbf{\Phi}}_d) = \mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\mathbf{\Phi}_d)$. Therefore, higher values of ρ imply that the two DRSs are closer, and hence, the estimates are more accurate. As for the measure defined by Xia, Tong, Li and Zhu (2002), let $\mathcal{S}(\hat{\mathbf{\Phi}}_q)$ denote the (estimated) DRS spanned by the columns of $p \times q$ matrix $\hat{\mathbf{\Phi}}_q$. The distance between $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\hat{\mathbf{\Phi}}_q)$ and $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\mathbf{\Phi}_d)$ can be measured by $m^2 = \left\|(I - \mathbf{\Phi}_d {\mathbf{\Phi}_d}^T)\hat{\mathbf{\Phi}}_q\right\|^2$ if $q < d$ and $m^2 = \left\|(I - \hat{\mathbf{\Phi}}_q \hat{\mathbf{\Phi}}_q^T)\mathbf{\Phi}_d\right\|^2$ if $q \geq d$. Here, smaller values of m^2 yield more accurate estimates.

We begin with two illustrative examples, where our only focus is accuracy (measured by the above distances) of estimates obtained using our estimation method. In each of our simulation study considered below, we specify either a linear or a nonlinear time series model.

n	ρ	m^2
100	0.8663	0.0188
200	0.9434	0.0047
300	0.9515	0.0024

Table 2.1: AR(2) model: Average values of accuracy measures ρ and m^2 based on 200 Monte Carlo replications.

We implement our computational algorithm described in Section 2.3.2 for samples of size n = 100, 200 and 300 drawn from each of these models. For each sample size, we perform 200 Monte Carlo replications of our algorithm, each yielding an estimate of $\mathbf{\Phi}_d$, for a specified value of p and d.

First, we consider a linear autoregressive model of order 2, AR(2), given by

$$x_t = 0.3x_{t-1} + 0.3x_{t-2} + \varepsilon_t,$$

where $p = 2$ and $d = 1$, and $\{\varepsilon_t\}$ is a sequence of independent standard normal random variables. Here, our interest is to estimate $\mathbf{\Phi}_1$. Table 2.1 gives the average values of accuracy measures ρ and m^2, respectively, based on 200 Monte Carlo replications for each sample size. Table 2.1 shows that the average values of ρ are in general close to 1, while those of m^2 are close to zero, implying that our estimates of $\mathbf{\Phi}_1$ are very accurate. Notice that the accuracy of our estimates, as measured in terms of ρ and m^2, increases with sample size, as expected.

Next, we consider the nonlinear time series model (Model 2.2) given by

$$\begin{aligned} x_t &= -1 - \cos((\pi/2)(x_{t-1} + 2x_{t-4})) \\ &\quad +0.2\varepsilon_t \exp(-(-2x_{t-1} + 2x_{t-2} - 2x_{t-3} + x_{t-4} - x_{t-5} + x_{t-6})^2), \end{aligned}$$

where $p = 6$, $d = 2$. This model is considerably more complicated than the AR(2) model above in that, in addition to a nonlinear mean function, the error term ε_t is also multiplied by a nonlinear function depending on the past of the series. Here, our interest is to estimate $\mathbf{\Phi}_2$. While we set the value of $d = 2$, we set the lag value $p = 6$ and 10, in order to study

n	lag p	ρ	m^2
100	10	0.7612	0.0922, 0.3613
100	6	0.8487	0.0540, 0.2523
200	10	0.7793	0.0849, 0.3374
200	6	0.8440	0.0654, 0.2547
300	10	0.7841	0.0853, 0.3267
300	6	0.8846	0.0630, 0.1810

Table 2.2: Model 2.2: Average values of accuracy measures ρ and m^2 based on 200 Monte Carlo replications.

the effect of using a wrong lag value of $p = 10$ on the accuracy of estimates. Table 2.2 gives the average values of ρ and m^2, respectively, based on 200 Monte Carlo replications for each sample size and lag value. Note that Table 2.2 gives two different m^2 values in each cell, which correspond to estimates $\hat{\Phi}_1$ and $\hat{\Phi}_2$, respectively. Once again, we see from Table 2.2 that the average values of ρ are close to 1 and m^2 values are close to zero, implying that our estimates of $\mathbf{\Phi}_2$ are reasonably accurate. Notice once again that the accuracy of our estimates, as measured in terms of ρ and m^2, increases with sample size. Also, accuracy of the estimates is better when correct lag is used in the estimation as opposed to using a wrong lag. Finally, accuracy of estimates for the linear time series model AR(2) given in Table 2.1 is in general better than those for the nonlinear time series model given in Table 2.2. Nevertheless, the accuracy results in Table 2.2 attests to the fact that our estimation procedure can perform reasonably well even when the conditional mean and the variance functions are nonlinear.

Next three examples deal with nonlinear time series models, where $p = 6$, but the number of dimensions vary from 1 to 3, that is, $d = 1, 2, 3$, respectively. The main aim here is to numerically investigate the performance of $\hat{d}_p$ and the *Shoulder Plot* proposed in Section 2.3.3 for inference about d and p, respectively. Therefore, in these examples, we not only

assess the accuracy of our estimation method but also use our iterative procedure described in Steps 1 and 2 of Section 2.3.3 to determine d and p.

Example 1:

Consider the model (Model 2.3)

$$x_t = -1 - \cos((\pi/2)(x_{t-3} + 2x_{t-6})) + 0.2\varepsilon_t,$$

where $p = 6$, $d = 1$ and $\{\varepsilon_t\}$ is a sequence of independent standard normal random variables. Here, our interest is to estimate $\mathbf{\Phi}_1$. Once again, Table 2.3 shows that the average values of ρ are very close to 1 and m^2 values are very close to zero, regardless of whether the lag p is 6 or 10. Here, the moderate sample performances are as good as those for the large sample ones. These imply that our estimates of $\mathbf{\Phi}_1$ are very accurate.

The above estimation was carried out after specifying $p = 6, 10$ and $d = 1$. Suppose we want to make inference about p and d based on samples drawn from Model 2.3, where the true dimension and lag are $d = 1$ and $p = 6$, respectively. For each sample size and lag $p = 6, 10$, we performed 200 Monte Carlo replications and, for each replication, we computed an estimate $\hat{d}_p$ defined in (2.4) using the threshold values $\tau_{p,n} = 0$, $\chi^2_p(.05)/(2n)$ and $\chi^2_p(.01)/(2n)$. From now on, we will refer to these three threshold values as "0-threshold", "0.05-threshold" and "0.01-threshold", respectively. We then tallied the estimated number of dimensions (out of 200 replications). These counts are reported in the last three columns of Table 2.3 for each sample size, lag value and threshold value, where f_i denotes the frequency of $\hat{d}_p = i$ and f_{i+} denotes the frequency of $\hat{d}_p \geq i$, out of 200 replications.

Table 2.3 shows that, for sample sizes $n = 100$ and 200, our estimator $\hat{d}_p$ using the 0.05-threshold and 0.01-threshold correctly estimates the true dimension, $d = 1$, a substantially higher percentage of times than the estimator using the 0-threshold. It is interesting to note that $\hat{d}_p$ using the 0.05-threshold and 0.01 threshold continues to perform well even when the lag is wrongly specified ($p = 10$). It is clear from Table 2.3 that, for $n = 100$ and 200, the performance of the estimator $\hat{d}_p$ using 0-threshold does not seem consistent with regard

n	lag p	ρ	m^2	0-threshold	0.05-threshold	0.01-threshold
100	10	0.9965	0.0069	f_1=114 f_{2+}=86	f_1=180 f_{2+}=20	f_1=197 f_{2+}=3
100	6	0.9985	0.0030	f_1 = 46 f_{2+}=154	f_1=174 f_{2+}=26	f_1=196 f_{2+}=4
200	10	0.9989	0.0021	f_1=94 f_{2+}=106	f_1=191 f_{2+}=9	f_1=197 f_{2+}=3
200	6	0.9995	0.0010	f_1=115 f_{2+}=85	f_1=194 f_{2+}=6	f_1=199 f_{2+}=1
300	10	0.9994	0.0011	f_1=165 f_{2+}=35	f_1=199 f_{2+}=1	f_1=200 f_{2+}=0
300	6	0.9997	0.00005	f_1=184 f_{2+}=16	f_1=200 f_{2+}=0	f_1=200 f_{2+}=0

Table 2.3: Model 2.3: Average values of accuracy measures ρ and m^2, and frequency of estimated dimension for 0-threshold, 0.05-threshold and 0.01-threshold, all based on 200 Monte Carlo replications. The true dimension is $d = 1$.

to identifying the true dimension correctly. When the sample size $n = 300$, however, our estimator $\hat{d}_p$ correctly estimates the true dimension a large percentage of times for all the three thresholds, with near-perfect or perfect performance by 0.05- and 0.01- threshold.

Next, we determine $\hat{p}$ using a *Shoulder Plot* described in Step 2 of Section 2.3.3. Here, we set $n = 300$, $p = 4, 5, 6$ and 7, and computed the finite sequence $\{\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,1})\}$ for 20 simulated data sets from Model 2.3. Using these values, we computed the average and the standard deviation of $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,1})$ values for each $p = 4, 5, 6, 7$. In Figure 2.1, we give a *Shoulder Plot* using the average values and average $\pm$ standard deviation values of $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,1})$. Clearly, Figure 2.1 indicates that the *Shoulder* is at $p = 6$. In fact, *Shoulder Plots* for 18 out of 20 simulated data sets (not given here) indicated that $\hat{p} = 6$. Similar results (not reported here) were also observed for $n = 100, 200$.

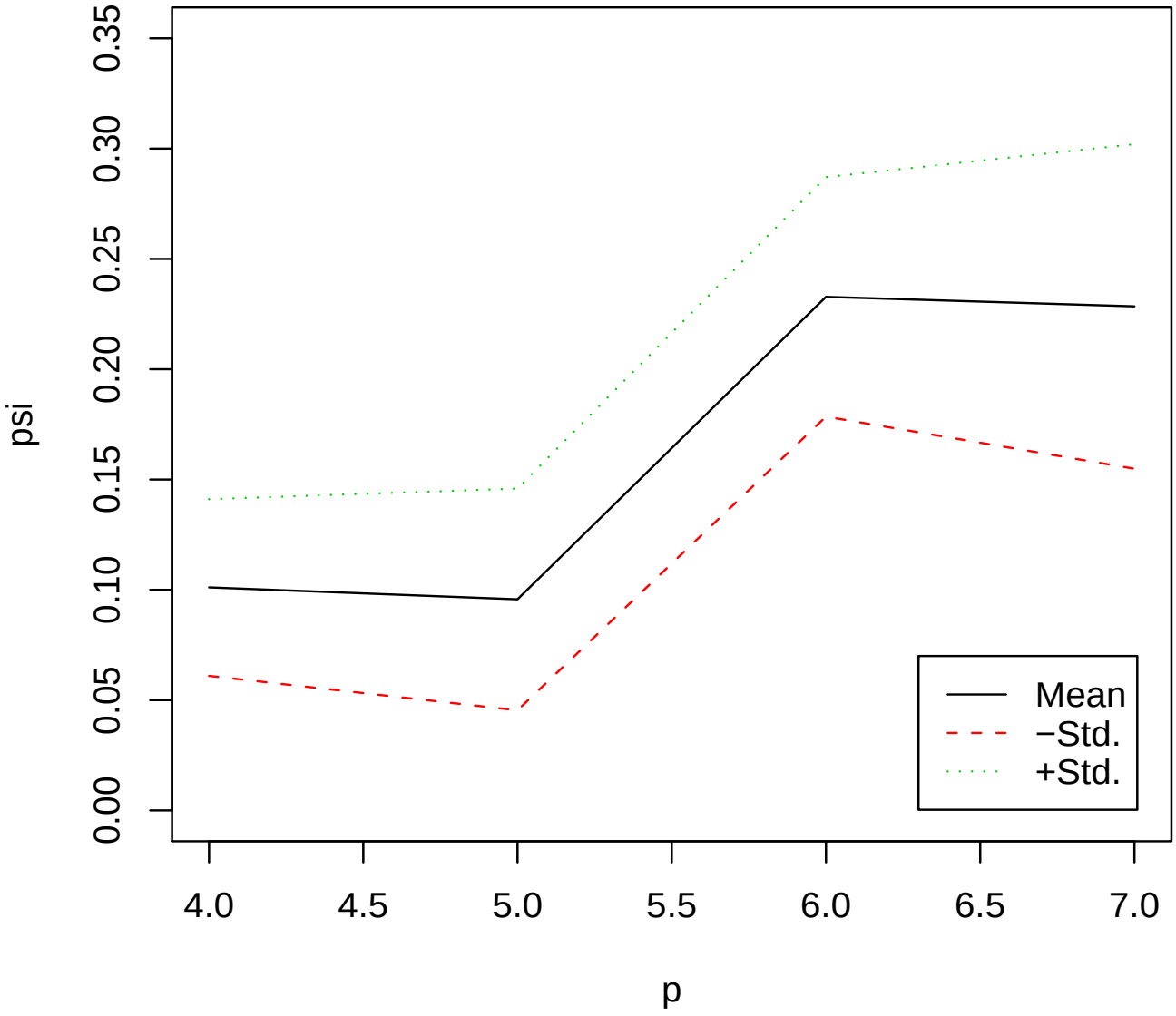

Figure 2.1: Example 1(Model 2.3): *Shoulder plot* of average values ("Mean"), average − standard deviation values ("-Std.") and average + standard deviation values ("+Std.") of $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,1})$ versus $p = 4, 5, 6, 7$ based on 20 simulated data sets, each with sample size $n = 300$.

n	lag p	ρ	m^2	0-threshold	0.05-threshold	0.01-threshold
100	10	0.9006	0.0977	$f_1 = 49$	f_1=137	f_1=155
			0.0954	f_2=67	f_2=54	f_2=45
				f_{3+}=84	f_{3+}=9	f_{3+}=0
100	6	0.9655	0.0366	f_1=26	f_1=41	f_1=50
			0.0314	f_2=52	f_2=148	f_2=145
				f_{3+}=122	f_{3+}=11	f_{3+}=5
200	10	0.9734	0.0288	f_1=60	f_1=85	f_1=113
			0.0241	f_2=80	f_2=83	f_2=71
				f_{3+}=60	f_{3+}=32	f_{3+}=16
200	6	0.9878	0.0137	f_1=23	f_1=34	f_1=37
			0.0106	f_2=82	f_2=147	f_2=149
				f_{3+}=95	f_{3+}=23	f_{3+}=17
300	10	0.9835	0.0177	f_1=58	f_1=71	f_1=87
			0.0151	f_2=85	f_2=91	f_2=84
				f_{3+}=57	f_{3+}=38	f_{3+}=29
300	6	0.9923	0.0090	f_1=14	f_1=21	f_1=21
			0.0063	f_2=93	f_2=156	f_2=159
				f_{3+}=93	f_{3+}=23	f_{3+}=20

Table 2.4: Model 2.4: Average values of accuracy measures ρ and m^2, and frequency of estimated dimension for 0-threshold, 0.05-threshold and 0.01-threshold, all based on 200 Monte Carlo replications. The true dimension is $d = 2$.

Example 2:

Consider the model (Model 2.4)

$$x_t|\mathbf{X}_{t-1} = -1 - \cos((\pi/2)(x_{t-1})) - \cos((\pi/2)(1/\sqrt{2})(x_{t-3} + 2x_{t-6})) + 0.2\varepsilon_t,$$

where $p = 6$, $d = 2$ and $\{\varepsilon_t\}$ is a sequence of independent standard normal random variables. Table 2.4 shows that for all sample sizes the average values of ρ are in general very close to 1 and m^2 are very close to zero, regardless of whether the lag p is 6 or 10. These imply that our estimates of $\mathbf{\Phi}_2$ are very accurate.

As for inference about d, Table 2.4 shows that, for all sample sizes and when the lag is correctly specified ($p = 6$), our estimator $\hat{d}_p$ using the 0.05-threshold and 0.01-threshold correctly estimates the true dimension, $d = 2$, about 73% to 80% of the times. However,

n	lag p	ρ	m^2	0-threshold	0.05-threshold	0.01-threshold
100	10	0.6505	0.2531	f_1=20	f_1=41	f_1=62
			0.1768	f_2=105	f_2=155	f_2=138
			0.2348	f_3=63	f_3=1	f_3=0
				f_{4+}=12	f_{4+}=0	f_{4+}=0
100	6	0.8519	0.1091	f_1=2	f_1=10	f_1=28
			0.0552	f_2=67	f_2=187	f_2=172
			0.1040	f_3=102	f_3=3	f_3=0
				f_{4+}=29	f_{4+}=0	f_{4+}=0
200	10	0.7936	0.1397	f_1=11	f_1=14	f_1=17
			0.0980	f_2=125	f_2=183	f_2=183
			0.1501	f_3=63	f_3=3	f_3=0
				f_{4+}=1	f_{4+}=0	f_{4+}=0
200	6	0.9299	0.0669	f_1=32	f_1=45	f_1=49
			0.0212	f_2=33	f_2=114	f_2=134
			0.0433	f_3=135	f_3=41	f_3=17
				f_{4+}=0	f_{4+}=0	f_{4+}=0
300	10	0.8631	0.0936	f_1=9	f_1=11	f_1=12
			0.0755	f_2=110	f_2=179	f_2=184
			0.0907	f_3=79	f_3=10	f_3=4
				f_{4+}=2	f_{4+}=0	f_{4+}=0
300	6	0.9637	0.0365	f_1=1	f_1=1	f_1=1
			0.0110	f_2=67	f_2=159	f_2=180
			0.0231	f_3=132	f_3=40	f_3=19
				f_{4+}=0	f_{4+}=0	f_{4+}=0

Table 2.5: Model 2.5: Average values of accuracy measures ρ and m^2, and frequency of estimated dimension for 0-threshold, 0.05-threshold and 0.01-threshold, all based on 200 Monte Carlo replications. The true dimension is $d = 3$.

Table 2.4 also shows that wrong specification of lag adversely affects the performance of $\hat{d}_p$, resulting in severe underestimation for 0.05- and 0.01-threshold. On the other hand, the 0-threshold does not perform well at all and, in fact, considerably overestimates the true dimension.

Example 3.

The next model is same as the one considered in Section 4.3, Example 3 of Xia, Tong, Li and Zhu (2002), who compare several estimation methods using m^2 distance defined above and estimate the dimension d using a method different from the one proposed here; see Table 2 and Section 2.2 of Xia, Tong, Li and Zhu (2002) for details. More specifically, we consider the nonlinear time series model (Model 2.5)

$$\begin{aligned} x_t &= -1 + (0.4)(1/\sqrt{5})(x_{t-1} + 2x_{t-4}) - \cos((\pi/2)(1/\sqrt{5})(x_{t-3} + 2x_{t-6})) \\ &\quad + \exp(-(1/\sqrt{15})^2(-2x_{t-1} + 2x_{t-2} - 2x_{t-3} + x_{t-4} - x_{t-5} + x_{t-6})^2) + 0.2\varepsilon_t \end{aligned}$$

where $p = 6$, $d = 3$ and $\{\varepsilon_t\}$ is a sequence of independent standard normal random variables.

Table 2.5 shows that, for large sample sizes and correct lag specification, accuracy of estimates of $\mathbf{\Phi}_3$ are better, as indicated by the large values of ρ and small values of m^2. Moreover, comparison of our m^2 values in Table 2.5 with those in Table 2 of Xia, Tong, Li and Zhu (2002) shows that the RMAVE method of Xia, Tong, Li and Zhu (2002) performs better than our estimation method. However, this is to be expected because their RMAVE method focuses on estimation of dimensions in the mean function and the nature of Model 2.5 helps their focus. On the other hand, their RMAVE method may not perform well for Model 2.2 above, where the error term is multiplied by a nonlinear function depending on the past of the series.

As for inference about d, Table 2.5 shows that, for all sample sizes and when the lag is correctly specified ($p = 6$), our estimator $\hat{d}_p$ using the 0-threshold correctly estimates the true dimension, $d = 3$, about 51% to 68% of the times. Here, it is important to point out that the performance of $\hat{d}_p$ using 0-threshold in correctly estimating the true dimension is slightly better than that of Xia, Tong, Li and Zhu's (2002) method for $n = 100$, but the latter method performs better than ours for $n = 200$ and 300, as shown in their Table 2. Table 2.5 also shows that wrong specification of lag adversely affects the performance of $\hat{d}_p$ resulting in severe underestimation for all the three thresholds. On the other hand, even

when the lag is correctly specified, the 0.05- and 0.01-threshold do not perform well at all for any sample size.

The last three examples seem to suggest that when $d = 1$ or 2, 0.05-threshold performs better, but when $d = 3$ the 0-threshold performs better. Thus, generally, we should use 0.05-threshold to estimate d, when we prefer smaller dimension; otherwise, use 0-threshold.

2.4.2 Wolf Yearly Sunspot Data

In the previous section we studied the performance of our estimation approach when data is drawn from a variety of nonlinear time series models. In this section, we employ our methods to analyze a real data. For this purpose, we revisit a classic time series known as *Wolf yearly sunspot numbers* for the years 1700 to 2001, giving a total 302 observations. Many scientists believe that sunspot numbers influence the weather on the Earth, which in turn impacts activities such as agriculture and telecommunications. This data has been extensively studied by many authors in time series literature and has been analyzed using various linear and nonlinear models; see, for example, Yule (1927), Bartlett (1950), Whittle (1954), Brillinger and Rosenblatt (1967), and Xia, Tong and Li (1999). More details about fitting a linear time series model to this data can be found in Example 6.2 of Wei (2006).

As discussed in Example 6.2 of Wei (2006), the sunspot series is stationary in the mean but not stationary in the variance. Using power transformation analysis it can be shown that a square root transformation applied to the data stabilizes its variance. Let z_t denote the sunspot number at time t and $x_t = \sqrt{z_t}$. Since the sample autocorrelation function x_t shows a sine-cosine wave and the sample partial autocorrelation function has relatively large spikes at lags 1, 2 and 9, Wei (2006) fits the following three models to x_t: (i) AR(2) model, (ii) AR(9) model, and (iii) AR(1,2,9) model, which depends only on lags 1, 2 and 9. See Series W2, Table 7.3 in Wei (2006) for more details on the estimates corresponding these models and estimates of error variance. Based on estimated error variance, Wei (2006) concludes that both AR(9) and AR(1,2,9) models are adequate for the data. Later, we will

p	d=1	d=2	d=3
2	0.7229	0.7533*	N/A
3	0.7259	0.7578*	0.7185
4	0.7258	0.7593*	0.7581
5	0.7255	0.7682*	0.7768
6	0.7377	0.7744*	0.7825
7	0.7590	0.7999*	0.8049
8	0.7634	0.8011*	0.8069
9	0.7846	0.8177*	0.8197
10	0.7843*	0.8146	0.8235
11	0.7837	0.8186*	0.8256
12	0.7818	0.8201*	0.8153

Table 2.6: Wolf yearly sunspot data: $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,d})$ values for $p = 2, \cdots, 12$, and $d = 1, 2$ and 3. For each p, $\hat{d}_p$ determined by 0.05-threshold is denoted by * in the table.

compute the forecasts from Wei's models (ii) and (iii) and use these as benchmarks to assess the performance of the model that we are about to propose. For model building we use the sunspot numbers for the years 1700 to 1991, yielding a sample of size $n = 292$. We then compute our model-based forecast of the sunspot numbers for the remaining years 1992 to 2001. The observed sunspot numbers for the years 1992 to 2001 will be used to assess all forecasts.

We begin the process of model building for the data (1700-1991) by first identifying the dimension d. For this, we set $p = 2, \cdots, 12$ and computed the $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,d})$ values for $d = 1, 2$ and 3 (see Step 1 of Section 2.3.3). Table 2.6 lists the $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,d})$ values for each p and d, except for the trivial case $p = 1$. Using (2.4) with 0.05-threshold and the values in Table 2.6 we obtained $\hat{d}_p = 2$ for all $2 \leq p \leq 12$, except when $p = 10$, for which $\hat{d}_p = 1$. Note that $\hat{d}_p$ for each p is indicated by an asterisk in Table 2.6. Since $\hat{d}_p = 2$ uniformly for all p, except for $p = 10$, we decided to estimate the unknown dimension $d = 2$.

Next, as discussed in Step 2 of Section 2.3.3, we created a *Shoulder Plot* using $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,2})$ values for $p = 2, \cdots, 12$ (see second column in Table 2.6). This *Shoulder Plot* is given in

Figure 2.2. It is clear from Figure 2.2 that the *Shoulder* is at $p = 9$. Our determination of $p = 9$ agrees with other approaches; for example, *R* software with command '*sunspot.ar <- ar(sunspot.year)*' determines lag $p = 9$ using the Akaike Information Criterion (AIC).

Having decided that $d = 2$ and $p = 9$, we now proceed to estimate the 9×2 matrix $\mathbf{\Phi}_2 = (\Phi_1, \Phi_2)$, whose columns form the basis of TSCS, $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\mathbf{\Phi}_2)$. We used our estimation method described in Section 2.3.2 and obtained the estimates $\hat{\Phi}_1$ and $\hat{\Phi}_2$. This enables us to work with a much smaller 2-dimensional vector $(\hat{\Phi}_1^T\mathbf{X}_{t-1}, \hat{\Phi}_2^T\mathbf{X}_{t-1})$ in predicting x_t instead of the 9-dimensional vector $\mathbf{X}_{t-1}$.

Next, we used trial-and-error approaches with graphical tools and plots of x_t against $\hat{\Phi}_1^T\mathbf{X}_{t-1}$ and $\hat{\Phi}_2^T\mathbf{X}_{t-1}$ to determine a time series model for the data. Nonlinear patterns of these plots containing cycles suggested that cosine functions are reasonable approximations. Moreover, we also determined that some of the estimated coefficients in the cosine functions are close to $\frac{\pi}{4}$, $\frac{\pi}{2}$, and π. In the end, all this leads us to the final nonlinear time series model given by

$$\begin{aligned} x_t &= 0.48 - 1.25\hat{\Phi}_1^T\mathbf{X}_{t-1} + 0.74\hat{\Phi}_2^T\mathbf{X}_{t-1} + 0.50\cos\left(\frac{\pi}{2}\hat{\Phi}_1^T\mathbf{X}_{t-1} + \pi\right) \\ &+ 0.25\cos\left(\frac{\pi}{2}\hat{\Phi}_2^T\mathbf{X}_{t-1} - \frac{\pi}{4}\right) + \epsilon_t. \end{aligned}$$

In order to assess the performance of our nonlinear time series model above, we computed model-based forecasts of sunspot numbers for the remaining years 1992 to 2001 using our model, and AR(9) and AR(1,2,9) models fitted in Wei (2006). These forecasts along with observed sunspot numbers for the years 1992 to 2001 are given in Table 2.7.

An overlay plot of forecast values from each of these models and the observed sunspot numbers is given in Figure 2.3. Moreover, for each of the three models, we computed the Mean Square Relative Error (MSRE)$= k^{-1}\sum_{t=1}^{k}(z_t - \hat{z}_t)^2/z_t$, where z_t is the observed sunspot number, $\hat{z}_t$ is its forecast value and k is the number of (future) observations. For $k = 10$, we give the MSRE values for each of the three models in Table 2.7. Note that our model produces an MSRE value which is almost half of those of Wei's models. Through this real data we have shown that our dimension reduction approach together with the estimation

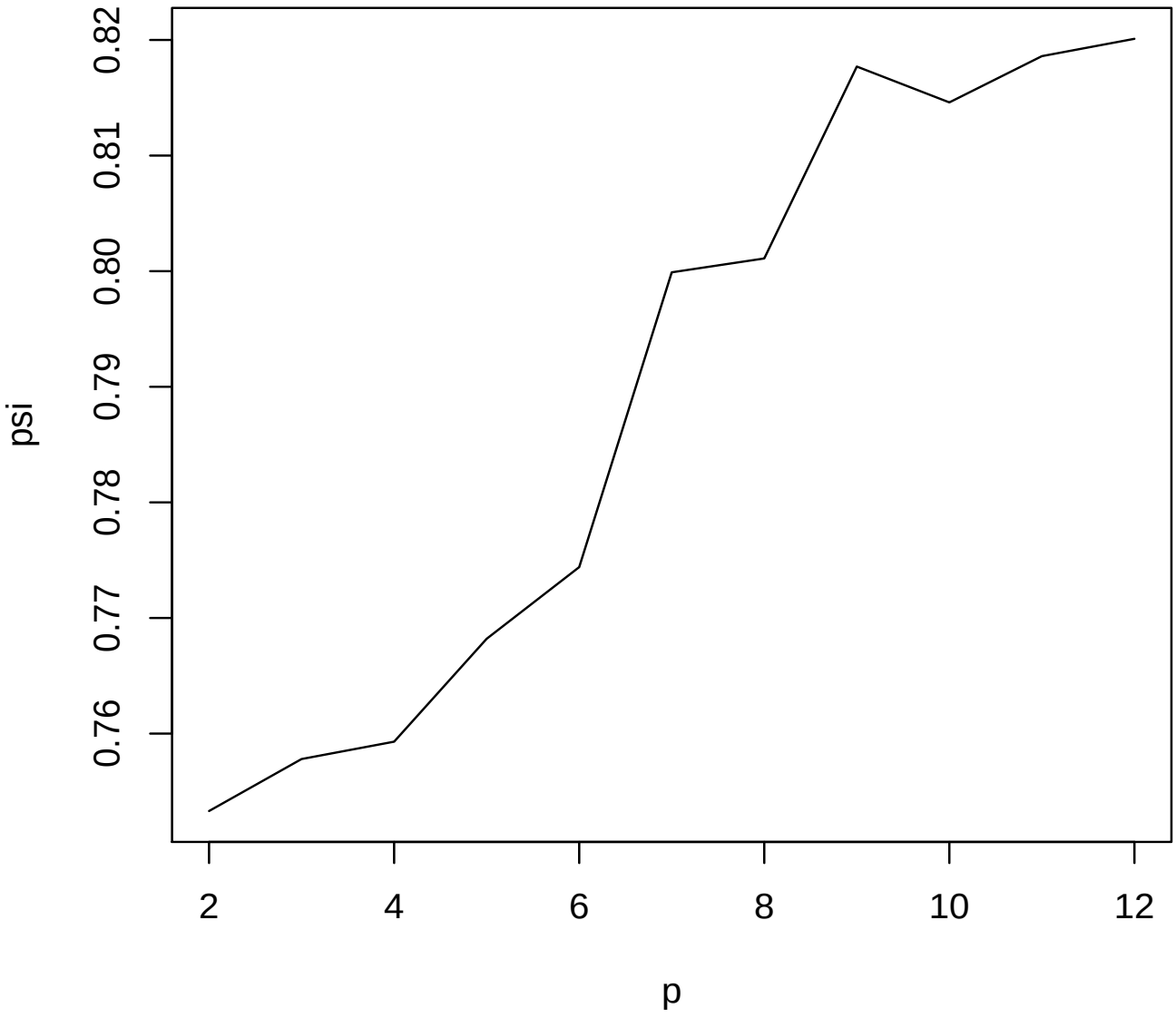

Figure 2.2: Wolf yearly sunspot data: *Shoulder plot* of $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,2})$ values (second column of Table 2.6) versus $p = 2, \cdots, 12$.

Year	Sunspot Number	AR(9)	AR(1,2,9)	Our model
1992	94.2996	121.8850	119.2840	102.1030
1993	54.6003	88.6637	83.9347	71.6920
1994	29.9001	46.9641	49.8700	42.4043
1995	17.5000	26.3525	27.6294	23.5257
1996	8.6001	17.5277	19.8872	19.7202
1997	21.4999	22.5212	28.3990	23.8431
1998	64.2995	45.4616	53.5825	54.7348
1999	93.3001	76.9934	84.6827	92.5245
2000	119.6000	106.5490	110.9600	119.0450
2001	111.0010	118.7780	120.6190	118.2990
MSRE		6.3191	6.2648	3.1838

Table 2.7: Observed sunspot numbers, forecasts from AR(9), AR(1,2,9) and our model, and MSRE for each model: Years 1992 - 2001.

method and graphical techniques lead us to a time series model which outperforms Wei's (2006) models for the sunspot data.

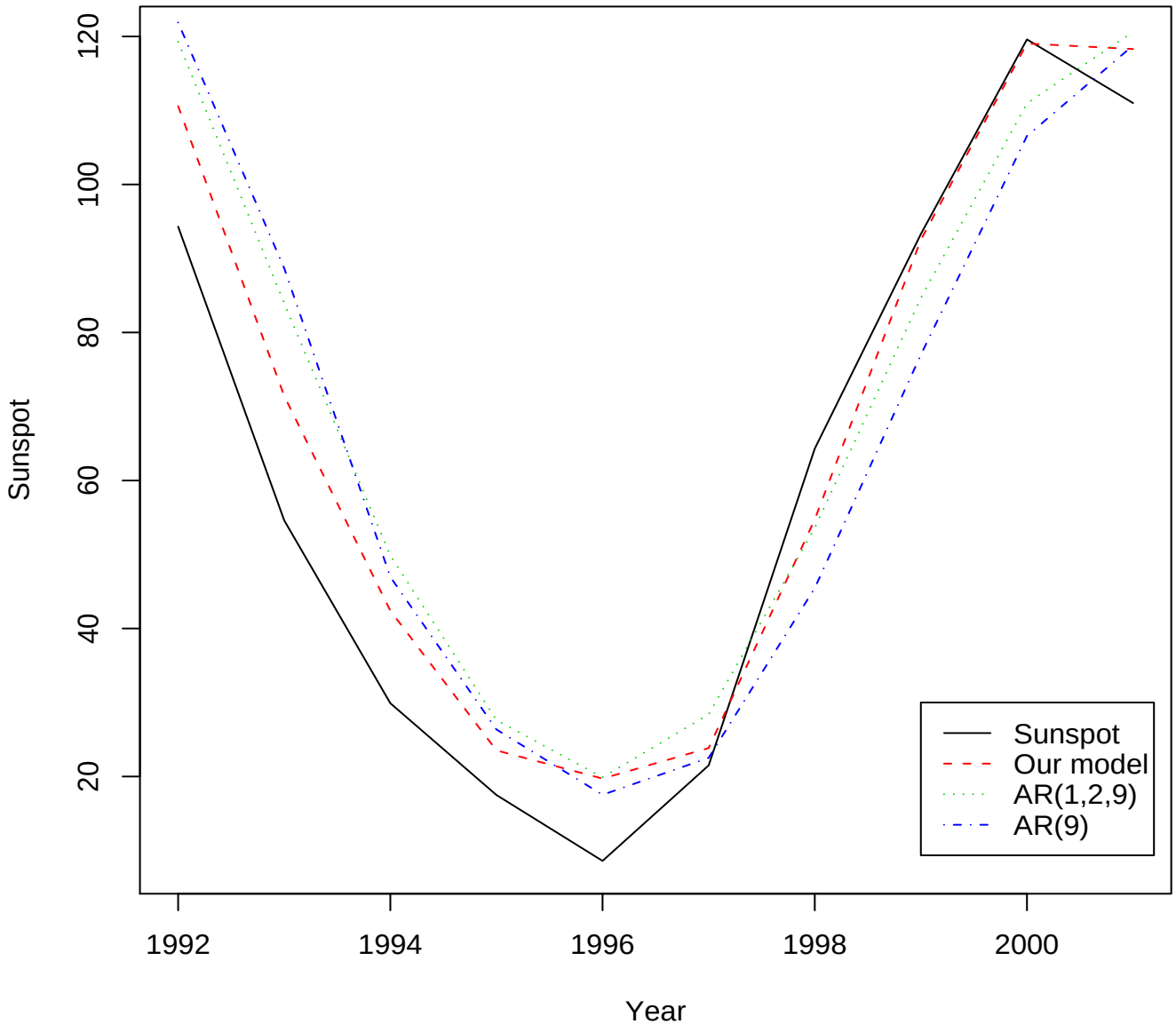

Figure 2.3: Wolf yearly sunspot data: Overlay plot of observed sunspot numbers (Sunspot) and forecast values from AR(9), AR(1,2,9), and our model: Years 1992-2001.

2.4.3 U.S. GNP Data

This gives seasonally adjusted quarterly growth rates of US Gross National Product (GNP) from 1947 to 1991, which is obtained from the Citibase database. Shummway and Stoffer (2000) conclude that a MA(2) model fits the data well, whereas Tiao and Tsay (1994) and Hall and Yao (2005) propose AR-type models. As mentioned in the introduction, Hall and Yao (2005) fix $p = 2$ and $d = 1$, and obtain an estimate $\hat{\Phi}_1 = (0.580, -0.815)^T$ of Φ_1. They conclude that the conditional distribution of $x_t|.58x_{t-1} - .815x_{t-2}$ provides a good approximation to that of $x_t|\{x_{t-1}, x_{t-2}\}$. Assuming their result, we used graphical methods to arrive at the following nonlinear mean function model for the data given by

$$\hat{x}_t = 0.74 + 0.02\gamma + 0.91\gamma^2,$$

where $\gamma = 100\hat{\Phi}_1^T\mathbf{X}_{t-1}$. We then added an error term to the above model and simulated a new time series. Using the methods described in Section 2.3.3, we determined that $\hat{d} = 1$ and $\hat{p} = 2$ for this new series, which shows that our methods correctly detected d and p.

2.4.4 U.S. Beer Production Data

This seasonal data (Wei 2006) concerns U.S. beer production (in millions of barrels) for 32 consecutive quarters from 1975 to 1982. Here, we illustrate that our method performs well when the data is seasonal and the sample size is small. We use the first 30 observations to build a model and then calculate model-based forecasts for the last two observations.

Using the methods described in Section 2.3.3, we obtained $\hat{d}_p = 1$ for each $2 \leq p \leq 9$ using the 0.05- and 0.01-threshold as you can see in Table 2.8. Moreover, the *Shoulder Plot* in Figure 2.4 clearly indicates that the *Shoulder* is at $p = 4$. Fixing $p = 4$ and $d = 1$, and using graphical techniques we arrive at the following model:

$$\hat{x}_t = 5.71 - 0.80\hat{\Phi}_1^T\mathbf{X}_{t-1} - 1.27\cos\left(\frac{\pi}{2}\hat{\Phi}_1^T\mathbf{X}_{t-1} - 2\pi\right).$$

To compare the performance of our model with the seasonal MA and AR models given in Wei (2006; see pages 178-182), we computed the Mean Square Error (MSE) $= k^{-1}\sum_{t=1}^{k}(x_t -$

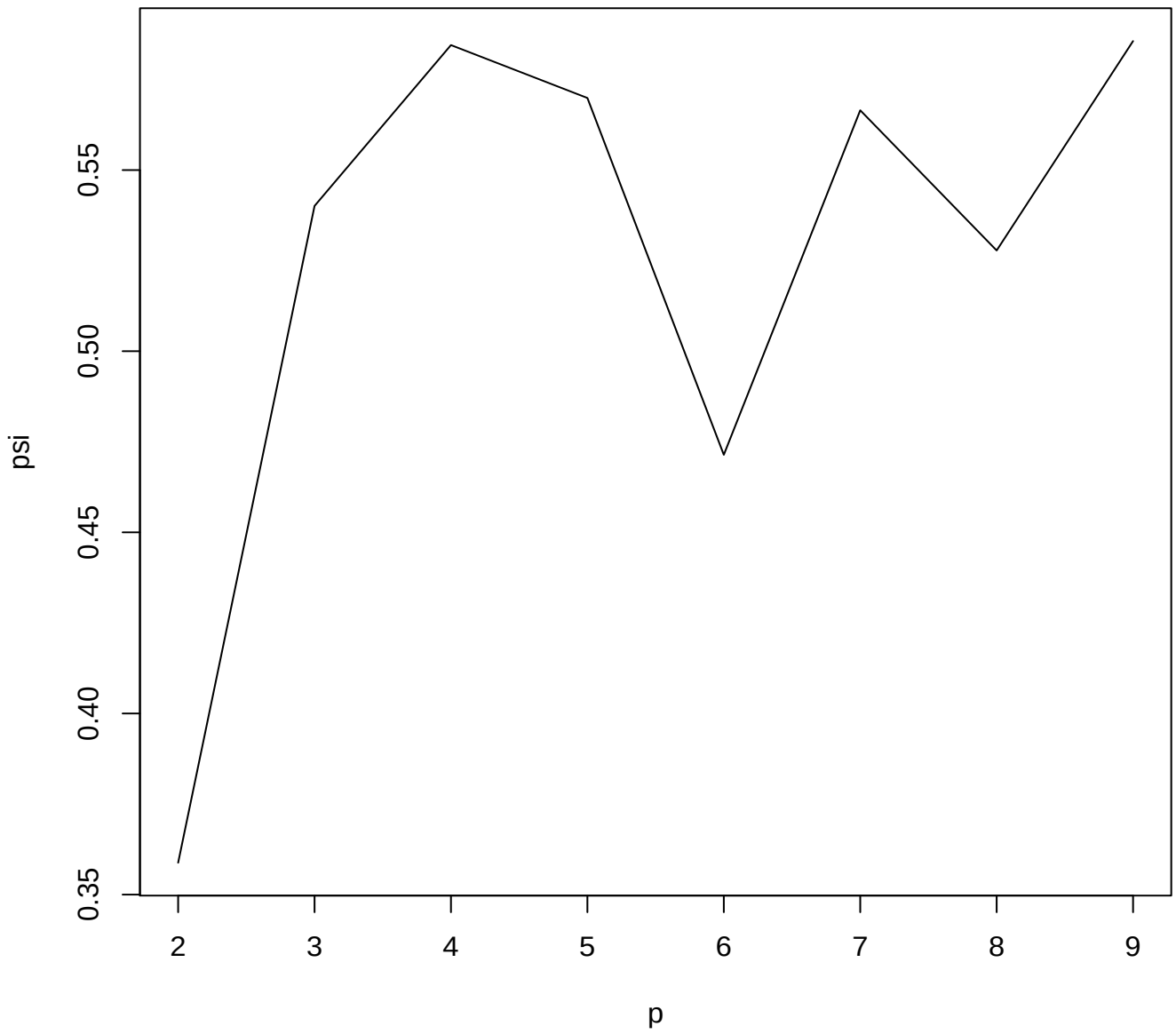

Figure 2.4: U.S. beer production data: *Shoulder plot* of $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,1})$ values versus $p = 2, \cdots, 9$.

p	d=1	d=2
2	0.3588*	0.4514
3	0.5401*	0.6677
4	0.5845*	0.6767
5	0.5699*	0.6535
6	0.4714*	0.7150
7	0.5665*	0.7183
8	0.5278*	0.7139
9	0.5856*	0.6930

Table 2.8: U.S. beer production data: $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{h}}_{p,d})$ values for $p = 2, \cdots, 9$, and $d = 1, 2$. For each p, $\hat{d}_p$ determined by 0.05 are denoted by *, in the table.

$\hat{x}_t)^2$ for the last two quarters of 1982, where x_t is the observed beer production, $\hat{x}_t$ is its forecast value and k is the number of (future) observations. MSE values for our model, the seasonal MA and AR models are 4.87, 8.41, and 8.50, respectively. Note that the MSE value for our model is about half of the other two models, indicating that our model performs better than the seasonal ARMA models.

2.5 Discussion

Literature has seen a proliferation of parametric and nonparametric methods for time series analysis; however, only few use a dimension reduction approach. In this chapter, we developed a new theory of dimension reduction in time series, which provides an initial phase when an adequate parsimoniously parametrized time series model is not yet available. Although the notion of TSCS is similar to that in regression, there are many differences due to the intrinsic nature of a time series. For instance, the fact that well-known dimension reduction methods such as Sliced Inverse Regression (Li, 1991) or Sliced Average Variance Estimation (Cook and Weisberg, 1991) are not relevant for time series (Xia, Tong, Li and Zhu, 2002) warrants a different treatment. Furthermore, p is usually known in regression, whereas it has

to be inferred from data for a time series. We believe that the sufficient dimension reduction approach proposed here may stimulate new ideas for modeling time series.

We proposed an estimating function and a sequential approach to estimate d and lag p. To estimate d, we proposed an estimator and suggested the use of either 0-threshold or 0.05-threshold. The choice of threshold certainly affects the value of $\hat{d}_p$, which increases as $\tau_{p,n}$ decreases. This choice, however, poses more challenges than the end of the story. For example, it is also possible to motivate a choice of threshold based on the AIC or the Schwarz Bayesian Criterion. Clearly, this raises an important question about the selection of threshold values; and more research needs to be done in this direction. As for the estimation of p, it is possible to define an estimator similar to $\hat{d}_p$. Nevertheless, the use of *Shoulder Plot* seems quite informative in our analysis because of its visual appeal and simplicity. Further investigation in this direction is also needed.

Overall, the theory of dimension reduction in time series poses many challenges, but a variety of encouraging results presented through our simulations seem to suggest that our method has great potential for providing a viable and meaningful alternative to traditional time series analysis. In fact, superior performance of our nonlinear time series model for Wolf yearly sunspot data, as compared to Wei's (2006) models, serves as a testament that our method is very useful in time series analysis. Also, for the seasonal U.S beer production data, our model performs better than those in Wei (2006). Finally, we consider estimation of TSCS and other related issues in this chapter. However, new dimension reduction methods such as central mean subspaces (Cook and Li, 2002), which focus on the mean function of time series, is yet to be developed. Research on the latter topic is the focus of next chapter.

2.6 Appendix

2.6.1 PROOF OF PROPOSITION 1

For notational simplicity, the single-index is introduced for this justification. We need to show that $\mathcal{S}(\delta)$ is a DRS when δ be a basis for $\mathcal{S}(\phi)$ and $\mathcal{S}(\alpha)$ so that $\phi = (\phi_1, \delta)$ and $\alpha = (\alpha_1, \delta)$ where $\alpha_1 \neq 0$ and $\phi_1 \neq 0$. Hence, for more general approach, let $\mathbf{K} = (\mathbf{K}_1, \mathbf{K}_2, \mathbf{K}_3)^T = (\phi_1^T \mathbf{X}_{t-1}, \alpha_1^T \mathbf{X}_{t-1}, \delta^T \mathbf{X}_{t-1})^T$. $\mathbf{K}$ has a density because $\mathbf{X}_{t-1}$ has a density and $(\phi_1, \alpha_1, \delta)$ is a full rank linear operator. Therefore, we can say the distribution of $(\mathbf{K}_1, \mathbf{K}_2)|(\mathbf{K}_3 = k_3)$ has a density and then

$$F_{x_t|\mathbf{K}} = F_{x_t|\mathbf{K}_1,\mathbf{K}_3} = F_{x_t|\mathbf{K}_2,\mathbf{K}_3}, \tag{2.5}$$

since $\mathcal{S}(\phi)$ and $\mathcal{S}(\alpha)$ are DRSs. Here, our goal is to show $F_{x_t|\mathbf{K}} = F_{x_t|\mathbf{K}_3}$.

Now, when k_3 is fixed, $(k_1, k_2')^T$ and $(k_1', k_2)^T$ are considered to be linked if either $k_1 = k_1'$ or $k_2 = k_2'$. Let take $(k_1, k_2')^T$, $(k_1, k_2)^T$, and $(k_1', k_2)^T$. By (2.5),

$$F_{x_t|\mathbf{K}_1=k_1',\mathbf{K}_3} = F_{x_t|\mathbf{K}_2,\mathbf{K}_3} \tag{2.6}$$

and

$$F_{x_t|\mathbf{K}_2=k_2',\mathbf{K}_3} = F_{x_t|\mathbf{K}_1,\mathbf{K}_3}. \tag{2.7}$$

By (2.5), (2.6), and (2.7),

$$F_{x_t|\mathbf{K}_1=k_1',\mathbf{K}_3} = F_{x_t|\mathbf{K}_2=k_2',\mathbf{K}_3} \tag{2.8}$$

Then,

$$F_{x_t|\mathbf{K}_1=k_1',k_2,k_3} = F_{x_t|k_1,\mathbf{K}_2=k_2',k_3}. \tag{2.9}$$

Provided there is at least one linked point like $(k_1, k_2)^T$ in this case, any two points can be chained together, which means (2.9) holds for any two points. Finally, $F_{x_t|\mathbf{K}_1,\mathbf{K}_2,\mathbf{K}_3}$ is a constant function of (k_1, k_2) and then the conclusion follows.

2.6.2 PROOF OF PROPOSITION 2

(i) If $\mathcal{S}(\mathbf{h}_1) \subseteq \mathcal{S}(\mathbf{h}_2)$, then it is possible to write $\mathbf{h}_1 = \mathbf{h}_2\mathbf{K}$ for some matrix $\mathbf{K}$. Therefore, with simple algebra,

$$\Psi(\mathbf{h}_2) - \Psi(\mathbf{h}_1) = \mathrm{E}\left[\mathrm{E}_{x_t|\mathbf{h}_2^T\mathbf{X}_{t-1}}\left\{\log\frac{p(x_t|\mathbf{h}_2^T\mathbf{X}_{t-1})}{p(x_t|\mathbf{K}^T\mathbf{h}_2^T\mathbf{X}_{t-1})}\right\}\right] \geq 0.$$

The inequality follows from a result on page 14 of Kullback (1959). Suppose $\mathcal{S}(\mathbf{h}_1) = \mathcal{S}(\mathbf{h}_2)$. Then $\mathbf{K}$ can be a nonsingular square matrix in the above argument, and the inequality will become an equality.

(ii) As in (i), we can write

$$\Psi(I) - \Psi(\mathbf{h}_1) = E\left[E_{x_t|\mathbf{X}_{t-1}}\left\{\log\frac{p(x_t|\mathbf{X}_{t-1})}{p(x_t|\mathbf{h}_1^T\mathbf{X}_{t-1})}\right\}\right] \geq 0$$

with equality if and only if $p(x_t|\mathbf{X}_{t-1}) = p(x_t|\mathbf{h}_1^T\mathbf{X}_{t-1})$. But, $x_t \perp\!\!\!\perp \mathbf{X}_{t-1}|\mathbf{h}_1^T\mathbf{X}_{t-1}$ if and only if $p(x_t|\mathbf{X}_{t-1}) = p(x_t|\mathbf{h}_1^T\mathbf{X}_{t-1})$. Hence the first assertion in (ii). From this and the definition of TSCS, $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\mathbf{\Phi}_d)$, for which the property $x_t \perp\!\!\!\perp \mathbf{X}_{t-1}|\mathbf{\Phi}_d^T\mathbf{X}_{t-1}$ holds, we have that $\Psi(I) = \Psi(\mathbf{\Phi}_d) \geq \Psi(\mathbf{h}_d)$. Now, if $\Psi(\mathbf{\Phi}_d) = \Psi(\mathbf{h}_d)$, then $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}(\mathbf{\Phi}_d) = \mathcal{S}(\mathbf{h}_d)$ by the uniqueness of TSCS. The other way conclusion follows from (i).

(iii) Since

$$\Psi(\mathbf{h}_1) = \mathrm{E}\left\{\log\frac{p(x_t, \mathbf{h}_1^T\mathbf{X}_{t-1})}{p(x_t)p(\mathbf{h}_1^T\mathbf{X}_{t-1})}\right\},$$

$\Psi(\mathbf{h}_1) \geq 0$ by a result of Kullback (1959, p.14).

Let $\boldsymbol{\alpha} = \arg\max_{\mathbf{h}_1}\Psi(\mathbf{h}_1)$, and, for any $p \times (q_2 - q_1)$ matrix $\boldsymbol{\beta}$, by using Kullback's result (1959, p. 14), we have

$$\begin{aligned} & \max\Psi(\mathbf{h}_2) - \max\Psi(\mathbf{h}_1) \geq \Psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \Psi(\boldsymbol{\alpha}) \\ = \; & \mathrm{E}_{\boldsymbol{\alpha}^T\mathbf{X}_{t-1}, \boldsymbol{\beta}^T\mathbf{X}_{t-1}}\left[\mathrm{E}_{x_t|\boldsymbol{\alpha}^T\mathbf{X}_{t-1}, \boldsymbol{\beta}^T\mathbf{X}_{t-1}}\left\{\log\frac{p(x_t|\boldsymbol{\alpha}^T\mathbf{X}_{t-1}, \boldsymbol{\beta}^T\mathbf{X}_{t-1})}{p(x_t|\boldsymbol{\alpha}^T\mathbf{X}_{t-1})}\right\}\right] \geq 0. \end{aligned}$$

However, equality cannot hold, unless $p(x_t|\boldsymbol{\alpha}^T\mathbf{X}_{t-1}, \boldsymbol{\beta}^T\mathbf{X}_{t-1}) = p(x_t|\boldsymbol{\alpha}^T\mathbf{X}_{t-1})$ for any $\boldsymbol{\beta}$; that is $x_t \perp\!\!\!\perp \boldsymbol{\beta}^T\mathbf{X}_{t-1}|\boldsymbol{\alpha}^T\mathbf{X}_{t-1}$, for any $\boldsymbol{\beta}$, and hence $x_t \perp\!\!\!\perp \mathbf{X}_{t-1}|\boldsymbol{\alpha}^T\mathbf{X}_{t-1}$. Thus, $\Psi(I) = \Psi(\boldsymbol{\alpha})$, by the definition of central subspace. This produces the contradiction that $d \leq (d-1)$. □

2.6.3 ASSUMPTION A1 AND LEMMA 1

The following assumptions apply to Lemma 1 stated below and Theorem 1 stated in Section 3.4. We first introduce some notations. Define the $\emptyset_{\mathbf{y}}^{\mathbf{x}}$-operator for functions $g : \mathbb{R}^k \to \mathbb{R}^1$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$ by

$$\emptyset_{\mathbf{y}}^{\mathbf{x}} g = \sum_{(\epsilon_1,...,\epsilon_k)\in\{0,1\}^k} (-1)^{\sum_{j=1}^k \epsilon_j} g\{\epsilon_1 x_1 + (1-\epsilon_1)y_1 + ... + \epsilon_k x_k + (1-\epsilon_k)y_k\}$$

(Reyni, 1962). Suppose $\mathcal{P}$ is the set of all finite partitions of $\mathbb{R}^k$ into rectangles and $\emptyset_{\mathbf{y}}^{\mathbf{x}} g$ is expressed by the corresponding limit if some components of $\mathbf{x}$ and $\mathbf{y}$ are positive or negative infinity. Then g is said to be of bounded variation if $\sup\{g(s), s \in \mathcal{P}\} < \infty$, where $g(s)$ is defined for $s = \sum_{i=1}^l [x_i, y_i)$ by $\sum_{i=1}^l |\emptyset_{y_i}^{x_i} g|$.

As in Sen (1974), we define a stationary ϕ-mixing sequence $\{\mathbf{h}^T\mathbf{X}_t, -\infty < t < \infty\}$ for any $p \times d$ matrix $\mathbf{h}$, on a probability space $(\Omega, \mathscr{A}, P)$ with each $\mathbf{h}^T\mathbf{X}_t$ having a continuous distribution F. That is, if $\mathscr{M}_{-\infty}^k$ and $\mathscr{M}_{k+n}^\infty$ are σ-fields generated by $\{\mathbf{h}^T\mathbf{X}_t, t \le k\}$ and $\{\mathbf{h}^T\mathbf{X}_t, t \ge k+n\}$, respectively, and if $A \in \mathscr{M}_{-\infty}^k$ and $B \in \mathscr{M}_{k+n}^\infty$, then for all k, $n \ge 0$ and $\phi_n \ge 0$, $|P(A \cap B) - P(A)P(B)| \le \phi_n P(A)$, where $\{\phi_n\}$ is independent of $\mathbf{h}$, $\{\phi_n\} \downarrow$ in n and $\lim_{n\to\infty} \phi_n = 0$. Furthermore, for each $m \ge 0$, we define $A_m(\phi) = \sum_{n=1}^\infty (n+1)^m \phi_n^{1/2}$ for $\{\phi_n\}$ defined above. Assume that $A_m(\phi) < \infty$ for some $m \ge 1$, as in Theorem 3.2 of Sen (1974). Since ϕ_n is independent of $\mathbf{h}$, the conclusion of Sen's Theorem 3.2 holds for $\mathbf{h}^T\mathbf{X}_t$ with the upper bound $C_\phi \lambda^{-2(m+1)}$ (for $\lambda \ge 1$), where $C_\phi (< \infty)$ only depends on $\{\phi_n\}$. Using this upper bound and same arguments as in the proof of Theorem 1(b) of Rüschendorf (1977), one can prove the following lemma.

Lemma 1 *Assume the conditions in A1 and that* $\sum_{n=1}^\infty \left(\frac{\gamma}{\sqrt{n}\, a_n^k}\right)^{2(m+1)} < \infty$, *for all* $\gamma \in \mathbb{R}_+$, *where* $\{a_n\}$ *is a sequence of bandwidths of the kernel density estimator* f_n *defined in Section 2.3.4. Suppose the kernel* K *in* f_n *is of bounded variation. Also, let the density functions satisfy the following conditions:* $p(x_t)$ *is uniformly continuous,* $p(\boldsymbol{h}^T\mathbf{X}_{t-1})$ *is uniformly continuous in* $\boldsymbol{h}$ *and* $\mathbf{X}_{t-1}$, *and* $p(\boldsymbol{h}^T\mathbf{X}_{t-1}, x_t)$ *is uniformly continuous in* $\boldsymbol{h}$, $\mathbf{X}_{t-1}$, *and* x_t, *where*

$\boldsymbol{h}^T\boldsymbol{h} = I$. Then the following results hold with probability one, as $n \to \infty$:

$$\sup_{x_t \in \mathbb{R}^1} |f_n(x_t) - p(x_t)| \to 0,$$

$$\sup_{\boldsymbol{h} \in \mathbb{R}^{p\times d}, \mathbf{X}_{t-1} \in \mathbb{R}^p} |f_n(\boldsymbol{h}^T\mathbf{X}_{t-1}) - p(\boldsymbol{h}^T\mathbf{X}_{t-1})| \to 0,$$

$$\sup_{\boldsymbol{h} \in \mathbb{R}^{p\times d}, \mathbf{X}_{t-1} \in \mathbb{R}^p, x_t \in \mathbb{R}^1} |f_n(\boldsymbol{h}^T\mathbf{X}_{t-1}, x_t) - p(\boldsymbol{h}^T\mathbf{X}_{t-1}, x_t)| \to 0.$$

2.6.4 PROOF OF THEOREM 1 AND LEMMA 2

Note that the constraint $\mathbf{h}^T\mathbf{h} = I$ does not guarantee that a matrix maximizing the objective function $\Psi(\mathbf{h})$ is unique. Nevertheless, the subspace corresponding to it is unique. Therefore, for identifiability, we may replace any basis matrix that maximizes the objective function by its orthogonal projection matrix, which is unique. Thus, without loss of generality and for the simplicity of our proof, we assume that the matrix solution is unique.

If $\hat{\mathbf{\Phi}}_n$ does not converge to $\mathbf{\Phi}_d$ with probability 1, there is a subsequence which is still indexed by n, and a $p \times d$ matrix $\mathbf{\Phi}_0$ satisfying $\mathbf{\Phi}_0^T\mathbf{\Phi}_0 = I$ and $\mathbf{\Phi}_0 \neq \mathbf{\Phi}_d$, such that $\hat{\mathbf{\Phi}}_n \to \mathbf{\Phi}_0$. Thus, for any $\epsilon > 0$, and large enough n, we have

$$f_n(x_t) = p(x_t) + \delta_{1,t}, \tag{2.10}$$

$$f_n(\hat{\mathbf{\Phi}}_n^T\mathbf{X}_{t-1}) = p(\hat{\mathbf{\Phi}}_n^T\mathbf{X}_{t-1}) + \eta_{2,t} = p(\mathbf{\Phi}_0^T\mathbf{X}_{t-1}) + \delta_{2,t}, \tag{2.11}$$

$$f_n(\hat{\mathbf{\Phi}}_n^T\mathbf{X}_{t-1}, x_t) = p(\hat{\mathbf{\Phi}}_n^T\mathbf{X}_{t-1}, x_t) + \eta_{3,t} = p(\mathbf{\Phi}_0^T\mathbf{X}_{t-1}, x_t) + \delta_{3,t}, \tag{2.12}$$

such that $|\delta_{k,t}| < \epsilon$ for all t and $k = 1, 2, 3$. Note that equation (2.10) and the first equalities in equation (2.11) and (2.12) follow from the conclusions of Lemma 1, whereas the uniform continuity conditions in Lemma 1 lead to the second equalities in equations (2.11) and (2.12). From these, we have

$$\log f_n(x_t) = \log p(x_t) + \log\left\{1 + \frac{\delta_{1,t}}{p(x_t)}\right\},$$

$$\log f_n(\hat{\mathbf{\Phi}}_n^T\mathbf{X}_{t-1}) = \log p(\hat{\mathbf{\Phi}}_0^T\mathbf{X}_{t-1}) + \log\left\{1 + \frac{\delta_{2,t}}{p(\mathbf{\Phi}_0^T\mathbf{X}_{t-1})}\right\},$$

$$\log f_n(\hat{\mathbf{\Phi}}_n^T\mathbf{X}_{t-1}, x_t) = \log p(\hat{\mathbf{\Phi}}_0^T\mathbf{X}_{t-1}, x_t) + \log\left\{1 + \frac{\delta_{3,t}}{p(\mathbf{\Phi}_0^T\mathbf{X}_{t-1}.x_t)}\right\},$$

Therefore, by the definition of κ_ι in Section 2.3.4 and that $\frac{\epsilon}{\iota} \to 0$, for $\hat{\mathbf{\Psi}}_n^\iota$ defined in the Theorem 1 we have that

$$\hat{\mathbf{\Psi}}_n^\iota(\hat{\mathbf{\Phi}}_n) = \frac{1}{n}\sum_{t=1}^{n} J(t \in \kappa_\iota) \log \frac{p(\mathbf{\Phi}_0^T \mathbf{X}_{t-1}, x_t)}{p(x_t)p(\mathbf{\Phi}_0^T \mathbf{X}_{t-1})} + o(1) = \bar{\Psi}_n^\iota(\mathbf{\Phi}_0) + o(1).$$

But,

$$\begin{aligned} \bar{\Psi}_n^\iota(\mathbf{\Phi}_0) - \Psi(\mathbf{\Phi}_0) &= \left\{ \frac{1}{n}\sum_{t=1}^{n} \log \frac{p(\mathbf{\Phi}_0^T \mathbf{X}_{t-1}, x_t)}{p(x_t)p(\mathbf{\Phi}_0^T \mathbf{X}_{t-1})} - \Psi(\mathbf{\Phi}_0) \right\} \\ &- \frac{1}{n}\sum_{t=1}^{n} J(t \in \kappa_\iota^c) \log \frac{p(\mathbf{\Phi}_0^T \mathbf{X}_{t-1}, x_t)}{p(x_t)p(\mathbf{\Phi}_0^T \mathbf{X}_{t-1})} \\ &= \tau_1 - \tau_2. \end{aligned}$$

From $\frac{n_\iota}{n} \to 0$ and the ergodic theorem, both τ_1 and τ_2 tend to 0 with probability one as $n \to \infty$. Hence, $\lim_{n\to\infty} \hat{\mathbf{\Psi}}_n^\iota(\hat{\mathbf{\Phi}}_n) = \Psi(\mathbf{\Phi}_0)$ with probability one. Since $\hat{\mathbf{\Psi}}_n^\iota(\hat{\mathbf{\Phi}}_n) \geq \hat{\mathbf{\Psi}}_n^\iota(\mathbf{\Phi}_d)$ by definition, taking limit on both sides we get $\Psi(\mathbf{\Phi}_0) \geq \Psi(\mathbf{\Phi}_d)$. On the other hand, by the definition of $\mathbf{\Phi}_d$, $\Psi(\mathbf{\Phi}_0) \leq \Psi(\mathbf{\Phi}_d)$ and therefore $\Psi(\mathbf{\Phi}_0) = \Psi(\mathbf{\Phi}_d)$. Due to the uniqueness, $\mathbf{\Phi}_0 = \mathbf{\Phi}_d$, which is a contradiction. Therefore, $\hat{\mathbf{\Phi}}_n \to \mathbf{\Phi}_d$ with probability 1.

Lemma 2 *Assume the conditions of Lemma 1 and Theorem 1. For each fixed p and k,* $\max_{\boldsymbol{h}_{p,k}} \hat{\mathbf{\Psi}}_n^\iota(\boldsymbol{h}_{p,k}) \to \max_{\boldsymbol{h}_{p,k}} \mathbf{\Psi}(\boldsymbol{h}_{p,k})$, *as* $n \to \infty$.

2.6.5 PROOF OF LEMMA 2

For simplicity, we assume that $\arg\max_{\mathbf{h}_{p,k}} \Psi(\mathbf{h}_{p,k})$ is unique. Then, by the arguments in the proof of Theorem 1 and its conclusion, we have that $\lim_{n\to\infty} \max_{\mathbf{h}_{p,k}} \hat{\mathbf{\Psi}}_n^\iota(\mathbf{h}_{p,k}) = \max_{\mathbf{h}_{p,k}} \mathbf{\Psi}(\mathbf{h}_{p,k})$, with probability one.

2.6.6 PROOF OF THEOREM 2

Let $c_k = \max_{\mathbf{h}_{p,(k+1)}} \mathbf{\Psi}(\mathbf{h}_{p,(k+1)}) - \max_{\mathbf{h}_{p,k}} \mathbf{\Psi}(\mathbf{h}_{p,k})$. By Proposition 2 (iii), it follows that $c_k > 0$ if $k < d$ and $c_k = 0$ if $k \geq d$. Hence, $d = \min\{k(\leq (p-1)) : c_k = 0\}$. Moreover, for each k, we have by Lemma 2 that $\hat{c}_k^\iota \to c_k$ as $n \to \infty$. Recall that $\hat{d}_p^\iota = \min\{k(\leq (p-1)) : \hat{c}_k^\iota \leq \tau_{p,n}\}$. Since $\tau_{p,n} \to 0$ as $n \to \infty$, it follows that $\hat{d}_p^\iota \to d$ as $n \to \infty$, with probability one.

2.7 References

[1] Barlett, M. S. (1950), "Periodogram analysis and continuous spectra," *Biometrika*, 37, 1–16.

[2] Brillinger, D. R. and Rosenblatt, M. (1967), "Asymptotic theory of kth order spectra," *Spectral Analysis of Time Series* (Ed. B. Harris), 153–188, New York: John Wiley.

[3] Cook, R. D. (1994), "On the interpretation of regression plots," *Journal of the American Statistical Association*, 89, 177–190.

[4] Cook, R. D. (1996), "Graphics for regressions with a binary response," *Journal of the American Statistical Association*, 91, 983–992.

[5] Cook, R. D. (1998a), *Regression Graphics: Ideas for studying regressions through graphics*, New York: Wiley.

[6] Cook, R. D. (1998b), "Principal Hessian directions revisited (with discussion)," *Journal of the American Statistical Association*, 93, 84–100.

[7] Cook, R. D. and Li, B. (2002), "Dimension reduction for the conditional mean in regression," *Annals of Statistics*, 30, 455–474.

[8] Cook, R. D. and Weisberg, S. (1991), "Discussion of 'sliced inverse regression' by K. C. Li," *Journal of the American Statistical Association*, 86, 328–332.

[9] Cover, T. M. and Thomas, J. A. (1991), *Elements of Information Theory*, New York: John Wiley & Sons.

[10] Fan, J., Heckman, M. E., and Wand, M. P. (1995). "Local polynomial kernel regression for generalized linear models and quasi-likelihood functions," *Journal of the American Statistical Association*, 90, 141–150.

[11] Fan, J., Yao, Q., and Tong, H. (1996). "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems," *Biometrika*, 83, 189–206.

[12] Gill, P. Murray, W., and Wright, M. H. (1981), *Practical Optimization*, New York: Academic Press.

[13] Hall, P. and Yao, Q. (2005), "Approximating conditional distribution functions using dimension reduction." *Annals of Statistics*, 33, 1404–1421.

[14] Hotelling, H. (1936), "Relations between two sets of variates," *Biometrika*, 28, 321–377.

[15] Kullback, S. (1959), *Information theory and statistics*, New York: Wiley.

[16] Li, K. C. (1991), "Sliced inverse regression for dimension reduction (with discussion)," *Journal of the American Statistical Association*, 86, 316–342.

[17] Li, K. C. (1992), "On principal Hessian directions for data visualization and dimension reduction: another application of Stein's Lemma," *Annals of Statistics*, 87, 1025–1039.

[18] Ng, S. and Perron, P. (2005). "A note on selection of time series models," *Oxford Bulletin of Economics and Statistics*, 67, 115–134.

[19] Reyni, A. (1962), "Wahrscheinlichkeitsrechnung," Berlin: VEB, Deutscher Verlag der Wisschenschaften.

[20] Rüschendorf, L. (1977), "Consistency of estimators for multivariate density functions and for the mode," *Sankhya* A, 39, 243–250

[21] Schott, J. R. (1994), "Determining the dimensionality in sliced inverse regression," *Journal of the American Statistical Association*, 89, 141-148.

[22] Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: John Wiley & Sons.

[23] Sen, P. K. (1974), "Weak convergence of multidimensional empirical processes for stationary ϕ-mixing processes," *Annals of Probability*, 2, 147–154.

[24] Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London and New York: Chapman and Hall.

[25] Shummway, R. H. and Stoffer, D. S. (2000), *Time Series Analysis and Its Applications*, New York: Springer-Verlag.

[26] Tiao, G. C. and Tsay, R. S. (1994), "Some advances in nonlinear and adaptive modeling in time series," *Journal of Forecasting*, 13, 109–131.

[27] Wei, W. W. S. (2006), *Time Series Analysis: Univariate and Multivariate Methods*, Boston: Pearson & Addison Wesley.

[28] Whittle, P. (1954), "A statistical investigation of sunspot observations with special reference to H.Alfven's sunspot model," *The Astrophysical Journal*, 120, 251–260.

[29] Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002), "An adaptive estimation of dimension reduction," *Journal of the Royal Statistical Society*, Ser. B, 64, 363–410.

[30] Ye, Z. and Weiss, R. E. (2003), "Using the bootstrap to select one of a new class of dimension reduction methods," *Journal of the American Statistical Association*, 98, 968–979.

[31] Yin, X. and Cook, R. D. (2005), "Direction estimation in single-index regressions," *Biometrika*, 92, 371–384.

[32] Yule, G. U. (1927), "On a method of investigating periodicities in disturbed series with special reference to wolfer's sunspot numbers," *Philosophical Transaction Royal Society London*, Ser. A, 226, 267–298.

Chapter 3

Time Series Central Mean Subspace [†]

[†]Park, J. H., Sriram, T. N., and Yin, X. To be submitted to *The Journal of Computational and Graphical Statistics*.

3.1 Introduction

In the previous chapter, we developed a sufficient dimension reduction method for time series, where our aim was to make inference about the conditional distribution of the series given the past. In many instances, time series analysis is concerned with inference about the conditional mean of the current observation given the past of the series, and less concerned with the other aspects of conditional distribution. In such instances, it may be useful to adapt our inquiry to fit that more specific objective.

Recently, Cook and Li (2002) developed dimension reduction methods that address inference about the conditional mean of response given the predictors. They introduced a notion of Central Mean Subspace (CMS), similar to the notion of Central Subspace, and an estimation methodology to estimate the CMS. Other dimension reduction methods for CMS available in the literature include principal Hessian direction (pHd; Li 1992), the Structure Adaptive Method (SAM; Hristache, Juditsky, Polzehl, and Spokoiny 2001), the Iterative Hessian Transformation (IHT; Cook and Li 2002), the Minimum Average Variance Estimation (MAVE; Xia, Tong, Li, and Zhu 2002). Of these, only the MAVE method developed by Xia, Tong, Li and Zhu (2002) is also applicable to dimension reduction in time series.

In this chapter, we focus on developing a notion of Time Series Central Mean Subspace (TSCMS) and propose a method to estimate it, when the lag and the dimension are known. While the development of our notion of TSCMS bears similarity to the one in Chapter 2, the method of estimating it differs from our approach in Chapter 2. In addition, we also discuss estimation of minimum dimension (when it exists) and lag. As in Chapter 2, we illustrate our method via simulations for a variety of linear and nonlinear time series models and through analysis of real data on Yearly number of lynx pelts.

The rest of the chapter is organized as follows. In Section 3.2, we develop a theory of dimension reduction in time series by introducing a notion of Central Mean Subspace in time series and study its properties. In Section 3.3, we discuss the estimation method for the Central Mean Subspace in time series when its dimension and lag of the series are known.

Monte Carlo simulations for a variety of linear and nonlinear time series models and a real data analysis are given in Section 3.4. A brief discussion of results obtained in this chapter is carried out Section 3.5. All the necessary proofs are given in Section 3.6.

3.2 Central Mean Subspace in Time Series

As before, let x_t denote the current observation of a time series and $\mathbf{X}_{t-1} = (x_{t-1}, ..., x_{t-p})^T$, where p is known. When focusing on the conditional mean of a series, dimension reduction hinges on finding a $p \times q$ matrix $\mathbf{\Phi}$, $q \leq p$, so that the $q \times 1$ vector $\mathbf{\Phi}^T\mathbf{X}_{t-1}$ includes all the information about x_t that is available from $\mathrm{E}(x_t|\mathbf{X}_{t-1})$. Note that this is considerably less restrictive than requiring that $\mathbf{\Phi}^T\mathbf{X}_{t-1}$ contains all the information about x_t that is available from $\mathbf{X}_{t-1}$, as in Chapter 2.

Definition 1 *If*

$$x_t \perp\!\!\!\perp \mathrm{E}(x_t|\mathbf{X}_{t-1})|\mathbf{\Phi}^T\mathbf{X}_{t-1},$$

then $\mathcal{S}(\mathbf{\Phi})$ *is mean dimension reduction subspace for the time series* x_t*.*

It is clear from the above definition that a time series dimension reduction subspace is necessarily a mean dimension reduction subspace, because $x_t \perp\!\!\!\perp \mathbf{X}_{t-1}|\mathbf{\Phi}^T\mathbf{X}_{t-1}$ implies $x_t \perp\!\!\!\perp \mathrm{E}(x_t|\mathbf{X}_{t-1})|\mathbf{\Phi}^T\mathbf{X}_{t-1}$. The following proposition gives equivalent conditions for the conditional independence in Definition 1.

Proposition 3 *The following three statements are equivalent:*

(i) $x_t \perp\!\!\!\perp \mathrm{E}(x_t|\mathbf{X}_{t-1})|\mathbf{\Phi}^T\mathbf{X}_{t-1}$.

(ii) $Cov(x_t, \mathrm{E}(x_t|\mathbf{X}_{t-1})|\mathbf{\Phi}^T\mathbf{X}_{t-1}) = 0$.

(iii) $\mathrm{E}(x_t|\mathbf{X}_{t-1})$ *is a function of* $\mathbf{\Phi}^T\mathbf{X}_{t-1}$.

Part (i) of Proposition 3 is exactly same as Definition 1. Part (ii) says there is no correlation between x_t and $\mathrm{E}(x_t|\mathbf{X}_{t-1})$ given $\mathbf{\Phi}^T\mathbf{X}_{t-1}$. Part (iii) is what might be suggested by

intuition as $\mathrm{E}(x_t|\mathbf{X}_{t-1}) = \mathrm{E}(x_t|\mathbf{\Phi}^T\mathbf{X}_{t-1})$. The above proposition says that any of the three conditions could be taken as the definition of a mean dimension reduction subspace. Similar to the TSCS defined in Chapter 2, we now define the smallest mean dimension reduction subspace for time series.

Definition 2 *Let $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})} = \cap\mathcal{S}_M$, where the intersection is over all mean dimension reduction spaces $\mathcal{S}_M$. If $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}$ is itself a mean dimension reduction subspace, then it is called the Time Series Central Mean Subspace (TSCMS).*

Note that TSCMS does not always exist, because the intersection of two mean dimension reduction subspaces is not necessarily a mean dimension reduction subspace. For example, if $p = 4$ and $\mathbf{X}_{t-1} = (x_{t-1}, x_{t-2}, x_{t-3}, x_{t-4})^T$, set $\mathrm{E}(x_t|\mathbf{X}_{t-1}) = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \phi_4 x_{t-4}$. Then the vector $(\phi_1, \phi_2, 0, \phi_4)^T$ forms a basis of TSCMS. On the other hand, when $x_t = \phi_1 x_{t-1}$, we have that $\mathcal{S}((1,0,0,0)^T)$, $\mathcal{S}((0,1,0,0)^T)$, and $\mathcal{S}((0,0,0,1)^T)$ are all minimum mean DRSs. But, there does not exist a TSCMS because $\cap\mathcal{S}_M$ is equal to the origin.

If $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}$ exists, then $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})} \subseteq \mathcal{S}_{x_t|\mathbf{X}_{t-1}}$ because the former is the intersection of a larger collection of subspaces. Hence, it may be possible to reduce the dimension from that of $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}$, if we are only concerned with $\mathrm{E}(x_t|\mathbf{X}_{t-1})$. Under mild conditions, as in Proposition 1 of Chapter 2 (Cook 1998, Page 108), it is also possible to prove the existence of the TSCMS. We assume for the remainder of this chapter that the TSCMS always exists.

The TSCMS is intended to play the same role when considering the conditional mean as the time series central subspace does when inquiring about the full conditional distribution of $x_t|\mathbf{X}_{t-1}$. Methods of estimating $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}$ are discussed in the next section.

3.3 Estimation of TSCMS

Having discussed some basic properties of the TSCMS, $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}$, we now turn our attention to finding population vectors in that subspace. In their pioneering article, Li and Duan (1989) demonstrated that a class of estimators, which includes the Ordinary Least Squares (OLS),

correctly estimate the dimension of the regression parameter regardless of the shape of the regression function, provided that the predictors satisfy a linearity condition. Recently, Cook and Li (2002) adopted this idea for the estimation of CMS in regression. For a description of this approach, consider an objective function of the form

$$R(a, \mathbf{b}) = \mathrm{E}[L(a + \mathbf{b}^T \mathbf{X}_{t-1}, x_t)],$$

where $a \in \mathbb{R}^1$, $\mathbf{b} \in \mathbb{R}^p$, and $L(\cdot, \cdot)$ is strictly a convex function of its first argument. The use of an objective function neither implies that any associated model is true nor that it provides proper fit of the data. Nevertheless, there is an association between the vectors derived from this objective function and $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}$, as shown below.

To this end, let the population minimizers be defined by

$$(\mathbf{\Phi}_0, \mathbf{\Phi}) = argmin_{a,\mathbf{b}} R(a, \mathbf{b}). \tag{3.1}$$

Let η be a basis matrix for $\mathcal{S}_{x_t|\mathbf{X}_{t-1}}$. As shown in Li and Duan (1989), if $\mathrm{E}(\mathbf{X}_{t-1}|\eta^T \mathbf{X}_{t-1})$ is linear in $\mathbf{X}_{t-1}$ and $\dim(\mathcal{S}_{x_t|\mathbf{X}_{t-1}}) = 1$, then $\mathbf{\Phi} \in \mathcal{S}_{x_t|\mathbf{X}_{t-1}}$. Now, using arguments in Cook (1998, page 143-147), it is possible to relax the above dimension restriction and show that $\mathbf{\Phi} \in \mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}$. However, if we restrict the objective function to

$$L(a + \mathbf{b}^T \mathbf{X}_{t-1}, x_t) = -x_t(a + \mathbf{b}^T \mathbf{X}_{t-1}) + \phi(a + \mathbf{b}^T \mathbf{X}_{t-1}) \tag{3.2}$$

based on the natural exponential family for some strictly convex function ϕ, then Φ always belongs to $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}$, as shown below in Theorem 3. From now on, we will refer to the above estimation method as the OLS method.

Theorem 3 *Let γ be a basis matrix for $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}$ and let $\mathbf{\Phi}$ be defined as in (3.1) with the exponential family objective function in (3.2). If $\mathrm{E}(\mathbf{b}^T \mathbf{X}_{t-1}|\gamma^T \mathbf{X}_{t-1})$ is a linear function in $\gamma^T \mathbf{X}_{t-1}$ for all $\mathbf{b} \in \Re^p$, then $\mathbf{\Phi} \in \mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}$.*

Proof of this theorem follows along the lines given in Section 3.6.2 (also see Cook and Li, 2002). As pointed out by Cook and Weisberg in their discussion of Li (1991), the most

important family of distributions satisfying the linearity condition in Theorem 3 is that of elliptically symmetric distributions, in particular multivariate normal distribution. In fact, if $\mathbf{X}_{t-1}$ is multivariate normal, then the linearity condition in Theorem 3 is satisfied. However, even if $\mathbf{X}_{t-1}$ is multivariate normal, x_t need not, in general, be a gaussian time series, as shown in Proposition 4 below.

Proposition 4 *Let* $(x_{t-1}, ..., x_{t-p})^T$ *has joint normal distribution, then* $(x_t, ..., x_{t-p+1})^T$ *has also joint normal distribution. However,* $(x_t, \mathbf{X}_{t-1})$ *has not necessarily normal distribution.*

In view of the above discussion and the discussion in Xia, Tong, Li and Zhu (2002, see page 365), the linearity condition in Theorem 3 is very restrictive for time series situation. Therefore, we propose a more suitable estimation methodology for time series, which is motivated by a method adopted in Xia, Tong, Li and Zhu (2002).

3.3.1 An Estimation Method

In their work on effective dimension reduction in regression, Xia, Tong, Li and Zhu (2002) suggested an adaptive approach based on conditional Minimum Average Variance Estimation (MAVE) method, which is also applicable to the time series context. Unlike the OLS approach, this method does not need strong assumptions on the probability structure of $\mathbf{X}_{t-1}$, nor does it need the linearity condition imposed in Theorem 3. In a related literature, Xia and An (1999) proposed an estimation method based on the idea of projection pursuit, assuming a projection pursuit autoregressive model for time series. In this section, we pursue an estimation method similar to the MAVE method of Xia, Tong, Li and Zhu (2002), but with a different intermediate approach.

As in Xia, Tong, Li and Zhu (2002), we minimize

$$\Psi(\mathbf{\Phi}) = \mathrm{E}(x_t - f(\mathbf{\Phi}^T \mathbf{X_{t-1}}))^2$$

with respect $\mathbf{\Phi}$, where $\mathrm{E}(x_t|\mathbf{X_{t-1}}) = f(\mathbf{\Phi}^T \mathbf{X_{t-1}})$. Since f is assumed to be unknown, we will estimate it using a Nadaraya-Watson estimator (Nadaraya, 1964 and Watson, 1964) defined

by

$$\hat{f}_\lambda(\mathbf{\Phi}^T\mathbf{x}_{t-1}) = \frac{\sum_{i=1}^n K\left(\frac{\mathbf{\Phi}^T\mathbf{x}_{t-1}-\mathbf{\Phi}^T\mathbf{x}_{i-1}}{\lambda}\right) x_i}{\sum_{j=1}^n K\left(\frac{\mathbf{\Phi}^T\mathbf{x}_{t-1}-\mathbf{\Phi}^T\mathbf{x}_{j-1}}{\lambda}\right)},$$

where λ is a bandwidth and K is an appropriate kernel density estimator, to be specified later. With these, we minimize a sample version of $\Psi(\mathbf{\Phi})$ defined by

$$\hat{\Psi}_n(\mathbf{\Phi}) = \sum_{t=1}^{n}(x_t - \hat{f}_\lambda(\mathbf{\Phi}^T\mathbf{x}_{t-1}))^2. \tag{3.3}$$

We will refer to $\hat{\Psi}_n$ as the Residual Sum of Squares (RSS). Although we adopt a similar estimation approach as in Xia, Tong, Li and Zhu (2002), we estimate the unknown mean function using a Nadaraya-Watson estimator whereas Xia, Tong, Li, Zhu (2002) estimate a local linear expansion of the mean function.

In our computations, we use a gaussian kernel for the one-dimensional case and a product gaussian kernel for the multi-dimensional case. Our experience indicates that gaussian kernels work well in the current context. More specifically, let G denote the univariate gaussian kernel, and $\mathbf{u} = (u_1, \cdots, u_k)^T$ be the $k \times 1$ random vector, for $k \geq 1$. Denote the ith observation by $\mathbf{u}_i = (u_{1i}, \cdots, u_{ki})^T$, then the k-dimensional kernel density estimate has the following form:

$$p_n(u_1, \cdots, u_k) = (n\prod_{j=1}^{k} a_{nj})^{-1} \sum_{i=1}^{n}\prod_{j=1}^{k} G\left(\frac{u_j - u_{ji}}{a_{nj}}\right), \tag{3.4}$$

where $a_{nj} = b_k s_j n^{-1/(4+k)}$ for $j = 1, \cdots, k$, and s_j is the corresponding sample standard deviation of u_j, which is updated during each iteration. Including s_j in the bandwidth term is not necessary; however, doing so usually improves the estimation. The constant b_k in a_{nj} is chosen as suggested in Silverman (1986, p. 87) or Scott (1992, p. 152). In all our computations, we use the function '*fmincon*' for minimization, the codes for which are available in *MATLAB*.

3.3.2 Estimation of Dimension d and Lag p

Our development of TSCMS assumes that the minimal dimension d is known. In practice, however, prior information on d may not be available. Here, we will develop a data-dependent

method for estimating d. Also, unlike in regression, there may not be any prior information available on the number of lags, p, and hence it will be useful to develop data-dependent methods to determine p. As mentioned in Chapter 2, in traditional time series analysis one usually uses autocorrelation and partial autocorrelation plots or estimation approaches to determine the lag p.

In time series analysis or, more generally, in any data analysis, several models may be appropriate for a given data set. Therefore, one needs a suitable criterion for model selection. Akaike (1973, 1974) introduced an information criterion known as Akaike's Information Criterion (AIC). It is well known that the AIC tends to overestimate the order of the autoregression (Shibata, 1976). Akaike (1978, 1979) also developed the Baysian information criterion (BIC), which is an extension of the AIC procedure; see Findley (1985) for a detailed discussion on properties of AIC. There are many other criteria for model selection in time series; see, for example, Parzen (1977), Hannan and Quinn (1979), Stone (1979) and Hannan (1980).

Recently, Xia, Tong, Li and Zhu (2002) proposed a Cross Validation (CV) method for estimating the unknown dimension d. But, they do not suggest any method to estimate p. Next, for our context, we propose an estimator of d and p using the Schwarz Baysian information Criterion (SBC) and the Residual Information Criteria (RIC; Shi and Tsai 2002). For an in-depth exposition of these model selection methods, see e.g., Schwarz (1978), Ng and Perron (2005) and Ni, Cook and Tsai (2005).

Note that if $p = 1$, then $d = 1$, and there is no need for dimension reduction. Thus, for a fixed value of lag $p(\geq 2)$, we determine $\hat{d}_p$ using the following SBC and RIC criteria;

$$SBC : \hat{d}_p = \min_{1\leq d\leq p} \{n \log(\hat{\mathbf{\Psi}}_n(\hat{\mathbf{\Phi}}_{p,d})/n) + dp \log(n)\} \tag{3.5}$$

$$RIC : \hat{d}_p = \min_{1\leq d\leq p} \{(n - dp) \log(\hat{\mathbf{\Psi}}_n(\hat{\mathbf{\Phi}}_{p,d})) + dp(\log(n) - 1) + 4/(n - dp - 2)\}. \tag{3.6}$$

For each lag p, we compute SBC and RIC, and obtain $\hat{d}_p$ and the associated $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{\Phi}}_{p,\hat{d}_p})$. Next, plot $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{\Phi}}_{p,\hat{d}_p})$ versus p, which we call an *Elbow Plot*. In such a plot, we will look for the

value of $\hat{p}$ at which $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{\Phi}}_{\hat{p},\hat{d}_p})$ is essentially the smallest. This usually creates an elbow-like situation at $p = \hat{p}$, hence the name *Elbow Plot*. This yields an estimate $\hat{p}$ of lag p. In Section 3.4, we use simulations and a real data set to illustrate how to determine the dimension and detect the lag.

3.4 Simulations and Data Analysis

In this section, we will carry out several simulation studies to demonstrate the performance of our method. All the simulation examples will focus only on the mean function. In addition, we will analyze a data on *Yearly number of lynx pelts sales* to illustrate the performance of our method on a real data. Furthermore, we also compare the forecasts based on our fitted model for the real data with those of another model available in the literature.

3.4.1 Simulations

In all our simulations, we use the measures proposed by Ye and Weiss (2003) and Xia, Tong, Li and Zhu (2002) to assess the accuracy of our estimates. We use the *vector correlation coefficient* (Ye and Weiss, 2003) defined by $\rho = \sqrt{|\hat{\mathbf{\Phi}}_d^T \mathbf{\Phi}_d \mathbf{\Phi}_d^T \hat{\mathbf{\Phi}}_d|}$, where $|\mathbf{A}|$ denotes the determinant of a matrix $\mathbf{A}$. Note that $0 \leq \rho \leq 1$, and when $\rho = 1$, $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}(\hat{\mathbf{\Phi}}_d) = \mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}(\mathbf{\Phi}_d)$. Therefore, higher values of ρ imply that the two spaces are closer, and hence, the estimates are more accurate. On the other hand, the method in Xia, Tong, Li and Zhu (2002) (similar to the one in Li, Zha and Chairomonte, 2005) measures the distance between $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}(\hat{\mathbf{\Phi}}_q)$ and $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}(\mathbf{\Phi}_d)$ using $m^2 = \left\|(I - \mathbf{\Phi}_d \mathbf{\Phi}_d{}^T)\hat{\mathbf{\Phi}}_q\right\|^2$ if $q < d$ and $m^2 = \left\|(I - \hat{\mathbf{\Phi}}_q \hat{\mathbf{\Phi}}_q^T)\mathbf{\Phi}_d\right\|^2$ if $q \geq d$. Here, smaller values of m^2 yield more accurate estimates.

In all our simulations, samples of size $n = 100$, 200 and 300 are considered. For each sample size, accuracy of estimates is based on 100 Monte Carlo replications. The error term $\{\varepsilon_t\}$ in all our models considered below is a sequence of independent standard normal random variables. In each of our simulation study, we first randomly select 100 initial values for model parameters and, for each initial set of values, we minimize the objective function $\hat{\Psi}_n(\mathbf{\Phi})$ in

n	lag p	ρ	m^2
100	2	0.9986	0.0029
200	2	0.9991	0.0017
300	2	0.9994	0.0011

Table 3.1: Model 3.1: Average values of accuracy measures ρ and m^2 based on 100 Monte Carlo replications.

(3.3) and compute the RSS $\hat{\Psi}_n(\hat{\mathbf{\Phi}})$. We will choose that initial value set for which the RSS value is the smallest.

Model 3.1:

Let

$$x_t = 1.56x_{t-1} - 0.56x_{t-2} + \varepsilon_t,$$

where $p = 2$ and $d = 1$. In this AR(2) model, the conditional mean function of the series is linear. We will now estimate $\mathbf{\Phi}_1$ and measure the accuracy of the estimate using the above mentioned methods. Table 3.1 gives the average values of accuracy measures ρ and m^2, respectively. It shows that the average values of ρ are in general close to 1, while those of m^2 are close to zero, implying that our estimates of $\mathbf{\Phi}_1$ are very accurate. Notice that the accuracy of our estimates increases with increasing sample size, as expected.

Model 3.2:

Let

$$\begin{aligned} x_t &= 0.5\{\cos(1.0)x_{t-1} - \sin(1.0)x_{t-2}\} \\ &\quad +0.4\exp[-16\{\cos(1.0)x_{t-1} - \sin(1.0)x_{t-2}\}^2] + 0.1\varepsilon_t, \end{aligned}$$

where $p = 2$ and $d = 1$. Note that the conditional mean function of the series is nonlinear. Table 3.2 gives the average values of ρ and m^2, respectively. Once again, the table shows that the estimates are accurate and the results attest to the fact that our estimation procedure performs reasonably well, even when the conditional mean is nonlinear.

n	lag p	ρ	m^2
100	2	0.9313	0.0817
200	2	0.9443	0.0622
300	2	0.9292	0.0800

Table 3.2: Model 3.2: Average values of accuracy measures ρ and m^2 based on 100 Monte Carlo replications.

Model 3.3:

Let

$$x_t = -1 - \cos((\pi/2)(x_{t-3} + 2x_{t-6})) + 0.2\varepsilon_t,$$

where $p = 6$ and $d = 1$. Table 3.3 shows once again that our estimates of $\mathbf{\Phi}_1$ are accurate. To detect the true dimension d (=1), we use the SBC and RIC criteria in (3.5) and (3.6), for each sample size and lag $p = 6$. In Table 3.3, we report f_i, the frequency of $\hat{d}_p = i$, based on 100 Monte Carlo replications using SBC and RIC. Here, f_{i+} denotes the frequency of $\hat{d}_p \geq i$. For sample sizes $n = 100$, 200 and 300, Table 3.3 shows that, in terms of correctly identifying d, both SBC and RIC perform well with RIC performing slightly better than SBC.

To make inference about p, given $d = 1$, we use the *Elbow Plot* defined above. For this, we compute the average and the standard deviation of $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{\Phi}}_{p,1})$ values for each $p = 4, 5, 6, 7$ based on 100 simulated data sets, each with $n = 300$. The *Elbow Plot* in Figure 3.1, giving the average values and average $\pm$ standard deviation values of $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{\Phi}}_{p,1})$ clearly indicates that the *Elbow* is at $p = 6$. In fact, 89 out of 100 *Elbow Plots* (not given here) indicated that $\hat{p} = 6$. Similar results (not reported here) were also observed for $n = 100, 200$.

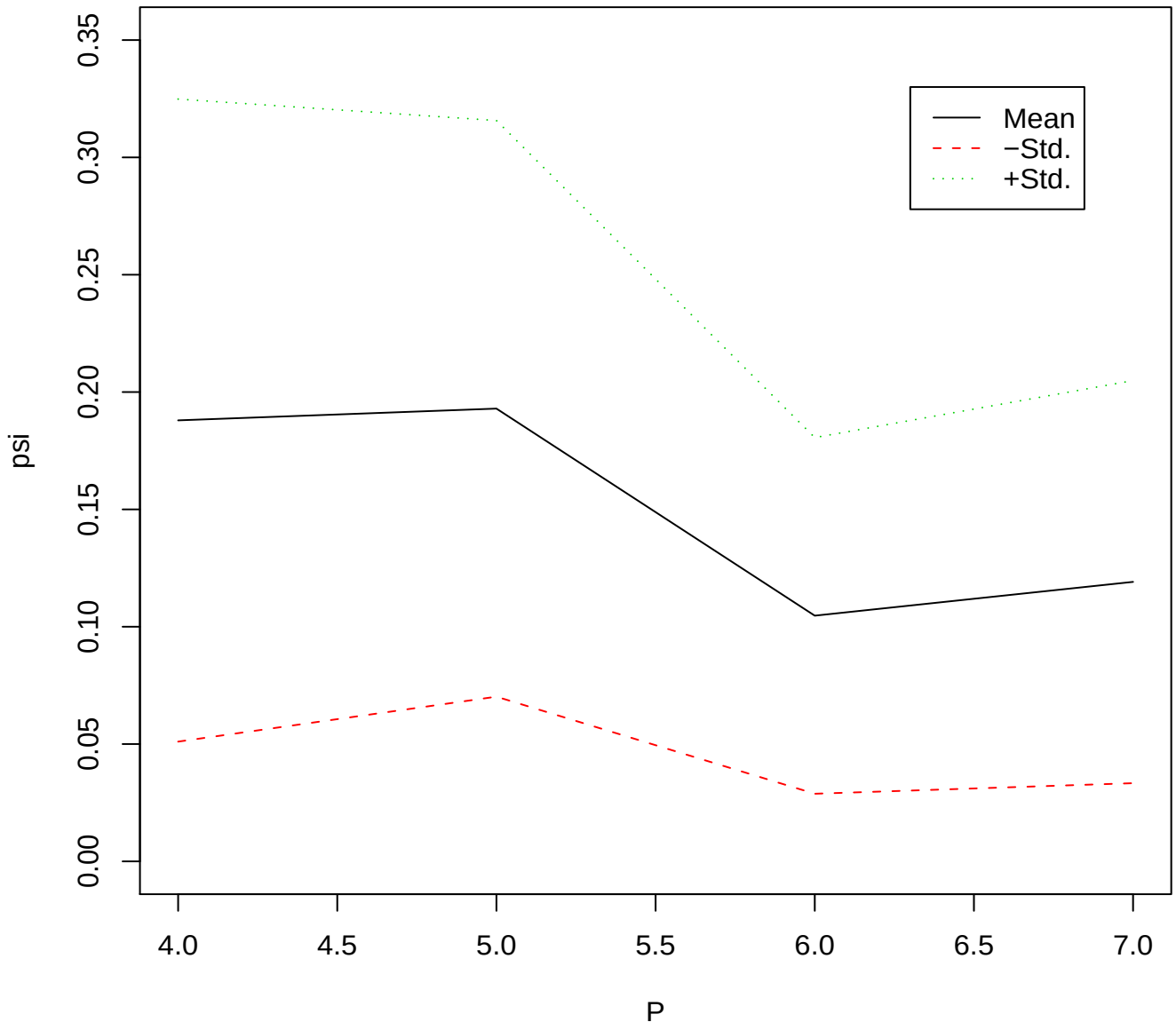

Figure 3.1: Model 3.3: *Elbow plot* of average values ("Mean"), average − standard deviation values ("-Std.") and average + standard deviation values ("+Std.") of $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{\Phi}}_{p,1})$ versus $p = 4, 5, 6, 7$ based on 100 simulated data sets, each with sample size $n = 300$.

n	lag p	ρ	m^2	SBC	RIC
100	6	0.9997	0.0223	$f_1 = 100^*$ f_{2+}=0	$f_1 = 100^*$ f_{2+}=0
200	6	0.9995	0.0010	$f_1 = 94^*$ f_{2+}=6	$f_1 = 100^*$ f_{2+}=0
300	6	0.9973	0.0054	$f_1 = 98^*$ f_{2+}=2	$f_1 = 100^*$ f_{2+}=0

Table 3.3: Model 3.3: Average values of accuracy measures ρ and m^2, and frequency of estimated dimension for SBC and RIC, all based on 100 Monte Carlo replications. The true dimension is $d = 1$.

Model 3.4:

Let

$$x_t = -1 - \cos((\pi/2)(x_{t-1})) - \cos((\pi/2)(1/\sqrt{5})(x_{t-3} + 2x_{t-6})) + 0.2\varepsilon_t,$$

where $p = 6$ and $d = 2$. Table 3.4 shows that the accuracy of the estimates of $\mathbf{\Phi}_2$ is reasonable. As for estimation of d, Table 3.4 shows that RIC correctly estimates the true dimension, $d = 2$, about 86% to 99% of the times, for all sample sizes and when the lag $p = 6$. However, SBC performs worse even for large sample sizes and considerably overestimates the true dimension, when the sample size $n \geq 200$. This seems rather counter-intuitive because SBC is generally more conservative.

Model 3.5:

Let

$$\begin{aligned} x_t &= -1 + (0.4)(1/\sqrt{5})(x_{t-1} + 2x_{t-4}) - \cos((\pi/2)(1/\sqrt{5})(x_{t-3} + 2x_{t-6})) \\ &\quad + \exp(-(1/\sqrt{15})^2(-2x_{t-1} + 2x_{t-2} - 2x_{t-3} + x_{t-4} - x_{t-5} + x_{t-6})^2) + 0.2\varepsilon_t, \end{aligned}$$

where $p = 6$ and $d = 3$. Table 3.5 shows that the accuracy of estimates of $\mathbf{\Phi}_3$ are better for large sample sizes and when the true lag $p = 6$. As for estimating d, Table 3.5 shows that both SBC and RIC correctly estimate the true dimension ($d = 3$) a higher percentage of

n	lag p	ρ	m^2	SBC	RIC
100	6	0.9075	0.0681	f_1=0	f_1=1
			0.0681	f_2=82*	f_2=99*
				f_{3+}=18	f_{3+}=0
200	6	0.9641	0.0301	f_1=0	f_1=0
			0.0301	f_2=47	f_2=99*
				f_{3+}=53	f_{3+}=1
300	6	0.9639	0.0287	f_1=0	f_1=0
			0.0287	f_2=18	f_2=86*
				f_{3+}=82	f_{3+}=14

Table 3.4: Model 3.4: Average values of accuracy measures ρ and m^2, and frequency of estimated dimension for SBC and RIC, all based on 100 Monte Carlo replications. The true dimension is $d = 2$.

n	lag p	ρ	m^2	SBC	RIC
100	6	0.7811	0.1278	f_1=1	f_1=13
			0.1278	f_2=59	f_2=87
			0.1278	f_3=23	f_3=0
				f_{4+}=17	f_{4+}=0
200	6	0.9465	0.0363	f_1=0	f_1=0
			0.0363	f_2=5	f_2=47
			0.0363	f_3=80*	f_3=53
				f_{4+}=15	f_{4+}=0
300	6	0.9683	0.0187	f_1=0	f_1=0
			0.0187	f_2=0	f_2=6
			0.0187	f_3=75*	f_3=94*
				f_{4+}=25	f_{4+}=0

Table 3.5: Model 3.5: Average values of accuracy measures ρ and m^2, and frequency of estimated dimension for SBC and RIC, all based on 100 Monte Carlo replications. The true dimension is $d = 3$.

time, especially as sample size gets larger. However, when the sample size is 100, both SBC and RIC underestimate the true dimension.

Results for models 3.3, 3.4 and 3.5 above seem to suggest that RIC performs better when the true dimension is smaller ($d = 1, 2$) while SBC criteria performs better when the true dimension is larger ($d = 3$).

3.4.2 Yearly number of lynx pelts sales data

In this section, we use a real data set to assess the performance of our method. The data set under discussion gives the yearly number of lynx sales for the years 1857 to 1911, consisting of 55 observations. Andrews and Herzberg (1985) report the data on the number of lynx pelts sold by Hudson's Bay Company in Canada. We begin with a brief background on the data set.

The Canada lynx *Lynx canadensis* is a beautiful wild felid (or cat) of the boreal forest. Like the cougar and the bobcat, the other two members of the cat family (Felidae) native to Canada, the Canada lynx (referred to as lynx) tends to be secretive and most active at night, and like them it is rarely seen in the wild. The Canada lynx, long haired and of lustrous flare when prime, is a valuable fur-bearer, though it is currently out of fashion and sells for a mere $60 a pelt (fur). Records from the early days of the fur trade are scarce, but for one year, 1763, there are records showing that 4,150 lynx pelts were exported from Canada to England, comprising a mere 2% of the furs in trade that year (Poland 1892). However, records of purchases of Canada lynx pelts by the Hudson's Bay Company during the 19th century attest to the growing popularity of cat pelts. Peak harvests were of the order of 80,000 pelts annually in the late 1800s, but they declined sharply after the turn of the century (Elton and Nicholson, 1942). In the early 1900s, approximately 64,000 bobcat and lynx pelts were sold annually in the United States, the world's largest fur market for much of this century (Osborn and Anthony, 1922). Following the Depression and World War II, the fur trade's source of supply underwent a major shift from mainly wild-trapped to

mainly ranched animals (IFTF 1989). Cats, however, are not ranched, and their proportion within the wild-caught minority of furbearers increased dramatically during the 1960s.

Wei (2006) analyzes this data set in Example 6.7 of his book, where he models the logarithm of the series. Based on traditional approach using ACF and PACF, Wei (2006) concludes that an AR(3) model is adequate for the data. For our data analysis, we consider the log transformed series of lynx pelt sales for the years 1857 to 1906 with $n = 50$. Our process begins with estimation of d and p, followed by the estimation of TSCMS. We then build a model, based on which we compute the lynx pelt sales number forecasts for the remaining five years: 1907 to 1911. We also fitted an AR(3) model for the year 1857 to 1906, and obtained forecasts for the remaining five years: 1907 to 1911.

To estimate d, we set $p = 2, 3, 4$, and 5 and compute SBC and RIC values for $d = 1, 2, 3$, and 4. Table 3.6 lists the SBC and RIC values for each p and d, except for the trivial case $p = 1$. Based on SBC and RIC criteria, we conclude that $\hat{d}_p = 1$ for all $2 \leq p \leq 5$. Note that $\hat{d}_p$ for each p is indicated by an asterisk in Table 3.6. Since $\hat{d}_p = 1$ uniformly for all p, we decided to use $d = 1$ in order to determine $\hat{p}$. The *Elbow Plot* in Figure 3.2, based on $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{\Phi}}_{p,1})$ values for $p = 2, 3, 4$, and 5, clearly indicates that the *Elbow* is at $p = 4$. From Table 3.6 and Figure 3.2, we conclude that the minimal dimension is $d = 1$ and the lag is $p = 4$ for our real data.

Setting $d = 1$ and $p = 4$, we use our estimation method in (3.3) to obtain an estimate, $\hat{\mathbf{\Phi}}_1$, of the 4×1 basis matrix $\mathbf{\Phi}_1$ in $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}(\mathbf{\Phi}_1)$. Then, we use the estimates and trial-and-error approaches with plots of x_t against $\hat{\mathbf{\Phi}}_1^T\mathbf{X}_{t-1}$ to build a time series model. Finally, all this led us to fit a linear time series model given by

$$\begin{aligned} \hat{x}_t &= 9.45 - 0.78\hat{\mathbf{\Phi}}_1^T\mathbf{X}_{t-1} \\ &= 9.45 - 0.57x_{t-1} - 0.14x_{t-2} + 0.19x_{t-3} + 0.48x_{t-4}. \end{aligned}$$

To compare the performance of our model with the AR(3) model, we use the forecasts for the lynx pelt sales numbers for the years 1907 to 1911 based on our model and AR(3) model, and compute the Mean Square Relative Error (MSRE)= $k^{-1}\sum_{t=1}^{k}(z_t - \hat{z}_t)^2/z_t$ and

p (SBC)	d=1	d=2	d=3	d=4
2	-1.2026^*	-1.1459		
3	-1.4718^*	-1.4458	-1.3713	
4	-1.6376^*	-1.5500	-1.3597	-1.1561
5	-1.5951^*	-1.4497	-1.2124	-0.8979
p (RIC)	d=1	d=2	d=3	d=4
2	-57.3648^*	-51.5332		
3	-68.4724^*	-61.0805	-50.6462	
4	-74.1494^*	-60.6710	-41.8626	-22.0403
5	-69.9898^*	-51.1338	-27.2237	-0.1745

Table 3.6: Yearly number of lynx pelts sales data: SBC and RIC values for $p = 2, 3, 4$ and 5, and $d = 1, 2, 3$ and 4. For each p, $\hat{d}_p$ determined by smallest value is denoted by * in the table.

Year	lynx pelts sales	AR(3)	Our model
1907	61478	51224.0745	54661.7679
1908	36300	32181.6533	34838.2186
1909	9704	16659.4772	18029.1375
1910	3410	9966.4659	9337.1632
1911	3774	8086.3704	6606.8272
MSRE		4939.3398	4077.1181
MARE		0.8125	0.6996

Table 3.7: Observed yearly number of lynx pelts sales, forecasts from AR(3) and our model, and MSRE and MARE for each model: Years 1907 - 1911.

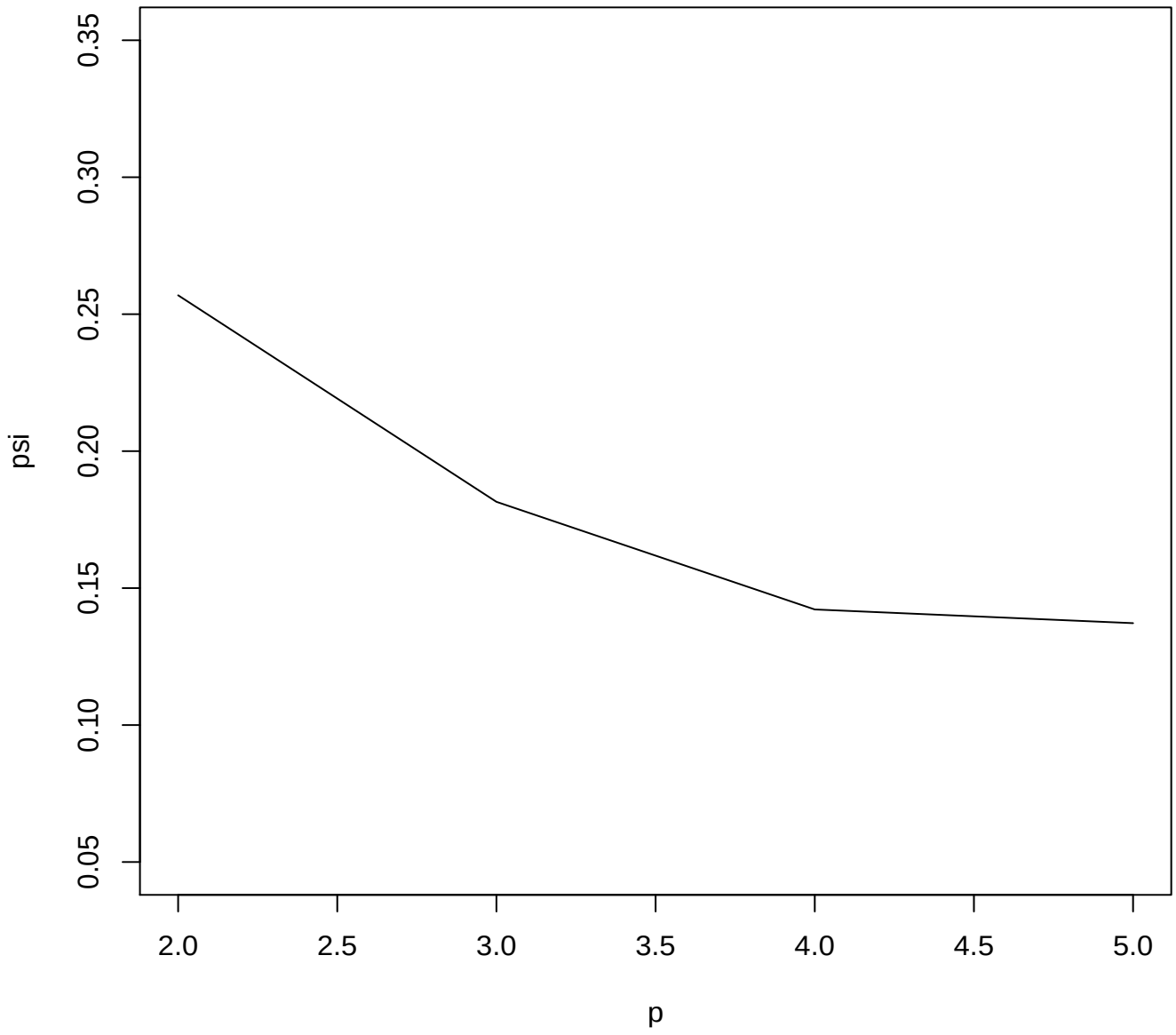

Figure 3.2: Yearly number of lynx pelts sales data: *Elbow plot* of $\hat{\mathbf{\Psi}}_n(\hat{\mathbf{\Phi}}_{p,1})$ values versus $p = 2, 3, 4, 5$.

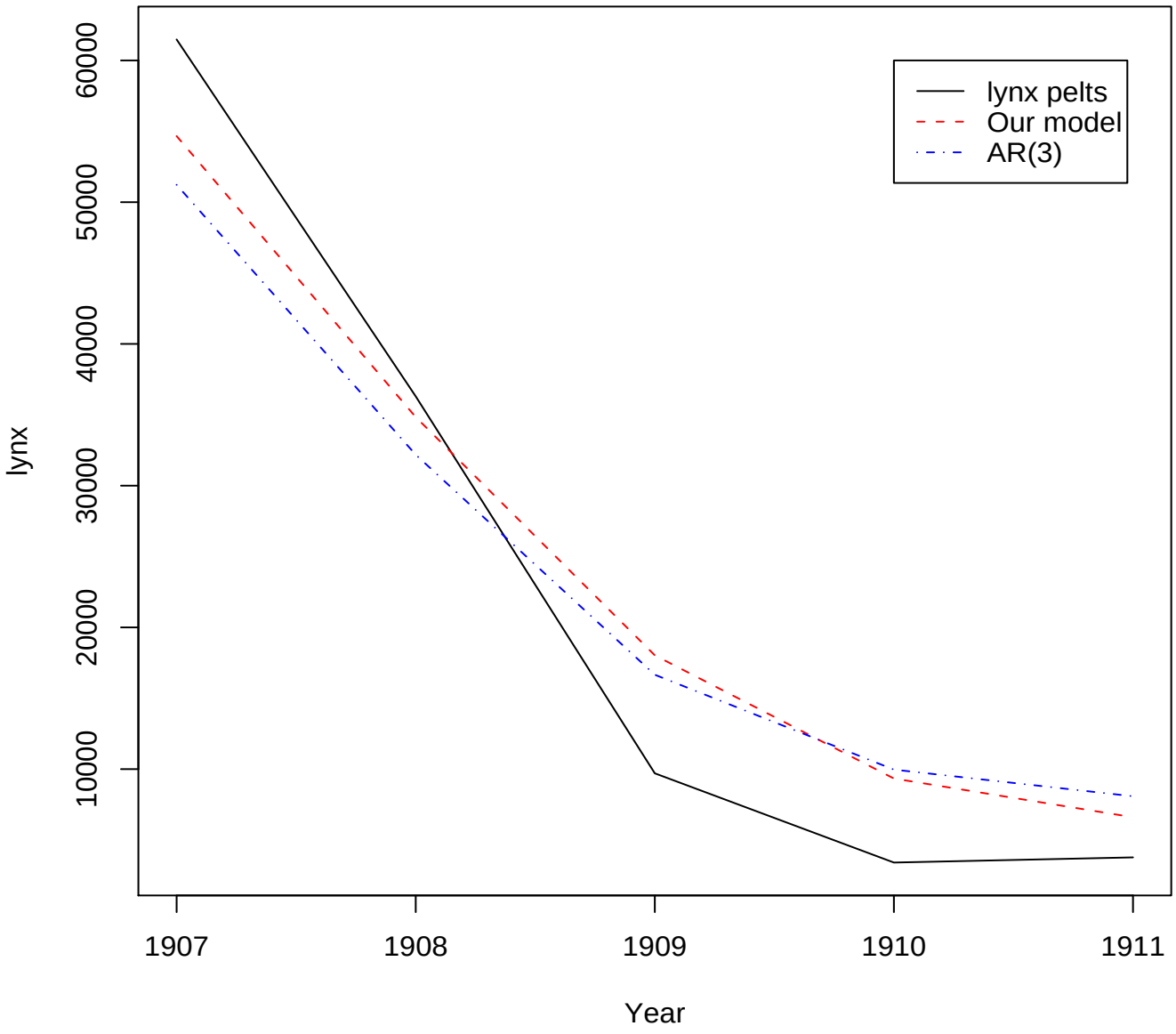

Figure 3.3: Yearly number of lynx pelts sales data: Overlay plot of observed lynx pelts sales number (lynx pelts) and forecast values from AR(3) and our model: Years 1907-1911.

the Mean Absolute Relative Error (MARE)= $k^{-1}\sum_{t=1}^{k}|z_t-\hat{z}_t|/z_t$, where z_t is the observed the lynx pelt number, $\hat{z}_t$ is its forecasted value and k is the number of (future) observations. As is seen from Table 3.7, the MSRE values for AR(3) and our model are 4939.3398 and 4077.1181, respectively, and the MARE values for AR(3) and our model are 0.8125 and 0.6996, respectively. Note that our model produces MSRE and MARE values, which are smaller than those of the AR(3) model. An overlay plot of forecast values from each of the above models and the observed lynx pelt sales numbers is given in Figure 3.3. This shows that the forecasts based on our model are better than those based on Wei's AR(3) model. All these show that our time series central mean subspace approach is a promising tool for time series analysis.

3.5 Discussion

In this chapter, we developed a new approach for dimension reduction in time series, which focuses only on the mean function of time series. After introducing the notion of TSCMS, we proposed an estimation method, which uses the Nadaraya-Watson kernel estimator (Nadaraya, 1964 and Watson, 1964) to estimate the mean function, and thereby estimate the TSCMS, when the lag p and the minimal dimension d are known. As in Chapter 2, we proposed a data dependent method based on criteria referred to as SBC and RIC in order to estimate the minimal dimension d, given lag p. We also proposed a graphical method to detect lag p, referred to as *Elbow Plot*.

Overall, the theory of mean dimension reduction subspace in time series poses many challenges, but a variety of encouraging results presented through our simulations seem to suggest that our method has great potential for providing a viable and meaningful alternative to traditional time series analysis, when only mean function is of interest. In fact, superior performance of our linear time series model for the yearly number of lynx pelts sales data, as compared to Wei's (2006) model, serves as a testament that our method is very useful in time series analysis.

It would be of interest to establish some theoretical properties of the estimates of TSCMS, as done in Chapter 2. This would entail development of new theoretical concepts. Hence, it will be pursued later. While we do not claim the superiority of our method over traditional time series methods, we do believe that our method will open new avenues in time series data analysis.

3.6 Appendix

3.6.1 PROOF OF PROPOSITION 3

It is obvious that part (i) implies part (ii). If $\mathrm{E}(x_t|\mathbf{X}_{t-1})$ is a function of $\mathbf{\Phi}^T\mathbf{X}_{t-1}$, then, given $\mathbf{\Phi}^T\mathbf{X}_{t-1}$, $\mathrm{E}(x_t|\mathbf{X}_{t-1})$ is a constant. Hence, it is independent of any other random variable. Therefore, clearly we have that part (iii) implies part (i). Now, we need to prove that part (ii) implies part (iii). From part (ii) we have that

$$\mathrm{E}[x_t\mathrm{E}(x_t|\mathbf{X}_{t-1})|\mathbf{\Phi}^T\mathbf{X}_{t-1}] = \mathrm{E}(x_t|\mathbf{\Phi}^T\mathbf{X}_{t-1})\mathrm{E}[\mathrm{E}(x_t|\mathbf{X}_{t-1})|\mathbf{\Phi}^T\mathbf{X}_{t-1}]. \tag{3.7}$$

Since

$$\begin{aligned}\mathrm{E}(x_t|\mathbf{\Phi}^T\mathbf{X}_{t-1}) &= \mathrm{E}[\mathrm{E}(x_t|\mathbf{X}_{t-1}, \mathbf{\Phi}^T\mathbf{X}_{t-1})|\mathbf{\Phi}^T\mathbf{X}_{t-1}] \\ &= \mathrm{E}[\mathrm{E}(x_t|\mathbf{X}_{t-1})|\mathbf{\Phi}^T\mathbf{X}_{t-1}],\end{aligned}$$

the right hand side of (3.7) is $\{\mathrm{E}[\mathrm{E}(x_t|\mathbf{X}_{t-1})|\mathbf{\Phi}^T\mathbf{X}_{t-1}]\}^2$. Similarly, the left hand side of (3.7) is

$$\begin{aligned}\mathrm{E}[x_t\mathrm{E}(x_t|\mathbf{X}_{t-1})|\mathbf{\Phi}^T\mathbf{X}_{t-1}] &= \mathrm{E}\{\mathrm{E}[x_t\mathrm{E}(x_t|\mathbf{X}_{t-1})|\mathbf{X}_{t-1}, \mathbf{\Phi}^T\mathbf{X}_{t-1}]|\mathbf{\Phi}^T\mathbf{X}_{t-1}\} \\ &= \mathrm{E}\{\mathrm{E}[x_t\mathrm{E}(x_t|\mathbf{X}_{t-1})|\mathbf{X}_{t-1}]|\mathbf{\Phi}^T\mathbf{X}_{t-1}\} \\ &= \mathrm{E}\{[\mathrm{E}(x_t|\mathbf{X}_{t-1})]^2|\mathbf{\Phi}^T\mathbf{X}_{t-1}\}.\end{aligned}$$

Therefore, $Var[\mathrm{E}(x_t|\mathbf{X}_{t-1})|\mathbf{\Phi}^T\mathbf{X}_{t-1}] = 0$, which implies that, given $\mathbf{\Phi}^T\mathbf{X}_{t-1}$, $\mathrm{E}(x_t|\mathbf{X}_{t-1})$ is a constant. Therefore, part (ii) implies part (iii) and hence the proposition.

3.6.2 PROOF OF THEOREM 3

Let γ be a basis for the central mean subspace:

$$\begin{aligned} R(a,\mathbf{b}) &= \mathrm{E}[-x_t(a+\mathbf{b}^T\mathbf{X}_{t-1})+\phi(a+\mathbf{b}^T\mathbf{X}_{t-1})] \\ &= \mathrm{E}[-\mathrm{E}(x_t|\gamma^T\mathbf{X}_{t-1})(a+\mathbf{b}^T\mathbf{X}_{t-1})+\phi(a+\mathbf{b}^T\mathbf{X}_{t-1})] \\ &\geq \mathrm{E}[-\mathrm{E}(x_t|\gamma^T\mathbf{X}_{t-1})(a+\mathbf{b}^T\mathrm{E}(\mathbf{X}_{t-1}|\gamma^T\mathbf{X}_{t-1}))+\phi(a+\mathbf{b}^T\mathrm{E}(\mathbf{X}_{t-1}|\gamma^T\mathbf{X}_{t-1}))] \\ &= \mathrm{E}[-x_t(a+\mathbf{b}^T\mathbf{P}_\gamma\mathbf{X}_{t-1})+\phi(a+\mathbf{b}^T\mathbf{P}_\gamma\mathbf{X}_{t-1})]. \end{aligned}$$

Since γ is a basis for $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}$, the second line is obvious. The inequality of third line is caused by convexity which is $\mathrm{E}[\phi(a+\mathbf{b}^T\mathbf{X}_{t-1})] \geq \phi(a+\mathbf{b}^T\mathrm{E}[\mathbf{X}_{t-1}|\gamma^T\mathbf{X}_{t-1}])$. The last line derives from the linearity of $\mathrm{E}(\mathbf{X}_{t-1}|\gamma^T\mathbf{X}_{t-1})$, that is, $\mathrm{E}(\mathbf{X}_{t-1}|\gamma^T\mathbf{X}_{t-1}) = \mathbf{P}_\gamma\mathbf{X}_{t-1}$, where $\mathbf{P}_\gamma$ is the projection on $\mathcal{S}_{\mathrm{E}(x_t|\mathbf{X}_{t-1})}$ regarding the usual inner product. Therefore, $R(a,\mathbf{b}) \geq R(a,\mathbf{P}_\gamma\mathbf{b})$ and the conclusion is the result of the unique $\mathbf{\Phi}$.

3.6.3 PROOF OF PROPOSITION 4

Here, consider a counterexample when $p = 2$. Let $(x_t, x_{t-1}, x_{t-2})^T$ and consider

$$x_t = \begin{cases} -x_{t-2} & -1 \leq x_{t-2} \leq 1 \\ x_{t-2} & otherwise \end{cases},$$

where $x_{t-1} \perp\!\!\!\perp x_{t-2}$, x_{t-1} and $x_{t-2} \sim N(0,1)$. Then, $x_t \sim N(0,1)$ by the Exercise 4.8(a) of Johnson and Wichern (2002).

From the Exercise 4.8(b) of Johnson and Wichern (2002), let x_t and x_{t-2} be a bivariate normal distribution. $\mathbf{K} = (x_{t-2}, x_t)^T \sim N(\mu, \Sigma)$ where $\mu = (0,0)^T$. Suppose $\mathbf{a} = (1,-1)^T$, then $\mathbf{a}^T\mathbf{K} = x_{t-2} - x_t$. Since $\mathbf{a}^T\mu = \mathbf{0}$ and

$$\mathbf{a}^T\Sigma\mathbf{a} = (1-1)\begin{pmatrix} 1 & \sigma_{21} \\ \sigma_{12} & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = 2 - \sigma_{21} - \sigma_{12},$$

$\mathbf{a}^T\mathbf{K} = x_{t-2} - x_t \sim N(0, 2 - 2\sigma_{21})$. Here, $P(x_{t-2} - x_t = 0) = 0$ by continuity. However, from initial assumption,

$$x_{t-2} - x_t = \begin{cases} 2x_{t-2} & -1 \leq x_{t-2} \leq 1 \\ 0 & otherwise \end{cases}$$

and $P(x_{t-2} - x_t = 0) = P(|x_{t-2}| > 1) = 0.3174$ from $x_{t-2} \sim N(0,1)$. Then, we directly got a contradiction that x_t and x_{t-2} do not have a bivariate normal distribution. The result leads that $(x_t, x_{t-1}, x_{t-2})^T$ does not have a joint normal distribution. Hence, the conclusion follows.

3.7 References

[1] Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," *Proceeding 2nd International Symposium on Information Theory* (Eds. B. N. Petrov and F. Csaki), 267–281, Akademiai Kiado, Budapest.

[2] Akaike, H. (1974), "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, AC-19, 716–723.

[3] Akaike, H. (1978), "A bayesian analysis of the minimum AIC procedure," *Annals of the Institute of Statistical Mathematics*, 30A, 9–14.

[4] Akaike, H. (1979), "A bayesian extension of the minimum AIC procedure of autoregressive model fitting," *Biometrika*, 66, 237–242.

[5] An, H. Z. and Huang, F. c. (1996), "The geometrical ergodicity of nonlinear autoregressive models," *Statistica Sinica*, 6, 943–956.

[6] Andrews, P. E., and Herzberg, A. M. (1985), *The Data: A Collection of Problems from Statistics*, Berlin: Springer-Verlag.

[7] Brockwell, P. J. and Davis, R. A. (1991), *Time Series: Theory and Methods*, New York: Springer-Verlag.

[8] Chan, K. S. and Tong, H. (1994), "A note on noise chaos," *Journal of the Royal Statistical Society*, Ser. B, 56, 301–311.

[9] Cook, R. D. and Li, B. (2002), "Dimension reduction for the conditional mean in regression," *Annals of Statistics*, 30, 455–474.

[10] Cook, R. D., and Ni, L. (2005), "Sufficient Dimension Reduction via Inverse Regression: A Minimum Discrepancy Approach," *Journal of the American Statistical Association*, 100, 410–428.

[11] Eubank, R. (1999), *Nonparametric regression and spline smoothing*, New York: Marcel Dekker.

[12] Findley, D. F. (1985), "On the unbiased property of AIC for exact or approximating linear stochastic time series models," *Journal of Time Series Analysis*, 6, 229–252.

[13] Johnson, R. A. and Wichern, D. W. (2002), *Applied Multivariate Statistical Analysis*, London: Prentice-Hall.

[14] Hannan, E. J. (1980), "The estimation of the order of an ARMA process," *Annals of Statistics*, 8, 1071–1081.

[15] Hannan, E. J. and Quinn, B. G. (1979), "The determination of the order of an autoregression," *Journal of the Royal Statistical Society*, Ser. B, 41, 190–195.

[16] Hristache, M., Juditsky, A., Polzehl, J., and Spokoiny, V. (2001), "Structure Adaptive Approach for Dimension Reduction," *The Annals of Statistics*, 29, 1537.1566.

[17] Hotelling, H. (1936), "Relations between two sets of variates," *Biometrika*, 28, 321–377.

[18] Li, B. Zha, H. and Chiaromonte, C. (2005). "Contour regression: a general approach to dimension reduction," *The Annals of Statistics*, 33, 1580–1616.

[19] Li, K. C. (1991), "Sliced inverse regression for dimension reduction (with discussion)," *Journal of the American Statistical Association*, 86, 316–342.

[20] Li, K. C. (1992), "On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Steins Lemma," *Journal of the American Statistical Association*, 87, 1025.1039.

[21] Li, K. C. and Duan, N. (1989), "Regression analysis under link violation," *The Annal of Statistics*, 17, 1009–1952.

[22] Nadaraya, E. (1964), "On estimating regression," *Theory of Probability and Its Applications*, 9, 141–142.

[23] Ng, S. and Perron, P. (2005), "A note on selection of time series models," *Oxford Bulletin of Economics and Statistics*, 67, 115–134.

[24] Ni, L., Cook, R. D. and Tsai, C. L. (2005), "A note on shrinkage sliced inverse regression," *Biometrika*, 92, 242–247.

[25] Parzen, E. (1977), "Multiple time series modeling: Determining the order of approximating autoregressive schemes," *Multivariate Analysis IV* (Ed. P. Krishnaiah), 283–295, North-Holland, Amsterdam.

[26] Schwarz, G. (1978), "Estimation the dimension of a model," *The Annal of Statistics*, 6, 461–464.

[27] Shi, P. and Tsai, C. L. (2002), "Regression model selection-a residual likelihood approach," *Journal of the Royal Statistical Society*, Ser. B, 64, 237–252.

[28] Shibata, R. (1976), "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, 63, 117–126.

[29] Stone, M. (1979), "comments on model selection criteria of Akaike and Schwarz," *Journal of the Royal Statistical Society*, Ser. B, 41, 276–278.

[30] Tjstheim, D. (1990), "Non-linear time series and Markov chain," *Advances in Applied Probability*, 22, 587–611.

[31] Xia, X. and An, H. Z. (1999), "Projection pursuit autoregression in time series," *Journal of Time Series Analysis*, 37, 1–16.

[32] Xia, Y., Tong, H., Li, W. K., and Zhu, L. X. (2002), "An adaptive estimation of dimension reduction," *Journal of the Royal Statistical Society*, Ser. B, 64, 363–410.

[33] Watson, G. (1964), "Smooth regression analysis," *Sankhya*, Ser. A, 26, 359–372.

[34] Ye, Z. and Weiss, R. E. (2003), "Using the bootstrap to select one of a new class of dimension reduction methods," *Journal of the American Statistical Association*, 98, 968–979.

[35] Yin, X. and Cook, R. D. (2002), "Dimension Reduction for the Conditional kth Moment in Regression," *Journal of the Royal Statistical Society*, Ser. B, 64, 159–175.

[36] Zhu, Y. and Zeng, P. (2006), "Fourier Methods for Estimating the Central Subspace and the Central Mean Subspace in Regression," *Journal of the American Statistical Association*, 101, 1638–1651.

Chapter 4

Conclusion

In this thesis, we developed a new theory of dimension reduction in time series, which provides an initial phase when an adequate parsimoniously parametrized time series model is not yet available. We introduced two notions such as TSCS and TSCMS for the purpose of dimension reduction in time series. While the notion TSCS captures information about x_t contained in $\mathbf{X}_{t-1}$ through minimal number of linear combinations of $\mathbf{X}_{t-1}$, the notion TSCMS captures information about x_t contained in $\mathrm{E}(x_t|\mathbf{X}_{t-1})$ through minimal number of linear combinations of $\mathbf{X}_{t-1}$. Due to these differences, we needed to adopt different estimation techniques for TSCS and TSCMS, respectively. Nevertheless, we illustrated extensively via Monte Carlo simulations and data analysis that our methods provide a viable alternative to the traditional time series methods.

We also considered issues intrinsic to the nature of a time series, such as estimation of lag p. To this end, we proposed use of a *Shoulder Plot* and an *Elbow Plot* to determine lag p in the case of TSCS and TSCMS, respectively. In addition, we also proposed estimators to make inference about the minimum dimension d associated with TSCS/TSCMS. In the case of TSCS, we theoretically showed strong consistency of our estimator of d, whereas we used SBC and RIC criteria for making inference about d in the case of TSCMS.

Overall, the theory of dimension reduction in time series poses many challenges, but a variety of encouraging results presented through our simulations seem to suggest that our method has great potential for providing a viable and meaningful alternative to traditional time series analysis. In the case of TSCS, superior performance of our nonlinear time series

model for Wolf yearly sunspot data, as compared to Wei's (2006) models, serves as a testament that our method is very useful in time series analysis. Also, for the seasonal U.S beer production data, our model performs better than those in Wei (2006). In the case of TSCMS, superior performance of our linear time series model for the yearly number of lynx pelts sales data, as compared to Wei's (2006) model, shows that our method is very useful in time series analysis. We believe that the sufficient dimension reduction approaches proposed in the thesis will stimulate new ideas for modeling time series.

Some future work may focus on different kernels. In this thesis, we only used gaussian kernel but other kernels such as Epanechnikov (Silverman, page 43) may be used in this direction. We used Nadaraya-Watson estimator in Chapter 3 but other estimators (Eubank, page 159) may be considered in the future research.