Smoothness Priors Analysis of Economic and Financial Time Series

Yoshinori Kawasaki

Contents

Chapter 1

Introduction

1.1 Overview

The aim of this thesis is to present new methods for modeling economic and financial time series using smoothness priors, and to demonstrate the usefulness of smoothness prior analysis in line with real problems. In model selection, this paper consistently employ the information criteria statistics. This chapter briefly reviews the contents of each chapter, and then survey the concept and history of smoothness priors analysis. Finally, we give a broad overview on the construction of various information criteria in the subsequent section.

The use of smoothness priors concept in time series analysis can be found in two major applications. The first one is so-called unobservable components model where a single time series is decomposed into several latent time series. In this problem, smoothness priors are the reflection of stochastic constraints on the unobserved components. The second problem is time-varying coefficients model where the coefficients (say, of a linear regression model) evolve smoothly as time elapses, so we formulate such smoothness by the stochastic constraints on the adjacent coefficients. For these problems, it has been a common knowledge since the late 1970's that the state-space formulation and the recursive formula like the Kalman filter are of advantage both in representation and estimation of the models. This thesis conforms with and tries to extend such a research stream.

From Chapter 2 to Chapter 5, we discuss the seasonal adjustment model which is one of the most important decomposition model in the economic time series analysis. Chapter 2 focuses on the 'optimality' of the seasonal adjustment procedures. If the sample power spectrum of seasonally adjusted series has troughs or dips at the seasonal frequency and its harmonics, the practitioners often make an assertion that the seasonal components are excessively removed from the original time series. It is known, however, that such seasonal dips are not the evidence of 'over-adjustment' but the characterization of a optimal seasonal adjustment procedure where

the optimality is defined in terms of minimum mean squared errors criterion. Laying stress on this point, Chapter 2 clarifies the mechanism of the appearance of seasonal dips. Given the minimum MSE seasonal adjustment procedure, the seasonal dips exist in theory, but it actually depends whether we could observe the dips. This paper digs up that the ratio of the dispersion of smoothness prior (on the seasonal component) and the observational noise variance determines the appearance and disappearance of seasonal dips. Published articles related to Chapter 2 are Kawasaki and Sato (1997a), Kawasaki (1997b).

Commonly, the (unobserved) seasonal component satisfies the relation that the sum of one period of seasonal components should be nearly zero, and is assumed to follow a zero mean random variable with finite variance. Let s_t be seasonal component, and s be the number of seasons in a period. Then what is usually assumed is $\sum_{j=0}^{s-1} s_{t-j} = v_t$ for some zero mean random variable v_t. It is easy to see that the characteristic equation of this seasonal component model has $s-1$ roots on the unit circle in the complex plane. Hence the standard seasonal component model can be regarded as the 'full' unit root model. Chapters 3, 4 and 5 extend the standard seasonal component model.

In Chapter 3, we propose the ARMA based seasonal component model while the standard seasonal component model is purely AR-based. Keeping the AR part unchanged, we add the MA part which is driven by a single extra parameter, and present a state-space representation for the model. Illustrative examples show that some flexibility has been brought to the seasonal component, whereas the seasonal pattern derived from the standard seasonal model is almost fixed as dummy variable. In many cases, the value of ABIC shows the superiority of the proposed model. To visually inspect whether the introduction of MA term is successful or not, a graphical representation for the estimated models is also proposed. Related publication to Chapter 3 is Kawasaki (2003).

In Chapter 4, a model selection approach to detect seasonal unit roots is proposed, and Chapter 5 examines its usefulness in terms of the empirical goodness of out-of-sample forecasting. Some roots of the characteristic equation of seasonal component model are allowed to be located outside the unit circle, that is, some of the cyclical elements that constitute seasonality can be stationary. We compare the information criterion statistics and decide whether a certain coefficient of seasonal polynomial is unity or less than unity. This procedure is motivated by the well-known shortcoming of widely used seasonal unit root test by Hylleberg et al. (1990). In general, unit root tests assume that the autocorrelation structure of the data generating process can be moderately approximated by a finite order AR process, but the tests deteriorate if the true innovation in the data generating process has a moving average term and its characteristic

roots lie near the unit circle. To make things worse, many empirical works suggest that ARIMA $(0,1,1) \times (0,1,1)_s$ model (so-called 'airline model') accounts very well for typical macroeconomic time series. This paper notes the similarities of autocorrelation functions between the structural time series model and the airline model, and compare the full unit root model and the partially stationary root model by such information criteria AIC and BIC. Our simulation results show that the procedure detects the true data generating process reasonably well unless the true parameters stay extremely close to unity. The published article related to Chapter 4 is Kawasaki and Franses (2003a).

The arguments in Chapter 4, basically designed for quarterly time series, are extended to accommodate monthly time series in Chapter 5. Because the direct extension of the quarterly procedure causes combinatorial explosion in monthly case, a synthesis of partial inference is proposed. Numerical experiments show that the synthesis of partial inference does not invoke serious deterioration of the detection probabilities compared to the simulation results from estimating all possible models. The usefulness of the proposed method is validated by the out-of-sample forecasting of industrial production series of various countries. Our empirical results show that the coefficients which corresponds to higher frequency seasonality tend to be judged to be less than unity. The models that allow partially stationary seasonality sometimes attain more than 20% reduction in MSE compared to the full unit root seasonal model. The related publication is Kawasaki and Franses (2003b).

In Chapter 6, we keep a little bit away from modeling seasonality, and discuss the various frameworks that enables us to extract latent common factor process from multiple time series. For this purpose, the theory and methods for principal component and factor analysis for multivariate time series are surveyed. This chapter is so to say a two dimentional plane, which is divided into four orthants by 'principal component analysis – factor analysis' coordinate and by 'frequency domain – time domain' coordinate. Principal component/factor analysis in frequency domain utilize the discrete Fourier transformation (DFT) of time series. Factor analysis in time domain generally gives explicit models to the latent factor process. On the other hand, a formal application of principal component analysis in time domain should be handled with care because such an approach discards the information contained in time series dependency, and is just focusing on the correlation of a multivariate data given at a fixd time. As an application, the structural time series approach is employed to build the bivariate model which aims to forecast inflation by extracting 'output gap' series as a common factor. Chapter 6 is based on the published article Kawasaki (2001).

Chapter 7 offers new models that enable efficient estimation of vector-valued regression

models with time-varying coefficients. The procedure is applied to forecasting excess stock returns. Unusual feature of the problem considered here is that the dimension of the observation vector (the number of stocks observed at a time) is much larger than the dimension of explanatory variables which is typically less than 10. We demonstrate that under the presence of such a big gap in the dimensions of state and observation the use of information or information/square root filter is indispensable. Furthermore, it turns out that special modeling is required to make it feasible the inversion of of the covariance matrix of observational error. This paper compares three different models in terms of information criteria, trading simulations, and the appropriateness of the interpretation of the estimated time-varying coefficients. What we call 'temporal effect model' is concluded to be the best. Chapter 7 is related to Kawasaki, Sato and Tachiki (1998, 2000).

Chapter 8 develops the estimation and evaluation procedure for nonlinear spline models with smoothness priors, which is applied to the estimation of yield curves from observed coupon bond prices. In yield curve estimation, even a small estimation error in the discount function leads to large error in the zero coupon yield curve and in the forward rate curve. To avoid this difficulty, spline bases are sometimes placed directly on the zero coupon yield or the forward rate. It is easy to see that this strategy is equivalent to fitting exponential spline models to the discount function. Some special cares must be taken to the determination of the magnitude of penalty term because the model assumed here is nonlinear with respect to the unknown parameters. We employ penalized likelihood approach and construct the generalized information criterion (GIC) to determine the roughness penalty and the number of basis functions. The tailor-made information criteria are derived for all the models considered in the article. The results of simulations and data analysis show that the exponential spline with integrated B-spline basis and McCulloch's natural cubic spline (both regularized) perform good. It is also shown that choosing the number of basis by GIC leads to better performance than placing abundant basis function to be controlled by the roughness penalty. Chapter 8 is based on Kawasaki and Ando (2002, 2003).

Chapter 9 discusses the modeling of intra-day periodicity in high frequent data in finance. Occurrence of financial transaction is said to have cluster effects as well as for the volatility. Autoregressive Conditional Duration models are often employed to investigate such a herding effect in transactions. Prior to the fitting of ACD model, a time of day function has to be estimated to eliminate the intra-day periodicity. As the most representative procedure, each duration time are classified into the equal interval bins of one day, and spline smoothing is performed for the average duration time of bins. Chapter 9 proposes a diagnostic procedure

to judge if the intra-day periodicity adjustment by the smoothing spline is appropriate or not. We employ the modeling of conditional intensity of the point process by the trigonometric series and the LaGuerre polynomials. If the spline based cure for the intra-day periodicity were appropriate, then the models with trigonometric series would be rejected. Our empirical analysis on Yen-Dollar exchange rate market reveals that considerable periodic components are left in the 'adjusted' point process. Chapter 9 is related to Kawasaki (2002).

1.2 Smoothness Priors Analysis

1.2.1 The Smoothness Priors Concept

The history of smoothness priors essentially starts with a problem addressed in Whittaker (1923). It was followed by Shiller (1973) and Akaike (1980a) in which the framework initiated by Shiller was continued. What was proposed in Akaike (1980a) is a quasi-Bayesian Gaussian disturbances linear regression, or least squares computations modeling framework. Stochastic difference equation constraints were placed on the prior distributions of the model parameters. The crucial computation was that of the likelihood of hyperparameters of those distributions. A considerable amount of other work was motivated by Akaike (1980a). Here we identify some of that work and some relationship of that work to other research as well as developments and extensions. The least squares computational framework of smoothness priors is also presented here.

The problem addressed by Whittaker (1923) in the estimation of a smooth trend, (the mean of a nonstationary mean time series), embedded in white noise was the first work in this subject. The term 'smoothness priors' is very likely due to Shiller (1973). Shiller (1973) did not appear to be aware of Whittaker's work. He modeled the distributed lag (impulse response) relationship between the input and output of economic time series under 'smoothness' constraints on the distributed lags expressed by a difference equation. A trade-off of the goodness-of-fit of the solution to the data and the goodness-of-fit of the solution to a smoothness constraint was determined by a single smoothness trade-off parameter. Shiller did not offer an objective method of choosing the smoothness tradeoff parameter. Akaike (1980a), continued the analysis initiated by Shiller. Akaike developed and exploited the concept of the likelihood of the Bayesian model and used a maximization of the likelihood procedure for determining the smoothness tradeoff parameter. (In Bayesian terminology, the smoothness tradeoff parameter is a "hyperparameter" , Lindley and Smith, 1972.) The smoothing problem context is understood to be common to a large variety of other statistical data analysis problems including density estimation and image

analysis, (Titterington 1985).

Following Akaike (1980a) and Kitagawa (1981), smoothness priors was primarily based on a normal distribution theory, linear model, stochastic regression treatment of stationary and nonstationary time series. Subsequently the terminology 'smoothness priors' has come into use in several papers by Kitagawa and Gersch (Gersch and Kitagawa 1983a, 1988, Kitagawa and Gersch 1984, 1985a, 1985b, Gersch 1992). In a very significant extension, Kitagawa (1987) showed a smoothness priors state space modeling of nonstationary time series in which neither the system noise or the the observation noise are necessarily Gaussian distributed. Among other papers, Kitagawa (1988, 1989, 1991, 1993, 1994, 1996), are further developments, extensions and applications of the not necessarily linear - not necessarily Gaussian state space modeling of time series.

The smoothness priors method is Bayesian. The Bayesianness provides a framework for doing statistical inference. A prior distribution on the model parameter is expressed in the form of a stochastic difference equation and parameterized by hyperparameters which in turn have a crucial role in the analysis. The maximization of the likelihood of a small number of hyperparameters permits the modeling of a time series with relatively complex structure and a very large number of implicitly inferred parameters. The crucial statistical ideas in smoothness priors are the likelihood of the Bayesian model and the use of likelihood as a measure of goodness-of-fit of the model.

1.2.2 Background, History and Related Work

A conceptual predecessor of smoothness priors can be seen in a smoothing problem posed by Whittaker (1923). In that problem the observations y_n, $n = 1,\ldots,N$ are given. They are assumed to consist of the sum of a "smooth" function f and observation noise or,

$$y_n = f_n + \varepsilon_n, \tag{1.1}$$

where $\varepsilon_n \sim N(0,\sigma^2)$. The problem is to estimate the unknown f_n, $n = 1,\ldots,N$. In a time series interpretation of this problem, f_n, $n = 1,\ldots,N$ is the trend of a nonstationary time series. A typical approach to this problem is to approximate f by a class of parametric polynomial regression models. The quality of the analysis is dependent upon the appropriateness if the assumed model class. A flexible model is desirable. In this context, Whittaker suggested that the solution balance a tradeoff of goodness-of-fit to the data and goodness-of-fit to a smoothness criterion. This idea was expressed by minimizing

$$\sum_{n=1}^{N}(y_n - f_n)^2 + \mu^2 \sum_{n=k+1}^{N}(\nabla f_n^k)^2 \tag{1.2}$$

for an appropriately chosen smoothness tradeoff parameter μ^2. In (1.2), ∇f_n^k expresses a $k-$th order difference constraint on the solution f, with $\nabla f_n = f_n - f_{n-1}$, $\nabla^2 f_n = \nabla(\nabla f_n)$, etc. Whittaker's original solution was not expressed in a Bayesian context. Whittaker and Robinson (1924) is a Bayesian interpretation of this problem. Greville (1957) showed that there is a unique solution to (1.2).

The properties of the solution to the problem (1.1)–(1.2) are apparent. If $\mu^2 = 0$, $f_n = y_n$ and the solution is a replica of the observations. As μ^2 becomes increasingly large, the smoothness constraint dominates the solution and the solution satisfies a $k-$th order constraint. For large μ^2 and $k = 1$, the solution is a constant, for $k = 2$, it is a straight line, etc. Whittaker left the choice of μ^2 to the investigator.

In a closely related direction, Schoenberg (1964) suggested an adoption of Whittaker's smoothing method to the fitting of a continuous function to the observed data points, with the data not necessarily evenly spaced. In that case the data model is,

$$y_i = f(x_i) + \varepsilon_i, \tag{1.3}$$

where the ε_i are as in (1.1) and f is assumed to be "smooth" on the interval $[a,b]$ and the observations are at the n points $x_1,\ldots,x_n$. An estimate of f is assumed to be the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int_a^b [f^{(m)}(x)]^2 dx, \tag{1.4}$$

with $f \in C^{m-1}$ and $\lambda > 0$. As in the regularly spaced data-discrete function solution problem, again we have a tradeoff between infidelity of the solution to the data, and the "roughness" of the solution as represented by the square integral of the m-th derivative. The nonnegative parameter λ controls the tradeoff. From this nonparametric function estimation interpretation, here too, the parameter of interest is a function.

Parzen (1961, 1963) in a development unrelated to smoothness priors, introduced a reproducing kernel Hilbert space (RKHS) approach to time series. De Boor and Lynch (1966) was a RKHS treatment of spline approximation. Kimmerdorf and Wahba (1970a, 1970b, 1971) exploited both developments and treated the general spline smoothing problem from an RKHS-stochastic equivalence perspective. A key result of Kimmerdorf and Wahba is that minimum norm interpolation and smoothing problems with quadratic constraints imply an equivalent Gaussian stochastic process. Their solutions are Bayesian estimates. Following from the minimum norm interpretation of the smoothing problem, an RKHS is the natural mathematical framework for smoothness priors. Weinert et al. (1980) exploited the equivalence to express spline smoothing algorithms in a computationally efficient state-space recursive computational

framework. Subsequently Wecker and Ansley (1983), and Kohn and Ansley (1988), generalized and realized state-space recursive computational algorithms and applied them to practical data analysis problems. Gersch (1992) includes a review of the RKHS state space approach to smoothness priors.

The Kimmerdorf and Wahba minimum norm-stochastic equivalence implies that the extensively studied signal extraction problem and the smoothing problem are equivalent problem statements. The significance of that result is that the smoothing problem context, and hence smoothness priors , is common to a large variety of other statistical data analysis problems, (i.e. smoothing problems), including density estimation, image restoration, and so on.

Smoothness priors also relates to the ill-posed problems or inverse problems and problems described as statistical regularization, Tikhonov (1963). In that context, scalar inverse problems are discretized and a quadratic regularization criterion is imposed in which, as in the Whittaker's problem, the solution balances a tradeoff between infidelity of the solution to the data and infidelity of the solution to the regularization criterion. (Nash and Wahba 1974 treat statistical regularization in the context of RKHS's.) Penalized likelihood methods, introduced in Good and Gaskins (1980) in the context of density estimation, have been used for example in regression for both Gaussian and non-Gaussian data (Wahba 1990, Silverman 1985, O'Sullivan et al. 1986, Gu 1990, Hastie and Tibshirani 1993, Green and Silverman 1994), density estimation (Good and Gaskins 1980, Leonard 1978, Silverman 1982, Gu and Qiu 1993, Gu 1993a), hazard rate function estimation (Anderson and Senthiselvan 1980, Gu 1993b), estimation for the intensity function of a Poisson process (Gu and Qiu 1993), and time varying coefficient modeling (Hastie and Tibshirani 1993), which relates to closely to the smoothness priors approach considered in this paper.

Further, the Bayesian framework easily provides estimates of the precision of the estimate. In regression for Gaussian data, the commonly used quadratic roughness penalty was shown by Wahba (1978) to be equivalent to a partially improper Gaussian prior, in the sense that the penalized likelihood estimator is identical to the mean of the corresponding Gaussian posterior. Gu (1992) showed that when the sampling likelihood is non-Gaussian, that under Wahba's prior, with appropriate approximations, the penalized likelihood estimator is the mean of the approximate posterior. Bayesian interpretations for the penalized likelihood models, including Bayesian confidence intervals, have for example also been given by Good and Gaskins (1980), Kohn and Ansley (1987), Leonard (1978), Nychka (1981), Silverman (1985), Wahba (1983) and Hastie and Tibshirani (1993). We also note that in a methodology that is relevant in high-dimensional regression problems, the relationship between penalized least squares and

estimation in linear additive models, (Buja et al. 1989), is the key tool in establishing many of the results.

1.2.3 Smoothness Priors Bayesian Modeling

The theoretical and computational approach in the seminal paper, Akaike (1980a) is described here. Consider the Gaussian disturbances stochastic linear regression model

$$y = X\theta + \varepsilon. \tag{1.5}$$

The dimensions of the matrices in (1.5) are y: $n \times 1$; X: $n \times p$; ε: $n \times 1$; θ: $p \times 1$. ε is normally distributed with mean 0 and covariance matrix $\sigma^2 I_n$, and θ is a normally distributed prior parameter vector independent of ε with mean 0 and covariance matrix $\lambda^{-2} D^{-1} D^{-T}$, D nonsingular. y is the vector of observed data, X and D are assumed known, ε is the observation noise vector, and σ^2 and λ^2, (where λ is referred to as a hyperparameter, Lindley and Smith 1972), are unknown. That is,

$$\begin{bmatrix} y \\ \theta \end{bmatrix} \sim N\left(\begin{bmatrix} X\theta \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 I & 0 \\ 0 & \lambda^{-2} D^{-1} D^{-T} \end{bmatrix} \right). \tag{1.6}$$

In this conjugate family Bayesian situation (Zellner 1971, Berger 1985), the mean of the posterior normal distribution of the parameter vector θ minimizes

$$||y - X\theta||^2 + \lambda^2 ||D\theta||^2. \tag{1.7}$$

If σ^2 were known, the computational problem in (1.7) could be solved by an ordinary least squares computation. The solution for θ, the posterior mean, is the minimizer of

$$\left|\left| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ \lambda D \end{bmatrix} \theta \right|\right|^2. \tag{1.8}$$

That solution is

$$\hat{\theta} = (X^T X + \lambda^2 D^T D)^{-1} X^T y \tag{1.9}$$

with the residual sum of squares,

$$\mathrm{SSE}(\hat{\theta}, \lambda^2) = y^T y - \hat{\theta}^T (X^T X + \lambda^2 D^T D)\hat{\theta}. \tag{1.10}$$

For a Bayesian smoothness priors interpretation of the problem, multiply (1.7) by $-1/(2\sigma^2)$ and exponentiate. Then, the θ that minimizes (1.7) also maximizes

$$\exp\left\{ -\frac{1}{2\sigma^2} ||y - X\theta||^2 \right\} \exp\left\{ -\frac{\lambda^2}{2\sigma^2} ||D\theta||^2. \right\} \tag{1.11}$$

From (1.11), the posterior distribution interpretation of the parameter vector θ is that it is proportional to the product of the conditional data distribution (likelihood), $p(y|X,\theta,\sigma^2)$, and a prior distribution, $\pi(\theta|\lambda^2,\sigma^2)$ on θ,

$$\pi(\theta|y,\lambda,\sigma^2) \propto p(y|X,\theta,\sigma^2)\pi(\theta|\lambda^2,\sigma^2). \tag{1.12}$$

The left hand side of (1.12) is a proper distribution. Consequently, the integration of the right hand side of (1.12) yields $L(\lambda^2,\sigma^2)$, the likelihood for the unknown parameters λ^2 and σ^2,

$$L(\lambda^2,\sigma^2) = \int_{-\infty}^{\infty} p(y|X,\theta,\sigma^2)\pi(\theta|\lambda^2,\sigma^2)d\theta. \tag{1.13}$$

I. J. Good (1965) referred to the maximization of (1.13) as a type II maximum likelihood method. A critically important result in Akaike (1980a) is that, since $\pi(\theta|y,\lambda,\sigma^2)$ is normally distributed, (1.13) can be expressed in the closed form,

$$L(\lambda^2,\sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}}|\lambda^2 D^T D|^{\frac{1}{2}}|X^T X + \lambda^2 D^T D|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\mathrm{SSE}(\hat{\theta},\lambda^2)\right\}. \tag{1.14}$$

The maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{N}\mathrm{SSE}(\hat{\theta},\lambda^2). \tag{1.15}$$

It is convenient to work with -2 log likelihood. Using (1.15) in (1.14) yields

$$-2\log L(\lambda^2,\sigma^2) = N\log(2\pi\hat{\sigma}^2) - \log|\lambda^2 D^T D| + \log|X^T X + \lambda^2 D^T D| + N. \tag{1.16}$$

A practical way to determine the value of λ^2 for which the -2 log likelihood is minimized, is to compute the likelihood for discrete values of λ^2 and search the discrete -2 log likelihood - hyperparameter space for the minimum. A precise estimate of λ^2 may be obtained by a numerical linear search algorithm.

Akaike (1980a) and Good and Gaskins (1980) are very likely the first practical uses of the likelihood of the Bayesian model and the likelihood of the hyperparameters as a measure of the goodness-of-fit of a model to data.

1.3 Construction of Information Criteria

1.3.1 Kullback-Leibler information number

Suppose we choose an appropriate one among various statistical models. Then, by what kind of criterion should we quantitatively assess the goodness-of-fit of the statistical models in question? Using the observed data $X_n = \{X_1,\ldots,X_n\}$, we estimate the parameters of the fitted model

to construct a statistical model $f(x|\hat{\theta})$, where $\hat{\theta} = \hat{\theta}(X_n)$ denotes an estimator of the unknown parameter vector. Let $g(x)$ be the density function of the true probabilistic model from which the data is generated. In fact, it is impossible to capture the probabilistic structure of this true model only through the finite number of observations. Hence we describe the data generating structure through a family of parametric model $\{f(x|\theta); \theta \in \Theta \subset R^p\}$ and approximate the true model $g(x)$ by the estimated statistical model $f(x|\hat{\theta})$.

This problem setting naturally leads to one idea that the goodness-of-fit of a statistical model $f(x|\hat{\theta})$ can be measured by its closeness to the true model $g(x)$ that generates the data. As a device to gauge the closeness of the two distributions, the Kullback-Leibler information number is often employed;

$$\begin{aligned} I(g(z), f(z|\hat{\theta})) &= E_G\left[\log \frac{g(Z)}{f(Z|\hat{\theta})}\right] \\ &= \int \log g(z) dG(z) - \int \log f(z|\hat{\theta}) dG(z), \end{aligned} \tag{1.17}$$

where $G(z)$ denotes the cumulative distribution function of the true model $g(z)$, and the expectation $E_G[\,\cdot\,]$ is taken with respect to the unknown distribution $G(z)$ with fixing $\hat{\theta} = \hat{\theta}(X_n)$. Here $Z = z$ stands for the randomly sampled another set of observation from the same true model $g(\cdot)$ so as to be independent of X_n. Then, the number $I\{g, f\}$ measures the averaged goodness of the prediction by $f(z|\hat{\theta})$ of another independent data set $Z = z$. $I\{g, f\} = 0$ if and only if $g = f$, and the smaller $I\{g, f\}$ the closer to the true model is the constructed statistical model. However, the Kullback-Leibler information number cannot be calculated because it depends on the knowledge of the true model $g(z)$ which is unknown in fact. The next subsection describes how the KL information number must be estimated.

1.3.2 Information Criteria

Because the first term in the right hand side of (1.17), $E_G[\log g(Z)]$, does not depend on the fitted model, the estimation of the Kullback-Leibler information number reduces to the second term

$$\int \log f(z|\hat{\theta}) dG(z) = \int g(z) f(z|\hat{\theta}) dz. \tag{1.18}$$

This term is referred to as the expected log-likelihood of the model $f(z|\hat{\theta})$. If we have an estimate of the expected log-likelihood, it can be employed as a model selection criterion. Though the expected log-likelihood again depends on the true model $g(z)$, it can be estimated by estimating the unknown probability distribution $G(z)$ from observations. usually the empirical distribution function $\hat{G}(z)$, of which probability function $\hat{g}(z) = 1/n$ endows equal probability

$1/n$ to each of n data points, is employed to estimate $G(z)$. Replacing the unknown probability distribution function $G(z)$ in (1.18) by its empirical distribution function $\hat{G}(z)$, we obtain

$$\int \log f(z|\hat{\theta}) d\hat{G}(z) = \frac{1}{n}\sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\theta}), \tag{1.19}$$

which turns out to be the log-likelihood of the model $f(x|\theta)$.

In this way, we can regard the log-likelihood as an estimate of the expected log-likelihood. But the log-likelihood itself is not constructed from the view point of prediction. To put it another way, in the real world we estimate the unknown distribution $G(z)$ *not* by the future realization Z_n that is independent of the observed data X_n, *but* by reusing the same data X_n that has been already used to estimate the unknown parameter θ. Hence, the log-likelihood (1.19) generally overestimates the expected log-likelihood. If we successfully estimate to what extent the log-likelihood is overestimated by the double use of the observed data, then we can obtain the unbiased estimator of the expected log-likelihood by subtracting the average bias from the log-likelihood. This quantity of average overestimation is defined by the bias which is brought by estimating the expected log-likelihood by the log-likelihood of the model $f(x|\theta)$, namely,

$$b(G) = E_X\left[\frac{1}{n}\sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\theta}) - \log f(z|\hat{\theta}) dG(z)\right], \tag{1.20}$$

where the expectation is taken with respect to the joint distribution of the sample, $\prod_{\alpha=1}^{n} G(x_\alpha)$.

Correction of this bias plays an essential role in the construction of information criteria. Generally speaking, information criteria are given as the log-likelihood corrected by estimating or approximating the bias in some way,

$$\begin{aligned} \mathrm{IC}_{EX} &= -2n\left\{\frac{1}{n}\sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\theta}) - hat b(G)\right\} \\ &= -2\sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\theta}) + 2n\hat{b}(G). \end{aligned} \tag{1.21}$$

However, it is difficult to give the exact bias $b(G)$ in (1.20) in a general framework. Hence we make use of the asymptotic expansion of the bias $b(G)$ with respect to the number of the observation n,

$$b(G) = \frac{1}{n}b_1(G) + \frac{1}{n^2}b_2(G) + O(n^{-3}), \tag{1.22}$$

and the estimate $\hat{b_1}(G)$ for the asymptotic bias is used for the bias correction of the log-likelihood. Replacing $\hat{b}(G)$ in (1.21) by the asymptotic bias $\hat{b_1}(G)/n$, an information criterion is constructed in the following form,

$$\mathrm{IC}_{AS} = -2\sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\theta}) + 2\hat{b_1}(G). \tag{1.23}$$

As just described, an information criterion is an estimate of the Kullback-Leibler information number (1.17). If we have many candidate models, the model with minimum information criterion statistic will be preferred. Bias correction term $\hat{b_1}(G)$ takes different form depending on model estimation procedure and on the relationship between the true model and the statistical model. In the next subsection, we state the model selection criterion when the model is estimated by the method of maximum likelihood, and subsequently in subsection we review the more general criterion in the sense that the model is not always estimated by the maximum likelihood method.

1.3.3 Model Selection Criterion Based on MLE

Let $f(z|\hat{\theta}_{ML})$ be the model given by the maximum likelihood estimation based on the data X_n and the family of parametric model $\{f(x|\theta); \theta \in \Theta \subset R^p\}$. Then an information criterion AIC is given as follows,

$$\text{AIC} = -2 \sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\theta}_{ML}) + 2p, \tag{1.24}$$

where p stands for the number of free parameters in the model. AIC comes out by approximating the asymptotic bias in (1.22) as $b_1(G) \approx p$. If the assumed family of parametric model $\{f(x|\theta), \theta \in \Theta\}$ contains the true model $g(x)$, or for some $\theta_0 \in \Theta$ there exists the distribution that satisfies $g(x) = f(x|\theta_0)$ of which distribution function $F(x|\theta_0) = F_{\theta_0}(x)$, then the asymptotic bias results in $b_1(F_{\theta_0})$, by which AIC is asymptotically derived.

According to the arguments in Akaike (1973, 1974), if the true model lies near the assumed parametric model, the bias of the log-likelihood of the model estimated by MLE can be well approximated by the number of free parameters contained in the model. This assumption has brought data analysts several merits. For example, we do not have to analytically evaluate the bias correction term, and it does not depend on the unknown distribution G. Because of this convenience, AIC has been widely used in various fields of scientific research, see Akaike and Kitagawa (1999).

Under the circumstance that the fitted model $f(x|\theta)$ does not necessarily contain the data generating distribution $g(x)$, the asymptotic bias in (1.23) and the corresponding information criterion takes a different form as follows (Takeuchi, 1976),

$$\text{TIC} - 2 \sum_{\alpha=1}^{n} \log f(x_\alpha|\hat{\theta}_{ML}) + 2\text{tr}(\hat{J}^{-1}(G)\hat{I}(G)), \tag{1.25}$$

where $\hat{J}$ and $\hat{I}$ are the estimates of the following $p \times p$ matrices J and I,

$$J(G) \quad = \quad -E_G\left[\frac{\partial^2 \log f(Z|\theta)}{\partial\theta\partial\theta'}\right],$$

$$I(G) \quad = \quad E_G\left[\frac{\partial \log f(Z|\theta)}{\partial \theta}\frac{\partial \log f(Z|\theta)}{\partial \theta'}\right]. \tag{1.26}$$

This information criterion comes from the asymptotic expansion $b(G) = n^{-1}\mathrm{tr}\{J(G)^{-1}I(G)\} + O(n^{-2})$. Suppose the data is generated from F_θ which is the distribution function of $f(z|\theta)$. Then, because we have $\mathrm{tr}\{J(F_\theta)^{-1}I(F_\theta)\} = p$ by the properties of Fisher information, TIC reduces to AIC.

Now, how should we construct information criteria that enables us to evaluate a wider class of models of which members are not necessarily estimated by MLE. GIC (Generalized Information Criterion) of Konishi and Kitagawa (1996) gave an answer to this general problem by the approach based on statistical functionals.

1.3.4 Generalized Information Criterion

Suppose data is generated from the unknown distribution $G(x)$ (or with density $g(x)$), and this true model $g(x)$ lies in the neighborhood of the fitted parametric model $\{f(x|\theta), \theta \in \Theta\}$. Then, the estimator of the parameter $\theta = (\theta_1, \ldots, \theta_p)'(\subset R^p)$ of the model is constructed dependent on the true distribution $G(x)$ (or $g(x)$) that generates the data, but it does not depend on $f(x|\theta)$. Suppose there exists a statistical functional $T_i(G)$ which maps a real valued function to Euclidean space, by which the estimator $\hat{\theta}_i$ of a parameter θ_i can be given as $\hat{\theta}_i = T_i(\hat{G})$ $(i = 1, \ldots, p)$ for the empirical distribution function $\hat{G}$ that summarizes the data. Let $T(G) = (T_1(G), \ldots, T_p(G))'$ be a p-dimensional functional vector of which i-th element is $T_i(G)$, then the estimator $\hat{\theta}$ can be rewritten as $\hat{\theta} = T(\hat{G})$. For example, suppose we adopt sample mean $\bar{X}_n$ to estimate the unknown mean μ of i.i.d. random variable. Then the functional that defines the population mean is $T_\mu(G) = \int x dG(x)$, and using this functional T_μ, the sample mean can be expressed as

$$T_\mu(\hat{G}) = \int x d\hat{G}(x) = n^{-1}\sum_{\alpha=1}^{n} X_\alpha.$$

Once the statistical functional $T_i(G)$ $(i = 1, \ldots, p)$ is given, we obtain the empirical influence function (derivative of the functional at $\hat{G}$),

$$T_i^{(1)}(x; \hat{G}) = \lim_{\varepsilon \to 0} \frac{T_i((1-\varepsilon)\hat{G} + \varepsilon\delta_x) - T_i(\hat{G})}{\varepsilon}, \tag{1.27}$$

where δ_x denotes the point mass probability distribution on x. This empirical influence function $T_i^{(1)}(z; \hat{G})$ plays an essential role in the construction of information criteria, thus we define the p-dimensional vector of empirical influence functions as

$$T^{(1)}(x; \hat{G}) = (T_1^{(1)}(x; \hat{G}), \ldots, T_p^{(1)}(x; \hat{G})) \tag{1.28}$$

of which i-th element is given by (1.27).

Now, the generalized information criteria (GIC) to evaluate the statistical model $f(z|\hat{\theta})$ ($\hat{\theta} = T(\hat{G})$) can be given as follows;

$$\mathrm{GIC} = -2\sum_{\alpha=1}^{n} \log f(X_\alpha|\hat{\theta}) + \frac{2}{n}\sum_{\alpha=1}^{n} \mathrm{tr}\left\{ T^{(1)}(X_\alpha;\hat{G}) \frac{\partial \log f(X_\alpha|\theta)}{\partial \theta'}\bigg|_{\theta=\hat{\theta}} \right\}, \tag{1.29}$$

where $\partial/\partial\theta = (\partial/\partial\theta_1, \ldots, \partial/\partial\theta_p)$. Rewriting the asymptotic bias used in GIC, we obtain

$$\sum_{i=1}^{p}\sum_{\alpha=1}^{n} T_i^{(1)}(X_\alpha;\hat{G}) \left[\frac{\partial \log f(X_\alpha|\theta)}{\partial \theta_i}\right]_{\theta=\hat{\theta}}.$$

This shows that the asymptotic bias in GIC is given as the sum of products of the empirical influence function of the estimator θ_i ($T_i^{(1)}(X_\alpha;\hat{G})$) and the score function of the estimated model ($\partial \log f(X_\alpha|\theta)/\partial\theta_i$). The influence function, which originates from robust statistics, is widely used to analyze the sensitivity of the estimator with respect to the small change of distribution. See Hampel et al. (1986) and Huber (1981). GIC enables the evaluation of statistical models estimated by various methods; it covers maximum likelihood estimator, robust estimator, penalized likelihood estimator and Bayes estimator. In this thesis, GIC is employed to evaluate nonlinear spline models with regularization, and its application to the yield curves estimation problem is discussed in Chapter 8.

1.3.5 Bayesian Approach: ABIC

Given the data $X_n = (X_1, \ldots, X_n)$, we have the likelihood $\prod_{\alpha=1}^{n} f(X_\alpha|\theta)$ for the parameter θ of the fitted model $\{f(x|\theta); \theta \in \Theta \subset R^p\}$. On top of this, in Bayesian approach, the parameter θ is modeled by the prior distribution $\pi(\theta|\omega)$ where ω, often referred to as hyperparameter, denotes a q-dimensional parameter vector ($q < p$) that specifies the prior distribution. Once the data is available, we obtain the posterior distribution of θ in the following equation,

$$\pi(\theta|X_n;\omega) = \frac{\prod_{\alpha=1}^{n} f(X_n|\theta)\pi(\theta|\omega)}{\int \prod_{\alpha=1}^{n} f(X_n|\theta)\pi(\theta|\omega)d\theta}. \tag{1.30}$$

Here we discuss the relation between the penalized likelihood and the marginalized likelihood. Let $\ell_\lambda(\theta)$ be the penalized likelihood which is formulated as follows,

$$\ell_\lambda(\theta) = \sum_{\alpha=1}^{n} \log f(X_\alpha|\gamma,\sigma) - \frac{n}{2}\lambda\gamma' K\gamma, \tag{1.31}$$

where $\theta = (\gamma,\sigma)'$, K is a non-negative definite matrix, and λ be the smoothing (or regularization) parameter. Note that γ in the penalty term can be regarded as a degenerated multivariate normal random vector, hence by adjusting the normalized constant $\ell_\lambda(\theta)$ can be rewritten as

$$\prod_{\alpha=1}^{n} f(X_\alpha|\gamma,\sigma)\frac{|n\lambda K|^{1/2}}{(2\pi)^{1/2}}\exp\left(-\frac{n}{2}\lambda\gamma' K\gamma\right) = \prod_{\alpha=1}^{n} f(X_\alpha|\gamma,\sigma)\pi(\gamma|\lambda). \tag{1.32}$$

Hence the penalized likelihood can be regarded as a Bayes model with the improper prior $\pi(\gamma|\lambda)$ which is controlled by hyperparameter $\lambda(>0)$. If λ and σ is given, maximization of penalized likelihood with respect to γ is equivalent to the maximization of the posterior distribution (1.30).

The most important problem is how should we determine the hyperparameter λ. In the context of the penalized likelihood approach, λ is determined by the model selection criterion, GIC. Akaike (1980a, 1980b) proposed the maximization of the marginalized likelihood

$$\int \prod_{\alpha=1}^{n} f(X_\alpha|\gamma,\sigma)\pi(\gamma|\lambda)d\gamma \tag{1.33}$$

to determine σ and λ. The estimated hyperparameters are often referred to as a MAP (maximum a posteriori) solution, so we denote them σ_{MAP} and λ_{MAP}. Once the MAP solution is found, we substitute σ and λ in (1.32) by their MAP estimates, then γ can be estimated so that (1.32) should be maximized. The criterion (1.33) is called ABIC (Akaike Bayesian Information Criterion), and its usefulness is widely recognized by many actual applications, see Bozdogan (1994), Kitagawa and Gersch (1996), Akaike and Kitagawa (1999).

Bibliography

[1] Kawasaki, Y. and Sato, S. (1997a), On the 'Optimality' of Seasonal Adjustment Procedures (in Japanese with English abstract), *Proceedings of the Institute of Statistical Mathematics*, Vol. 45, No. 2, 245–263.

[2] Kawasaki, Y. (1997b), Comment on T. Kimura's "Some Practical Issues Involved with the Seasonal Adjustment" (in Japanese), *Proceedings of the Institute of Statistical Mathematics*, Vol. 45, No. 2, 207–211.

[3] Kawasaki, Y. (2003), A Structural Time Series Model Facilitating Flexible Seasonality, *Journal of the Japanese Society of Computational Statistics (to appear)*.

[4] Kawasaki, Y. and Franses, P. H. (2003a), Detecting Seasonal Unit Roots in A Structural Time Series Model, *Journal of Applied Statistics*, Vol. 30, No. 4, 373–387.

[5] Kawasaki, Y. and Franses, P. H. (2003b), Do Seasonal Unit Roots Matter for Forecasting Monthly Industrial Productions?, *Journal of Forecasting (in press)*.

[6] Kawasaki, Y. (2001), Principal Component and Factor Analysis for Multiple Time Series (in Japanese with English abstract), *Proceedings of the Institute of Statistical Mathematics*, Vol. 49, No. 1, 109–131.

[7] Kawasaki, Y., Sato, S. and Tachiki, S. (1998) Smoothness Prior Approach to Estimate Large Scale Multifactor Models, *ISM Research Memorandum No. 714*, The Institute of Statistical Mathematics, Tokyo.

[8] Kawasaki, Y., Sato, S. and Tachiki, S. (2000), Vector-Valued Multiple Regression Model with Time Varying Coefficients and Its Application to Predict Excess Stock Returns, *Proceedings of IEEE/IAFE/INFORMS Conference on Computational Intelligence for Financial Engineering*, pp. 162–165.

[9] Kawasaki, Y. and Ando, T. (2002), Nonlinear Regression Models with Regularization and Their Application to Yield Curve Estimation (in Japanese with English abstract), *Proceedings of the Institute of Statistical Mathematics*, Vol. 50, No. 2, 149–164.

[10] Kawasaki, Y. and Ando, T. (2003), Estimating Term Structure Using Nonlinear Splines: A Penalized Likelihood Approach, Unpublished Manuscript.

[11] Kawasaki, Y. (2002), Modeling Periodicity in High Frequent Financial Data, in *Modeling Seasonality and Periodicity: Proceedings of the 3rd International Symposium on Frontiers of Time Series Modeling*, ISM Report on Research and Education No. 13, 239-251.

Chapter 2

On the 'Optimality' of Seasonal Adjustment Procedures

2.1 Introduction

Modeling seasonality is one of the important problem in economic time series analysis. Above all, the methods in which a deconvolution into several components including seasonality is considered is called a seasonal adjustment method. A demand for such a decomposition springs up from some practical motivations. For example, if an economist tries to forecast the middle- or long-range trend of economy, then it is natural for him to want to remove seasonal periodic pattern prior to describing his economic outlook. Unfortunately, there seems to be no universal criterion to decide the best seasonal adjustment method, because what is the best way for an empirical analyst entirely depends on the nature of his decision problem, on his own loss function, and on to what extent he expects the precision of the seasonal adjustment procedure should be. In that sense, an empirical researcher who uses seasonally adjusted series implicitly accepts an notion of the optimality in a seasonal adjustment procedure whether apparent or opaque. Generally speaking, seasonal adjustment procedures can be classified into two classes, model based methods and procedural methods. In a model based method, it is common to consider a criterion to choose the best model, or a criterion to choose the best parameter in a class of parametric models, and we seek the best seasonal adjustment model under a certain statistical criterion. The goodness of time series model is generally assessed in terms of the goodness of one-step ahead prediction, which reduces to minimum mean squared error (minimum MSE hereafter) under some mild assumptions.

On the other hand, long before the statistical modeling and the model selection had been widely accepted, the desirable conditions which seasonally adjusted series are expected to satisfy were raised by some researchers mainly in terms of their spectral properties. See Nerlove

(1964) for example. He asserts that the sample power spectrum of the seasonally adjusted series should have neither peaks nor dips at the seasonal frequencies, and that the dips at the seasonal frequencies are the evidence of excessive removal of spectral component, what is called an over adjustment. However, it has been turned out that such 'seasonal dips' observed for the sample power spectrum of seasonally adjusted series inevitably associated with the minimum MSE criterion. Interestingly, it is demonstrated by the same author in Grether and Nerlove (1970). One of the aims of this chapter is to look back at this old problem from a modern point of view. It just so happens that many empirical researchers are interested in the seasonal adjustment methods triggered by the release of X-12-ARIMA by the Bureau of Census, U. S. Department of Commerce. (See Findley et al. (1996).) Hence it is timely to revisit the past arguments on the optimality of seasonal adjustment procedures.

This chapter is organized as follows. In section 2.2, we review the historical background on how the optimality criteria raised by Nerlove (1964) had been dismissed by Grether and Nerlove (1970), and also characterize the time invariant filters which satisfy the minimum MSE criterion. In section 2.3, we give an answer to an essential question, "Why do seasonal dips appear or disappear from time to time in actual data analysis?" The key is to understand the role of the magnitude of the variance of irregular component. In the second half of section 2.3, we generate some artificial economic time series which we believe the representative cases as regards real economic data, and we apply DECOMP (Akaike et al., 1985) and X-12-ARIMA, which is respectively the representative of the model based seasonal adjustment methods and the moving average based seasonal adjustment procedure. We deduce some caveats from a practical point of view. Section 2.4 concludes.

2.2 Minimum MSE Criterion and Seasonal Dips

2.2.1 Seasonal Adjustment and Predictability

If we restrict the class of models to linear and Gaussian, minimum MSE criterion is essentially the same as the maximum likelihood principle. Likelihood of a time series model can be interpreted as the accumulation of the one-step ahead prediction error. This notion of predictability is important to both model based and moving average based seasonal adjustment methods. In a model based seasonal adjustment, the notion of predictability is used to exclude the overparametrized models among the candidate models estimated via the method of maximum likelihood. In other words, a model which shows good fit within sample is not necessarily the best forecast model because there may be a danger of overfitting. Such a parameter redun-

dancy will be eliminated by model selection criteria or by statistical tests. It should be noted that despite we are primarily interested in the unobservable signal (seasonal component) buried in the observations, it is through predictability that the goodness of fit within sample is guaranteed (in terms of maximum entropy).

In moving average based seasonal adjustment procedures, a model with high prediction accuracy is essentially needed in a different sense from above. This is because there will be missing values at the both ends of the series after the moving average procedure is performed. To interpolate these missing values, a time series model is fitted to the data, and the forecasted and back-casted values obtained from the model are used to augment the terminal missing observations. Especially, augmenting the future observations is most important because we are usually interested in the seasonally adjusted values of the recent or current observations which are heavily dependent on the accuracy of the forecasts. Hence it can be explained that X-12-ARIMA proposed by Findley et al. (1995) robustified the Census X-11 method by enriching the options that help us to build time series models with more predictability. Their effort toward the robustification could be possibly influenced by the robust seasonal adjustment procedure found in Kitagawa (1989), though the Census Bureau adopted a completely different methodology. Although the core part of X-12-ARIMA procedure is still the moving-average method, the procedure as a whole will be more modeling-oriented than before.

2.2.2 Nerlove's Spectral Criteria

As is stated at the beginning, seasonal adjustment can be viewed as a decomposition of time series into several unobserved components. BAYSEA (Akaike and Ishiguro, 1980) and DECOMP (Kitagawa, 1981) are the examples of the realization of this recognition. Even for the Census X-11 method, some researchers have been revealed that there are explicit statistical models behind the procedure. See Wallis(1974), Cleveland and Tiao(1976), Ozaki and Thomson (1994). Excluding the use of nonlinear filter, the seasonal adjustment methods in practical use can be roughly classified into two classes, namely, the additive model or the multiplicative model which results in the additive model after the log transformation. According to the arguments found in Grether and Nerlove (1970), the specification (or the belief) that each component has its distinct characteristics and the sum of them is equal to the observation have been rooted in this research topic around the beginning of the 20th century. From a purely economic point of view, it is ideal if we could explain the seasonal fluctuation by some causal relationship, but it would be too demanding under the limitation of the quality of actual data. Hence, when it comes to seasonal adjustment, the arguments are usually focused technically on the removal of

yearly periodic components.

If we want to discuss the goodness of estimation and prediction in a statistical problem, the explicit formulation of statistical models premises transparent arguments. However in seasonal adjustment, the comparison of methods have not been done in such a model-based way. It is not until 1970's that the notion of predictability had been introduced and had become popular in the context of seasonal adjustment. Before that, various desirable conditions on the seasonally adjusted series had been proposed and examined most-to-least. Especially, Nerlove (1964) raised important questions about the goodness of the seasonal adjustment methods. Among the criteria Nerlove posed, this chapter concerns the following two.

1. The sample spectrum of a seasonally adjusted series should not have any peaks at the seasonal frequency and its harmonics.

2. The sample spectrum of a seasonally adjusted series should not have any dips (or troughs) at the seasonal frequency and its harmonics.

Suppose we have monthly time series. Then the seasonal frequency is $1/12 \times 2\pi = \pi/6$, and its harmonics are $k/12 \times 2\pi$ ($k = 2, \ldots, 6$). Arranged in ascending order, these frequencies correspond to from one year cycle to 2 months cycle, and the cycles less than 2 months are folded back to the $[0, \pi]$ interval at the Nyquist frequency π. In this article, we use the word 'seasonal frequencies' to mean the seasonal frequency and its harmonics. The above two criteria essentially address the same thing, but only the latter had been regarded as a problematic phenomenon often found in actual seasonal adjustment. Namely, the dips at the seasonal frequencies called 'seasonal dips' had been discussed back to back the suspicion of over-adjustment. So to say, the first criterion addresses the aim of seasonal adjustment, and the second one assures us the appropriateness of the seasonal adjustment. But later in Grether and Nerlove (1970), Nerlove himself showed that this apparently acceptable criterion is incompatible with the minimum mean squared error estimation of the seasonal component. At present, the various criteria on the spectral properties of the seasonally adjusted series posed by Nerlove (1964) are meaningless. For example, see Hylleberg (1986, Chapter3). But the questions made by Nerlove should be (and actually have been) appreciated because it aroused many arguments on the optimality of the seasonal adjustment procedures.

In the section 2.2.4 and later, we discuss the properties of the minimum MSE seasonal adjustment filter. Before it, we make a small remark on the word 'filter' because it is so commonly used that its meaning heavily depends on the context. Sometimes it means a sort of estimation scheme, and sometimes means the weight function implied by a estimation scheme. As the ter-

minology in state space modeling, filter means the estimation problem in which the present state vector is estimated based on the information up to present. Hence the signal extraction problem that will be stated in the next subsection is also the filtering problem. On the other hand, when we are dealing with a linear time invariant filter like a finite window moving average, the word 'filter' simply refers to the weight function or the coefficients. In this article we sometimes use 'filter' sometimes to stand for the weight function and sometimes to the rational function that yields the weight function, but no confusion will occur.

2.2.3 Signal Extraction

As the preparation for the next subsection, this subsection describe the classical solution for the signal extraction problem, which mostly depends on Whittle (1963). Suppose we can observe the stationary time series $\{x_t\}$ where the subscript t stands for time. We are originally interested in the unobserved signal y_t, but we can observe it only after it is contaminated with the noise δ_t, hence it is assumed that $x_t = y_t + \delta_t$. Here we consider the least squares estimation problem of the signal y_t using the future values of x_t as well as its past values. That is, our aim is to estimate by least squares the double-sided symmetric infinite filter γ_s in

$$\begin{aligned} \hat{y_t} &= \sum_{-\infty}^{\infty} \gamma_s x_{t-s} \\ \gamma(z) &= \sum \gamma_s z^s. \end{aligned}$$

For the simplicity of the arguments, we assume the spectral density is continuous. Then from the properties of the least squares estimators, we obtain

$$\begin{aligned} 0 &= \mathrm{cov}(y-\hat{y}, x_{t-j}) \\ &= \mathrm{cov}(y, x_{t-j}) - \sum \gamma_k \Gamma_{j-k} \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{ij\lambda} \left[f_{yx}(\lambda) - \gamma(e^{-i\lambda}) f_{xx}(\lambda) \right] d\lambda \qquad (j = 0, \pm 1, \pm 2, \ldots), \end{aligned}$$

where Γ_{j-k} stands for the autocovariance function of x_t, f_{xx} is the spectral density function of x_t, and f_{yx} is the cross spectral density of x_t and y_t. For the above equation to hold for all j, it is necessary to hold for all λ that

$$\gamma(e^{-i\lambda}) = \frac{f_{yx}(\lambda)}{f_{xx}(\lambda)}.$$

Hence putting $e^{-i\lambda} = z$ in $f_{xx}(\lambda)$ and so on, we rewrite the above equation in terms of covariance generating functions, we obtain

$$\gamma(z) = \frac{g_{yx}(z)}{g_{xx}(z)}. \tag{2.1}$$

In brief, the double-sided filter we pursue can be obtained by the fraction of the polynomials though we do not state the algorithm. The algorithm is detailed in Whittle (1963) and Nerlove, Grether and Carvalho (1979). Conversely, putting $z = e^{-i\lambda}$ in $g_{xx}(z)$ yields the spectral density function of x_t by applying the Wiener-Khintchin formula. In that sense, the autocovariance functions can be identified with the spectral density functions. Throughout this article, the autocovariance function are denoted as $g_{xx}(z)$, and the function of λ obtained by putting $z = e^{-i\lambda}$ will be denoted as $f_{xx}(\lambda)$, etc.

As stated at the beginning of this subsection, what is meant by the signal extraction is rather general. However, in the context of seasonal adjustment, the signal extraction sometimes suggests the specific method called Unobserved Components ARIMA (UCARIMA) model where the trend and seasonal component are extracted by the filter given by (2.1) based on the fitted seasonal ARIMA model. In the same manner, the one-sided infinite filter and the asymmetric filter can be obtained. In addition, the argument here can be extended to the case of nonstationary time series, see Bell (1984) and Maravall (1985). For the practical implementation, see Burman (1980). In practice, at least in Burman (1980), many forecasts and backcasts are augmented prior to the application of a long-term double-sided moving average filter.

2.2.4 What Is Behind the Seasonal Dips

Based on the arguments in previous subsections, we observe that the seasonal dips are inevitable in linear model-based seasonal adjustment methods. The data generating process that Grether and Nerlove (1970) prepared for their simulation studies are the followings,

$$
\begin{aligned}
X_t &= T_t + S_t + I_t && (2.2)\\
T_t &= \frac{\varepsilon_t + 0.8\varepsilon_{t-1}}{(1-0.95L)(1-0.75L)} && (2.3)\\
S_t &= \frac{\nu_t + 0.6\nu_{t-1}}{1-0.9L^{12}} && (2.4)\\
I_t &= \eta_t, && (2.5)
\end{aligned}
$$

where $\{\varepsilon_t\}$, $\{\nu_t\}$, $\{\eta_t\}$ are mutually uncorrelated white noise sequence. T_t, S_t and I_t usually called the trend-cycle component, the seasonal component and the irregular component respectively. Sometimes the cycle component is specified separately, but at least in this chapter no cyclic component will be introduced explicitly. Hence just for simplicity, we label the trend-cycle component as the trend component. Besides, the irregular component is sometimes referred to as the observational noise because the irregular component corresponds to the observational noise in the state-space formulation of the seasonal adjustment models. These words will be concomitantly used because they will cause no confusion in what follows.

Suppose we want to extract the trend component T_t from X_t. The following argument holds in parallel for other components. In the seasonal adjustment in UCARIMA, at first ARIMA model has to be fitted to the original series X_t of which moving average representation gives the denominator in the equation (2.1). Because the innovation processes $\{\varepsilon_t\}$, $\{\nu_t\}$ and $\{\eta_t\}$ are assumed to be mutually uncorrelated white noise sequence, the covariance generating function g_{XT} can be reduced to the autocovariance function of each component such as g_{TT}, hence the a.c.g.f. of X_t can be expressed as the sum of the a.c.g.f of each component, that is,

$$g_{XX}(z) = g_{TT}(z) + g_{SS}(z) + g_{II}(z)$$

For the detailed expression of g_{XX}, see Grether and Nerlove (1970). If we can determine the variance of the innovation process of each component, by (2.1) $\gamma(z) = g_{XT}(z)/g_{XX}(z) = g_{TT}(z)/g_{XX}(z)$ gives the optimal signal extraction filter (or the symmetric weight function) in the sense of minimum MSE. Hence, according to the same argument, we can construct the prediction filter of any step ahead forecast.

So far in our example, we have complete knowledge on the true data generating process. Hence for simulated data, we are virtually in an ideal situation that we can construct the optimal seasonal adjustment filter (in minimum MSE sense) without the effect of observational error. The surprising conclusion that Grether and Nerlove(1970) deduced is that we observe the seasonal dips in the sample power spectrum of seasonally adjusted series *in spite of* this ideal situation. This fact can be understood in the following manner. Using the signal extraction filter $\gamma(z)$, and by the relation $f_{\hat{T}\hat{T}}(\lambda) = \gamma(e^{i\lambda})\gamma(e^{-i\lambda})f_{XX}(\lambda)$, the spectral density of the trend component can be given as

$$f_{\hat{T}\hat{T}}(\lambda) = \left\{ \frac{f_{TT}(\lambda)}{f_{XX}(\lambda)} \right\}^2 f_{XX}(\lambda) \tag{2.6}$$

or as

$$f_{\hat{T}\hat{T}}(\lambda) = \frac{f_{TT}(\lambda)}{1 + \dfrac{f_{SS}(\lambda)}{f_{TT}(\lambda)} + \dfrac{f_{II}(\lambda)}{f_{TT}(\lambda)}}. \tag{2.7}$$

Bearing in mind the typical shape of the (pseudo) power spectrum of an economic time series, the second term in the denominator of (2.7), $f_{SS}(\lambda)/f_{TT}(\lambda)$, take a large value around the seasonal frequencies. Hence, according to Grether and Nerlove (1970), the seasonal dips of the sample spectrum of the seasonally adjusted series is not the evidence of overadjustment as pointed out in Nerlove (1964), but the characterization of the minimum MSE-optimal seasonal adjustment filter.

Sims (1978) offers another helpful point of view. For simplicity, suppose we consider only two components in our decomposition problem (2.2), namely the seasonal and the non-seasonal component. In other words, we do not distinguish between I_t and T_t, hence we set $T_t + I_t = Y_t$.

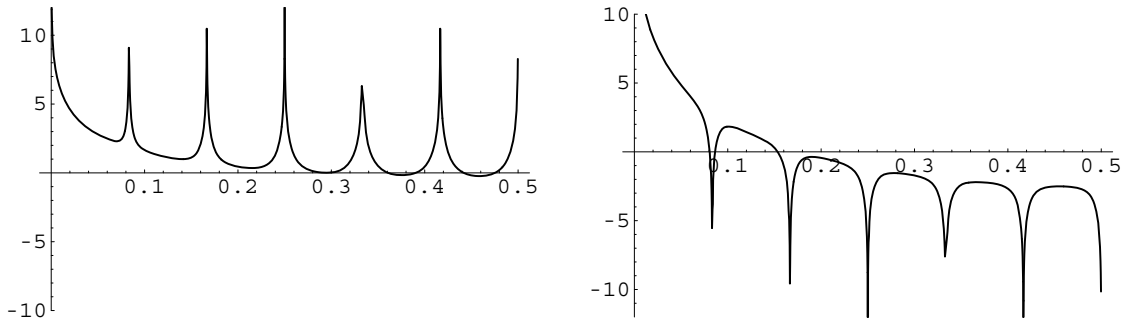
By taking logarithm of both sides of (2.6), we obtain

$$\log f_{\hat{Y}\hat{Y}}(\lambda) = 2\log f_{YY}(\lambda) - \log f_{XX}(\lambda). \tag{2.8}$$

Let us consider the behavior of (2.8) around the seasonal frequencies. The power spectrum of the trend component is usually assumed to be a smooth function in λ, having a peak at the zero frequency, and its power decreasing as λ tends to π. On the other hand, as the power spectrum of the irregular component is uniformly distributed in the frequency domain, we conclude that $f_{YY}(\lambda)$ should be a smooth function in λ, and f_{YY} does not have any peak or dip at seasonal frequencies. As an natural consequence, the peaks at seasonal frequencies in the sample spectrum of the original time series are inevitably transformed into the dips in the sample spectrum of the seasonally adjusted series. If we allow the sample spectrum of the seasonally adjusted series to have peaks at seasonal frequencies, it is theoretically possible to remove the seasonal dips from $f_{\hat{Y}\hat{Y}}$, but no one would support this idea because this contradicts the first principle posed by Nerlove (1964).

Figure 2.1 shows the plot of the sample spectrum of the original series (Figure 2.1(a)) and the seasonally adjusted series (Figure 2.1(b)) from the Wiener-Kolmogorov filter explained in section 2.2.3. The models are defined in the next section by the equations (2.9) through (2.12), and the innovation variances of the seasonal and non-seasonal component are set to be equal. Without loss of generality, the variance of the observational noise is set to 0. Figure 2.1 (a) and 2.1 (b) show that the seasonal dips in the spectrum of the seasonally adjusted series are the mirror image of the seasonal peaks in the spectrum of the original series. As will be seen later in (2.11), the commonly used seasonal component models imply that the infinite height of peaks at the seasonal frequencies. But the equation (2.8) manifests that the seasonal dips are inevitable whether the peak is finite or infinite. That is to say, even if the roots of characteristics polynomial of the seasonal component model are all located outside the unit circle, the seasonal dips always exist because they are just the turned-over peaks of the spectrum of the original series. Therefore it is a wrong idea that introducing stationary roots in the seasonal polynomial might lead to the removal of seasonal dips. (For example, see Kimura (1997).)

We make an additional remark on the seasonal peaks. Even if the true model implies the divergence of the spectrum at the seasonal frequencies, the height of the peak in the sample version are always underestimated compared to its theoretical level. This can be confirmed



(a) (b)

Figure 2.1: Plot of theoretical log-spectrum of (a) original series and (b) adjusted series.

easily in the following way. See Burman (1980). As regards $T_t + I_t = Y_t$, by the assumption of the mutually uncorrelated noise, we obtain $f_{XX}(\lambda) = f_{YY}(\lambda) + f_{SS}(\lambda)$. Hence the (pseudo) spectrum of the estimated seasonal component is given by

$$f_{\hat{S}\hat{S}}(\lambda) = \left\{ \frac{f_{SS}(\lambda)}{f_{XX}(\lambda)} \right\}^2 f_{XX}(\lambda) = \frac{\{f_{SS}(\lambda)\}^2}{\{f_{SS}(\lambda) + f_{YY}(\lambda)\}^2} f_{XX}(\lambda)$$

and in the same way, the (pseudo) spectrum of the estimated seasonally adjusted series is given by

$$f_{\hat{Y}\hat{Y}}(\lambda) = \frac{\{f_{YY}(\lambda)\}^2}{\{f_{SS}(\lambda) + f_{YY}(\lambda)\}^2} f_{XX}(\lambda).$$

It is clear that we have $f_{\hat{S}\hat{S}}(\lambda) + f_{\hat{Y}\hat{Y}}(\lambda) < f_{XX}(\lambda)$. Therefore, even when the sample power spectrum of the original series does not show the strong evidence on the divergence of the seasonal peaks, it is still legitimate to employ the seasonal model with infinite peaks at the seasonal frequencies, and unit roots in the seasonal component model do not cause the seasonal dips.

To summarize the arguments made so far, the seasonal dips inevitably exist unless we abandon the minimum MSE criterion. Nevertheless, empirical researches are still found here and there that compare the various seasonal adjustment procedures in terms of seasonal dips, but they do not account for the reason why the seasonal dips appear and disappear from time to time. We clarify the reason in the next section.

2.3 Seasonal Dip and Noise Level

2.3.1 Effect of the Variance of Observational Noise

In this section we switch from the signal extraction approach to the Bayesian seasonal adjustment approach. Focusing on a simplified version of DECOMP, we try to reach the deeper

understanding on the appearance of the seasonal dips mainly from the numerical aspects. In BAYSEA and DECOMP it is customary to do model selection by AIC, and if we restrict the class of the models to linear-Gaussian, the maximization of Kullback-Leibler information number is equivalent to the minimum MSE. Furthermore, given the same component models, the signal extraction and the Bayesian approach only differ in the determination of the innovation variance of each component. Therefore, in terms of the interpretation of the estimated results, the findings derived in the previous section apply to the model-based Bayesian seasonal adjustment. A delicate issue shared by the ARIMA model based and the Bayesian seasonal adjustment is that the spectral densities for the trend and the seasonal components are just formally defined because the components model usually contains unit roots in the associated polynomials. On the other hand, it should be noted that the classical solution described in the previous section is extended to the case of the nonstationary time series model, see Bell (1984). Our simulation and empirical analysis are based on the original source of DECOMP. For the details of the modeling and the algorithm employed in DECOMP, see Kitagawa (1981, 1986), and see Kitagawa (1997) for its recent development and involved issues.

In this section, we prepare the following standard settings for the trend and the seasonal component.

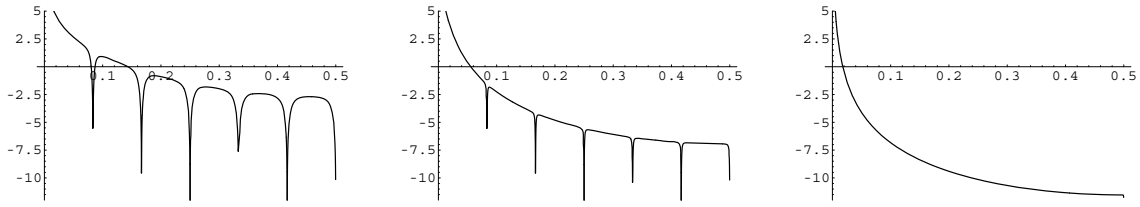
$$
\begin{aligned}
X_t &= T_t + S_t + I_t && (2.9)\\
T_t &= \frac{\varepsilon_t}{(1-L)^2} && (2.10)\\
S_t &= \frac{v_t}{1+\cdots+L^{11}} && (2.11)\\
I_t &= \eta_t. && (2.12)
\end{aligned}
$$

Again, we assume the innovation process $\{\varepsilon_t\}$, $\{v_t\}$ and $\{\eta_t\}$ are mutually uncorrelated white noise sequence, with their variance τ_1^2, τ_2^2 and σ^2, respectively. At first, we observe how the pseudo spectrum of the 'optimal' seasonally adjusted series changes according to the magnitude of σ^2. For the set of equations (2.9) through (2.12), we define the relation corresponding to (2.8) as follows,

$$
\log f_{\hat{T}\hat{T}}(\lambda) = \log f_{TT}(\lambda) - \log\left\{1 + \frac{\tau_2^2}{\tau_1^2}\frac{|1+e^{-i\lambda}+\cdots+e^{-i\lambda\times 11}|^{-2}}{|1-e^{-i\lambda}|^{-4}} + \frac{\sigma^2}{\tau_1^2|1-e^{-i\lambda}|^{-2}}\right\}. \tag{2.13}
$$

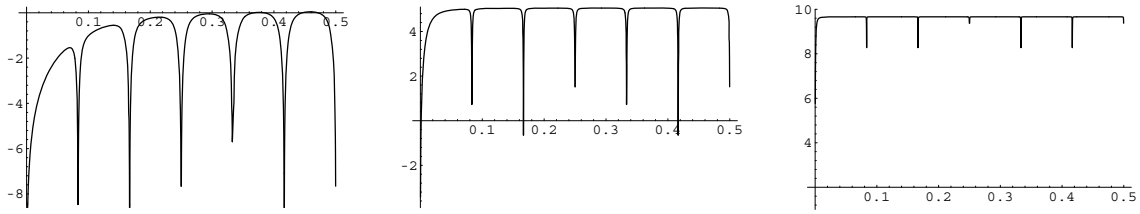
Similarly, we can define the relation corresponding to (2.8) on the spectrum of the observational noise.

$$
\log f_{\hat{I}\hat{I}}(\lambda) = \log f_{II}(\lambda) - \log\left\{1 + \frac{1}{\sigma^2}\left(\tau_1^2|1-e^{-i\lambda}|^{-4} + \tau_2^2|1+e^{-i\lambda}+\cdots+e^{-i\lambda\times 11}|^{-2}\right)\right\} \tag{2.14}
$$



(a) (b) (c)

Figure 2.2: Plot of theoretical spectrum of trend component when (a) $\sigma^2 = 1$, (b) $\sigma^2 = 10^2$, (c) $\sigma^2 = 10^5$.



(a) (b) (c)

Figure 2.3: Plot of theoretical spectrum of irregular component when (a) $\sigma^2 = 1$, (b) $\sigma^2 = 10^2$, (c) $\sigma^2 = 10^5$.

In fact, $\hat{I}_t$ is estimated as a by-product of the estimation of the trend and the seasonal component, hence the above formula is of no practical use and is just considered for the discussion of the spectral properties of the estimated observational noise based on the 'optimal' seasonal adjustment method. In that sense, $f_{\hat{I}\hat{I}}$ will not be purposely estimated.

Figure 2.2 and Figure 2.3 exhibit the behavior of the (pseudo) power spectrum from the 'optimal' seasonal adjustment when the magnitude of the variance of the observational noise varies. In all the cases, the innovation variances of the trend component and of the seasonal component are set to be equal value, 1.0, and only the variance of the observation noise differs. Same in Figure 2.2 and Figure 2.3, σ^2's are set to 1.0 in (a), $\sigma^2 = 100$ in (b), and $\sigma^2 = 10000$ in (c). It should be noted that such an absolute scale as 100 or 10000 does not matter but the relative variance like τ_1^2/σ^2 or τ_2^2/σ^2 is essentially important. In fact, the relative variance can take the seemingly extreme value like 10^{-5} if the trend and the seasonal component are very stable and do not change over time where τ_1^2 and/or τ_2^2 would be very small. The extreme is the case of the deterministic trend and/or the dummy seasonality, which should be discussed in the numerical example 4 in the next subsection.

The message from Figure 2.2 and 2.3 is very clear. As we increase the level of the power spectrum of the observational noise incrementally, $0.45(\approx \log(1/2\pi))$, $5.06(\approx \log(10^2/2\pi))$, $9.66(\approx \log(10^5/2\pi))$, the seasonal dips observed in the spectrum of the irregular component

gradually become invisible in Figure 2.3. Similarly in the (pseudo) spectrum of the trend component, the seasonal dips become imperceptible as σ^2 gets larger (Figure 2.2). We reach the same conclusion by (2.8), (2.13) and (2.14). If σ^2 gets larger, the contribution of the term from which the seasonal dips arise will be relatively weakened. The same mechanism applies not only for the dips at the seasonal frequencies but at the zero frequency which takes place as a result of the trend removal by differencing. The dip at zero frequency in Figure 2.3(a) is mainly caused by the term $\tau_1^2/(\sigma^2|1-e^{-i\lambda}|^4)$ in equation (2.14). As σ^2 becomes large, this 'trend-dip' also becomes invisible while in reality it certainly exists.

From above considerations, we can illustrate the presumable cases where the seasonal dips do not come to the surface.

- The seasonal dips in the power spectrum of the irregular component become invisible (not vanish) as σ^2 becomes large because in the second term of the right hand side of (2.14) the argument of log tends to 1.

- As regards the pseudo spectrum of the trend component, the same holds. In the argument of log in the second term of the right hand side of (2.13), the variance of the observational noise dominates other terms, which results in the subtraction of a large constant from $f_{TT}(\lambda)$. This constant basically depends on λ, but the difference between the seasonal and the non-seasonal frequencies become little as σ^2 gets large. At this moment, $f_{\hat{T}\hat{T}}(\lambda)$ loses substantial power except at zero frequency compared to the true pseudo spectrum, but the seasonal dips are not exposed.

- In the pseudo spectrum of the trend component, if the innovation variance of the seasonal component (τ_2^2) is much smaller than the variance of the trend component (τ_1^2), it is expected that the seasonal dips will become imperceptible. This could happen when the lower order trend model is fitted and its goodness of fit is poor. In such a case, the estimated trend is very noisy, and τ_1^2 naturally gets larger.

As just described, the seasonal dips are inconspicuous or hard to spot when the variance of irregular component (σ^2) is large relative to τ_1^2 and τ_2^2. Conversely, if we force the seasonally adjusted series to be contaminated with the white noise of substantial variance level, we could manage to minimize the saliency of the seasonal dips. But it is not obvious how this ad hoc, excessive-noise-oriented procedure can be justified.

2.3.2 Numerical Examples

So far we have characterized the model-based seasonal adjustment methods. In fact, what is most commonly used method is no doubt the moving average based one which has been developed and improved by the Census Bureau of the U. S. Department of Commerce. In this subsection, we apply DECOMP and X-12-ARIMA to the artificially generated data with various specification of the observational and the system noise, and compare the performance of the two methods. Though the experiments here is far from exhaustive, the results highlight their idiosyncracies in connection with the relationship between the seasonal dips and the relative noise level discussed in the previous sections.

The experimental data are generated in the following manner. In all the experiments the length if time series is set to 150. The generated time series are supposed to be captured by the set of models from (2.9) to (2.12). Hence the trend component is assumed to be generated by the linear deterministic time trend

$$T_t = 6 + \frac{8}{150}t,$$

and the seasonal component is generated by

$$S_t = K(4CP_t - CP_t^{SA}),$$

where $K = 6.94 \times 10^{-4}$, and let CP_t and CP_t^{SA} be the original series of the national consumption expenditure and its seasonally adjusted annual rate series respectively, with their sample span from 1959 Q1 to 1996 Q2. The irregular component I_t is assumed to obey identically, independently distributed with N(0,σ^2). Here we consider three cases of $\sigma^2 = 49,\ 1,\ 10^{-6}$, which will be referred to the Case 1, Case 2 and Case 3. Only for the Case 2, the experiment with deterministic seasonality is conducted, which will be referred to the Case 4. The sum of the generated trend and the observational noise is called 'true seasonally adjusted series' as a matter of convenience.

When fitting DECOMP, the order of trend is fixed to 2, the order of seasonal is set to 1 while the number of seasons are 4, and we exclude the cyclical component expressed by the stationary AR model. As regards X-12-ARIMA, we apply the default setting which will be the typical choice of empirical researchers. Multiplicative seasonality is always assumed, and the results in this chapter basically remains the same even if we employ the additive seasonality.

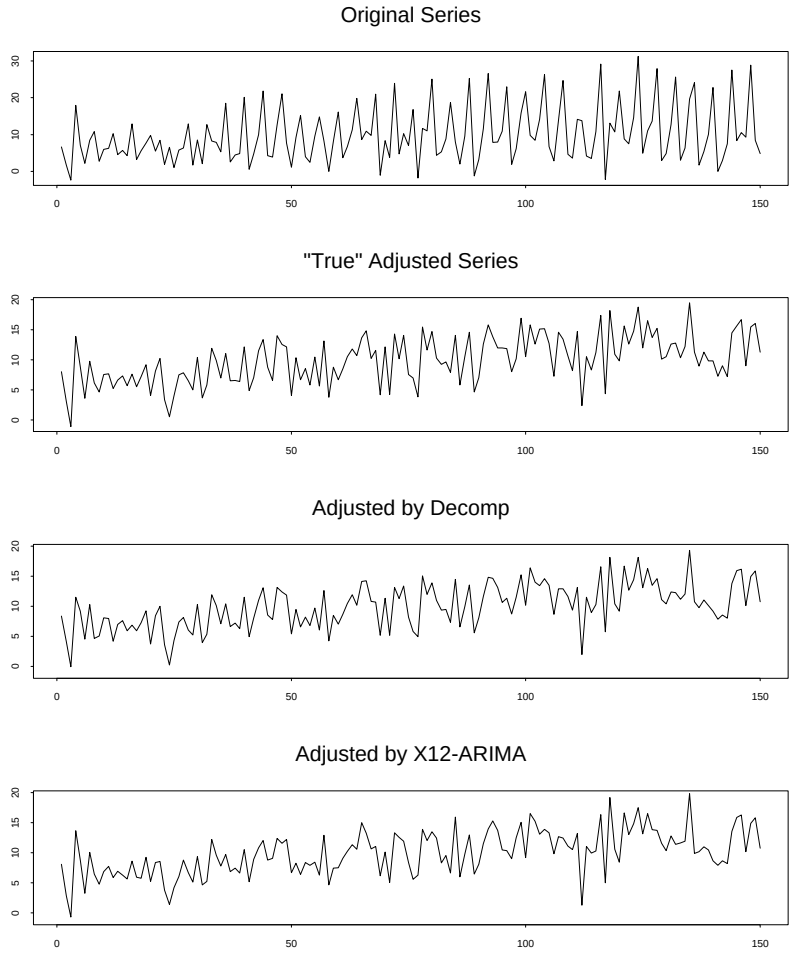
Figures 2.4 through 2.7 consist of 5 panels headed (a) through (e) in order. These are the plots of (a) the original series X_t, (b) 'true' seasonally adjusted series, (c) trend + irregular in DECOMP, (d) seasonally adjusted series by X-12-ARIMA, and (e) is the simultaneous plot

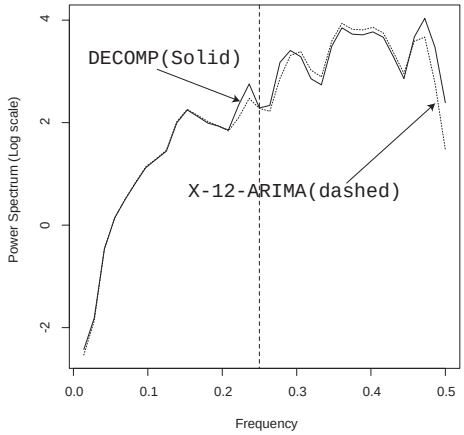
of the log of the power spectrum of (c) and (d) after first differencing. In smoothing the periodograms, Akaike window (Akaike and Nakagawa, 1972) is employed here because we are interested in the sharpness of the peaks and dips more than the overall property on the entire frequency domain.

Now we will examine the results of our numerical experiments. Figure 2.4 reports the results of case 1 where the true data generating process contains excessive observational noise ($\sigma^2 = 49$). According to the knowledge we obtained so far, we expect that the seasonal dips will be inconspicuous though they really exist, which is in fact confirmed by the graphs in Figure 4. The estimated spectra of the first difference of the adjusted series almost look alike, and the seasonal dips are not noticeable. Apparently from the plot of seasonally adjusted series (panel (c) and (d)), we may conclude that the seasonal adjustment is successfully done both in DECOMP and in X-12-ARIMA.

In case 2 (Figure 2.5), we contemplate the intermediate case in the sense that the variance of the observational noise is comparable to that of the innovation. The panel (e) of Figure 2.5 is typically found found in the empirical research that asserts 'the seasonal dips are the evidence of the over adjustment.' (See Kimura (1997) for example.) Certainly, the sample spectrum of the adjusted series by X-12-ARIMA does not have the seasonal dips while the counterpart of DECOMP has. But, what we should care more than the dips is that the seasonally adjusted series given by X-12-ARIMA has a higher level of observational noise over the whole frequency domain. As the plot of the spectrum is drawn in the log scale, even a small leads to the big difference in the actual time series plot, which is clear in the plot of adjusted series (panel (c) and (d) in Figure 2.5). Seemingly, it is difficult to tell which is closer to the 'true' data generating process. The result of DECOMP looks too smooth while the result brought by X-12-ARIMA is too wiggly, so one might say both has failed. At least from this experiment, one may well have an impression that DECOMP tends to yield smoother adjusted series and X-12-ARIMA tends to produce noisy adjusted series. Interestingly, the next experiment supports this conjecture.

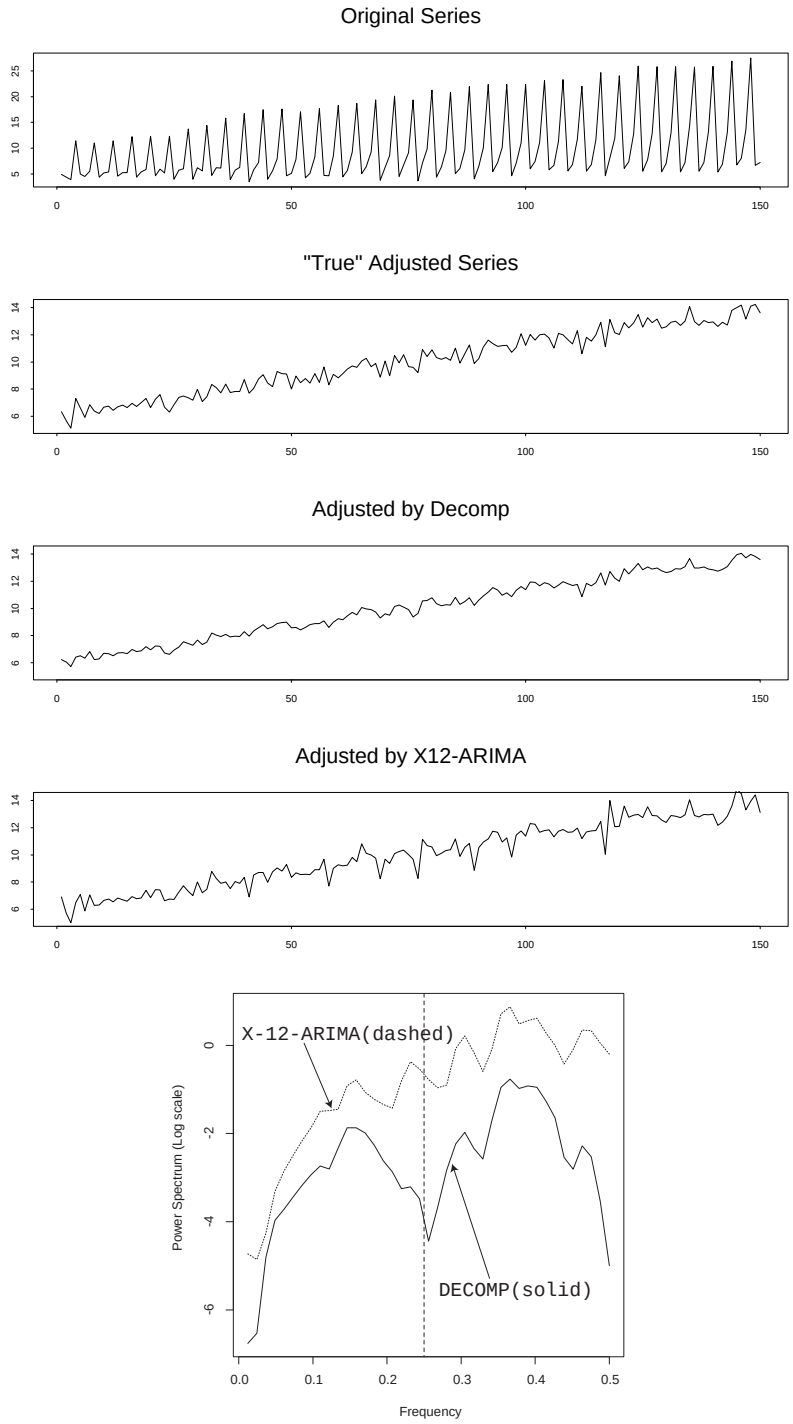
Case 3 exhibited in Figure 2.6 is the case of small observational noise. From the panel (a), we recognize that the signal to be detected is clear and seemingly little observational noise is imposed. This is case is frequently observed in quarterly economic time series. Seasonally adjusted series looks like a deterministic function of t (panel (b) of Figure 2.6) because the true observational noise is too small compared to the variation of the signal. The results of the adjustment are drastically different between DECOMP and X-12-ARIMA. In the panel (e), the levels of the irregular components are distantly away from each other, and the time series plot of the adjusted series (panel (c) and (d)) let us convince that the conjecture in the previous





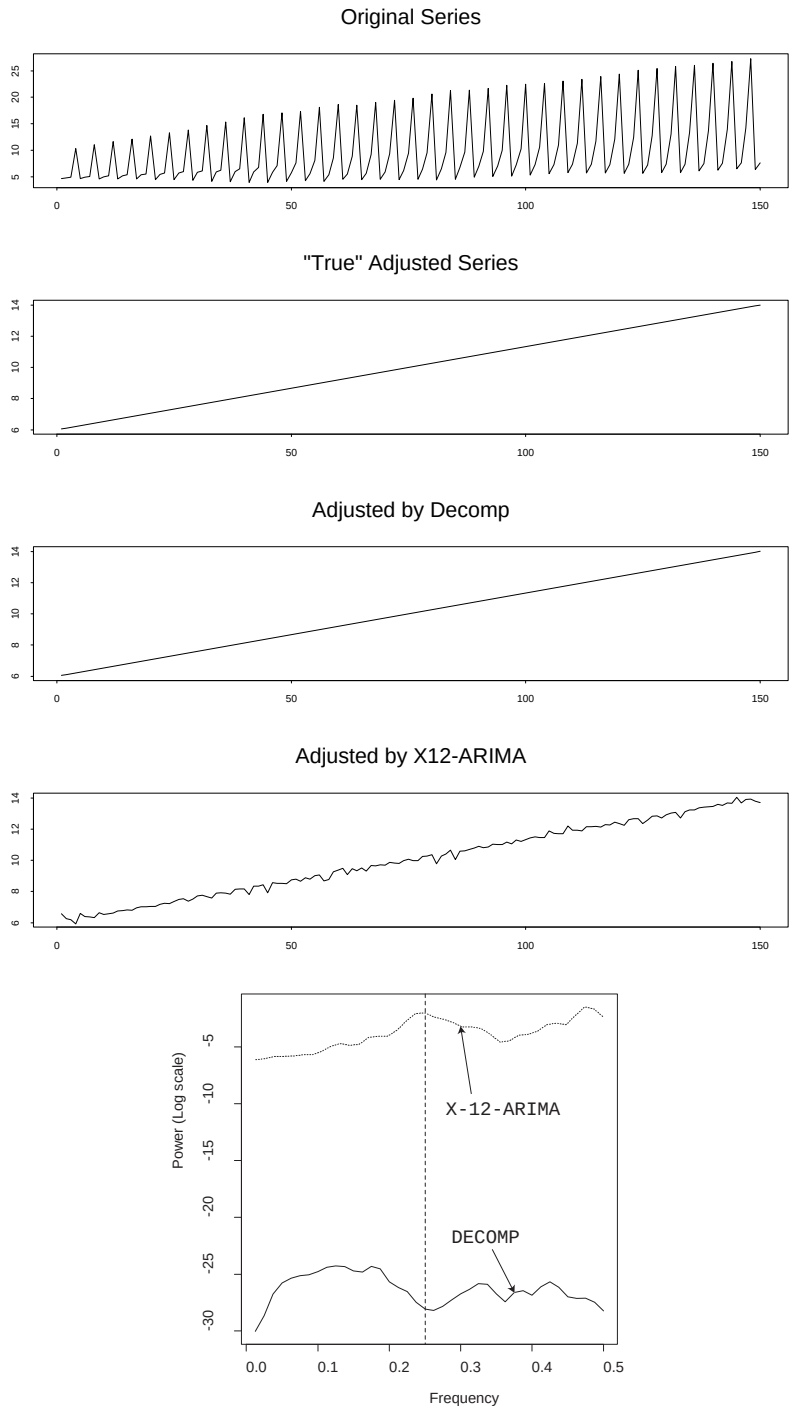
(e) Sample power spectra of two adjusted series

Figure 2.4: Case 1: Excessive observational noise ($\sigma^2 = 49$)



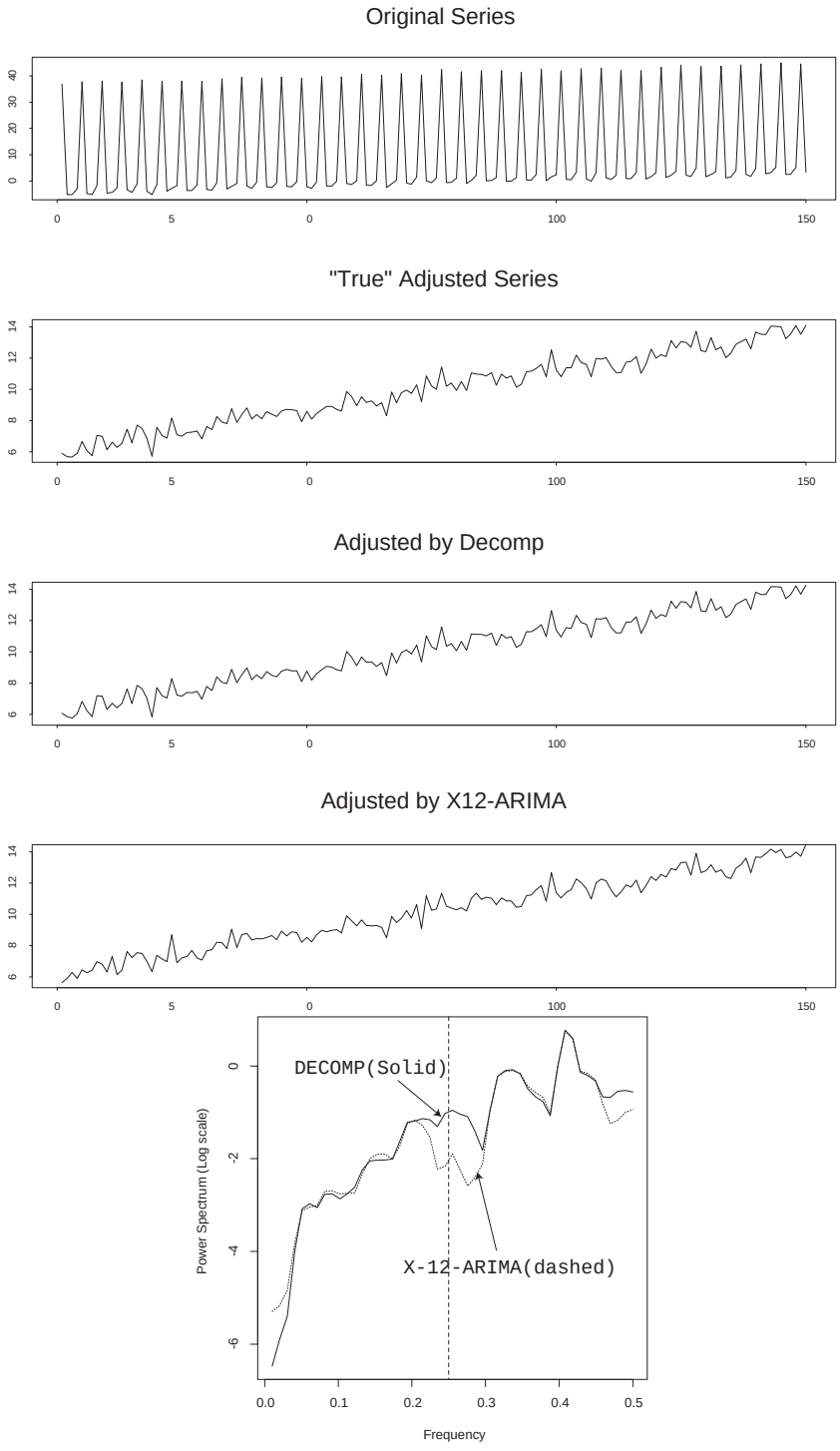
(e) Sample power spectra of two adjusted series

Figure 2.5: Case 2: Intermediate level of observational noise ($\sigma^2 = 1$)



(e) Sample power spectra of two adjusted series

Figure 2.6: Case 3: Almost no observational noise ($\sigma^2 = 10^{-6}$)



(e) Sample power spectra of two adjusted series

Figure 2.7: Case 4: Deterministic seasonality ($\sigma^2 = 1$)

experiment may hold. The results could be changed if we put various options in applying X-12-ARIMA, but this conversely tells us one practical lesson on using X-12-ARIMA under default settings.

As is already mentioned in the previous subsection, the theory suggests that the seasonal dips are likely to come to the surface in such a instance like Case 3. What is important from a practical viewpoint is how large the level of the)pseudo) power spectrum of the seasonally adjusted series is. According to the panel (e) of Figure 2.6, the variation in the first difference of the adjusted series by DECOMP is nearly 0, which shows DECOMP gives the correct answer in this case. Even if the negative autocorrelation is detected in the irregular component or in the first difference of the adjusted series, such correlation affects neither the prediction accuracy nor the seasonal adjustment. Therefore, it is strongly recommended to draw the sample power spectrum of the adjusted series in fist difference. Watching the level of the power spectrum, one may well discuss whether or not the seasonal dips are harmful in light of the prediction accuracy he expects.

Finally, as one of the practically important cases, we consider the case of deterministic seasonality as Case 4. Because $\tau_2^2 \approx 0$ relatively inflates the magnitude of the observational noise, this Case 4 can be identified with the large observational noise case though the actual variance is 1.0. In Figure 2.7 (e), the dips does not seem to be exposed as is suggested by the theory and the results in Case 1. (It is difficult to judge whether the depressed area around $\pi/2$ should be regarded as the seasonal dips or not.)

It is interesting that this case belongs to the same category as the case of excessive observational noise (Case 1) while the plot of the original series gives us the impression that Case 4 would be similar to Case 2 and case 3. Compare Figure 2.7 (a) with the panel (a) of Figure 2.4, 2.5 and 2.6. As is clear in this example, the clear and periodic seasonal pattern does not always lead to the elicitation of the seasonal dips.

2.3.3 Discussion

In the experiments conducted in section 2.3.2, we had some anticipation on the behavior of the adjusted series brought by DECOMP, and actually the results are along what the theory built in section 2.3.1 suggests. As regards X-12-ARIMA, it is not certain to what extent the analogies from the arguments in the signal extraction applies. X-12-ARIMA can be identified with X-11 apart from the preadjustment procedures. Shiskin and Plewes(1978) reports an empirical finding that X-11 sometimes yields noise excessive seasonal adjustment. On the other hand, there are some research that tries to approximate X-11 filter by the signal extraction filter based on

ARIMA models. See Cleveland and Tiao (1976) for example. Unless we explicitly change the type of moving average filter by choosing other options, the application of X-11 filter means that the decomposition of time series is done with a fixed variance ratio. Investigating the properties of the approximate X-11 filter, some researchers suggest that the irregular component estimated by X-11 filter has larger variability than that estimated by the model based method. See Chapter 6.2.3 in Harvey (1989). Under the assumption that the ARIMA model based approximation of X-11 filter is legitimate to some extent, we may conclude that the seasonal dips are much less found in X-12-ARIMA adjustment than other model based adjustment just because X-12-ARIMA tends to yields the noise-contaminated adjusted series.

Based on the arguments made so far, it is meaningless to raise the issue of seasonal dips as an criterion to compare the seasonal adjustment methods. (Actually it has been meaningless since the paper by Grether and Nerlove (1970).) If someone happens to see the empirical research in which the seasonal dips are treated in connection with the legitimacy of seasonal adjustment, it is advised to check the noise level of the irregular component, as is done in Figure 2.5 (e). One must be cautious particularly in case the spectrum of the adjusted series *in level* is reported because the difference in the noise levels of the irregular components between competing models will be disguised due to the problem of drawing scale.

In this chapter, we have shown that the construction of seasonal adjustment filter by the minimum MSE criterion inevitably cause the seasonal dips, and described the situations where the seasonal dips should come to the surface. It is felt that the absence or the presence of the seasonal dips seemingly has nothing to do with the quality of the seasonal adjustment procedures.

An article by Ansley and Wecker (1984) is interesting because it offers some speculations on if we can consider another statistical criterion that is compatible with no seasonal dips. Their idea is deceptively easy; if we take the square root of the optimal filter in (2.6), then we have no seasonal dips in the (pseudo) power spectrum of the adjusted series. Namely, taking $(f_{TT}/f_{XX})^{1/2}$ as a new filter, we obtain

$$f_{\hat{T}\hat{T}}(\lambda) = \left\{ \left(\frac{f_{TT}(\lambda)}{f_{XX}(\lambda)} \right)^{\frac{1}{2}} \right\}^2 f_{XX}(\lambda) = f_{TT}(\lambda).$$

Attention must be paid not to be mislead by the fact here. The point of Ansley and Wecker (1984) is *not* that we can make the seasonal adjustment filter that produce no seasonal dips *but* that they investigated the negative effect of the filter $(f_{TT}/f_{XX})^{1/2}$. To summarize the points, at first the new filter is inferior to (2.6) in the sense of minimum MSE. They present a measure to gauge the loss in MSE by requiring the estimated spectrum to coincide the true spectrum under

the condition that the true data generating process is known. The second, more suggestive result is that only one component can be forced to have the same spectrum that the true component model has, but it is impossible to attain the same results for other components simultaneously. In other words, if we construct the filter so that $f_{\hat{Y}\hat{Y}} = f_{YY}$, then $f_{\hat{S}\hat{S}}$ gets the short end of stick. Moreover, it is not obvious how the Ansley-Wecker filter can be implemented in time domain.

To add some comments on the Ansley-Wecker filter, some empirical researchers quote this paper asserting that “it is possible to derive the minimum MSE filter imposing the restriction that we have no seasonal dips", but this is completely misleading. In the first place, the Ansley-Wecker filter is not the minimum MSE filter as they have shown by themselves. In the second place, the Ansley-Wecker filter is derived as a sufficient condition for the filter without seasonal dips, hence they did not derived the necessary condition that a filter with no seasonal dips should satisfy. It is an interesting future topic if some filter can be derived as a necessary condition for no seasonal dips.

2.4 Conclusion

Within the class of linear Gaussian model, if we form the optimal (time-invariant) seasonal adjustment filter by the minimum MSE criterion, the seasonal dips are inevitable. Namely, there is a trade-off between the presence of the dips and the goodness of the estimation and the prediction. If the seasonal dips are inconspicuous in the sample (pseudo) power spectrum of the adjusted series, it is doubted that the dips are buried in the relatively excessive observational noise. This is theoretically transparent at least in the framework of the model based seasonal adjustment methods like the signal extraction and the Bayesian method. Things are going more or less same in X-12 ARIMA because X-11 procedure can be well approximated by the seasonal ARIMA model and the structural time series model. As a practical matter, practitioners should care that X-12 may provide the noisy seasonal adjustment. Conversely, if this observation really gets the point, the assertion that “the seasonal dips are seldom found in the adjusted series by X-12-ARIMA" ironically makes sense, but it reveals that X-12-ARIMA sacrifice the statistical goodness of fit and predictability. From the numerical experiments conducted varying the variance of the observational noise and the seasonal pattern, it is concluded that the absence or the presence of the seasonal dips seemingly has nothing to do with the quality of the seasonal adjustment procedures. Moreover, the experiments here successfully characterize the hidden features of DECOMP and X-12-ARIMA. Reading a monograph like Hylleberg (1986) gives us an impression that the issue of the seasonal dips seems to be definitely-settled matter, while

the conclusion of Grether and Nerlove (1970) still sounds paradoxical. Can we find another optimality in statistical sense that proves to be seasonal-dip-free? Such a new direction of research is desirable, and it is the time to reject the looming empirical research reporting the absence or presence of the seasonal dips.

Chapter 3

A Structural Time Series Model Facilitating Flexible Seasonality

3.1 Introduction

Time series with trend and seasonal components is an important generalization of nonstationary mean time series. Such time series occur for example in meteorological, oceanographic and economic studies. Trend and seasonal, given time index t, are the unobserved components to be squeezed out of the original observation. This ill-posed nature requires the introduction of reasonable stochastic constraints on the unobserved components, which leads to some Bayesian treatment. One of the earliest works of such Bayesian modeling is perhaps Harrison and Stevens (1971). Akaike (1980) presented a sophisticated methodology on Bayesian seasonal adjustment method which admits computationally straightforward penalized least squares methods. In estimating such a decomposition problem in a Bayesian framework, a Markovian representation via state space form is very useful. Kitagawa (1981), Gersch and Kitagawa (1983) and Kitagawa and Gersch (1984) extended the ideas in Akaike (1980) to state space formulations and incorporated seasonal decomposition features. Harvey and Todd (1983) and Harvey (1984, 1985) went side by side along with this line of research though a Bayesian point of view is not very much stressed. West *et al.* (1985) and West and Harrison (1986) developed a fully Bayesian framework while Akaike's methodology (and his follower's too) can be regarded as empirical- or quasi-Bayesian approach. In this paper we consider the modeling of nonstationary mean time series with trend and seasonal by state space methods.

The economic application of seasonal adjustment time series have provoked an extensive literature and a variety of software. Popular software products based on state space modeling are Kitagawa's DECOMP in TIMSAC 84 (Akaike *et al.* , 1985) and STAMP (Koopman *et al.* , 2000) initially developed by Harvey. Empirical researchers using such softwares sometimes

complain that the estimated seasonal patterns are very steady considering the use of the stochastic seasonality. In a state space modeling, the most commonly used seasonal component model is the stochastic dummy seasonality, that is, the sum of seasonal factors within a period follows zero mean white noise. Frequently observed stiff seasonal patterns are rather due to this stochastic dummy specification, not to the state space formulation itself.

One way to increase the seasonal variability is to introduce a fat-tailed distribution for the innovation of seasonal component. In Kitagawa (1989), the spline based numerical integration developed in Kitagawa (1987) is applied for seasonal adjustment, while Kitagawa (1994) proposes a new implementation for the Gaussian-sum smoother. Monte Carlo filter presented by Kitagawa (1996) is also applicable to a non-Gaussian seasonal adjustment. A method proposed by Shepherd and Pitt (1997) is also applicable though their presentation puts emphasis on non-Gaussian measurement. However, these techniques are more likely to be instrumental to depict the abrupt change in the seasonal pattern or to detect the outliers automatically, or when the measurement distribution is primarily non-Gaussian. In addition, there still remains a computationally intensive task, and often a user-friendly tool is not available for practitioners to perform it.

The aim of this article is to present a new model to increase the flexibility and variability of seasonal pattern within a framework of structural time series model. We still stay at the linear Gaussian assumption, hence the estimation is quite easy and fast. The empirical analysis performed in this article demonstrates that the proposed method is richly expressive in estimating the seasonal component, and that such a decomposition is supported in terms of one-step ahead prediction, too.

This paper is organized as follows. In section 3.2 we briefly review the framework of basic structural model for time series with trend and seasonality. A state space form for the BSM and its estimation algorithm are also reviewed in subsection 3.2.2 and 3.2.3. In section 3.3, a parsimonious modeling to produce flexible seasonality is presented. After introducing some existing methods in subsection 3.3.1, a model with seasonal summation driven by a finite order moving average process with a single unknown parameter is proposed in subsection 3.3.2. A state space form for the proposed model are described in section 3.3.4. Section 3.4 details the results of the empirical analyses on 11 economic time series of U.K. and Japan. Section 3.5 concludes this article.

3.2 Modeling Trend-Seasonality

3.2.1 Basic Structural Model

As a basis of the discussion of this article, this section explains a popular model called basic structural model. Modeling trend-seasonality with state-space form have been explored since the end of 1970's. Trend and seasonal are regarded as unobservable components, and for each unobservable component a stochastic model is assumed. One of the most popular specification is a set of the equations described as follows.

$$y_t = \mu_t + \gamma_t + \varepsilon_t \tag{3.1}$$

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t \tag{3.2}$$

$$\beta_t = \beta_{t-1} + \zeta_t \tag{3.3}$$

$$\gamma_t = -\sum_{j=1}^{s-1} \gamma_{t-j} + \omega_t \tag{3.4}$$

In above equations, we assume that each of ε_t, η_t, ζ_t, ω_t follows zero mean normal distributions but with different variance; σ_ε^2, σ_η^2, σ_ζ^2 and σ_ω^2 respectively. This set of equations is often referred to as Harvey's basic structural model (BSM hereafter), see Harvey (1989, p.47). Equation (3.1) is called observational (or measurement) equation. This reflects our observation that the salient features of economic time series are trend μ_t and seasonality γ_t, and the rest is regarded as irregular component ε_t.

Trend component consists of two latent variables μ_t and β_t, which is respectively referred to 'stochastic level' and 'stochastic slope'. The equation (3.2) plus (3.3) is called the local linear trend model. The name comes from the fact that the drift term β_t plays a role of a linear trend rather than a constant in (3.2). On the other hand, it is also possible to consider the following trend model in stead of (3.2) plus (3.3);

$$\mu_t = 2\mu_{t-1} - \mu_{t-2} + \eta_t. \tag{3.5}$$

If we rewrite (3.5) as $\mu_t = \mu_{t-1} + (\mu_{t-1} - \mu_{t-2}) + \eta_t$, it is easily understood that (3.5) is a special case of the local linear trend model in the sense that the stochastic slope β_t is also driven by the same process η_t rather than by a different process ζ_t. From now on, trend model is fixed to (3.5) in this article, and the seasonal adjustment model (3.1) together with (3.5) and (3.4) will be referred to the BSM again, as this will make no confusion here.

3.2.2 State Space Form

In order to facilitate the introduction of extended models in the next section, we sum up the basic outline of the state space representation of the BSM and its estimation. Due to the assumption of no correlation among innovation and noise process, the state space representation can be built up as a composition of small state space models for the individual components. To save space, we assume $s = 4$ just for the presentation purpose. From equation (3.5) and (3.4), it turns out that the essential quantity that determines the present distribution of μ_t and γ_t will be given by a vector

$$\alpha_{t-1} = (\mu_{t-1}, \mu_{t-2}, \gamma_{t-1}, \gamma_{t-2}, \gamma_{t-3})' \tag{3.6}$$

where the prime ($'$) denotes the transpose of a vector or a matrix. By setting submatrices as follows,

$$T_1 = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix}, \; T_2 = \begin{bmatrix} -1 & -1 & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \; R_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \; R_2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

new matrices T and R are defined as

$$T = \begin{bmatrix} T_1 & O \\ O & T_2 \end{bmatrix}, \quad R = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix}, \quad \eta_t = \begin{bmatrix} \eta_t \\ \omega_t \end{bmatrix}. \tag{3.7}$$

Then the transition of the state vector can be written in a matrix notation as

$$\alpha_t = T\alpha_{t-1} + R\eta_t. \tag{3.8}$$

As we observe that the measurement equation (3.1) just extracts and adds the components μ_t and γ_t, defining $z' = (1,0,1,0,0)$ yields the relation between the observation and the state as

$$y_t = z'\alpha_t + \varepsilon_t. \tag{3.9}$$

Now the BSM is put in a state space form by (3.8) and (3.9). Note that the specification of α_t, T, R and z' described above is not a unique one because the transformations of these vectors and matrices by a regular square matrix still give rise to the same state space model.

3.2.3 Model and State Estimation

Let a_{t-1} denote the minimum mean squared error (MMSE) estimator of α_{t-1} based on the observations up to time $t-1$. Let P_{t-1} denote the $m \times m$ covariance matrix of the estimation error, i.e.

$$P_{t-1} = \mathrm{E}[(\alpha_{t-1} - a_{t-1})(\alpha_{t-1} - a_{t-1})'].$$

Given a_{t-1} and P_{t-1}, the MMSE estimator of α_t and the covariance matrix of the estimation error is given by

$$\begin{aligned} a_{t|t-1} &= Ta_{t-1} \\ P_{t|t-1} &= TP_{t-1}T' + RQR' \end{aligned}$$

where $Q = \mathrm{diag}(\sigma_\eta^2, \sigma_\omega^2)$. These two equations are known as the *prediction equations*.

Once the new observation, y_t, becomes available, the estimator of α_t, $a_{t|t-1}$, can be updated. The *updating equations* are given by the following two equations,

$$\begin{aligned} a_t &= a_{t|t-1} + P_{t|t-1}z' f_t^{-1}(y_t - z'a_{t|t-1}) \\ P_t &= P_{t|t-1} - P_{t|t-1}z' f_t^{-1} z P_{t|t-1} \end{aligned}$$

where $f_t = z'P_{t|t-1}z + \sigma_\varepsilon^2$. Repetition of prediction and updating constitutes so-called the Kalman filter.

Unless $\sigma_\varepsilon^2 = 0$, the estimation problem of a state space model is double-folded. Given the unknown hyperparameters $\psi = (\sigma_\varepsilon^2, \sigma_\eta^2, \sigma_\omega^2)'$, running Kalman filter and fixed interval smoother yields the estimates of unobservable components $\{\hat{\mu}_t\}_{t=1}^T$, $\{\hat{\gamma}_t\}_{t=1}^T$ and hence $\{\hat{\varepsilon}_t\}_{t=1}^T$. The vector of unknown parameters, ψ, can be estimated by the maximum likelihood method. The likelihood function for a time series can be decomposed into the product of the density functions of one step ahead prediction error $v_t = y_t - z'a_{t|t-1}$. The variance of observation noise σ_ε^2 usually can be concentrated out of the likelihood function. Let $\psi^* = (\sigma_\eta^2, \sigma_\omega^2)'$, then

$$\log L_c(\psi^*) = -\frac{1}{2}\left\{ T\log 2\pi\tilde{\sigma}^2(\psi^*) + \sum_{t=1}^{T} \log f_t + T \right\}$$

must be maximized with respect to the unknown parameters ψ^*, while $\tilde{\sigma}^2(\psi^*)$ is given by

$$\tilde{\sigma}^2(\psi^*) = \frac{1}{T}\sum_{t=1}^{T} \frac{v_t^2}{f_t}.$$

Model comparison will be done based on AIC (Akaike, 1973). As regards the initial state settings, we employ the 'large κ approximation' (Harvey, 1989, p.121). The specific value for κ employed here will be stated in section 3.4 in conjunction with the scale of the time series. Once the unknown hyperparameters are estimated, then the unobserved components are estimated by the fixed interval smoother. For the algorithm of the fixed interval smoother, see Anderson and Moore (1979, p.187–190), Harvey (1989, p.154) or Kitagawa and Gersch (1996, p.58).

3.3 Parsimonious Modeling toward Flexible Seasonality

In this section, three seasonal component models which will be compared in the real data analysis section 3.4 are presented. In the subsection 3.3.1, the standard model (the BSM) and its existing modification are presented. In the section 3.3.2, a parsimonious modeling of MA driven seasonal summation will be introduced.

The basic motivation of this paper is to examine the appropriateness of a new seasonal model which increases the variability of the seasonal component. However, it is *not* because we believe that larger variance of the seasonal component is desirable. What is sought in this paper is, at first, to prepare the framework which allows us to estimate more flexible seasonal component than the BSM, and secondly, to investigate through empirical analysis whether such a model is really favored or not in terms of a model selection criterion. If we estimate a seasonal ARIMA model and try to decompose it to do seasonal adjustment, it is known that there is no unique solution for such a decomposition. Then we must place an arbitrary assumption on the allocation of the variance contributions among trend, seasonal and other components under consideration. For example, Box et al. (1978) asserts that the variance of the seasonal component should be minimized. It should be noted, however, that we do not have to employ such an arbitrary criterion because we only have to determine the hyperparameters by the maximum likelihood method and compare the candidate model by AIC statistic.

3.3.1 Driving Noise of Seasonal Summation

Let s be the number of seasons observed in a period. For economic time series, the length of a period is usually one year and the cases of $s = 4$ and 12 draw great deal of attention. Now, define the seasonal summation operator $S(L)$ by

$$S(L) = 1 + L + \cdots + L^{s-1}. \tag{3.10}$$

Then a concise expression $S(L)\gamma_t = \omega_t$ can be given to (3.4). This model is often referred to as 'dummy seasonality', see Harvey (1989). To avoid a possible confusion with deterministic dummy seasonality, we call this model the stochastic dummy seasonality. The model (3.4) can be regarded as a stochastic constraint on seasonal component such that the sum of s-consecutive seasonal factors will follow zero mean independent random variable. Though the seasonal component γ_t can vary as time evolves, the seasonal pattern cannot change very much if the estimated dispersion parameter $\hat{\sigma}_\omega^2$ is very small. As is already defined in section 3.2.1, the second order difference equation (3.5) plus (3.10) will be referred to as the BSM in the data analysis section 3.4.

There have been several researches to allow more flexibility for the seasonal component. One idea is to employ trigonometric seasonal specification, see Hannan, Terrell and Tuckwell (1970) and Harvey (1989, p.41–42). But a scepticism may be cast on this model that a more emphasis is put on the evolution of separate seasons than on the serial association of consecutive seasons. Another obvious drawback is that this approach requires many additional hyperparameters while we cannot always expect the gain in fitting accuracy. In such a case, though it depends on the situation, the number of hyperparameters may be reduced by the equality/zero constraints on some of the dispersions of seasonal components.

Kitagawa and Gersch (1984, p.386) introduce a higher order seasonal polynomial such that $S^2(L)\gamma_t = \omega_t$. This type of modeling can cope with gradual change in seasonal pattern while its difficulty is that the state dimension of the model becomes larger. This extension has been already implemented in the software DECOMP in TIMSAC-84, and is also available on Web-Decomp. (As regards Web-Decomp, see Sato (1997) and the web site he maintains, `http://ssnt.ism.ac.jp/inets2/title.html`. If it is not active any more, visit `http://www.ism.ac.jp/~sato/` and find the entrance for Web-Decomp.)

Seasonal model (3.4) can be viewed as the seasonal summation is driven by a white noise process. One idea to bring more variability to the stochastic dummy seasonality is to replace the white noise by the 'colored' (i.e., autocorrelated) noise,

$$(1-\Phi L)S(L)\gamma_t = \omega_t. \tag{3.11}$$

Provided that the variance σ_ω^2 is the same, the unconditional variance of the seasonal summation, $\sigma_\omega^2/(1-\Phi)^2$ is greater than σ_ω^2 as long as $\Phi < 1$. Ozaki and Thomson (1992) call (3.11) the pink-noise driven seasonal component model because the innovation process $\omega_t/(1-\Phi L)$ has much power at lower frequencies that reminds us of the infrared ray. Throughout this article, the second order stochastic trend (3.5) plus (3.11) will be referred to as the BSM-AR.

3.3.2 Seasonal Summation Driven by MA

To increase the seasonal summation variability, it appears more direct to introduce a finite order MA process on the right hand side of (3.10). Let $\Theta(L)$ denote a certain form of polynomial of the backward shift operator L, then the seasonal component model can be written as $S(L)\gamma_t = \Theta(L)\omega_t$. What we concern about is not a formal extension of models but a practical guideline to specify $\Theta(L)$.

A motivation for MA-driven seasonal model is found, for example, in the past effort to build a time series model which is expected to play the same role as a conventional seasonal adjust-

ment procedure. From 1970's to early 80's, many researchers sought unobserved components models that approximate Census X-11 seasonal adjustment procedure. See Cleveland and Tiao (1976), Wallis (1982), Burridge and Wallis (1984), for example. Some authors proposed decomposition methods based on seasonal ARIMA model, Burman (1980) and Hillmer and Tiao (1982) to name a few. As Burridge and Wallis (1984) pointed out, models with a predominantly autoregressive specification generate long signal-extraction filters that are not able to approximate the relatively rapid decline of the X-11 filter coefficients. Because the seasonal model in the BSM involves only AR polynomial in (3.10), the corresponding filter coefficients cannot be ignored even in the remote lags. In this context, there is a clear preference for moving average dominated specifications in modeling seasonality. In general, many articles in this period recommend the inclusion of a seasonal moving average model. For example, Burridge and Wallis (1984) proposed the following component model (for monthly economic time series) which will correspond to the symmetric filter in Census X-11 procedure,

$$S(L)\gamma_t = (1 + 0.71L^{12} + 1.00L^{24})\omega_t.$$

Another motivation emanates from accumulated experiences on seasonal ARIMA model fitting to seasonal time series. Since Box and Jenkins (1976), there have been many empirical works to show that ARIMA(0,1,1)$\times$(0,1,1)$_s$ accounts for a wide variety of economic time series with seasonality. This econometric folklore tells us that, for a detrended series $\bar{y}_t$,

$$(1 - L^s)\bar{y}_t = (1 - \Theta L^s)\varepsilon_t,\ |\Theta| < 1 \tag{3.12}$$

fits reasonably. Seasonal differencing operator $1 - L^s = (1 - L)(1 + L + \cdots + L^{s-1})$ involves the usual differencing operator $1 - L$. To avoid common factor between trend and seasonal component model, only the summation type operator is considered when modeling seasonality in a structural time series model. Hence the equation (3.12) suggests the following seasonal component model,

$$S(L)\gamma_t = (1 + \Theta L + \Theta^2 L^2 + \cdots + \Theta^{s-1}L^{s-1})\omega_t. \tag{3.13}$$

where $|\Theta| < 1$. Provided that the variance σ_ω^2 is the same, the unconditional variance of the seasonal summation, $(1 + \Theta + \Theta^2 + \cdots + \Theta^{s-1})\sigma_\omega^2$ is greater than σ_ω^2 unless $\Theta = 0$. Throughout this article, the second order stochastic trend (3.5) plus (3.13) will be referred to as the BSM-MA. Originally, the words BSM, BSM-AR and BSM-MA refer to a set of equations. From now, these words will be used as if they point only to the seasonal component models in some cases , but this will not invite misunderstanding.

In stead of (3.13), we could start with a more general model such as

$$S(L)\gamma_t = (1+\theta_1 L+\theta_2 L^2+\cdots+\theta_{s-1}L^{s-1})\omega_t. \tag{3.14}$$

There are a couple of reasons why we do not employ this general form. Firstly, (3.14) is exhaustive in the number of parameters. Second, even if we employ the 'general-to-specific' modeling strategy, there is another parameterization to take over (3.14) that gives us a more intuitively natural interpretation of the parameters, and that shows a more reasonable way to put restrictions on the parameters. Finally, such a specification enables us to understand the offset effect of the BSM-MA, which will be clarified in the next subsection.

3.3.3 Pseudo-Spectrum Offset around Seasonal Frequencies

In this subsection, we shed another light on the role of MA term in seasonal component model. It is known that the simultaneous use of AR and MA operators can mimic a 'line spectrum' when the zeros of both operators have a common argument. Whether it gives rise to a peak or a trough depends on the magnitude relation of the modulus of roots. If the modulus of AR roots are greater/smaller than those of MA roots, then the spectrum has peaks/troughs. Let us consider the power spectrum $f(\lambda)$ $(0 \leq \lambda \leq \pi)$ of ARMA(2,2) process given by

$$f(\lambda) \propto \left| 1-\sum_{j=1}^{2}\Theta_j e^{-2\pi i j\lambda} \right|^2 \Big/ \left| 1-\sum_{j=1}^{2}\Phi_j e^{-2\pi i j\lambda} \right|^2,$$

where Φ_1 and Φ_2 are fixed to $0.99\sqrt{2}$ and $-(0.99)^2$ respectively, and $\Theta_1 = \Theta\sqrt{2}$ and $\Theta_2 = -\Theta^2$ vary dependent on the parameter Θ. Because we only consider $|\Theta| < 1$ cases, the power spectrum $f(\lambda)$ has its peak at $\lambda^* = 0.083$.

Four panels in Figure 3.1 shows how the shape of $\log f(\lambda)$ changes as Θ tend to unity. The left-upper panel ($\Theta = 0$) corresponds to the BSM, in which case the power spectrum is widely spread around $\lambda^* = 0.083$. It is easily seen that the peak becomes sharper and more concentrated around λ as Θ tends to unity. In addition to that, the level of power spectrum except for the peak frequency gets flatter as $\Theta \to 1$. In other words, simultaneous use of AR and MA polynomials with common argument offsets the power spectrum at all frequencies but the common argument (in this example λ^*=0.083). As a result, the shape of log-power spectrum closely resembles that of a line spectrum apart from constant.

If we turn to the seasonal component models (3.4), (3.11) and (3.13), the power spectrum cannot be defined any more because the process is not stationary due to the unit roots contained in $S(L) = 0$. Even for such a case, a formally defined spectrum called pseudo-spectrum is often

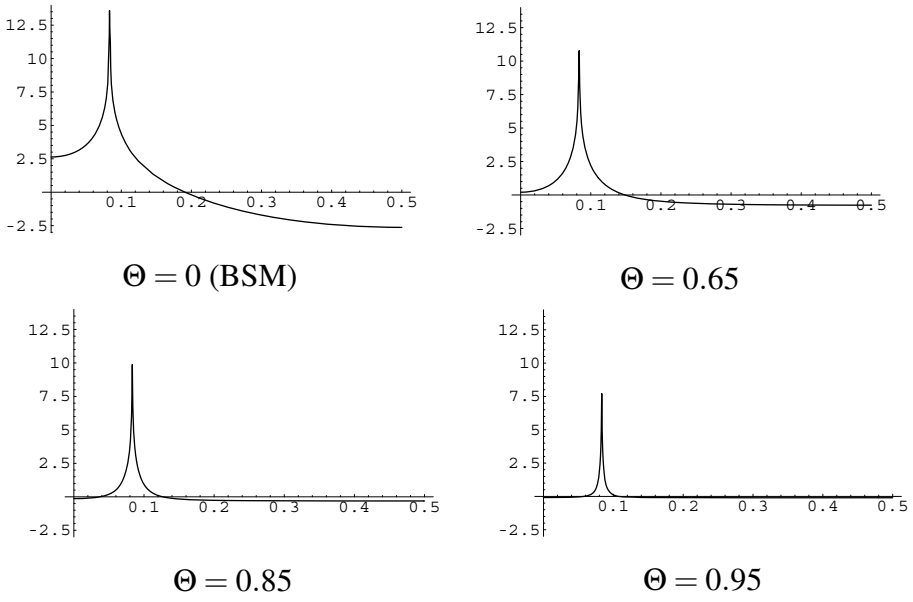


Figure 3.1: Log-spectrum shape of ARMA(2,2) processes.

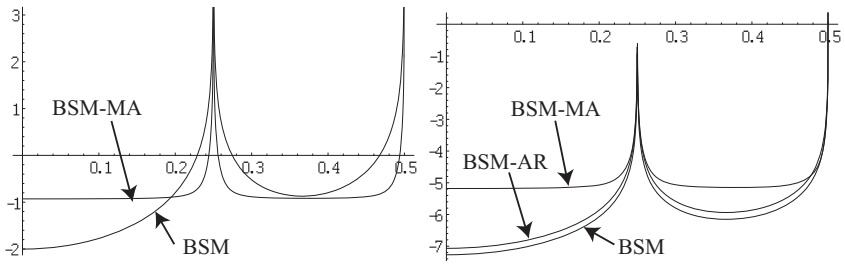


Figure 3.2: Left: Log of pseudo-spectra of the BSM and BSM-MA (with $\Theta = 0.9$), Right: Estimated pseudo-spectra of seasonal component for private sector consumption in Japan (PC-SMP).

considered to characterize a time series. See Harvey (1989, p.64). As an example, the log of pseudo-spectra of seasonal components for both BSM and BSM-MA with $\Theta = 0.9$ are drawn in the left panel of Figure 3.2. For simplicity, we assign the same value to σ_ω^2 in both models, and the root of AR polynomial is slightly pitched outside the unit circle just for drawing this figure. (Theoretically, two peaks at $\lambda = 0.25, 0.5$ should be infinite.) We observe that the seasonal component of the BSM accounts for not only the power at seasonal frequencies but also the substantial portion of the power at the neighboring frequencies. Considering the offset effect mentioned above, there is a possibility that BSM-MA may prevent the BSM from excessively removing the frequency components neighboring the seasonal frequencies. In addition, the seasonal component of BSM-MA with $\Theta = 0.9$ has more power at lower frequencies than the BSM, which is also due to the offset effect. Meanwhile, the stochastic dummy model loses its power at lower frequencies and enhances the power at higher frequencies.

In terms of the smoother mean, however, the most crucial factor is the level of the power spectrum, σ_ω^2. Technically, $\sigma_\omega^2 \approx 0$ is also required for $f(\lambda)$ to be close to a line spectrum. So it is legitimate to say that the simultaneous use of AR and MA terms with a common argument will produce an almost deterministic periodicity contaminated with a white noise of which level completely depends on the time series in question. If $\Theta \to 1$ and $\sigma_\omega^2 \to 0$ simultaneously, the estimated seasonal factor should be almost deterministic. This is the case of 'cancellation', see Box and Jenkins (1976, p. 248). Then, removing the seasonality in advance by deterministic dummy variables would be appropriate. If $\Theta \to 1$ but $\sigma_\omega^2 >> 0$, then the seasonal factor will be interpreted as the deterministic dummy variables on which a white noise process is superimposed. The right panel of Figure 3.2 shows the log of the estimated power spectra of the seasonal components of BSM, BSM-AR and BSM-MA for Japanese private sector total consumption. As is common with many series analyzed in the section 3.4, the power level of the seasonal component is much increased and flattened by employing BSM-MA.

ARMA modeling with common argument and different modulus can be introduced to all the seasonal frequencies of the stochastic dummy model. Let us denote the fundamental seasonal frequency and its harmonics by $\lambda_j = 2\pi j/s$ for $j = 1, \ldots, [s/2]$ where

$$[s/2] = \begin{cases} s/2 & \text{for } s \text{ even} \\ (s-1)/2 & \text{for } s \text{ odd} \end{cases}$$

The seasonal summation operator can be written as the product of the full complement of trigonometric operators, i.e.,

$$S(L) = \prod_{j=1}^{[s/2]} \gamma_j(L)$$

where

$$\gamma_j(L) = 1 - (2\cos\lambda_j)L + L^2, \quad j = 1, \ldots, [s/2]$$

when s is odd. When s is even, $\gamma_j(L)$ is defined as above for $j = 1, \ldots, s/2-1$, while for $j = s/2$ it is

$$\gamma_{s/2}(L) = 1 + L.$$

For monthly data the seasonal summation operator can be factorized into the six trigonometric operators, see Harvey (1989, p.22) for example. Let us take a look at one of the trigonometric AR polynomials, $1 - \sqrt{3}L + L^2$, which corresponds to 12 months period. Hence introducing MA term polynomial $1 - \sqrt{3}\Theta_1 L + \Theta_1^2 L^2$ with $-1 < \Theta_1 < 1$ leads to a sharp peak at 'once-in-a-year' frequency, $\lambda_1 = \pi/6$. Allowing MA term at all the seasonal frequencies, the stochastic dummy seasonal component model can be extended to take the following form,

$$\begin{aligned} S(L)\gamma_t &= (1 - \sqrt{3}\Theta_1 L + \Theta_1^2 L^2)(1 - \Theta_2 L + \Theta_2^2 L^2)(1 + \Theta_3^2 L^2) \\ &\times (1 + \Theta_4 L + \Theta_4^2 L^2)(1 + \sqrt{3}\Theta_5 L + \Theta_5^2 L^2)(1 + \Theta_6 L)\omega_t \end{aligned}$$

where the $|\Theta_j|$'s are all expected to be less than unity. If the six parameters $\Theta_1, \ldots, \Theta_6$ are estimated freely, it means that peak properties can differ by the seasonal frequencies. But it may cause too much flexibility to obtain a slight gain in accounting for the process variation. Thus in this article we assume $\Theta_1 = \cdots = \Theta_6 = \Theta$ which reduces to BSM-MA, (3.13). This equality constraint can be rephrased that the offset effect is expected all alike for the seasonal frequencies, and the roots of the MA polynomial are pitched outside the unit circle at an equal distance.

3.3.4 State Space Representation for BSM-MA

We close this section with a remark on a state space representation of BSM-MA model. For the BSM, the state vector (3.6) consists of the unobserved components and their lagged variables. As regards the transition matrix for the seasonal component (the T_2 block of matrix T in (3.7)), the first row essentially corresponds to the seasonal component model and other rows merely shift the time index. The same holds for BSM-AR, too. However, such a simple construction cannot be extended straightforward if the moving average terms are incorporated in the model.

A state space representation for (3.13) will be given as follows. Let $\tilde{\gamma}_{t+i|t-1}$ be a predictor of γ_{t+i} based on the observation up to time $t-1$, and on the innovations up to t, namely,

$$\tilde{\gamma}_{t+i|t-1} = -\sum_{j=i+1}^{s-1} \gamma_{t+i-j} - \sum_{j=i}^{s-1} \Theta^j \omega_{t+i-j}.$$

It can be readily verified that the following recursive relationship holds,

$$\gamma_t = \gamma_{t-1} + \tilde{\gamma}_{t|t-2} + \omega_t \tag{3.15}$$

$$\tilde{\gamma}_{t+i|t-1} = -\gamma_{t-1} + \tilde{\gamma}_{t+i|t-2} - \Theta^i \omega_t,\ i = 1, \ldots, s-1. \tag{3.16}$$

Let us define the state vector as

$$\alpha_t = (\gamma_t, \tilde{\gamma}_{t+1|t-1}, \ldots, \tilde{\gamma}_{t+s-1|t-1})'$$

where s denotes the number of seasons in a period. Then a set of s-equations given by (3.15) and (3.16) constitute a state space representation together with the state α_t, and the submatrices T_2 and R_2 in (3.7) should be replaced by the followings,

$$\tilde{T}_2 = \begin{bmatrix} -1 & 1 & & \\ -1 & & 1 & \\ \vdots & & & \ddots \\ -1 & & & 1 \\ 0 & & & 0 \end{bmatrix}, \quad \tilde{R}_2 = \begin{bmatrix} 1 \\ -\Theta \\ -\Theta^2 \\ \vdots \\ -\Theta^{s-1} \end{bmatrix}.$$

Note that the dimension of the state for the usual stochastic dummy seasonal model (3.4) and the seasonal summation driven by AR (3.11) is $s-1$ while it increases just by one for the model (3.13). This predictor-based state space representation will be attributed to Akaike (1974).

3.4 Real Data Analysis

In this section, we analyze 11 time series with trend and seasonality. First 6 series are Japanese economic time series; consumption in private sector (to be abbreviated to PCSMP, and its span analyzed is from 1980Q1 to 2002Q4), machinery order (MORDER, 1987:04–2002:12), money supply (M2CD, 1980:01–2002:12), new car registration (NEWCAR, 1980:01–2002:12), industrial production (IIP, 1980:01–2002:12), and Tokyo district sales of department stores (TDS, 1980:01–2002:12). The remaining 5 series are taken from the textbook of Harvey (1989); coal, gas and electricity demand of other final users (UKCOAL, UKGAS, UKELEC, 1960Q1–1986Q4), car drivers killed or seriously injured (CDKSI, 1969:01–1982:12), and international airline passengers (AIRLINE, 1949:01–1960:12) of which original source is Box and Jenkins (1976). All the series are log-transformed. In the preliminary analysis, it is verified that AIC statistic corrected by the determinant of Jacobian matrix supports the log-transformation for each series. Because of the log transformation, the scale of the original series sticks around from 5 to 15. Hence, as for the initialization of the Kalman filter, we assume the diagonal matrix for P_0 of which elements are all set to 10^4 in all the models. The first element of the initial

state mean is replaced by the sample mean which is computed using the first quarter of the time series. The rest of the initial state element are assumed to be 0.

3.4.1 Preprocessing

As regards PCSMP and TDS, the effects of the introduction of consumption tax (1989:04) and its rise (1997:04) are removed prior to the model comparison. The unusual increase in March is the consequence of spending rush ahead of the consumption tax introduction or its hike while the atypical depress in April is the counteraction to the March's rush. From the visual inspection of seasonal factors obtained from the BSM, there seems to exist two outliers for PCSMP (1997Q1, 1997Q2) and four outliers for TDS (1989:03, 1989:04, 1997:03, 1997:04). Just for confirmation, each series was analyzed by Web-Decomp with the level-2 outlier detection option. Though only 1997Q2 of PCSMP is judged to be outlier-free, we treat the all six observations as outliers.

After the locations of outlier are identified, the preadjustment will be performed in the following manner. Firstly, we create the season-wise series from the original time series. In other words, the observations for specific season (for example, only Q1, only January, etc.) are collected to form another time series. Secondly, for this season-wise series, the data judged as outliers are treated as missing values. We fit first or second order trend model to the season-wise series, and the obtained smoother mean for the missing values replace the outliers. A list of corrections for original data follows. For PCSMP, 72000.8 $\rightarrow$ 68552.8 (1997Q1), 67146.6 $\rightarrow$ 68360.0 (1997Q2). As for TDS, 287.883 $\rightarrow$ 227.255 (1989:03), 175.539 $\rightarrow$ 203.041 (1989:04), 271.883 $\rightarrow$ 210.733 (1997:03) and 166.127 $\rightarrow$ 182.729 (1997:04).

3.4.2 Overview of Results

The estimation results are summarized in Table 3.1 and Table 3.2. The BSM-MA attains the minimum AIC for all series but UKCOAL, and the improvements in AIC are sometimes substantial. In every case, the estimated innovation variance of the seasonal component, σ^2_ω, is larger than those estimated in the BSM and the BSM-MA. It can be interpreted that MA-driven seasonal summation has successfully brought more flexibility than the BSM and the BSM-AR. Moreover, the BSM-MA is also supported from a predictive point of view, i.e., in terms of minimum AIC.

The estimated AR parameters of BSM-AR, $\hat{\Phi}$'s, are generally small. At least within the worked examples in this paper, they rarely exceed 0.01. This even helps to increase the power of the seasonal component. A typical result of BSM-AR is seen in the case of PCSMP for

Table 3.1: Estimation results for Japanese macroeconomic data

PCSMP	$\hat{\sigma}_\eta^2$	$\hat{\sigma}_\omega^2$	$\hat{\sigma}_\varepsilon^2$	Φ or Θ	AIC
BSM	0.16×10^{-5}	0.86×10^{-5}	0.35×10^{-4}	—	-508.60
BSM-AR	0.16×10^{-5}	0.53×10^{-5}	0.39×10^{-4}	0.83×10^{-2}	-509.70
BSM-MA	0.15×10^{-5}	0.68×10^{-4}	0.31×10^{-6}	0.84	-523.88
MORDER	$\hat{\sigma}_\eta^2$	$\hat{\sigma}_\omega^2$	$\hat{\sigma}_\varepsilon^2$	Φ or Θ	AIC
BSM	0.31×10^{-4}	0.28×10^{-2}	0.29×10^{-5}	—	-306.73
BSM-AR	0.30×10^{-4}	0.28×10^{-2}	0.29×10^{-5}	0.63×10^{-2}	-303.16
BSM-MA	0.20×10^{-4}	0.40×10^{-2}	0.31×10^{-5}	0.89	-329.02
M2CD	$\hat{\sigma}_\eta^2$	$\hat{\sigma}_\omega^2$	$\hat{\sigma}_\varepsilon^2$	Φ or Θ	AIC
BSM	0.24×10^{-5}	0.88×10^{-6}	0.10×10^{-5}	—	-2144.16
BSM-AR	0.24×10^{-5}	0.85×10^{-6}	0.99×10^{-6}	0.18×10^{-1}	-2154.06
BSM-MA	0.14×10^{-5}	0.28×10^{-5}	0.85×10^{-6}	0.85	-2166.41
NEWCAR	$\hat{\sigma}_\eta^2$	$\hat{\sigma}_\omega^2$	$\hat{\sigma}_\varepsilon^2$	Φ or Θ	AIC
BSM	0.44×10^{-5}	0.31×10^{-3}	0.87×10^{-3}	—	-706.40
BSM-AR	0.44×10^{-5}	0.31×10^{-3}	0.87×10^{-3}	0.23×10^{-6}	-702.40
BSM-MA	0.40×10^{-5}	0.16×10^{-2}	0.26×10^{-5}	0.92	-802.69
IIP	$\hat{\sigma}_\eta^2$	$\hat{\sigma}_\omega^2$	$\hat{\sigma}_\varepsilon^2$	Φ or Θ	AIC
BSM	0.12×10^{-4}	0.85×10^{-4}	0.12×10^{-6}	—	-1324.49
BSM-AR	0.12×10^{-4}	0.86×10^{-4}	0.12×10^{-6}	0.36×10^{-2}	-1320.90
BSM-MA	0.90×10^{-5}	0.11×10^{-3}	0.11×10^{-6}	0.28	-1325.01
TDS	$\hat{\sigma}_\eta^2$	$\hat{\sigma}_\omega^2$	$\hat{\sigma}_\varepsilon^2$	Φ or Θ	AIC
BSM	0.14×10^{-5}	0.11×10^{-3}	0.14×10^{-3}	—	-1113.81
BSM-AR	0.14×10^{-5}	0.86×10^{-4}	0.15×10^{-3}	0.66×10^{-2}	-1126.18
BSM-MA	0.12×10^{-5}	0.38×10^{-3}	0.44×10^{-5}	0.89	-1168.32

Table 3.2: Estimation results for UK data in Harvey (1989)

UKCOAL	$\hat{\sigma}_\eta^2$	$\hat{\sigma}_\omega^2$	$\hat{\sigma}_\varepsilon^2$	Φ or Θ	AIC
BSM	0.72×10^{-5}	0.67×10^{-9}	0.17×10^{-1}	—	-35.81
BSM-AR	0.72×10^{-5}	0.33×10^{-7}	0.17×10^{-1}	0.57×10^{-7}	-31.81
BSM-MA	0.74×10^{-5}	0.73×10^{-4}	0.17×10^{-1}	0.98	-31.83
UKGAS	$\hat{\sigma}_\eta^2$	$\hat{\sigma}_\omega^2$	$\hat{\sigma}_\varepsilon^2$	Φ or Θ	AIC
BSM	0.84×10^{-5}	0.41×10^{-2}	0.95×10^{-3}	—	-134.07
BSM-AR	0.85×10^{-5}	0.41×10^{-2}	0.92×10^{-3}	0.17×10^{-2}	-130.11
BSM-MA	0.64×10^{-5}	0.75×10^{-2}	0.20×10^{-5}	0.64	-157.36
UKELEC	$\hat{\sigma}_\eta^2$	$\hat{\sigma}_\omega^2$	$\hat{\sigma}_\varepsilon^2$	Φ or Θ	AIC
BSM	0.88×10^{-5}	0.48×10^{-3}	0.12×10^{-2}	—	-235.75
BSM-AR	0.89×10^{-5}	0.49×10^{-3}	0.12×10^{-2}	0.43×10^{-2}	-232.21
BSM-MA	0.85×10^{-5}	0.24×10^{-2}	0.73×10^{-5}	0.78	-258.52
CDKSI	$\hat{\sigma}_\eta^2$	$\hat{\sigma}_\omega^2$	$\hat{\sigma}_\varepsilon^2$	Φ or Θ	AIC
BSM	0.22×10^{-5}	0.24×10^{-8}	0.44×10^{-2}	—	-213.44
BSM-AR	0.19×10^{-5}	0.11×10^{-3}	0.41×10^{-2}	0.21×10^{-8}	-204.67
BSM-MA	0.21×10^{-5}	0.48×10^{-2}	0.12×10^{-5}	0.99	-302.70
AIRLINE	$\hat{\sigma}_\eta^2$	$\hat{\sigma}_\omega^2$	$\hat{\sigma}_\varepsilon^2$	Φ or Θ	AIC
BSM	0.29×10^{-3}	0.28×10^{-3}	0.14×10^{-5}	—	-391.64
BSM-AR	0.20×10^{-3}	0.29×10^{-3}	0.32×10^{-3}	0.15×10^{-1}	-348.26
BSM-MA	0.88×10^{-5}	0.94×10^{-3}	0.13×10^{-5}	0.94	-445.99

example. In the right panel of Figure 3.2, the log-spectrum of BSM-AR is drawn slightly above the BSM, and the estimated seasonal innovation variance increased from 0.53×10^{-5} to 0.86×10^{-5}. As a result, the plots of their smoother mean are almost indistinguishable from each other. Ozaki (1997) also reports similar results on BSM-AR which he refers to 'dynamic BAYSEA' model.

On the other hand, the estimated MA parameter $\hat{\Theta}$ for PCSMP is 0.84, and the estimated seasonal innovation variance is 0.68×10^{-4}. Accordingly, the level of log-power spectrum in the right panel of Figure 3.2 is flattened and pushed upward in comparison with those of the BSM and the BSM-AR.

Two panels in the bottom of Figure 3.3 show the seasonal components estimated by the BSM and the BSM-MA, and the right-upper panel shows their difference. Four panels in Figure 3.4 show the annual plot of every quarter. We observe that these annual plots of BSM-MA swings around those of the BSM and BSM-AR. Figure 3.4 clearly exhibits the difference between the seasonal component of the BSM (thin solid line with symbol +) and of BSM-MA (thick solid line). The BSM and BSM-AR (thin solid line without symbol in Figure 3.4) produce quite similar results in both figures. Other successful cases exhibit the similar features, but the graphs

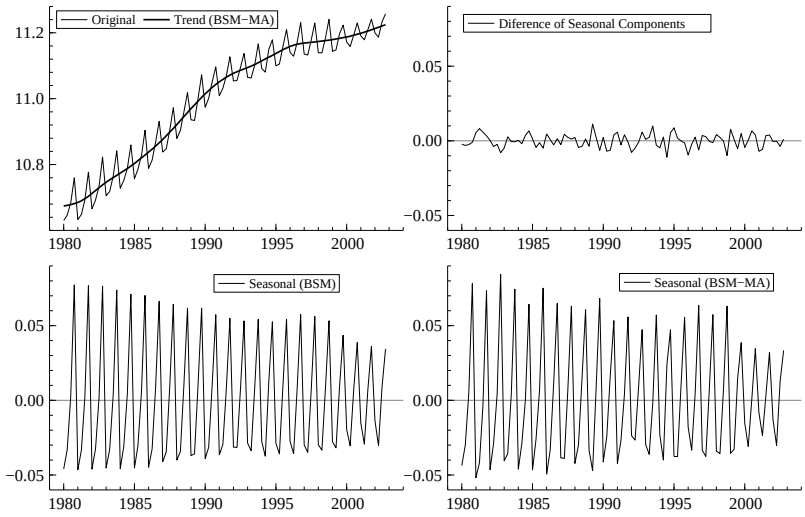


Figure 3.3: Original plus trend (upper-left), seasonal factors for CDKSI (bottom) and the difference of two seasonal factors (upper-right) for PCSMP.

are omitted for the reason of space.

3.4.3 Noteworthy Exceptions

Table 3.1 and 3.2 show that for most of time series considered here the seasonal summation driven by MA model improves the simple modeling by the BSM, except UKCOAL. What is striking is the decrease in the AIC statistic in the case of CDKSI. From Table 3.2, AIC of BSM-MA, -302.70 is much smaller than that of the BSM, -213.44. In terms of information criterion, BSM-MA is overwhelmingly superior to the BSM. Nonetheless, once we give a glance at over the right-lower panel of Figure 3.6, a doubt comes up if we should accept the difference of AIC at its face value. To put it plainly, the seasonal component of BSM-MA appears to be just the subtraction of trend component from the original series. On the other hand, considering the unstable seasonality in the original time series, the seasonal pattern estimated by the BSM looks too regular to be plausible. When it comes to seasonal adjustment, neither the BSM nor BSM-MA gives a satisfactory solution.

Let s_t and $\bar{s}_t$ be the seasonal component derived from the BSM and the BSM-MA respectively. Then $\sigma^* = \sqrt{\mathrm{Var}[(s_t - \bar{s}_t)^2]}$ is the standard deviation of the perturbation introduced by MA term. The key feature of the CDKSI case is that σ^* is very large in comparison with the maximum amplitude of the seasonal pattern of the BSM, namely $\max s_t - \min s_t$.

Turning to the case of UKCOAL, we know from Table 3.2 that the BSM is better than the BSM-MA by the minimum AIC criterion. Figure 3.5 show the results of the UKCOAL

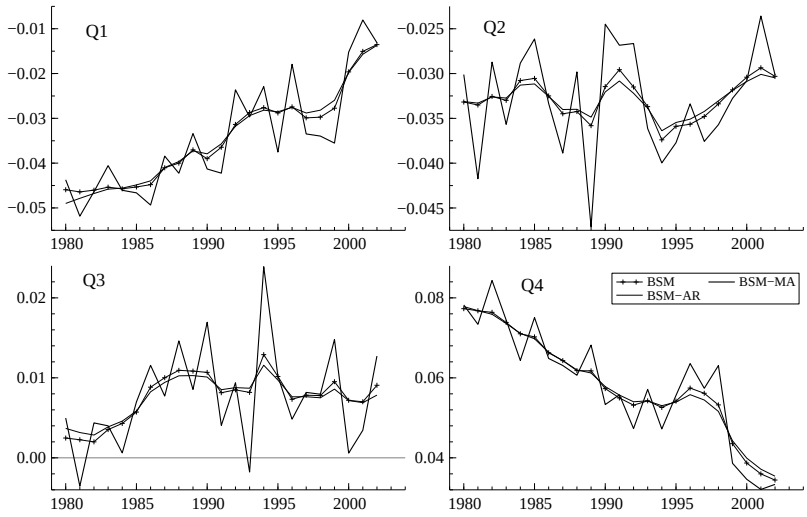


Figure 3.4: Annual plot of every quarters of the estimated seasonal components in the PCSMP case.

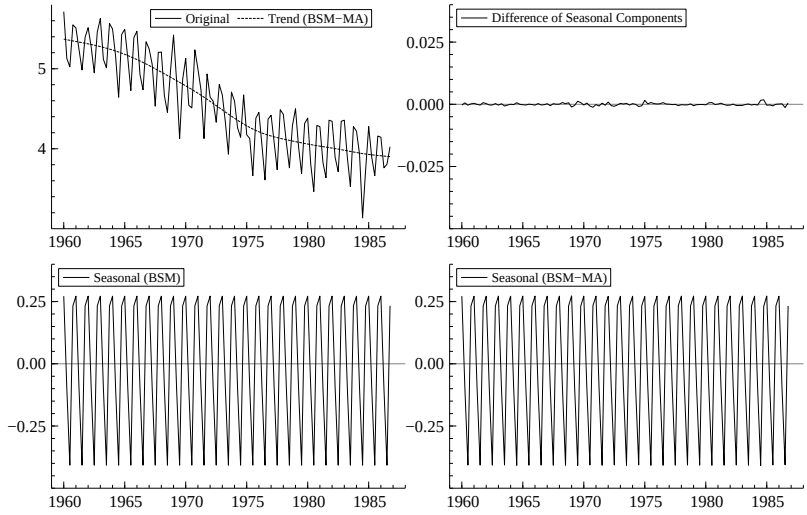


Figure 3.5: Original plus trend (upper-left), seasonal factors for CDKSI (bottom) and the difference of two seasonal factors (upper-right) for UKCOAL.

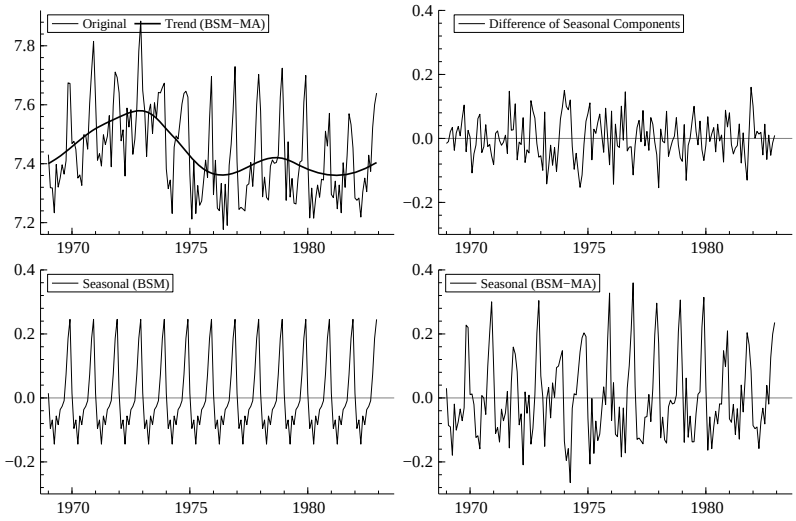


Figure 3.6: Original plus trend (upper-left), seasonal factors for CDKSI (bottom) and the difference of two seasonal factors (upper-right) for CDKSI.

case. The lower two panels exhibit that the seasonal components are indistinguishable from one another , and almost deterministic whichever model to be employed as the seasonal component. The difference of seasonal factors of the BSM and the BSM-MA (the right-upper panel of Figure 3.5) manifests that the BSM-MA could not introduce any additional variability into the seasonal component. To conclude, the key feature of the UKCOAL case is, σ^* is too small relative to $\max s_t - \min s_t$.

Another significant feature shared by both CDKSI and UKCOAL cases is that the estimated seasonal MA parameter, $\hat{\Theta}$, is extremely close to 1. But from the right-lower panel of Figure 3.6, we cannot say this is the case of polynomial cancellation because the estimated seasonal pattern is far from deterministic. Looking at Figure 3.6, an idea easily comes up with us that the wild seasonal component may be regarded as the nearly periodic seasonal pattern laid over the zero mean stationary process. As for UKCOAL, the estimated seasonal patterns are nearly deterministic nevertheless the original series does not exhibit such a steady seasonality. Hence in any case, it is suspected that some important component may be lacking in the model specification.

3.4.4 Including a Cyclical Component

Upon the observations in the previous subsection, we add a cyclical component to the BSM-MA model and see its impact on the AIC values, on the seasonal moving average parameter (Θ) and on the seasonal pattern. We assume that the cyclical component can be expressed by a finite

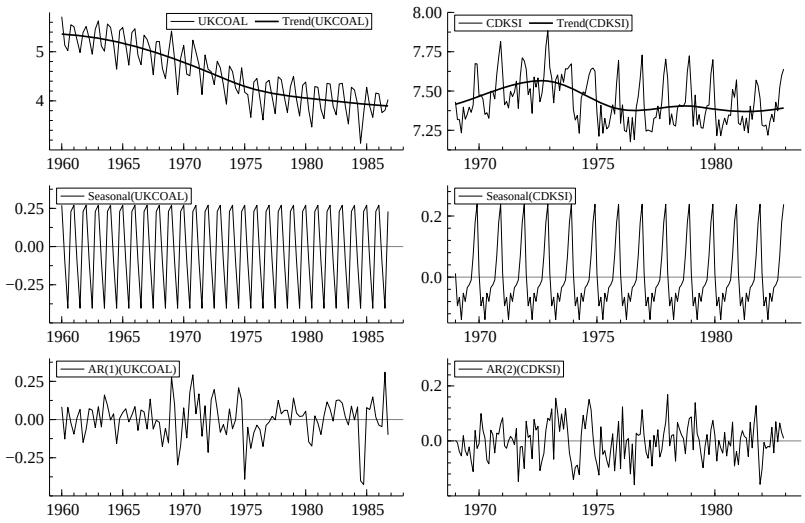


Figure 3.7: Original plus trend (upper), seasonal (middle) and AR component (bottom) for UKCOAL (left column) and CDKSI (right column).

order (up to 4-th order here) stationary autoregressive process,

$$\psi_t = \sum_{i=j}^{m} \rho_j \psi_{t-j} + \kappa_t,$$

and that the time series can be decomposed as

$$y_t = \mu_t + \gamma_t + \psi_t + \varepsilon_t.$$

As the state space representation for this decomposition is obvious, the reader is invited to visit the textbook like Harvey (1989), Kitagawa and Gersch (1996).

For UKCOAL, it turned out that the BSM-MA with the first order stationary AR component attains the minimum AIC, -62.68. This is much smaller than the AIC value for the BSM-MA, -31.83. The estimated trend, seasonal and cyclical component are plotted in the left column of Figure 3.7. The seasonal MA parameter $\hat{\Theta}$ is 0.99, which is extremely close to unity. As we recall, however, the BSM attained the minimum AIC at least in the analysis without a cyclical component, so we had better compare the BSM and the BSM with a cyclical component model. The AIC of the latter is -66.68 which is also much smaller than the AIC of the BSM, -35.81. To conclude, the best model for UKCOAL is the BSM with a cyclical component expressed by AR(1). If we fit the MA driven seasonal model (3.13), the estimated $\hat{\Theta}$ is close to 1. This suggests the seasonality in UKCOAL is almost deterministic, and the simple seasonal summation (3.4) with very small σ_ω^2 suffices.

Now we turn to the CDKSI case. After the model estimation and selection, we find including a second order stationary AR to the BSM-MA improves the AIC value, $-302.70 \rightarrow -304.46$.

The seasonal MA parameter for CDKSI is substantially diminished from 0.99 to 0.28. We doubt if the MA parameter is really needed, hence we fit the BSM including a AR(2) component. Contrary to our expectation, the AIC statistic of the model is -269.39, which is inferior to the BSM-MA with a cyclical component. Thus, the best model among the models we tried here is the BSM-MA with a cyclical component model expressed by a stationary AR of order 2. Three panels in the right column of Figure 3.7 show the estimated components for CDKSI. Paralleling the AR component with the difference of seasonal components in Figure 3.6, the fluctuations brought by the MA process are almost captured by the cyclical component.

3.4.5 A Graphical Representation

In the subsection 3.4.3, it is remarked that the appropriateness of the flexibility brought by MA term should be determined in connection with the range of seasonal pattern of the time series. Thus we introduce a simple measure on the pertinence of the seasonal component model, and propose a graphical representation. Let $R = \max s_t - \min s_t$. What appears to be essential is the ratio, R/σ^*. Considering that R is a sort of range and σ^* is the standard deviation. It seems more natural to consider the length of interval, such as $\pm 2\sigma^*$ or $\pm 3\sigma^*$ for example. Here we adopt $\pm 3\sigma^*$ interval, and define the following quantity to measure the impact of the fluctuation brought by MA term on the seasonal pattern,

$$M = \log_{10}\{R/6\sigma^*\}.$$

If $\bar{s}_t$ is very wild, then the argument of the log function will be close to unity, so M is close to 0. If s_t and $\bar{s}_t$ are alike, then small value of σ^* will lead to large M. (If σ^* happens to be zero, then we discard the measure M. Such a case does not interest us at all because the MA term is not effectively working and the BSM is obviously better.) As is already pointed out in section 3.4.3, another key quantity is $\hat{\Theta}$, the estimated MA parameter. Therefore, the 2-dimensional plot of (M, $\hat{\Theta}$) is expected to give some information on the modeling of the time series of interest.

Figure 3.8 shows the graphical layout of the BSM-MA models applied to the 11 time series in this paper. The horizontal axis denotes the measure M defined above, and the vertical axis indicates the seasonal MA parameter, Θ. Two points connected with an arrow mean that the AIC statistic is improved by including a cyclical component, and subsequently the graphical layout of (M,Θ) is changed. It is striking that $\hat{\Theta}$'s are diminished and M's are centered around 1.00 ± 0.25 after the cyclical component is added to the model.

We observe that the two exceptional cases (CDKSI and UKCOAL) are located at the upper-left and the upper-right in this model map. $M \approx 0$ means the seasonal variability increased

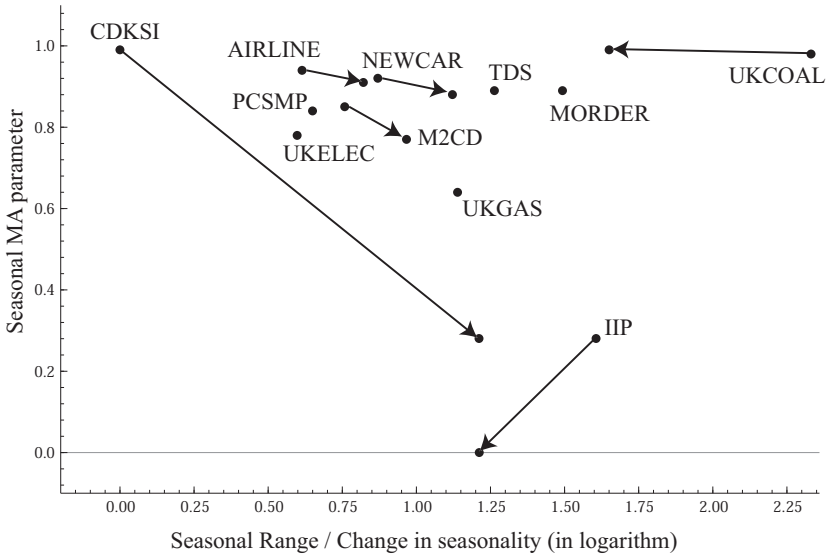


Figure 3.8: Graphical layout of the estimated models.

by the BSM-MA model is almost comparable with the seasonal range of original time series. On the other hand, large M indicates that the BSM-MA cannot introduce any flexible seasonal pattern compared to the BSM. In some cases this suggests that the seasonality for the time series of interest is almost deterministic, which sounds reasonable for the UKCOAL case. From our empirical analysis, it is inferred that $\hat{\Theta} \approx 1$ suggests the 'cancellation' is occurring on the seasonal component model (e.g. UKCOAL), or the possible misspecification as in the CDKSI case.

3.5 Conclusion

This article proposed a parsimonious modeling of flexible seasonality within a framework of structural time series models. The basic idea is to drive the seasonal summation by a moving average process with just one parameter, which has been referred to the BSM-MA throughout this article. A state space representation for the model is also given. Compared to the simple seasonal summation (BSM) and the AR-driven seasonal summation (BSM-AR), the BSM-MA attains the minimum AIC in 10 out of 11 cases. In all successful cases, the estimated seasonal innovation variance is larger than those estimated by the BSM and BSM-AR, which leads to the increase of the power spectrum of the seasonal component. Annual plot of every quarters or months reveals the wiggly movement introduced by moving average terms. A close examination of the UKCOAL and CDKSI cases provides us some information. Adequacy of the additional seasonal perturbation brought by the BSM-MA depends on how big it is relative to the range of

seasonal pattern. Hence the log of the ratio of the maximum amplitude of seasonal pattern to the interval length of the ± 3-standard deviations obtained by the seasonal difference between the BSM and the BSM-MA is introduced as a measure, M. Both too small and too large M suggests the possible misspecification. Using M and Θ, a graphical representation for the estimated models is also proposed, which serves to mark out the seemingly unsuccessful cases. Even for such cases, the decomposition including a cyclical component is proved to amend the existing models from as is shown in our empirical analysis.

Chapter 4

Detecting Seasonal Unit Roots in A Structural Time Series Model

4.1 Introduction and motivation

The structural time series model is often found to be a useful tool for describing and forecasting economic time series, see Kitagawa and Gersch (1984), Kitagawa (1981), Harvey (1985) and Harvey (1989) for a few of the earlier references. Its usefulness seems to be based mainly on the fact that the model does not require many parameters, while it can adapt to the observed data with great flexibility. A structural time series model can be written in ARMA format, and then it can be seen that it often assumes several unit roots in the AR polynomial and some near unit roots in the MA polynomial. If a structural time series model is considered for quarterly time series, it can be written as a model which is close to the, what is called, airline model, see Harvey (1984) and Maravall (1985). This airline model, which has been introduced in Box and Jenkins (1970), assumes a first order and an annual differencing filter, which amounts to two nonseasonal unit roots and three seasonal unit roots, see Hylleberg et al. (1990) and Franses (1996) for details on the terminology.

A practically relevant question for modeling and forecasting concerns the issue of the number of unit roots in a given time series. Based on purely autoregressive models, Hylleberg et al. (1990) have been the first to propose tests for seasonal unit roots in economic data. Since then, many new methods and alternative versions of the original procedure have been proposed. A drawback of most of these methods is, as the simulation results in Ghysels et al. (1994) indicate, that they have serious size distortions in case the data are generated by MA models with near unit roots. In other words, when the data would have been generated by an airline model, and hence by a model like a structural time series model, these tests may be of little use. In sum, the best strategy seems to be to start with the structural time series model itself, and to consider

the presence of seasonal unit roots within that framework. It is the aim of the present paper to put forward such a method, where we use the familiar model selection criteria, to evaluate its empirical performance, and to apply it to a wide range of macroeconomic time series.

The outline of this paper is as follows. In section 4.2, we highlight some features of the structural time series model and the airline model. In section 4.3, we present an outline of our model selection approach. After evaluating the performance of AIC and BIC using Monte Carlo simulations, we formulate a practical decision rule based on both criteria. In section 4.4, we apply this rule to a set of 22 macroeconomic time series variables. In section 4.5, we conclude our paper with some remarks.

4.2 Models

A version of a structural time series model for y_t, $T = 1,2,\ldots,T$ is given by a set of three equations, that is,

$$y_t = \mu_t + s_t + w_t, \quad w_t \sim \text{NID}(0,\sigma_w^2) \tag{4.1}$$

$$(1-L)^2\mu_t = u_t, \quad u_t \sim \text{NID}(0,\sigma_u^2) \tag{4.2}$$

$$(1+L+L^2+L^3)s_t = v_t, \quad v_t \sim \text{NID}(0,\sigma_v^2) \tag{4.3}$$

where the error processes w_t, u_t and v_t are also mutually independent. This approach to modelling nonstationary trending and seasonal time series became popular through the work of Kitagawa (1981), Kitagawa and Gersch (1984) and Harvey (1985), inter alia. If we adopt a first order random walk plus variable drift for the trend model instead of (4.2), it is appropriate to refer to the set of equations as the basic structural model (BSM) after Harvey (1989, p.172). Throughout this paper, however, we assume a second order random walk for the trend model. A second order random walk model seems consistent with the decomposition based on the airline model because the real positive unit root of the seasonal difference operator is usually absorbed into the lag polynomial of trend model to ensure the uniqueness of the decomposition. We refer to (4.1) – (4.3) as the Kitagawa-Gersch type structural model (KG-SM) after Kitagawa and Gersch (1984).

Substituting (4.2) and (4.3) in (4.1) yields that y_t can be described by

$$(1-L)(1-L^4)y_t = \zeta_t \tag{4.4}$$

where ζ_t is a moving average process of order 5 [MA(5)]. If this MA(5) process can be decomposed as $(1-\theta L)(1-\Theta L^4)\varepsilon_t$, where $\{\varepsilon_t\}$ is standard white noise process with variance σ_ε^2, the

Model			ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
Airline Model							
$\theta = 0.8$	$\Theta = 0.8$		-0.488	—	0.238	-0.488	0.238
$\theta = 0.8$	$\Theta = 0.4$		-0.488	—	0.168	-0.345	0.168
$\theta = 0.4$	$\Theta = 0.8$		-0.345	—	0.168	-0.488	0.168
$\theta = 0.4$	$\Theta = 0.4$		-0.345	—	0.119	-0.345	0.119
Structural Model							
$\sigma_u^2 = 1.0$	$\sigma_v^2 = 0.25$	$\sigma_w^2 = 20.0$	-0.444	0.026	0.246	-0.468	0.234
$\sigma_u^2 = 0.1$	$\sigma_v^2 = 0.25$	$\sigma_w^2 = 1.5$	-0.468	0.057	0.203	-0.380	0.190
$\sigma_u^2 = 0.4$	$\sigma_v^2 = 0.25$	$\sigma_w^2 = 2.0$	-0.342	0.095	0.216	-0.360	0.180
$\sigma_u^2 = 0.35$	$\sigma_v^2 = 0.25$	$\sigma_w^2 = 1.5$	-0.331	0.107	0.208	-0.337	0.169

Table 4.1: Autocorrelation functions of airline and structural models

resultant model is called the airline model, see Box and Jenkins (1970). The autocovariances γ_k, $k = 0, 1, 2, \ldots$, for ζ_t are

$$\gamma_0 = 4\sigma_u^2 + 6\sigma_v^2 + 4\sigma_w^2 \tag{4.5}$$

$$\gamma_1 = 3\sigma_u^2 - 4\sigma_v^2 - 2\sigma_w^2 \tag{4.6}$$

$$\gamma_2 = 2\sigma_u^2 + \sigma_v^2 \tag{4.7}$$

$$\gamma_3 = \sigma_u^2 + \sigma_w^2 \tag{4.8}$$

$$\gamma_4 = -2\sigma_w^2 \tag{4.9}$$

$$\gamma_5 = \sigma_w^2 \tag{4.10}$$

$$\gamma_j = 0 \quad \text{for } j = 6, 7, \ldots. \tag{4.11}$$

One can see that only γ_0 is slightly different from that of BSM because we have one parameter less (the dispersion of the time varying drift) and we have a second order trend model here, but the rest of the autocorrelation function is the same as that of BSM, see Harvey (1989, p.56). In the same fashion as in Maravall (1985), TABLE 4.1 compares the autocorrelation functions of the two models for some typical values of the parameters. The only theoretical difference is that ρ_2 is positive in the structural model while in the airline model it is exactly zero. In practice, ρ_2 can be very small.

When $\sigma_u^2 = 0$ and $\sigma_v^2 = 0$, it is easy to see that (4.4) reduces to the airline model with $\theta = 1$ and $\Theta = 1$. Strictly speaking, when $\sigma_u^2 \neq 0$, the ζ_t process in (4.4) is invertible. This can be easily understood from the fact that for ζ_t in (4.4) holds that

$$\sum_{k=1}^{\infty} \rho_k = -\frac{1}{2} + \frac{8\sigma_u^2}{\gamma_0} \tag{4.12}$$

that is, the theoretical autocorrelations only sum to $-\frac{1}{2}$ when $\sigma_u^2 = 0$. One can observe from (4.5) – (4.11) and from TABLE 4.1 that the autocorrelation function of (4.4) can come close to that of the airline model with θ and Θ close to unity. Hence, the KG-SM is flexible enough to generate a wide range of time series data, amongst which are those that one may want to describe by an airline model.

In this paper we investigate the presence of seasonal unit roots in the Kitagawa-Gersch type model. The model of our interest is (4.1) – (4.3) and our alternative model is (4.1) and (4.2) with

$$(1+aL)(1+bL^2)s_t = v_t \tag{4.13}$$

where $0 < a \leq 1$ and $0 < b \leq 1$. If $0 < a < 1$, there is no seasonal unit root -1, and if $0 < b < 1$, there are no seasonal unit roots i and $-i$.

4.3 Detecting Seasonal Unit Roots

In this section we put forward a method to examine if the structural time series model in (1) - (3) assumes too many seasonal unit roots.

4.3.1 Model Selection Approach

Throughout this section there are four models to be compared. They share equations (4.1) and (4.2), that is, we assume that all models consist of three components and that the trend component μ_t and the observational noise component w_t are common. Only the seasonal components differ according to the number of assumed seasonal unit roots. In sum, we consider

$$\text{Model 0} \quad : \quad (1+aL)(1+bL^2)s_t = v_t \tag{4.14}$$

$$\text{Model 1} \quad : \quad (1+L)(1+bL^2)s_t = v_t \tag{4.15}$$

$$\text{Model 2} \quad : \quad (1+aL)(1+L^2)s_t = v_t \tag{4.16}$$

$$\text{Model 3} \quad : \quad (1+L)(1+L^2)s_t = v_t \quad . \tag{4.17}$$

In models 0,1 and 2, a and b are unknown hyperparameters, so these and σ_v^2 have to be estimated. Strictly speaking, model 0, for example, consists of (4.1), (4.2) and (4.14), but there will be no confusion even if we simply refer to equation (4.14) as model 0 instead of mentioning the whole set of equations. Model 3 is the KG-SM itself as discussed in the previous section. Therefore we will refer to Model 0 through 2 as the extended Kitagawa-Gersch type Structural Models (EKG-SM), which seems to make sense because Model 3 is a restricted version of the other models.

The state space form immediately follows if we define the state vector α_t as $\alpha_t = (\mu_t, \mu_{t-1}, s_t, s_{t-1}, s_{t-2})$. Then, it is easy to see that the transition equation ((4.18) below) and the measurement equation ((4.19) below) are explicitly defined as follows.

$$\alpha_t = T\alpha_{t-1} + R\eta_t \tag{4.18}$$

$$y_t = Z\alpha_t + \varepsilon_t \tag{4.19}$$

where T, R and Z are defined as

$$T = \left(\begin{array}{cc|ccc} 2 & -1 & & O & \\ 1 & 0 & & & \\ \hline & & -a & -b & -ab \\ & O & 1 & 0 & 0 \\ & & 0 & 1 & 0 \end{array}\right), \qquad R = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

$$Z = (1,0,1,0,0).$$

The mean of η_t is assumed to be zero, and the covariance matrix of η_t, denoted by Q, is the diagonal matrix with elements σ_u^2 and σ_v^2 by assumption, see (4.2) and (4.3). The system matrix T defined above is for Model 0. By restricting a and/or b to 1, we have state space forms for Model 1, Model 2 and Model 3.

When we assume the Gaussian distribution for every noise process (see (4.1)–(4.3)), the above modelling approach invariably requires the Kalman filter, see Anderson and Moore (1979) and Harvey (1989), inter alia. In short, one should repeat the *prediction* step

$$\alpha_{t|t-1} = T\alpha_{t-1|t-1} \tag{4.20}$$

$$P_{t|t-1} = TP_{t-1|t-1}T' + RQR' \tag{4.21}$$

where $P_{t-1|t-1}$ denotes the covariance matrix of the state estimation error, and *update* recursively;

$$\alpha_{t|t} = \alpha_{t|t-1} + P_{t|t-1}Z'F_{t|t}^{-1}(y_t - Z\alpha_{t|t-1}) \tag{4.22}$$

$$P_{t|t} = P_{t|t-1} - P_{t|t-1}Z'F_{t|t}^{-1}ZP_{t|t-1}, \tag{4.23}$$

where $F_{t|t} = ZP_{t|t-1}Z' + \sigma_w^2$. Given the initial conditions $\alpha_{1|0}$ and $P_{1|0}$, the log-likelihood function can be expressed in the prediction error decomposition form,

$$\log L = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log F_{t|t} - \frac{1}{2}\sum_{t=1}^{T} v_t F_{t|t}^{-1} v_t \tag{4.24}$$

where $v_t = y_t - Z\alpha_{t|t-1}$. The unknown hyperparameters of the model are usually estimated by the method of maximum likelihood using numerical optimization techniques like BFGS formula.

As for the initialization of the Kalman filter, we employ the 'large κ' approximation for the diffuse prior, that is, we assume the diagonal form of covariance matrix $P_{1|0} = \kappa I$ with large κ, see Harvey (1989, p.121). Throughout this paper, κ is set to 10^4, which seems quite a reasonable choice for the sake of numerical stability. Strictly speaking, we could perform different initializations for the four seasonal models under consideration. We concern, however, only the non-stationary and near non-stationary cases, which gives us an impression that the same 'large κ' approximation does not bring enormous bias in the simulations and empirical analysis in this paper. For the initial state vector $\alpha_{1|0}$, the elements related to the seasonal component are set to zero. On the other hand, we estimate the trend initial condition using the first quarter of the whole sample period so that the initial trend value must not be too far away from the actual observation. Because of this treatment, we regard $\alpha_{1|0}$ as a part of the hyperparameters of the model though the elements of $\alpha_{1|0}$ are not the objects of numerical optimization. It should be noted, however, that this does not affect the model selection procedure stated below as long as we employ the same specification for $\alpha_{1|0}$ and $P_{1|0}$.

Our basic strategy to determine the number of seasonal unit roots in y_t is as follows. For given data, we estimate all the four models (4.14) through (4.17). If we introduce extra parameters which are unnecessary (a and/or b), this will be reflected in the values of information criteria. On the other hand, if seasonal unit roots do not exist, the information criterion statistics of incorrectly restricted models (Model 3 or even Model 1 and Model 2) should be inferior. Although many variants of information criteria are proposed since Akaike (1973), the most popular criteria in practice seem to be AIC and BIC,

$$\begin{aligned} \text{AIC} &= -2\hat{\ell} + 2k \\ \text{BIC} &= -2\hat{\ell} + k \log T \end{aligned}$$

where $\hat{\ell}$ denotes the maximized log-likelihood, and k is the number of estimated unknown parameters, see Akaike (1973) and Schwarz (1978), respectively. It is not clear a priori if we should rely on a single information criterion statistic in our decision. Hence it is worthwhile to investigate how AIC and BIC behave in relevant situations. The corresponding simulation results are reported in the next subsection.

4.3.2 Monte Carlo Design

In this subsection, we describe the Monte Carlo design. When both AIC and BIC select the same model by minimum AIC and BIC respectively, we select that model. If two criteria give a split decision, essentially the case could not be conclusive, but still we can get some information about the nature of the time series under investigation, as we will demonstrate below.

First, we give our Monte Carlo design in detail. All the programs used in this paper are coded in FORTRAN and are run on various computers. All data are generated by the KG model or the EKG models based on their state space representations. The parameter values used in the simulations are as follows. For the trend component model, we use the second order random walk model throughout this paper, with initial values (11.56628, 11.57554). The initial state mean vector for quarterly seasonal component is given by $(-0.00168,\ 0.00289,\ -0.09478)$. The trend dispersion parameter is set to 0.234×10^{-7}, seasonal dispersion to 0.110×10^{-7} and the observational noise variance equals 0.195×10^{-3}. This implies that the S/N ratio defined by σ_v^2/σ_w^2 is 0.562×10^{-4}, so this is the case of excess observational noise relative to signals (or seasonal component).

The simulation results reported in this section concern two issues. First, we change the sample size of the simulated series while keeping the hyperparameter values at the values just stated above. Secondly, upon fixing the sample size to 100, we only alter the dispersion parameter of the seasonal component. Hence we try different levels of the signal-to-noise ratio by changing σ_w^2.

4.3.3 Sample Size and S/N Ratio

In the first set of simulations, we change the sample size from 50 to 160, whereas the two seasonal coefficients (a and b) vary from 1.000 to 9.700 by 0.001. In unreported preliminary research, we examined lower values of a and b like 0.8 or 0.7. However it appears that only the cases close to unit roots and the unit roots case itself are worth reporting here. In fact, the further away from 1 the stationary roots of the seasonal polynomial are, the AIC and BIC invariably detect the true data generating process.

For each set of (a, b), we have 100 replications. Figure 4.1 and Figure 4.2 show the surface plot of the selection probability when we use AIC and BIC respectively. In Figure 4.1 and Figure 4.2, the upper panel shows the results for $T = 50$ to 80, and the last one shows for $T = 130$ to $T = 160$. The left foreground corner of the plot box corresponds to the parameter grid $(a,b) = (0.970, 0.970)$, and the right background corner to the grid $(a,b) = (1.000, 1.000)$. There are deep troughs around the 'near non-stationary' grids. The selection probability improves as the

Table 4.2: Sample Sizes and Correct Detection Frequencies

(A) Detection Frequencies by AIC for Selected Grids

	$T=60$				$T=80$			
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.70	0.70	0.72	0.66	0.79	0.82	0.64	0.68
0.99	0.43	0.58	0.54	0.56	0.72	0.81	0.76	0.70
0.98	0.61	0.72	0.69	0.56	0.82	0.94	0.88	0.78
0.97	0.64	0.74	0.73	0.64	0.78	0.91	0.85	0.86
	$T=100$				$T=120$			
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.74	0.82	0.78	0.73	0.70	0.78	0.76	0.71
0.99	0.79	0.94	0.91	0.89	0.84	0.96	0.97	0.88
0.98	0.80	0.98	0.92	0.92	0.82	0.98	0.96	0.94
0.97	0.83	0.99	0.96	0.92	0.86	1.00	0.98	0.91
	$T=140$				$T=160$			
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.66	0.77	0.83	0.74	0.67	0.85	0.89	0.86
0.99	0.83	0.99	0.98	0.97	0.75	1.00	1.00	0.98
0.98	0.81	0.99	0.95	0.92	0.85	1.00	0.96	0.93
0.97	0.83	0.99	0.99	0.91	0.82	1.00	0.98	0.93

(B) Detection Frequencies by BIC for Selected Grids

	$T=60$				$T=80$			
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.89	0.58	0.60	0.65	0.95	0.86	0.60	0.65
0.99	0.28	0.29	0.19	0.30	0.63	0.50	0.43	0.41
0.98	0.51	0.39	0.43	0.35	0.79	0.73	0.68	0.61
0.97	0.55	0.42	0.40	0.40	0.80	0.70	0.61	0.62
	$T=100$				$T=120$			
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.94	0.93	0.84	0.79	0.92	0.95	0.86	0.76
0.99	0.79	0.71	0.68	0.68	0.85	0.91	0.86	0.76
0.98	0.87	0.89	0.77	0.85	0.94	0.97	0.88	0.86
0.97	0.91	0.93	0.74	0.75	0.94	1.00	0.94	0.80
	$T=140$				$T=160$			
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.95	0.92	0.87	0.77	0.94	0.98	0.99	0.96
0.99	0.98	0.97	0.97	0.83	0.96	1.00	0.99	0.91
0.98	0.98	0.99	0.92	0.91	0.98	0.99	0.96	0.86
0.97	0.94	0.98	0.97	0.88	0.91	0.99	0.94	0.86

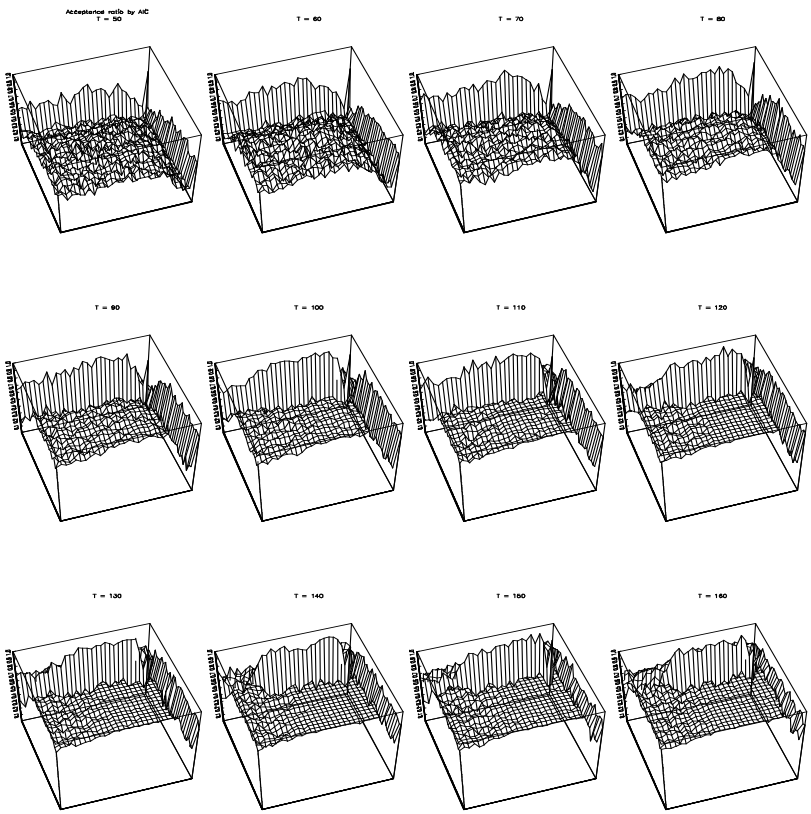


Figure 4.1: Detection frequency surface for AIC

sample size increases. The surface is gradually moving upward.

TABLE 4.2 corresponds to a selected set of simulations. In this table, only results for 16 out of 961 parameter grids are reported for 1.000, 0.990, 0.980 and 0.970. When the sample size is very small like 60, the detection probability of BIC is lower than that of AIC. On the other hand, if there is at least one seasonal unit root, BIC performs better provided that we have enough observations. Having said so, if the true model (with one or two seasonal unit roots) is very close to the three unit roots model, it is very difficult for BIC to detect the true model even in large samples. When the sample is large enough, the surface shape around the grid (1.000, 1.000) is somewhat smoother for AIC.

Now we turn to the effects of the signal-to-noise ratio on the detection frequencies by AIC and BIC. Fixing the sample size to 80, we change the variance of the observational noise σ_w^2

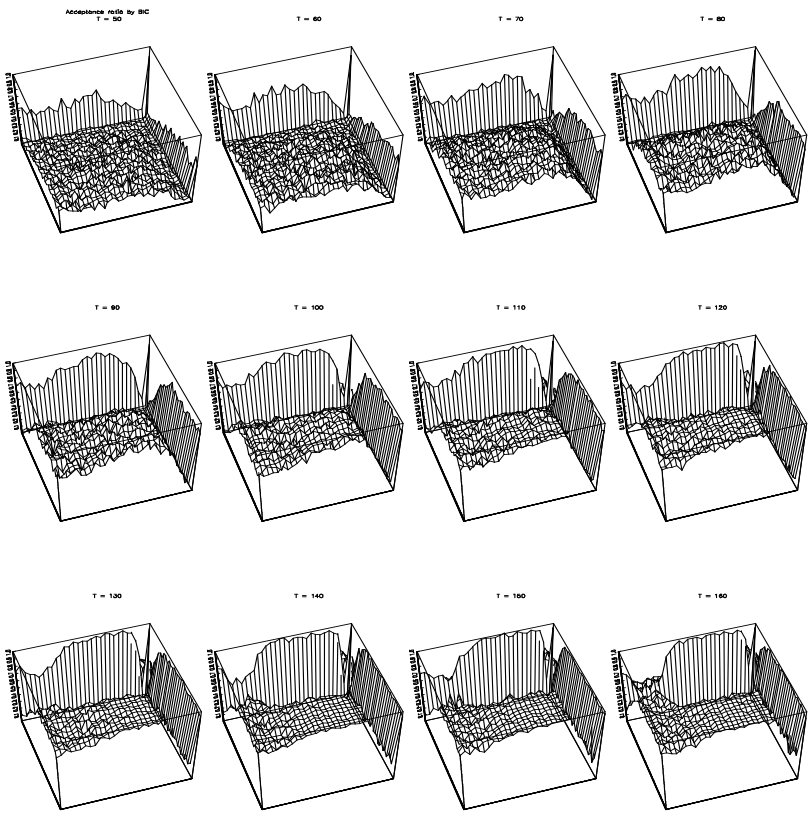


Figure 4.2: Detection frequency surface for BIC

from 10^{-3} to 10^{-7}. Let q be the S/N ratio defined by $q = \sigma_v^2/\sigma_w^2$. Keeping the dispersion parameter of the seasonal component the same (10^{-5}), we obtain five different S/N ratios, namely $q = 0.1 \times 10^{-2}$, 0.1×10^{-1}, $\ldots$, 0.1×10^{2}. The dispersion parameter of the trend component σ_u^2 is also fixed to 10^{-5}, so this is the case of almost deterministic trend plus deterministic seasonality. The same initial state as in TABLE 4.2 is used in this experiment.

Again, the two seasonal coefficients (a and b) vary from 1.000 to 9.700 by 0.001. For each set of (a, b), we have 100 replications. To save space, only 4 cases except for $q = 1$ are reported. Similar to TABLE 4.2, only selected parameter grids are tabulated. From TABLE 4.3, it can be easily perceived that the detection frequencies both by AIC and BIC are getting better as the S/N ratio becomes large. Because the large S/N ratio implies that the signal to be detected is clear relative to the observational noise, it is intuitively easy to understand these simulation

Table 4.3: S/N Ratios and Correct Detection Frequencies

	AIC				BIC			
				$q = 0.1 \times 10^{-2}$				
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.86	0.88	0.89	0.88	0.98	0.96	0.99	0.97
0.99	0.49	0.53	0.52	0.56	0.20	0.20	0.19	0.18
0.98	0.67	0.68	0.66	0.64	0.49	0.43	0.40	0.47
0.97	0.73	0.81	0.84	0.79	0.73	0.65	0.71	0.67
				$q = 0.1 \times 10^{-1}$				
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.86	0.85	0.87	0.88	0.97	0.97	0.94	0.98
0.99	0.86	0.93	0.97	0.94	0.86	0.88	0.91	0.85
0.98	0.84	0.94	0.88	0.88	0.76	0.85	0.78	0.76
0.97	0.76	0.84	0.85	0.86	0.59	0.57	0.63	0.65
				$q = 0.1 \times 10$				
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.95	0.94	0.89	0.93	0.99	0.95	0.90	0.97
0.99	0.98	0.97	1.00	0.97	1.00	0.97	1.00	0.97
0.98	0.87	0.95	0.89	0.90	0.95	0.95	0.89	0.90
0.97	0.94	0.90	0.89	0.92	0.96	0.90	0.89	0.92
				$q = 0.1 \times 10^{2}$				
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.96	0.92	0.87	0.97	1.00	0.93	0.87	0.97
0.99	0.97	0.96	1.00	0.98	1.00	0.96	1.00	0.98
0.98	0.90	0.95	0.88	0.91	0.97	0.95	0.88	0.91
0.97	0.94	0.92	0.88	0.92	0.97	0.92	0.88	0.92

results.

To summarize this subsection, the larger the sample size, the easier it is to detect the true seasonal unit root structure both by AIC and BIC. Especially when we have enough observations such as 120 or more, we can rely on BIC. When T is large, the 'power' of the model selection procedure seems satisfactory. In finite small sample cases, such as $T = 60$ or 80, BIC performs poorly. BIC has the tendency to put too much penalty on the models with stationary seasonal roots even if they correspond with the true DGP. On the other hand, the performance of AIC is generally satisfactory, although in large samples AIC prefers, as is widely recognized, overparameterized models when there exist non-stationary roots. The large the signal-to-noise ratio parameter seems to help both AIC and BIC.

Table 4.4: Detection Frequencies of Combined Decision Rule, Compared to Those of AIC and BIC ($T = 60, 80, 100, 120$)

$T = 60$					$T = 80$			
Combined Decision Rule								
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.77	0.43	0.66	0.72	0.85	0.69	0.86	0.88
0.99	0.53	0.30	0.43	0.46	0.77	0.54	0.78	0.73
0.98	0.52	0.33	0.35	0.48	0.67	0.52	0.66	0.70
0.97	0.54	0.33	0.44	0.58	0.67	0.50	0.70	0.68
AIC								
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.68	0.56	0.69	0.75	0.72	0.70	0.83	0.84
0.99	0.61	0.46	0.62	0.57	0.81	0.73	0.88	0.90
0.98	0.62	0.43	0.53	0.62	0.71	0.71	0.80	0.81
0.97	0.61	0.51	0.65	0.67	0.64	0.67	0.83	0.81
BIC								
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.86	0.30	0.56	0.65	0.92	0.63	0.81	0.80
0.99	0.40	0.17	0.26	0.29	0.70	0.33	0.55	0.49
0.98	0.45	0.13	0.22	0.29	0.59	0.31	0.38	0.43
0.97	0.51	0.18	0.32	0.46	0.61	0.33	0.52	0.53
$T = 100$					$T = 120$			
Combined Decision Rule								
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.87	0.77	0.90	0.95	0.82	0.85	0.99	0.94
0.99	0.82	0.78	0.90	0.85	0.88	0.90	0.98	0.95
0.98	0.79	0.84	0.84	0.82	0.83	0.88	0.89	0.95
0.97	0.72	0.63	0.68	0.67	0.72	0.85	0.83	0.79
AIC								
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.72	0.77	0.80	0.87	0.65	0.84	0.92	0.86
0.99	0.84	0.89	0.97	0.94	0.77	0.99	1.00	0.98
0.98	0.77	0.95	0.94	0.91	0.80	0.98	0.99	1.00
0.97	0.73	0.82	0.80	0.80	0.70	0.92	0.95	0.92
BIC								
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.92	0.67	0.85	0.92	0.87	0.84	0.94	0.97
0.99	0.77	0.55	0.66	0.68	0.83	0.80	0.90	0.83
0.98	0.66	0.59	0.67	0.62	0.77	0.63	0.70	0.81
0.97	0.56	0.43	0.60	0.55	0.62	0.61	0.72	0.63

TABLE 4 (*Continued*)
Detection Frequencies of Combined Decision Rule
Compared to Those of AIC and BIC ($T = 140$, $T = 160$)

	$T = 140$				$T = 160$			
Combined Decision Rule								
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.87	0.91	0.98	0.99	0.85	0.93	0.97	0.94
0.99	0.94	1.00	0.97	0.96	0.94	0.98	1.00	0.99
0.98	0.92	0.94	0.99	0.92	0.83	0.99	0.98	0.94
0.97	0.73	0.87	0.85	0.87	0.82	0.91	0.89	0.83
AIC								
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.64	0.85	0.90	0.90	0.67	0.85	0.90	0.83
0.99	0.83	1.00	1.00	1.00	0.83	1.00	1.00	1.00
0.98	0.82	0.99	1.00	0.99	0.79	1.00	1.00	0.98
0.97	0.70	0.95	0.94	0.97	0.81	0.96	0.94	0.88
BIC								
$a\backslash b$	1.00	0.99	0.98	0.97	1.00	0.99	0.98	0.97
1.00	0.93	0.93	0.99	1.00	0.92	0.98	0.99	0.99
0.99	0.90	0.96	0.91	0.88	0.96	0.94	0.99	0.92
0.98	0.89	0.80	0.86	0.76	0.80	0.89	0.87	0.86
0.97	0.61	0.71	0.72	0.71	0.72	0.71	0.77	0.71

4.3.4 Practical Decision Rule and its Performance

In empirical analysis, it may occur that both criterion disagree on the number of seasonal unit roots. Therefore, before we turn to the results of the empirical analysis on empirical data, we define our model selection strategy here as follows.

- Rule (C) : If AIC and BIC agree, we choose that model.
- Rule (S) : If AIC and BIC do not agree, then we compare the values of AIC and BIC for the best two models. The criterion with the largest difference between these two models gets priority, and the final model is selected by this criterion.

It is worth performing simulations to see if this rule works well. TABLE 4.4 gives the results for $T = 60$ and $T = 80$. For each sample size, the selection probability table (4×4) by our combined decision rule is reported in the top panel, which is followed by the results of AIC and BIC.

All the results are obtained by new simulation runs, and the S/N ratio is set to $q = 0.56 \times 10^4$, whereas $q = 0.56 \times 10^{-4}$ in TABLE 4.2. The reason is that in empirical analysis, we often see large values of S/N ratio. (See also section 4.4 below.)

Our combined decision rule seems to give quite satisfactory detection rates, though are not always better than the results obtained for AIC or BIC separately. Our procedure decides somewhat in between both information criteria, but the discrepancy among these procedures is contracting when T gets large.

The combined decision rule proposed here seems to have two advantages. AIC does not have very high detection probability when we have three seasonal unit roots, in which case our combined rule definitely improves AIC, though not so sharp as BIC does. On the other hand, BIC performs poorly when there is no seasonal unit root, and especially when the sample size is very small. In those cases, our procedure amends the weakness of BIC.

4.4 Empirical Analysis

In this section we illustrate our approach for a selected set of 22 quarterly macroeconomic time series variables, namely 8 UK series previously analyzed in Osborn (1990), 6 US time series in Franses (1996a) and GNP data of 8 countries in Hylleberg et al. (1993). See these studies for a description of the data. In these studies, the data are all analyzed using the AR–based HEGY method. The simulation results in Ghysels et al. (1994) suggests that airline type of models make the HEGY tests to reject seasonal unit roots too frequently. Hence one would expect that if the structural time series model is an adequate modeling device, our method would yield more seasonal unit roots. To allow for a comparison, we include HEGY results as they are obtained in the aforementioned studies.

The results are reported in TABLE 4.5, 4.6 and 4.7. For each series, we report AIC and BIC values for the all four models (4.14) – (4.17). In the third panel of each table, the number of seasonal unit roots by our method (upper) and the result of the above mentioned studies (lower) are shown. When we cannot obtain clear-cut answers, we resort to the Rule (S) defined in subsection 4.3.4, and those less-conclusive cases are shown with a dagger like $3^{\dagger}$. These cases will be discussed later in detail in subsection 4.4.2.

The S/N ratio q is reported in the last row of each panel. Only the q of the model selected by the EKG-SM method is reported. Except for UK import data (UKIMPORT), every series has a large q value.

4.4.1 Conclusive cases

For 12 out of 22 time series, AIC and BIC find the same number of seasonal unit roots. Among those 12 cases, our EKG-SM-based method and the AR-based HEGY method find agree-

Table 4.5: Empirical Results : UK Data

		UKEXPORT	UKGDP	UKIMPORT	UKINVPUB
AIC values	Model 3	-418.33	-619.91^{*}	-397.39	-228.34^{*}
	Model 2	-416.59	-618.76	-408.00^{*}	-224.34
	Model 1	-418.40^{*}	-617.91	-395.44	-226.34
	Model 0	-417.16	-616.76	-406.39	-224.34
BIC values	Model 3	-397.95^{*}	-599.52^{*}	-377.00	-209.57^{*}
	Model 2	-393.26	-595.46	-384.70^{*}	-202.88
	Model 1	-395.10	-594.61	-372.13	-204.89
	Model 0	-390.96	-590.54	-380.18	-200.20
Decision on the Number of Seasonal Unit Roots					
EKG-SM		$3^{\dagger}$	3	2	3
HEGY-AR		0	0	0	3
q		0.892×10^{-2}	0.421×10^{-1}	0.105×10^{-5}	0.289×10^{4}

		UKINVTOT	UKNONDUR	UKTOTCO	UKWORKFO
AIC values	Model 3	-438.53	-767.39^{*}	-676.55	-1035.62
	Model 2	-440.58^{*}	-766.01	-676.91^{*}	-1049.92^{*}
	Model 1	-436.58	-765.39	-674.61	-1033.62
	Model 0	-438.58	-764.01	-674.96	-1048.42
BIC values	Model 3	-418.19^{*}	-747.00^{*}	-656.16^{*}	-1015.23
	Model 2	-417.28	-742.71	-653.61	-1026.62^{*}
	Model 1	-413.28	-742.09	-651.31	-1010.32
	Model 0	-412.37	-737.80	-648.75	-1022.21
Decision on the Number of Seasonal Unit Roots					
EKG-SM		$2^{\dagger}$	3	$3^{\dagger}$	2
HEGY-AR		0	3	3	0
q		0.643×10^{-1}	0.129	0.714×10^{-1}	0.584×10^{-1}

ment for only 5 time series. For UKINVPUB, UKNONDUR, USNONDUR, GERMAN and NETHER, the EKG-SM-based method gives the same answer as the HEGY method, and *all of these cases are the three unit root cases.* For 7 other cases, namely for UKGDP, UKIMPORT, UKWORKFO, USDUR, UK, ITALY and SWEDEN, AIC and BIC detect the same number of seasonal unit root(s), and *they find more seasonal unit roots as the HEGY does in most cases.* One exception of these seven cases is GNP of ITALY. While the HEGY method finds three seasonal unit roots, AIC and BIC conclude this series has just one seasonal unit root. In fact, the estimated annual seasonal coefficient for ITALY is 0.968 which is a little bit away from unity. As the estimated biannual seasonal coefficient is 0.999, this supports the existence of biannual unit root for the GNP of ITALY.

Table 4.6: Empirical Results : US Data

		USCONSTO	USDUR	USINDPRO
AIC values	Model 3	-949.63	-451.51^{*}	-592.07
	Model 2	-951.58	-450.20	-592.90^{*}
	Model 1	-951.98	-450.56	-590.12
	Model 0	-955.07^{*}	-449.24	-590.94
BIC values	Model 3	-927.28^{*}	-429.16^{*}	-572.10^{*}
	Model 2	-926.04	-424.65	-570.08
	Model 1	-926.44	-425.02	-567.30
	Model 0	-926.33	-420.50	-565.27
Decision on the Number of Seasonal Unit Roots				
EKG-SM		$0^{\dagger}$	3	$3^{\dagger}$
HEGY-AR		1	1	1
q		0.469×10^{-1}	0.184×10^{-1}	0.109×10^{2}

		USMONEY	USNONDUR	USSERVI
AIC values	Model 3	-775.45	-951.37^{*}	-1100.97
	Model 2	-773.57	-950.45	-1101.48
	Model 1	-776.39^{*}	-950.46	-1101.15
	Model 0	-774.44	-949.58	-1101.96^{*}
BIC values	Model 3	-755.06^{*}	-929.02^{*}	-1078.62^{*}
	Model 2	-750.27	-924.90	-1075.93
	Model 1	-753.09	-924.91	-1075.61
	Model 0	-748.29	-920.85	-1073.22
Decision on the Number of Seasonal Unit Roots				
EKG-SM		$3^{\dagger}$	3	$3^{\dagger}$
HEGY-AR		0	3	0
q		0.983×10^{-1}	0.107	0.169

4.4.2 Examination of Split-Decision Cases

The 10 remaining cases, in which AIC and BIC find different numbers of seasonal unit roots, were classified by the Rule (S) defined in subsection 4.3.4. Though our combined decision rule seems to work reasonably well, we will discuss those less conclusive cases in detail in this subsection.

From the results presented in TABLE 4.5, 4.6 and 4.7, these 10 cases can be classified into two groups. The first group consists of UKINVTOT, UKTOTCO, USMONEY, CANADA, UKEXPORT and USINDPRO. The results for these 6 time series are clearly suggestive on the existence or non-existence of unit root at a certain seasonal frequency.

For UKINVTOT the estimated annual coefficient is $\hat{b} = 0.999$, which strongly supports the existence of annual seasonal unit roots. In the same way, $(\hat{a}, \hat{b}) = (0.992, 0.999)$ in UKTOTCO

Table 4.7: Empirical Results : GNP Data

		JAPAN	UK	ITALY	TAIWAN
AIC values	Model 3	-443.34	-606.06^{*}	-239.70	-418.19
	Model 2	-442.61	-605.28	-237.73	-417.19
	Model 1	-446.50	-604.06	-246.06^{*}	-422.95
	Model 0	-446.63^{*}	-603.28	-244.08	-424.65^{*}
BIC values	Model 3	-425.92	-585.93^{*}	-225.04	-399.48
	Model 2	-422.70	-585.28	-220.97	-395.80
	Model 1	-426.59^{*}	-581.06	-229.31^{*}	-401.57^{*}
	Model 0	-424.23	-577.40	-225.24	-400.60
Decision on the Number of Seasonal Unit Roots					
EKG-SM		$1^{\dagger}$	3	1	0
HEGY-AR		3	1	3	3
q		0.240×10	0.362×10^{-1}	0.466×10^{-1}	0.232

		NETHER	GERMANY	CANADA	SWEDEN
AIC values	Model 3	-144.71^{*}	-523.36	-522.09	-328.45^{*}
	Model 2	-136.45	-527.33^{*}	-523.06^{*}	-322.45
	Model 1	-144.17	-525.62	-520.89	-322.89
	Model 0	-136.56	-524.64	-521.08	-320.88
BIC values	Model 3	-132.22^{*}	-504.40	-503.38^{*}	-308.61^{*}
	Model 2	-122.18	-505.65^{*}	-501.69	-304.35
	Model 1	-129.89	-503.95	-498.70	-304.79
	Model 0	-120.50	-500.26	-497.02	-300.52
Decision on the Number of Seasonal Unit Roots					
EKG-SM		3	2	$3^{\dagger}$	3
HEGY-AR		3	2	3	2
q		0.427	0.241×10^{2}	0.142×10^{2}	0.171×10^{3}

shows clear evidence of annual seasonal unit root while the estimated biannual coefficient lies in a difficult region. The results for USMONEY presents a strong evidence for biannual unit root together with the estimated coefficient $\hat{a} = 0.999$, whereas the HEGY method finds no seasonal unit roots. For CANADA, the estimated coefficients in Model 0 are given as $(\hat{a}, \hat{b}) = (0.936, 1.000)$ which strongly supports the annual non-stationarity. Model 2 and Model 3 are very close to each other both in terms of AIC and BIC. In UKEXPORT, introducing annual stationary roots improves the log-likelihood only by a small amount; 1.03 in Model 1 compared to Model 3. This leads to very close values of AIC between Model 1 and Model 3, and finally BIC decides because it puts heavier penalty, $\ln(136) \approx 4.92 > 2$ for an extra parameter. At least, the biannual non-stationarity is strongly suggested for this series.

USINDPRO differs from the above 5 time series in a sense that the conclusion in this paper is

totally different from that of HEGY test. Both information criteria and the estimated coefficients $(\hat{a},\hat{b}) = (0.973, 0.996)$ support the existence of the annual unit roots while the conclusion of HEGY is one unit root at biannual frequency.

Among the remaining 4 series, that is, TAIWAN, JAPAN, USCONSTO and USSERVI, the first two series need some remarks on the estimated parameters, although still partially conclusive results could be deduced. For GNP series of JAPAN and TAIWAN, the problem is that the estimated dispersion parameters are not stable among the fitted models. For JAPAN, estimated $\hat{\sigma}_v^2$ values in Model 3, Model 2 and Model 1 are 10 to 20 times larger than in Model 0. Except for this difficulty, the results suggest the annual non-stationarity for GNP of JAPAN. In contrast, only Model 3 has the large dispersion parameter $\hat{\sigma}_v^2$ in TAIWAN. But it can be seen from TABLE 4.7 that both AIC and BIC support small number of seasonal unit roots, and the estimated coefficients for Model 0 $(\hat{a},\hat{b}) = (0.968, 0.939)$ look supportive for the combined decision rule here.

USCONSTO and USSERVI are the cases for which the combined decision rule seems not to work well. For both series, the gain in the log-likelihood by introducing extra parameters is not very large, and is almost equal to the penalty term given by either AIC or BIC, which results in an extreme split-decision; Model 0 and Model 3. For USCONSTO, the estimated coefficients $(\hat{a},\hat{b}) = (0.995, 0.994)$ for Model 0 suggest that the true parameters might lie in a near non-stationary region in the parameter space, whereas the number of seasonal unit roots are determined to be 0. To the contrary, as the estimated coefficients of Model 0 in USSERVI are $(\hat{a},\hat{b}) = (0.965, 0.977)$, the conclusions of HEGY and AIC look reasonable, while the number of seasonal unit roots are determined to be 3 by the combined decision rule here.

4.5 Conclusion

In this paper we proposed a model selection approach to detect the number of seasonal unit roots, especially for series which may well be described by the airline model in nearly overdifferenced situations. We considered a version of structural time series models which has close relation with the airline model, and extended it to concern various seasonal unit root hypothesis. We investigated the detection frequencies of AIC and BIC under various parameter grids, sample sizes and S/N ratios. Our Monte Carlo simulations suggest that our method performs well. Even in cases extremely close to the unit root case, the combined decision rule based on AIC and BIC still provides a useful strategy to determine the number of seasonal unit roots.

Our empirical analysis shows that our method and the HEGY test can lead to substantially

different results, where our method tends to find more seasonal unit roots. This can be interpreted as follows. Because the estimated dispersion parameter of the seasonal component is very small for many series, it is expected that the airline model with nearly overdifferencing will fit well for these series. In such a case, the AR based HEGY test is known to over-reject the null hypothesis of unit root. Hence, it is natural that we find more seasonal unit roots than would find upon using the HEGY method.

Chapter 5

Do Seasonal Unit Roots Matter for Forecasting Monthly Industrial Productions?

5.1 Introduction

This paper is concerned with out-of-sample forecasting of monthly total industrial production series for OECD countries, where the data have not been seasonally adjusted. Indeed, typical properties of industrial production series for such industrialized countries are an upward-moving trend, which suggests the usefulness of analyzing growth rates, and pronounced seasonal variation. In some cases one may assume that this seasonal variation is approximately constant over time, that is, one may consider a model for the growth rates which includes constant seasonal intercepts. In other cases, seasonality seems to change over time, and then one may want to consider models with seasonal unit roots, see, for example, Franses (1996a) for a review of various time series models for seasonal data. Whatever choice one makes, it has become well recognized that properly taking care of seasonality improves the out-of-sample forecasting quality of the model, see for example Osborn et al. (1999) for a recent illustration.

As several industrial production data display patterns that seem to correspond with slowly changing trends and seasonality, one may want to consider the class of structural time series models, see Harvey (1989) . An important property of these models is that these models can be written as ARIMA type time series models with near-unit roots in the MA polynomial. It is well known that these models allow for substantial flexibility in describing and forecasting economic time series data. For example, in Thury and Witt (1998) it is shown that a structural time series model yields rather accurate forecasts for industrial production.

A key property of most structural time series models, and in particular of the often applied Basic Structural Model [BSM] , see Harvey (1989) , is that they impose eleven seasonal unit

roots in case of monthly data. See also Proietti (2000) for a recent survey of various specifications for the seasonal component in a structural time series model. The assumption of many unit roots allows for substantial changes in the seasonal pattern over time. The obvious question is now whether this assumption allows for too much flexibility, and hence whether perhaps a smaller number of seasonal unit roots would be more appropriate. In this paper we address this issue, and we put it into a forecasting perspective, that is, we examine whether a smaller number of seasonal unit roots (if so indicated according to within-sample analysis) yields better forecasts. In order to detect seasonal unit roots, we use the method recently developed in Kawasaki and Franses (1999). This method starts off with a structural time series model, and it uses model selection criteria to indicate the number of seasonal unit roots. This method is particularly useful for the present purposes as the simulation results in Ghysels et al. (1994) show that seasonal unit root tests which are based on a purely autoregressive model, like those developed in Hylleberg et al. (1990), fail dramatically in case the data have near-unit root MA structures.

The outline of our paper is as follows. In Section 2 we discuss and extend the model selection approach in Kawasaki and Franses (1999). As their method was developed for quarterly data, we extend it to monthly data in the present paper. In Section 3 we apply the method to monthly industrial production series for sixteen OECD countries. In this section our main interest concerns the relevance of within-sample analysis for out-of-sample forecasting. In section 4, we conclude with some remarks.

5.2 Detecting seasonal unit roots

In this section we provide an outline of the model selection method to detect the appropriate number of seasonal unit roots in a structural time series model. First, we briefly discuss some aspects of the basic structural model. Next, we provide the relevant details concerning the application of the method to monthly data.

5.2.1 Basic Structural Model

One often considered version of a structural time series model for a monthly time series $\{y_t\}$ contains the following three equations, that is,

$$y_t = \mu_t + s_t + w_t, \quad w_t \sim \text{NID}(0, \sigma_w^2) \tag{5.1}$$

$$(1-L)^2 \mu_t = u_t, \quad u_t \sim \text{NID}(0, \sigma_u^2) \tag{5.2}$$

$$S(L)s_t = (1+L+\cdots+L^{10}+L^{11})s_t = v_t, \quad v_t \sim \text{NID}(0,\sigma_v^2) \tag{5.3}$$

where L is the usual lag operator and where the error processes w_t, u_t and v_t are also mutually independent. This model turns out to be useful for describing and forecasting time series with slowly evolving trends and seasonality, see Kitagawa (1981), Kitagawa and Gersch (1984) and Harvey (1985), among others.

If one would adopt a first order random walk plus variable drift for the trend model instead of (5.2), one usually refer to the resultant set of equations as the basic structural model (BSM) after Harvey (1989, p.172), see also Proietti (2000). Throughout this paper, however, we assume a second order random walk for the trend model. Note that (5.2) can be written as

$$\mu_t = \mu_{t-1} + (\mu_{t-1} - \mu_{t-2}) + u_t$$

and hence it can be easily seen that the second order random walk model is a slightly restricted version of that of BSM, in the sense that the time varying drift in the BSM ($\beta_{t-1} = \mu_{t-1} - \mu_{t-2}$) is driven by the same noise u_t. The main reason to adopt (5.2) in this paper is that (1) and (3) can be summarized as a seasonal ARIMA model for y_t with a structure that comes close to the well-known airline model, denoted as ARIMA $(0,1,1)\times(0,1,1)_{12}$, see Harvey (1989, p.56), Maravall (1985) and Kawasaki and Franses (1999). This airline model is often found useful to describe seasonal economic time series.

The state space form of (1) and (3) immediately follows from the definition of the state vector α_t as $\alpha_t = (\mu_t, \mu_{t-1}, s_t, \ldots, s_{t-11})$. The transition equation and the measurement equation are then defined by

$$\alpha_t = T\alpha_{t-1} + R\eta_t \tag{5.4}$$

$$y_t = Z\alpha_t + \varepsilon_t \tag{5.5}$$

respectively, where T, R and Z are defined as

$$T = \left(\begin{array}{cc|ccc} 2 & -1 & \multicolumn{3}{c}{\multirow{2}{*}{O}} \\ 1 & 0 & & & \\ \hline \multicolumn{2}{c|}{\multirow{3}{*}{O}} & -1 & \cdots & -1 \quad -1 \\ & & 1 & \cdots & 0 \quad 0 \\ & & \vdots & \ddots & \vdots \quad \vdots \\ & & 0 & \cdots & 1 \quad 0 \end{array}\right), \qquad R = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{pmatrix}$$

$$Z = (1,0,1,0,\ldots,0).$$

The mean of $\eta_t = (u_t, v_t)'$ is assumed to be zero, and the covariance matrix of η_t, denoted by Q, is the diagonal matrix with elements σ_u^2 and σ_v^2 by assumption, see (5.2) and (5.3).

When we assume the Gaussian distribution for every noise process (see (5.1)–(5.3)), the above modelling approach invariably requires the Kalman filter, see Anderson and Moore (1979) and Harvey (1989), inter alia. In short, one should repeat the *prediction* step

$$\alpha_{t|t-1} = T\alpha_{t-1|t-1} \tag{5.6}$$

$$P_{t|t-1} = TP_{t-1|t-1}T' + RQR' \tag{5.7}$$

where $P_{t-1|t-1}$ denotes the covariance matrix of the state estimation error, and *update* recursively;

$$\alpha_{t|t} = \alpha_{t|t-1} + P_{t|t-1}Z'F_{t|t}^{-1}(y_t - Z\alpha_{t|t-1}) \tag{5.8}$$

$$P_{t|t} = P_{t|t-1} - P_{t|t-1}Z'F_{t|t}^{-1}ZP_{t|t-1}, \tag{5.9}$$

where $F_{t|t} = ZP_{t|t-1}Z' + \sigma_w^2$. Given the initial conditions $\alpha_{1|0}$ and $P_{1|0}$, the log-likelihood function can be expressed in the prediction error decomposition form,

$$\log L = -\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T}\log F_{t|t} - \frac{1}{2}\sum_{t=1}^{T} v_t F_{t|t}^{-1} v_t \tag{5.10}$$

where $v_t = y_t - Z\alpha_{t|t-1}$. The unknown hyperparameters of the model are usually estimated by the method of maximum likelihood using numerical optimization techniques like BFGS formula.

As for the initialization of the Kalman filter, we employ the 'large κ' approximation for the diffuse prior, that is, we assume the diagonal form of covariance matrix $P_{1|0} = \kappa I$ with large κ, see Harvey (1989, p.121). Throughout this paper, κ is set to 10^4, which seems quite a reasonable choice for the sake of numerical stability. Strictly speaking, we could perform different initializations for the four seasonal models under consideration. We concern, however, only the non-stationary and near non-stationary cases, which gives us an impression that the same 'large κ' approximation does not bring enormous bias in the simulations and empirical analysis in this paper. For the initial state vector $\alpha_{1|0}$, the elements related to the seasonal component are set to zero. On the other hand, we estimate the trend initial condition using the first quarter of the whole sample period so that the initial trend value must not be too far away from the actual observation. Because of this treatment, we regard $\alpha_{1|0}$ as a part of the hyperparameters of the model though the elements of $\alpha_{1|0}$ are not the objects of numerical optimization. It should be noted, however, that this does not affect the model selection procedure stated below as long as we employ the same specification for $\alpha_{1|0}$ and $P_{1|0}$.

Table 5.1: Stationary seasonal polynomials

non-stationary		stationary	frequency
$1+L$	$\longrightarrow$	$1+c_1L$	π (2 months)
$1+L^2$	$\longrightarrow$	$1+c_2^2L^2$	$\frac{\pi}{2}$ (4 months)
$1-\sqrt{3}L+L^2$	$\longrightarrow$	$1-\sqrt{3}c_3L+c_3^2L^2$	$\frac{\pi}{6}$ (12 months)
$1+\sqrt{3}L+L^2$	$\longrightarrow$	$1+\sqrt{3}c_4L+c_4^2L^2$	$\frac{5\pi}{6}$ (2.4 months)
$1-L+L^2$	$\longrightarrow$	$1-c_5L+c_5^2L^2$	$\frac{\pi}{3}$ (6 months)
$1+L+L^2$	$\longrightarrow$	$1+c_6L+c_6^2L^2$	$\frac{2\pi}{3}$ (3 months)

5.2.2 Seasonal Roots

The lag polynomial in equation (5.3) can be factorized as

$$\begin{aligned} S(L) &= (1+L)(1+L^2)(1+L^4+L^8) \\ &= (1+L)(1+L^2)(1-L^2+L^4)(1+L^2+L^4) \\ &= (1+L)(1+L^2)(1-\sqrt{3}L+L^2)(1+\sqrt{3}L+L^2) \\ &\quad \times(1-L+L^2)(1+L+L^2)\,. \end{aligned} \tag{5.11}$$

Each factor has its zero(s) on the unit circle in the complex plane. Each of them corresponds to a certain seasonal frequency less than or equal to the Nyquist frequency. Here we introduce additional parameters $c_1,\ldots,c_6$ to allow stationary roots at seasonal frequencies. Therefore we assume all c_i's ($i=1,\ldots,6$) lie between 0 and 1 so that the zeros of polynomials lie outside the unit circle. The way we parameterize is summarized in Table 5.1.

The new parameters $c_1,\ldots,c_6$ generate a family of $2^6=64$ different models, while the new seasonal component model will be given as

$$\left(1+\sum_{i=1}^{11} g_iL^i\right)s_t = v_t \tag{5.12}$$

where g_i's can be expressed in terms of $c_1,\ldots,c_6$, see Appendix. In accordance, the matrix T in equation (4) will be generalized as follows.

$$T = \left(\begin{array}{cc|ccc} 2 & -1 & & O & \\ 1 & 0 & & & \\ \hline & & -g_1 & \cdots & -g_{10} \quad -g_{11} \\ & O & 1 & \cdots & 0 \quad\ 0 \\ & & \vdots & \ddots & \vdots \quad\ \vdots \\ & & 0 & \cdots & 1 \quad\ 0 \end{array}\right)$$

Our basic strategy to determine the number of seasonal unit roots in y_t is as follows. For given data, we estimate all the models and choose a model that attains minimum information

criterion. Although many variants of information criteria are proposed since Akaike (1973), the most popular criteria in practice seem to be AIC and BIC,

$$\begin{aligned} \text{AIC} &= -2\hat{\ell} + 2k \\ \text{BIC} &= -2\hat{\ell} + k \log N \end{aligned}$$

where $\hat{\ell}$ denotes the maximized log-likelihood, k is the number of estimated unknown parameters, and N denotes the sample size used in estimating parameters, see Akaike (1973) and Schwarz (1978), respectively. Given the simulation evidence concerning quarterly series in Kawasaki and Franses (1999), we can expect a high detection rate. These simulations are also suggestive concerning the cases when the method does not work very well. These concern the situation when $q = \sigma_v^2/\sigma_w^2$, is small, that is when the signal-to-noise ratio between the dispersion parameter of the seasonal component and the observation error variance is small, when the sample size is smaller than fifteen years of seasonal data, and when the root of the true seasonal polynomial is extremely close to the unity.

5.3 Forecasting Monthly Industrial Production

In this section we use the method outlined in the previous section to determine the number of seasonal unit roots in monthly industrial production series. Upon the indication of a model selection criterion, we then examine if imposing seasonal unit roots or stationary seasonal roots leads to better out-of-sample forecasts.

5.3.1 Data

We apply our methodology to monthly industrial production index series of sixteen OECD countries, that is, Austria, Belgium, Canada, Germany, Finland, France, Greece. Italy, Japan, Luxembourg, the Netherlands, Norway, Portugal, Spain, the U.K. and the U.S.A. All series are transformed by the natural logarithmic function. The sample period and sample size for each country are reported in Table 5.2. Although most series were available from January 1960 onwards, we decided to analyze only the data from January 1975 onwards. The main reason is that we found that for many series the characteristics have changed dramatically, and hence we encountered problems while fitting constant parameter models. For Greece and Portugal we decided to start even later due to instabilities. For Germany we take only the observations after the re-unification for obvious reasons. Hence, except for Germany, we have more than 20 years of data for all countries.

Table 5.2: Data description

Country	Sample period	Sample size
Austria	Jan. 1975 – Aug. 1998	284
Belgium	Jan. 1975 – Sep. 1998	285
Canada	Jan. 1975 – Sep. 1998	287
Germany	Jan. 1991 – Sep. 1998	93
Spain	Jan. 1975 – Sep. 1998	287
Finland	Jan. 1975 – Sep. 1998	285
France	Jan. 1975 – Sep. 1998	285
U.K.	Jan. 1975 – Oct. 1998	286
Greece	Jan. 1980 – Sep. 1998	225
Italy	Jan. 1975 – Sep. 1998	285
Japan	Jan. 1975 – Oct. 1998	286
Luxembourg	Jan. 1975 – Jul. 1998	282
Netherlands	Jan. 1975 – Sep. 1998	285
Norway	Jan. 1975 – Oct. 1998	286
Portugal	Jan. 1976 – Sep. 1998	273
U.S.A.	Jan. 1975 – Oct. 1998	286

Table 5.3: Estimated Trade-off Parameters

Country	$\hat{q}$	Country	$\hat{q}$
Austria	0.783×10^{-1}	Japan	0.160
Belgium	0.295×10^{-1}	Luxembourg	0.786×10^{-1}
Canada	0.116×10^{2}	Netherlands	0.487×10^{-1}
Finland	0.314×10^{2}	Norway	0.487×10^{-1}
France	0.287×10^{2}	Portugal	0.195
Germany	0.740×10^{3}	Spain	0.381
Greece	0.677×10^{-2}	U.K.	0.213×10^{3}
Italy	0.668	U.S.A.	0.128×10

Table 5.4: RMSPE reduction in static forecasting based on AIC

Country	RMSPE	Reduction(%)	12	6	4	3	2.4	2
Austria	0.02593	△0.119						◇
Belgium	0.04798	△0.370	●					
Canada	0.00680	2.704			●		○	○
Germany	0.02397	△14.628					○	
Spain	0.02613	△1.711	◇					
Finland	0.03085	△3.970	○	○	○	○	○	◇
France	0.02435	△1.878	○	○	○	◇	○	
U.K.	0.02890	△5.197	○	◇	○	○	○	◇
Greece	0.03107	0						
Italy	0.03298	0						
Japan	0.02099	△4.948	○					
Luxembourg	0.03639	1.239	○					
Netherlands	0.02356	△23.926	●	●	◇	●	○	◇
Norway	0.02844	△11.751	○	○	○	○	○	○
Portugal	0.02422	△8.009	○	◇				
U.S.A.	0.00751	1.660		◇	○		○	●

Table 5.3 reports the estimated trade-off parameters ($\hat{q}$) when we fit the BSM for all 16 countries. Though the BSM might not the best model for each country's industrial production, the estimated trade-off still sheds some lights on the validity of the model selection procedure used here. The bigger the trade-off, the better the procedure works. In another words, the signal to be detected is relatively clear and less buried in the noise. Except Greece, the smaller $\hat{q}$'s are of the order 10^{-1}. The aforementioned simulation results suggest that for these cases one can expect a high detection rate when the sample size gets large. The length of most time series analyzed is about for 24 years, which would correspond to 100 observations on a quarterly series. The simulation results in Kawasaki and Franses (1999) suggest that for $q = 10^{-1}$ and $T = 80$ the correct detection rates vary from 0.76 to 0.99, depending on which roots are truly equal to unity.

5.3.2 Out-of-sample forecasting

In this subsection we report on the out-of-sample forecasting performance of the various models. We hold out the last 48 observations for forecasting evaluation for all the countries except for Germany. For Germany, we consider only 12 such observations due to shortage of data. We estimate the parameters of all 64 models using the effective sample size T'. The model selected by AIC or BIC is used to generate a single one-step-ahead forecast. Next, we add an observa-

Table 5.5: RMSPE reduction in static forecasting based on BIC

Country	RMSPE	Reduction(%)	12	6	4	3	2.4	2
Austria	0.02596	0						
Belgium	0.04815	0						
Canada	0.00662	0						
Germany	0.02466	△12.174					◇	
Spain	0.02658	0						
Finland	0.03111	△3.157	○	○	○	○	◇	◇
France	0.02416	△2.653	○	◇				
U.K.	0.02918	△4.290	◇				○	
Greece	0.03170	0						
Italy	0.03298	0						
Japan	0.02099	△4.948	○					
Luxembourg	0.03594	0						
Netherlands	0.02435	△21.346			●		○	◇
Norway	0.02844	△11.751	○	○	○	○	○	○
Portugal	0.02497	△5.916	○					
U.S.A.	0.00739	0						

tion, we estimate the parameters again, and we compare all the models using AIC and BIC and generate a one-step-ahead forecast using the by then preferred model. This is repeated until the last observation. We calculate the root mean squared prediction error according to

$$\text{RMSPE} = \sqrt{\sum_{i=1}^{48} (\hat{y}_{T'+i} - y_{T'+i})^2 / 48}.$$

and only for 12 forecasts for Germany. As we are interested in comparing models with all seasonal unit roots imposed and models with some or all stationary seasonal roots, we compute the reduction in RMSPE given the decision based on AIC and BIC. The results are summarized in Table 5.4 and or for AIC and BIC, respectively. A Δ in the third column denotes a reduction in RMSPE. For example, in Table 5.4, the stationary seasonal root models selected by AIC implies a 24% reduction for the series for the Netherlands. A zero in the third column of these two tables means that AIC or BIC suggests that the model with only seasonal unit roots gets preferred. In Table 5.4 and or we also report on the stability of finding stationary seasonal roots. A ○ indicates that a stationary root is always supported at a specific seasonal frequency. A white ◇ shows frequent acceptance of a stationary root, or acceptance for a lengthy period, while a • entails that acceptance is rare. Clearly, the results in the last six columns of Tables 5.4 and 5.5 suggest that the finding of stationary roots can be unstable over time. This suggests that it is worthwhile to repeat estimation and model selection anytime a new observation becomes

available. In sum, the results in Tables 5.4 and 5.5 show that seasonal stationary roots can lead to (sometimes substantially) smaller root mean squares prediction errors as compared with a unit roots model. Additionally, both AIC and BIC seem to be useful for determining the number of stationary seasonal unit roots in the seasonal component polynomial. In terms of RMSPE, the use of AIC seems to lead to slightly better results than using BIC in most countries, although for Canada, Luxembourg, and the U.S.A. it yields worse results. With BIC there is a tendency to select the model with all seasonal unit roots more often, also for these three countries, and hence this strategy does not lead to forecast deterioration. As for the stability of the seasonal polynomial structure, we find more $\diamond$ and $\bullet$ in Table 5.4 than in Table 5.5. This corresponds with the fact that AIC usually allows for more parameters than BIC does.

5.4 Concluding remarks

In this paper we proposed a model selection approach to detect the number of seasonal unit roots for monthly time series within the context of a structural time series model. The usual basic structural model assumes that there are eleven such seasonal unit roots. We proposed to use AIC and BIC to examine if one or more of these are in fact stationary. Additional to the within-sample analysis, we examined whether the introduction of stationary roots improves one-step-ahead out-of-sample forecasting. We considered monthly industrial production series for sixteen OECD countries, for which we generally found that when AIC or BIC indicate that a smaller number of seasonal unit roots should be assumed and hence that some roots are stationary, the corresponding model also gives more accurate forecasts. As for the difference between AIC and BIC, we can conclude that AIC generally attains higher reductions in the root mean squared prediction errors, but BIC seems a safer choice as it did not lead to worse forecasts.

Appendix

We give the result of expansion of the following polynomial

$$(1+c_1L)(1+c_2^2L^2)(1-\sqrt{3}c_3L+c_3^2L^2)(1+\sqrt{3}c_4L+c_4^2L^2)(1-c_5L+c_5^2L^2)(1+c_6L+c_6^2L^2)$$

which finally constitutes the left hand side of equation (12). These results can be easily verified by a language like Mathematica.

$$\begin{aligned} g_1 &= -c_1+\sqrt{3}c_3-\sqrt{3}c_4+c_5-c_6 \\ g_2 &= -c_2^2-c_3^2-c_4^2-c_5^2+c_1(\sqrt{3}c_3-\sqrt{3}c_4+c_5-c_6) \end{aligned}$$

$$
\begin{aligned}
& +c_5c_6 - c_6^2 + c_4(\sqrt{3}c_5 - \sqrt{3}c_6) + c_3(3c_4 - \sqrt{3}c_5 + \sqrt{3}c_6) \\
g_3 \;=\; & c_4^2(c_5 - c_6) + c_3^2(-\sqrt{3}c_4 + c_5 - c_6) + c_2^2(\sqrt{3}c_3 - \sqrt{3}c_4 + c_5 - c_6) \\
& -c_5^2c_6 + c_5c_6^2 + c_4(-\sqrt{3}c_5^2 + \sqrt{3}c_5c_6 - \sqrt{3}c_6^2) \\
& +c_3(\sqrt{3}c_4^2 + \sqrt{3}c_5^2 - \sqrt{3}c_5c_6 + \sqrt{3}c_6^2 + c_4(-3c_5 + 3c_6)) \\
& +c_1(-c_2^2 - c_3^2 - c_4^2 - c_5^2 + c_5c_6 - c_6^2 + c_4(\sqrt{3}c_5 - \sqrt{3}c_6) + c_3(3c_4 - \sqrt{3}c_5 + \sqrt{3}c_6)) \\
g_4 \;=\; & -c_5^2c_6^2 + c_4^2(-c_5^2 + c_5c_6 - c_6^2) + c_4(-\sqrt{3}c_5^2c_6 + \sqrt{3}c_5c_6^2) \\
& +c_3^2(-c_4^2 - c_5^2 + c_5c_6 - c_6^2 + c_4(\sqrt{3}c_5 - \sqrt{3}c_6)) \\
& +c_2^2(-c_3^2 - c_4^2 - c_5^2 + c_5c_6 - c_6^2 + c_4(\sqrt{3}c_5 - \sqrt{3}c_6) + c_3(3c_4 - \sqrt{3}c_5 + \sqrt{3}c_6)) \\
& +c_3(\sqrt{3}c_5^2c_6 - \sqrt{3}c_5c_6^2 + c_4^2(-\sqrt{3}c_5 + \sqrt{3}c_6) + c_4(3c_5^2 - 3c_5c_6 + 3c_6^2)) \\
& +c_1(c_4^2(c_5 - c_6) + c_3^2(-\sqrt{3}c_4 + c_5 - c_6) + c_2^2(\sqrt{3}c_3 - \sqrt{3}c_4 + c_5 - c_6) \\
& -c_5^2c_6 + c_5c_6^2 + c_4(-\sqrt{3}c_5^2 + \sqrt{3}c_5c_6 - \sqrt{3}c_6^2) \\
& +c_3(\sqrt{3}c_4^2 + \sqrt{3}c_5^2 - \sqrt{3}c_5c_6 + \sqrt{3}c_6^2 + c_4(-3c_5 + 3c_6))) \\
g_5 \;=\; & -\sqrt{3}c_4c_5^2c_6^2 + c_4^2(-c_5^2c_6 + c_5c_6^2) \\
& +c_3^2(c_4^2(c_5 - c_6) - c_5^2c_6 + c_5c_6^2 + c_4(-\sqrt{3}c_5^2 + \sqrt{3}c_5c_6 - \sqrt{3}c_6^2)) \\
& +c_3(\sqrt{3}c_5^2c_6^2 + c_4^2(\sqrt{3}c_5^2 - \sqrt{3}c_5c_6 + \sqrt{3}c_6^2) + c_4(3c_5^2c_6 - 3c_5c_6^2)) \\
& +c_2^2(c_4^2(c_5 - c_6) + c_3^2(-\sqrt{3}c_4 + c_5 - c_6) - c_5^2c_6 + c_5c_6^2 \\
& +c_4(-\sqrt{3}c_5^2 + \sqrt{3}c_5c_6 - \sqrt{3}c_6^2) \\
& +c_3(\sqrt{3}c_4^2 + \sqrt{3}c_5^2 - \sqrt{3}c_5c_6 + \sqrt{3}c_6^2 + c_4(-3c_5 + 3c_6))) \\
& +c_1(-c_5^2c_6^2 + c_4^2(-c_5^2c_6 + c_5c_6^2) + c_4(-\sqrt{3}c_5^2c_6 + \sqrt{3}c_5c_6^2) \\
& +c_3^2(-c_4^2 - c_5^2 + c_5c_6 - c_6^2 + c_4(\sqrt{3}c_5 - \sqrt{3}c_6)) \\
& +c_2^2(-c_3^2 - c_4^2 - c_5^2 + c_5c_6 - c_6^2 + c_4(\sqrt{3}c_5 - \sqrt{3}c_6) + c_3(3c_4 - \sqrt{3}c_5 + \sqrt{3}c_6)) \\
& +c_3(\sqrt{3}c_5^2c_6 - \sqrt{3}c_5c_6^2 + c_4^2(-\sqrt{3}c_5 + \sqrt{3}c_6) + c_4(3c_5^2 - 3c_5c_6 + 3c_6^2))) \\
g_6 \;=\; & -c_4^2c_5^2c_6^2 + c_3(3c_4c_5^2c_6^2 + c_4^2(\sqrt{3}c_5^2c_6 - \sqrt{3}c_5c_6^2)) \\
& +c_3^2(-c_5^2c_6^2 + c_4^2(-c_5^2 + c_5c_6 - c_6^2) + c_4(-\sqrt{3}c_5^2c_6 + \sqrt{3}c_5c_6^2)) \\
& +c_2^2(-c_5^2c_6^2 + c_4^2(-c_5^2 + c_5c_6 - c_6^2) + c_4(-\sqrt{3}c_5^2c_6 + \sqrt{3}c_5c_6^2) \\
& +c_3^2(-c_4^2 - c_5^2 + c_5c_6 - c_6^2 + c_4(\sqrt{3}c_5 - \sqrt{3}c_6)) \\
& +c_3(\sqrt{3}c_5^2c_6 - \sqrt{3}c_5c_6^2 + c_4^2(-\sqrt{3}c_5 + \sqrt{3}c_6) + c_4(3c_5^2 - 3c_5c_6 + 3c_6^2))) \\
& +c_1(-\sqrt{3}c_4c_5^2c_6^2 + c_4^2(-c_5^2c_6 + c_5c_6^2) + c_3^2(c_4^2(c_5 - c_6) - c_5^2c_6 + c_5c_6^2 \\
& +c_4(-\sqrt{3}c_5^2 + \sqrt{3}c_5c_6 - \sqrt{3}c_6^2)) + c_3(\sqrt{3}c_5^2c_6^2 + c_4^2(\sqrt{3}c_5^2 - \sqrt{3}c_5c_6 + \sqrt{3}c_6^2) \\
& +c_4(3c_5^2c_6 - 3c_5c_6^2)) + c_2^2(c_4^2(c_5 - c_6) + c_3^2(-\sqrt{3}c_4 + c_5 - c_6)
\end{aligned}
$$

$$
\begin{aligned}
& -c_5^2c_6+c_5c_6^2+c_4(-\sqrt{3}c_5^2+\sqrt{3}c_5c_6-\sqrt{3}c_6^2) \\
& +c_3(\sqrt{3}c_4^2+\sqrt{3}c_5^2-\sqrt{3}c_5c_6+\sqrt{3}c_6^2+c_4(-3c_5+3c_6)))) \\
g_7 \;=\;& \sqrt{3}c_3c_4^2c_5^2c_6^2+c_3^2(-\sqrt{3}c_4c_5^2c_6^2+c_4^2(-c_5^2c_6+c_5c_6^2)) \\
& +c_2^2(-\sqrt{3}c_4c_5^2c_6^2+c_4^2(-c_5^2c_6+c_5c_6^2)+c_3^2(c_4^2(c_5-c_6)-c_5^2c_6+c_5c_6^2 \\
& +c_4(-\sqrt{3}c_5^2+\sqrt{3}c_5c_6-\sqrt{3}c_6^2)) \\
& +c_3(\sqrt{3}c_5^2c_6^2+c_4^2(\sqrt{3}c_5^2-\sqrt{3}c_5c_6+\sqrt{3}c_6^2)+c_4(3c_5^2c_6-3c_5c_6^2))) \\
& +c_1(-c_4^2c_5^2c_6^2+c_3(3c_4c_5^2c_6^2+c_4^2(\sqrt{3}c_5^2c_6-\sqrt{3}c_5c_6^2)) \\
& +c_3^2(-c_5^2c_6^2+c_4^2(-c_5^2+c_5c_6-c_6^2)+c_4(-\sqrt{3}c_5^2c_6+\sqrt{3}c_5c_6^2)) \\
& +c_2^2(-c_5^2c_6^2+c_4^2(-c_5^2+c_5c_6-c_6^2)+c_4(-\sqrt{3}c_5^2c_6+\sqrt{3}c_5c_6^2) \\
& +c_3^2(-c_4^2-c_5^2+c_5c_6-c_6^2+c_4(\sqrt{3}c_5-\sqrt{3}c_6)) \\
& +c_3(\sqrt{3}c_5^2c_6-\sqrt{3}c_5c_6^2+c_4^2(-\sqrt{3}c_5+\sqrt{3}c_6)+c_4(3c_5^2-3c_5c_6+3c_6^2)))) \\
g_8 \;=\;& -c_3^2c_4^2c_5^2c_6^2+c_2^2(-c_4^2c_5^2c_6^2+c_3(3c_4c_5^2c_6^2+c_4^2(\sqrt{3}c_5^2c_6-\sqrt{3}c_5c_6^2)) \\
& +c_3^2(-c_5^2c_6^2+c_4^2(-c_5^2+c_5c_6-c_6^2)+c_4(-\sqrt{3}c_5^2c_6+\sqrt{3}c_5c_6^2))) \\
& +c_1(\sqrt{3}c_3c_4^2c_5^2c_6^2+c_3^2(-\sqrt{3}c_4c_5^2c_6^2+c_4^2(-c_5^2c_6+c_5c_6^2)) \\
& +c_2^2(-\sqrt{3}c_4c_5^2c_6^2+c_4^2(-c_5^2c_6+c_5c_6^2) \\
& +c_3^2(c_4^2(c_5-c_6)-c_5^2c_6+c_5c_6^2+c_4(-\sqrt{3}c_5^2)+\sqrt{3}c_5c_6-\sqrt{3}c_6^2) \\
& +c_3(\sqrt{3}c_5^2c_6^2+c_4^2(\sqrt{3}c_5^2-\sqrt{3}c_5c_6+\sqrt{3}c_6^2)+c_4(3c_5^2c_6-3c_5c_6^2)))) \\
g_9 \;=\;& c_2^2(\sqrt{3}c_3c_4^2c_5^2c_6^2+c_3^2(-\sqrt{3}c_4c_5^2c_6^2+c_4^2(-c_5^2c_6+c_5c_6^2))) \\
& +c_1(-c_3^2c_4^2c_5^2c_6^2+c_2^2(-c_4^2c_5^2c_6^2+c_3(3c_4c_5^2c_6^2+c_4^2(\sqrt{3}c_5^2c_6-\sqrt{3}c_5c_6^2)) \\
& +c_3^2(-c_5^2c_6^2+c_4^2(-c_5^2+c_5c_6-c_6^2)+c_4(-\sqrt{3}c_5^2c_6+\sqrt{3}c_5c_6^2)))) \\
g_{10} \;=\;& -c_2^2c_3^2c_4^2c_5^2c_6^2+c_1c_2^2(\sqrt{3}c_3c_4^2c_5^2c_6^2+c_3^2(-\sqrt{3}c_4c_5^2c_6^2+c_4^2(-c_5^2c_6+c_5c_6^2))) \\
g_{11} \;=\;& -c_1c_2^2c_3^2c_4^2c_5^2c_6^2
\end{aligned}
$$

Chapter 6

Principal Component and Factor Analysis for Multiple Time Series

6.1 Introduction

Principal component analysis and factor analysis are very popular data reduction technique used in various fields of science. Although they are suitable for analyzing a cross-sectional data set where several items are observed by several individuals, both principal component and factor analysis are not actively pursued for time series data. The reason might be called for the most important premise that the sample is drawn identically and independently from a multivariate (sometimes Gaussian, additionally) distribution, which is completely irrelevant for time series settings.

As an example, let us take a look at factor analysis with single factor. Suppose p-variate observation x_i $(i = 1, \ldots, p)$ can be expressed by a single common factor z and idiosyncratic factor ε_i as

$$x_i(t) = \lambda_i z(t) + \varepsilon_i(t), \quad t = 1, \ldots, n$$

where the unobservable factor $z(t)$ and $\varepsilon_i(t)$ are mutually uncorrelated for all i and t, and the variance of $z(t)$ is normalized to be unity. Suppose the suffix t stands for individuals. Then the assumption of no correlation in $z(t)$ may be justified according to circumstances. However, if $x_i(t)$ is time series, $x_i(t)$ is in general correlated with $x_i(t+s)$, the assumption of no correlation does not hold at all. A formal application of the theory based on i.i.d. sampling to time series settings will inevitably invite misleading results.

The aim of this chapter is to survey the past research on principal component and factor analysis for multiple time series. Under what kind of assumption, or by what kind of approach are they justifiable? We clean-up the preceding research by two cut surface. In the first approach, the discrete Fourier transform of original time series which becomes asymptotically indepen-

dent is applied to the classical framework of principal component and factor analysis. These frequency domain approach is discussed in section 6.2. The second approach gives a model for unobservable factor process, which will be called a 'dynamic factor model' in this monograph. In section 6.3, we introduce three approaches for estimation of dynamic factor models in time domain. Section 6.4 raises a cautionary note on the use of principal component analysis in time domain — more precisely speaking, the formal use of PCA neglecting the possible lag structure in the latent factor process — and illustrates such danger by a real data analysis. Section 6.5 concludes this chapter.

6.2 Frequency Domain Approach

The first approach makes use of the data transformation which enables direct application of classical theory of principal component and factor analysis. It is discrete Fourier transform (DFT hereafter) that is employed here. The theory of time series analysis based on DFT of original time series has a longer history than that of time domain approach familiarized by Box and Jenkins (1970). In that sense, the methods reviewed in this section are classical but it is worth while mentioning that they play very important roles in deriving the theoretical results in stationary time series analysis, beyond the case of principal component analysis. The point is that the correlation among neighboring observations in time domain is converted to the heteroscedasticity of the distribution of DFT which is independent, at least asymptotically, with respect to frequency.

Suppose we observe multiple stationary time series $x_t (t = 1, \ldots, n)$. Then for $\nu_k = k/n$ $(k = 0, 1, \ldots, n)$, the DFT of x_t

$$X(\nu_k) = n^{-1/2} \sum_{t=1}^{n} x_t \exp(-2\pi i t \nu_k)$$

asymptotically follows a multivariate normal distribution. (See Theorem 4.4.1 in Brillinger (1981).) Then, the spectral density matrix $f_x(\nu)$ plays the same role as the covariance matrix does for classical PCA and FA. Only $\hat{f}_x(\nu)$ is needed to perform principal component analysis, while asymptotic normality gives theoretical justification of maximum likelihood estimation in factor analysis.

Subsequently in section 6.2.1, we summarize the results described in Brillinger (1981) concerning the PCA for stationary multiple time series. The results concerning factor analysis in frequency domain stated in section 6.2.2 are taken from Priestly et al. (1974), Priestly and Subba Rao (1975), Geweke (1977) among others. Section 6.2.3 introduces a couple of interesting applications exploiting frequency domain approach.

6.2.1 Principal Component Analysis in Frequency Domain

Classical Principal Component Analysis

Keeping the time series matters at bay, we start with i.i.d. situation under which we have p-variate data $x = (x_1, \ldots, x_p)'$. Now we want to transform x into a scalar so that it contains as much information on x as possible. Hence we search a vector c which maximize the variance of the following linear combination

$$y = c'x = c_1 x_1 + \cdots + c_p x_p, \tag{6.1}$$

where we impose a restriction $c'c = 1$. Let Σ_x be the covariance matrix of x. Rewriting the problem here, we are interested in the vector $c(\neq 0)$ that satisfies

$$\max c'\Sigma_x c. \tag{6.2}$$

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ be the eigenvalues of the matrix Σ_x arranged in descending order, and denote the corresponding eigenvectors normalized to have norm 1 as $e_1, e_2, \ldots, e_p$. Then the solution to (6.2) is given by $c = e_1$, and the linear combination $y_1 = e_1'x$ attains maximum variance $\mathrm{var}(y_1) = \lambda_1$, namely,

$$\max c'\Sigma_x c = e_1'\Sigma_x e_1 = \lambda_1.$$

The scalar y_1 is referred to as the first principal component. Similarly, we have the second principal component $y_2 = c'x$ so that c should maximize $\mathrm{var}(y_2)$ under the conditions $\mathrm{cov}(y_1, y_2) = 0$ and $c'c = 1$. Repeating the same argument, we can define up to the p-th principal component. The share of the variance of each principal component in the total variance reflects its importance. By $\mathrm{tr}(\Sigma_x) = \lambda_1 + \cdots + \lambda_p$, the contribution ratio of k-th principal component can be defined as $\mathrm{var}(y_k)/\mathrm{tr}(\Sigma_x) = \lambda_k / \sum_{j=1}^{p} \lambda_j$. Actual data analysis will be made not on the population covariance matrix Σ_x but on the sample covariance matrix $S_x = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'$ based on repeated measurement $x_1, \ldots, x_n$

PCA for Time Series

Let us go back to time series. Suppose we observed p-variate stationary time series $x_1, \ldots, x_t$. Mean is assumed to be zero without loss of generality, and the existence of $p \times p$ spectral density matrix $f_x(\nu)$ is also assumed. Note that $f_x(\nu)$ is complex-valued nonnegative definite Hermitian matrix. On the analogy of the arguments in usual principal component analysis ((6.1) and (6.2)), for a fixed ν, we consider the complex-valued univariate series $y_t(\nu) = c(\nu)^* x_t$ given by a complex vector $c(\nu)$. (Here c^* denotes the transpose of complex conjugate of c.) Under the

restriction $c(\nu)^* c(\nu) = 1$, we want to find a weighting vector $c(\nu)$ so that the spectral density at ν should be maximized. Because the spectral density matrix of $y_t(\nu)$ at the frequency ν can be expressed as $f_y(\nu) = c(\nu)^* f_x(\nu) c(\nu)$, the problem here is to find a complex vector $c(\nu)$ ($\neq 0$) that satisfies

$$\max c(\nu)^* f_x(\nu) c(\nu). \tag{6.3}$$

Let $\lambda_1(\nu) \geq \lambda_2(\nu) \geq \ldots \geq \lambda_p(\nu)$ be the eigenvalues of matrix $f_x(\nu)$, and $e_1(\nu), e_2(\nu), \ldots, e_p(\nu)$ be the corresponding eigenvectors of which norm is normalized to one. Then, the solution to (6.3) is given by $c(\nu) = e_1(\nu)$, and $y_t(\nu) = e_1(\nu)^* x_t$ attains the maximum variance, namely,

$$\max c(\nu)^* f_x(\nu) c(\nu) = e_1(\nu)^* f_x(\nu) e_1(\nu) = \lambda_1(\nu).$$

Changing notation from $y_t(\nu)$ to $y_{t1}(\nu)$, we call $y_{t1}(\nu)$ the first principal component a frequency ν. Similarly $y_{tk}(\nu)$ ($k = 2, \ldots, p$) can be defined like the usual principal component analysis. Further, the argument so far holds for all ν.

Data Reduction and PCA Series

The following problem may give us a motivation to devise principal component and factor analysis in frequency domain (Brillinger, 1981). Suppose at a remote place multivariate time series data is observed and collected. From the observation station the data is somehow transmitted, but the number of the channels for transmission is limited. Then we have to reduce the original data to a lower dimension data, say to a univariate process y_t, with a minimum loss of information contained in the original data x_t,

$$y_t = \sum_{j=-\infty}^{\infty} c_{t-j}^* x_j,$$

where the filter $\{c_j\}$ is absolutely summable ($\sum_{j=-\infty}^{\infty} |c_j| < \infty$). Conversely, the data receiver recovers the original data from y_t by a absolute summable filter $\{b_j\}$,

$$\hat{x}_t = \sum_{j=-\infty}^{\infty} b_{t-j} y_j.$$

In order to approximate x_t by $\hat{x}_t$ as accurate as possible, we determine the filter so that

$$\mathrm{E}\{(x_t - \hat{x}_t)^* (x_t - \hat{x}_t)\}$$

should be minimized. Let $b(\nu)$ and $c(\nu)$ be Fourier transform of $\{b_j\}$ and $\{c_j\}$ respectively, that is,

$$c(\nu) = \sum_{j=-\infty}^{\infty} c_j \exp(-2\pi i j \nu)$$

or conversely,

$$c_j = \int_{-1/2}^{1/2} c(\nu) \exp(2\pi i j \nu) d\nu. \tag{6.4}$$

(Similarly we define $b(\nu)$ and $\{b_j\}$.) According to the Theorem 9.3.1 of Brillinger (1981), the solution to this problem is to choose $c(\nu)$ so as to satisfy (6.3) and to take $b(\nu) = \overline{c(\nu)}$. To put it another way, the solution is given by $c(\nu) = e_1(\nu)$ and $b(\nu) = \overline{e_1(\nu)}$. Filter, or a set of weight coefficients, is given by Fourier inverse transform (6.4). The series obtained in this way, y_{t1} say, is to be called the first principal component series. In the same fashion y_{t2} through y_{tp} will be defined.

Statistical Inference

In fact $f_x(\nu)$ must be estimated from the sample path $x_1, \ldots, x_n$. To begin with, we calculate the periodogram matrix $I_n(\nu_j)$ from the data, and smooth them by an appropriate lag window to obtain $\hat{f}_x$,

$$\hat{f}_x(\nu_j) = \sum_{\ell=-(L_n-1)/2}^{(L_n-1)/2} h_{\ell,n} I_n(\nu_j + \ell/n).$$

Here the number of the terms of weight coefficient L_n is an odd number dependent on the sample size n, and is assumed to satisfy $L_n/n \to 0$ as $n \to \infty$. Spectral window is symmetric with respect to ℓ, each weight $h_{\ell,n}$ is positive and they sum up to 1. Now define the sum of squared weight coefficients by $\eta_n^{-2} = \sum_{\ell=-(L_n-1)/2}^{(L_n-1)/2} h_{\ell,n}^2$. Under some regularity conditions, the joint distribution of the eigenvalues and eigenvectors

$$\left(\eta_n \left[\hat{\lambda}_1(\nu_j) - \lambda_1(\nu_j) \right] / \lambda_1(\nu_j), \quad \eta_n \left[\hat{e}_1(\nu_j) - e_1(\nu_j) \right]' \right)'$$

converges to zero mean multivariate normal distribution as $n \to \infty$, after multiplied by the normalizing constant $\sqrt{n/L_n}$. The eigenvalues converge to the standard normal, and the eigenvalues and the eigenvectors are asymptotically independent. The asymptotic covariance matrix of the eigenvectors takes the form of

$$\Sigma_{e_1}(\nu_j) = \eta_n^{-2} \lambda_1(\nu_j) \sum_{\ell=2}^{p} \lambda_\ell(\nu_j) \{\lambda_1(\nu_j) - \lambda_\ell(\nu_j)\}^{-2} e_\ell(\nu_j) e_\ell(\nu_j)^*.$$

This results only tells about the distribution of the eigenvector corresponding to the first principal component, which turns out to depend on other eigenvalues and eigenvectors. It is also possible to give a confidence interval to the elements of $\hat{e}_1$. Setting $\hat{e}_1(\nu) = (\hat{e}_{11}(\nu), \ldots, \hat{e}_{1p}(\nu))'$ and let $s_j^2(\nu)$ be the j-th diagonal element of $\hat{\Sigma}_{e_1}(\nu)$, then

$$\frac{2|\hat{e}_{1j}(\nu) - e_{1j}(\nu)|^2}{s_j^2(\nu)} \quad (j = 1, \ldots, p)$$

asymptotically follows χ^2 distribution with degree of freedom 2.

6.2.2 Factor Analysis in Frequency Domain

Classical Factor Analysis

Same as in the previous section, we begin with the case where the repeated measurement from i.i.d. distribution is possible. Let x be a $p \times 1$ random vector with its mean zero and the covariance matrix Σ_x. Factor analysis assumes the observation x can be explained by a small number of unobserved common factors $z = (z_1, \ldots, z_q)'$. To put it another way, the following model is assumed,

$$x = \mathscr{B}z + \varepsilon.$$

Here $\mathscr{B}$ denotes a $p \times q$ factor loading matrix, z stands for a $q \times 1$ factor vector with $\mathrm{E}(z) = 0$ and $\mathrm{E}(zz') = I_q$. The error term ε is assumed to be independent of the factor, and its covariance matrix is diagonal given by $D = \mathrm{diag}(d_1^2, \ldots, d_p^2)$. The difference from multiple regression model is that z is not observed. This model of factor analysis can be restated in terms of the covariance structure of x as follows.

$$\Sigma_x = \mathscr{B}\mathscr{B}' + D. \tag{6.5}$$

That is, the covariance matrix of x is the sum of the symmetric nonnegative definite matrix with rank $q(\leq p)$ and the nonnegative definite diagonal matrix.

Given the sample $x_1, \ldots, x_n$, there are a couple of estimation methods to obtain the parameters of factor analysis model. The most feasible one is perhaps the principal component method. Let S_x be the sample covariance matrix, and $(\hat{\lambda}_i, \hat{e}_i)$ be the eigenvalues and eigenvectors calculated from S_x, where $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$. The principal component method throws away $\hat{\lambda}_{q+1}, \ldots, \hat{\lambda}_p$, and takes

$$\hat{\mathscr{B}} = (\sqrt{\hat{\lambda}_1}\hat{e}_1; \sqrt{\hat{\lambda}_2}\hat{e}_2; \cdots; \sqrt{\hat{\lambda}_q}\hat{e}_q).$$

Let $\hat{\delta}_j^2$ be the j-th diagonal element of $S_x - \hat{\mathscr{B}}\hat{\mathscr{B}}'$. Then $\hat{D}$ can be estimated by $\hat{D} = \mathrm{diag}(\hat{d}_1^2, \ldots, \hat{d}_p^2)$. If D is a scalar multiple of identity matrix, factor analysis reduces to principal component analysis, see Okamoto (1986, p. 24) for example. Though it is not always the case, this fact gives some justification to the principal component method.

The second method is to employ the method of maximum likelihood under the additional assumption that z and ε follow the multivariate normal distribution. Apart from constant, the main terms of log-likelihood function becomes

$$-2\ln L(\mathscr{B}, D) = n \ln |\Sigma_x| + \sum_{j=1}^{n} x_j \Sigma_x^{-1} x_j.$$

Note that $\mathscr{B}$ and D are incorporated into the log-likelihood through $\Sigma_x = \mathscr{B}\mathscr{B}' + D$. However, there remains some indeterminacy due to factor rotation, hence the maximum likelihood estimator is not uniquely derived. Usually the uniqueness of the solution is ensured by, for example, restricting $\mathscr{B}D\mathscr{B}'$ to a diagonal form. Maximization of the likelihood function is done by a numerical optimization procedure.

FA for Time Series

Along with the case of principal component analysis, let x_t be a p-variate stationary time series with its spectral density matrix $f_x(\nu)$. Remember this $f_x(\nu)$ plays the same role as the covariance matrix Σ_x does in the i.i.d. situation. Then by analogy with the ordinary factor analysis (6.5), we formulate the model for time series as follows,

$$f_x(\nu) = \mathscr{B}(\nu)\mathscr{B}(\nu)^* + D(\nu)$$

where $\mathscr{B}(\nu)$ is a $p \times q$ complex-valued matrix, $\mathrm{rank}(\mathscr{B}(\nu)) = q \leq p$, and $D(\nu)$ is a real-valued nonnegative definite diagonal matrix.

The simplest specification is that the multiple time series is essentially determined by a single latent factor. More specifically, we consider

$$x_{tj} = c_j s_{t-\tau_j} + \varepsilon_{tj}, \quad j = 1, \ldots, p \tag{6.6}$$

for $x_t = (x_{t1}, \ldots, x_{tp})$. Here $c_j \geq 0$ and τ_j respectively stand for the element-wise factor loading and the phase shift. We assume that the common factor s_t and the idiosyncratic factor $\varepsilon = (\varepsilon_{t1}, \ldots, \varepsilon_{tp})'$ are independent, and that the spectral density matrix of ε_t, denoted by $D_\varepsilon(\nu)$, is diagonal. Then by the DFT of x_{tj}

$$X_j(\nu) = n^{-1/2} \sum_{t=1}^{n} x_{tj} \exp(-2\pi i t \nu),$$

the frequency domain counterpart of the model (6.6) is derived as follows,

$$X_j(\nu) = a_j(\nu) X_s(\nu) + X_{\varepsilon_j}(\nu), \tag{6.7}$$

where $a_j(\nu) = c_j \exp(-2\pi i \tau_j \nu)$, and $X_s(\nu)$ and $X_{\varepsilon_j}(\nu)$ are the DFT of the common factor s_t and the idiosyncratic factor ε_{tj} respectively. Rewriting (6.7) element-wise, the complex version of a classical single factor analysis model is obtained as

$$\begin{pmatrix} X_1(\nu) \\ \vdots \\ X_p(\nu) \end{pmatrix} = \begin{pmatrix} a_1(\nu) \\ \vdots \\ a_p(\nu) \end{pmatrix} X_s(\nu) + \begin{pmatrix} X_{\varepsilon_1}(\nu) \\ \vdots \\ X_{\varepsilon_p}(\nu) \end{pmatrix}.$$

To put it more concisely, we have

$$X(\nu) = a(\nu)X_s(\nu) + X_\varepsilon(\nu). \tag{6.8}$$

From (6.8) we have the following relationship on the spectral density of x_t,

$$f_x(\nu) = b(\nu)b(\nu)^* + D_\varepsilon(\nu),$$

where $b(\nu)$ is a $p \times 1$ complex valued vector that satisfies $b(\nu)b(\nu)^* = a(\nu)f_s(\nu)a(\nu)^*$.

Statistical Inference

Let us extend the single factor model presented in the preceding section as follows;

$$x_t = \sum_{j=-\infty}^{\infty} \Lambda_j s_{t-j} + \varepsilon_t, \tag{6.9}$$

where $\{\Lambda_j\}$ is a $p \times q$ real-valued factor loading matrix, s_t is a q-variate stationary process which corresponds to the common factors. All the factors are assumed to be mutually independent, that is, the $q \times q$ spectral matrix of s_t becomes a diagonal matrix $f_s(\nu) = \mathrm{diag}(f_{s1}(\nu), \ldots, f_{sq}(\nu))$. On the other hand, ε_t is assumed to be white noise and be independent of s_t. Under these assumptions, the $p \times p$ spectral matrix of ε_t becomes diagonal such that $D_\varepsilon(\nu) = \mathrm{diag}(f_{\varepsilon 1}(\nu), \ldots, f_{\varepsilon q}(\nu))$. In addition, the matrix norm of Λ_j is assumed to be finite. Then Λ_j is well-defined by

$$\Lambda(\nu) = \sum_{t=-\infty}^{\infty} \Lambda_t \exp(-2\pi i t \nu), \tag{6.10}$$

and in addition define $\mathscr{B}(\nu)$ by $\mathscr{B}(\nu) = \Lambda(\nu)f_s^{1/2}(\nu)$, then the spectral density matrix of x_t satisfies the following relationship,

$$f_x(\nu) = \Lambda(\nu)f_s(\nu)\Lambda(\nu)^* + D_\varepsilon(\nu) = \mathscr{B}(\nu)\mathscr{B}(\nu^*) + D_\varepsilon(\nu). \tag{6.11}$$

Putting $f_s(\nu) = I_q$ for all ν for the sake of model identifiability, we obtain $\mathscr{B}(\nu) = \Lambda(\nu)$. Nevertheless, the arbitrariness due to factor rotation still remains unsolved.

To estimate the unknown parameters in model (6.11), the principal component method can be employed similarly as in the previous section. Let $\hat{f}_x(\nu)$ be the estimates of $f_x(\nu)$, and let $(\hat{\lambda}_j(\nu), \hat{e}_j(\nu))$, $j = 1, \ldots, p$, be the set of eigenvalues and eigenvectors of $\hat{f}_x(\nu)$, in descending order of $\hat{\lambda}_j$. Then, same as in the ordinary principal component analysis, we can estimate the matrix $\mathscr{B}$ by

$$\hat{\mathscr{B}}(\nu) = (\sqrt{\hat{\lambda}_1(\nu)}\hat{e}_1(\nu); \sqrt{\hat{\lambda}_2(\nu)}\hat{e}_2(\nu); \cdots; \sqrt{\hat{\lambda}_q(\nu)}\hat{e}_q(\nu)).$$

Let $\hat{f}_{\varepsilon j}(\nu)$ be the j-th diagonal element of $\hat{f}_x(\nu) - \hat{\mathscr{B}}(\nu)\hat{\mathscr{B}}(\nu)^*$. Then the spectral density matrix of he idiosyncratic factor can be estimated by $\hat{D}_\varepsilon(\nu) = \mathrm{diag}(\hat{f}_{\varepsilon 1}(\nu), \ldots, \hat{f}_{\varepsilon p}(\nu))$.

We can employ the MLE instead of the principal component method. Let $X(\nu_j)$ be the DFT of $x_1, \ldots, x_p$ evaluated at frequency $\nu_j = j/n$, and let $X_s(\nu_j)$ and $X_\varepsilon(\nu_j)$ be the DFT of the common factor and the idiosyncratic factor respectively. Then under appropriate conditions, we have

$$X(\nu_j + \ell/n) \approx \Lambda(\nu_j) X_s(\nu_j + \ell/n) + X_\varepsilon(\nu_j + \ell/n) \tag{6.12}$$

for $\ell = 0, \pm 1, \pm 2, \ldots, \pm(L_n - 1)/2$. Note that $\Lambda(\nu_j)$ is given by (6.10). The approximation in (6.12) is asymptotically valid as $n \to \infty$ in the neighborhood of ν_j, thus the sample analogue of the frequency domain counterpart of (6.9) can be constructed at least locally.

Because $\{X(\nu_j + \ell/n); \ell = 0, \pm 1, \pm 2, \ldots, \pm(L_n - 1)/2\}$ asymptotically and independently follows a multivariate complex normal distribution with its mean zero and the covariance $f_x(\nu_j)$ under mild conditions, -2 times the approximate log-likelihood can be given as

$$-2 \ln L(\mathscr{B}(\nu_j), D_\varepsilon(\nu_j)) = n \ln |f_x(\nu_j)| + \sum_{\ell=-(L_n-1)/2}^{(L_n-1)/2} X^*(\nu_j + \ell/n) f_x^{-1}(\nu_j) X(\nu_j + \ell/n).$$

Unknown parameters are implicitly built in the log-likelihood through $f_x(\nu_j) = \mathscr{B}(\nu_j)\mathscr{B}(\nu_j)^* + D_\varepsilon(\nu_j)$. We call attention about the MLE proposed here. In the first place, because the distribution of $X(\nu_j + \ell/n)$ is *asymptotically* normal with its covariance $f_x(\nu_j)$, $X(\nu_j + \ell/n)$ for finite n cannot be regarded as the exact independent samples from a normal distribution with its covariance $f_x(\nu_j)$. Hence it should be noted that the MLE here is not identical to the MLE of ordinary factor analysis that is based on the exact, small sample distributional assumption. Secondary, the asymptotic normality of DFT used here is the property not at a fixed Fourier frequency but around a fixed Fourier frequency.

6.2.3 Applications

This section introduces some interesting applications of the principal component/factor analysis of time series in frequency domain. At first we take a look at the numerical examples found in Geweke (1977) and Brillinger (1981) that are the pioneering works in this field. The example in Brillinger's book might not be very suggestive in implications for a specific science, Geweke's example, business cycle index, has been a hot topic and has been studies in the context of dynamic factor analysis of time series. On the other hand, principal component/factor analysis are convenient tools if data analysts are interested in a certain frequency prior to the modeling. As such examples, we review Shumway and Stoffer (2000), Young and Pedregal (1999).

Geweke (1977) is frequently referred work in the field of frequency domain factor analysis of time series because of its theoretical contribution. In this paper, Geweke fits a single-factor model to the layoff rate, manhours, quit rate, and industrial production, and discuss the appropriateness of the model. The latent factor here can be interpreted as 'demand'. Generally in economic time series, nonstationary component such as trend and seasonality is dominant, hence a care must be taken on the use of the theory and method based on stationarity assumption. To avoid this problem, Geweke (1977) carefully removed the frequency bands that obviously correspond to nonstationary component, and subsequently performed factor analysis. As an another application of the framework of Geweke (1977), we mention Geweke and Singleton (1981).

In Chapter 9 of Brillinger (1981), we find an example of frequency domain factor analysis on the monthly average temperature time series of 14 cities, one of which is New Haven and the rest are all European cities. The sample period is 21 years, or 252 time points. Plotting spectral densities of principal component series generally shows its peak at zero frequency and a slow decay toward higher frequencies. The level of spectral density is highest in the first principal component, and the level gradually comes down from then on. The results of analysis suggest that the first principal component can be interpreted as the average temperature of 13 European cities, and the second principal component seems to stand for Hew Haven. The difference in the level of spectrum enables us to detect alien data. Though there was no specific frequency that attracts our interest prior to the analysis, this example suggests the possibility of clustering of multivariate time series.

Examples of principal component/factor analysis in frequency domain are not so rich in numbers as in time domain. One of the reason may be partly accounted for by the fact that it cannot be an attractive tool unless the importance of certain period or frequency is obvious in a specific problem. For example, if we are interested in the existence or non-existence of a common trend in multiple time series, then the specific frequency of interest is zero. If we want to investigate the reaction of brain to the stimulation from outside, then naturally we are interested in the frequency that is determined by the repetition cycle of the experiments. Alternatively, beginning with questioning the existence of common factors, subsequently we may be interested in the dominant frequency component of the common factor which was not clear prior to the analysis. Such a periodicity of interest is very clear in the following two examples.

Shumway and Stoffer (2000) analyzed fMRI(functional magnetic resonance imaging) data by principal component/factor analysis in frequency domain. In the experiment five experimen-

tal subjects had his/her hand periodically brushed. The stimulus was applied for 32 seconds and then stopped for 32 seconds. The sampling rate was one observation every two seconds for 256 seconds. While this experiment is repeated, fMRI data was collected at various locations in the brain like cortex, thalamus and cerebellum. The data is averaged over the subjects, hence the data is multivariate with respect to the locations of the brain. During the 256 seconds of experiment, there are four cycles of stimulus and rest, so the frequency of our interest is $\nu = 4/128$. As is anticipated, the spectral density of the first principal component series has its peak at $\nu = 4/128$. After constructing the confidence interval for the principal component vector at $\nu = 4/128$, they conclude that one of the eight measurement locations did not response to the brush stimulus.

Young and Pedregal (1999) provides a very interesting analysis on U.S. macroeconomic time series; unemployment rate, gross domestic product, consumption, government expenditure, and private sector investment. All the data is quarterly, seasonally adjusted, and transformed into growth rate. The aim of research is to investigate the influence from governmental expenditure and private sector investment to unemployment rate. Spectral analysis of individual series reveals the existence of 4-years or 8-years cycle. By plotting the first and second principal component of $\hat{f}_x(\nu)$, taking frequencies ν as the horizontal axis, it turns out that the first principal component has its peak at 4-years cycle, while the second around 8-years cycle. The first principal component vector at the 4-years-cycle frequency has contribution on all the series except government expenditure. Conversely, the second principal component vector at the 8-years-cycle frequency has most of its influence on government expenditure only. Together with the value of coherency, they suggest that government expenditure is rather exogenous to other economic time series.

Finally we remark that a special attention has been paid in economics to judge if a multiple time series has a common trend component or not. To put it another way, research interest is concentrated on zero frequency, which seems to be suitable for the framework presented here. In fact, most of the procedures proposed are built in time domain approach. Thus the problem of estimating common trend will be deferred in the next section, and we just mention Phillips and Ouliaris (1988) here.

6.3 Factor Analysis in Time Domain

The central idea of the procedures presented in the previous section is that the DFT of time series ensures the asymptotic independence of the periodograms, which subsequently makes it

possible to apply the classical framework of principal component/factor analysis. This section deals with a different and more direct approach where we give explicit models to the latent factor process. Dynamic factor model is a generic term to refer to such models. The aim of this section is to introduce several methods related to dynamic factor model, and state their estimation, applications, and their mutual relationship.

Dynamic factor models has been developed side-by-side with their applications to the problems in psychology. Along with such a historical context, we begin with extending the original factor analysis model from the viewpoint of covariance structure analysis in section 6.3.1. On the other hand, from the viewpoint of time series modeling, dynamic factor models are closely related to the unobservable components model and the structural time series model which has been intensively developed since the late of 1970's. Actually, dynamic factor models can be interpreted as a special case of them, see section 6.3.2. The advantage of this approach is that we can assume a wider class of models than the class of polynomial models, even for a nonstationary factor case. At the same time, it is possible to build the dynamic factor model based on multivariate ARIMA model which will be reviewed in section 6.3.3. Section 6.3.4 addresses reduced rank regression model and error correction model, where an explicit model is not always given.

6.3.1 Dynamic Factor and Covariance Structure Modeling

Historical Overview in Psychometrics

A formal application of factor analysis to the multivariate longitudinal data observed in psychological experiment has been criticized since Holtzman (1962), Anderson (1963). One of the important criticism is as follows; the latent factor will affect the original time series not only coincidentally but with some time lags. On the other hand, it is difficult to apply the method presented in section 6.2.1 in psychology because the length of time series is too short to validate the use of discrete Fourier transform. Length of economic time series is also limited, though not severe as in experimental data in psychology. The data analyzed in Brillinger (1981) is climate data, which is rich in sample size compared to the most cases in social sciences. Because the estimate of spectral density $\hat{f}_x(\nu)$ is defined only in the neighborhood of ν, insufficient sample size leads to the limitation of frequencies to be considered, and it becomes difficult to obtain a good estimate of $\hat{f}_x(\nu)$ because the smoothing by lag window will be difficult. On the bandwidth selection, there seems to be no general rule because it depends on if the researcher is interested in all the frequency band or in rather limited band. For example Tukey (1978, p.26–27) gives some suggestions.

Covariance Structure Modeling

It is Molenaar (1985) who proposed the framework of dynamic factor model taking the above issues into account. His idea looks direct. To begin with, we write the ordinary factor analysis model with n-variate p-factors as

$$x(t) = \mathscr{B}z(t) + \varepsilon(t), \quad t = 0, 1, \ldots, n. \tag{6.13}$$

In the context of factor analysis, $z(t)$, $\mathscr{B}$ and $\varepsilon(t)$ are referred to as common factor, factor loading and idiosyncratic factor respectively. Covariance of the common factors at different time is assumed to be of the form,

$$\mathrm{cov}(z(t), z(t-u)) = \Xi(u).$$

It is also assumed that there is no correlation among idiosyncratic errors, and the autocorrelation of single idiosyncratic error is the function of time difference (u) only, namely,

$$\mathrm{cov}(\varepsilon(t), \varepsilon(t-u)) = D(u) = \mathrm{diag}(d_1(u), \ldots, d_p(u)).$$

Now we extend (6.13) so that it includes higher lags up to s,

$$x(t) = \sum_{u=0}^{s} \mathscr{B}(u)z(t-u) + \varepsilon(t), \quad t = 0, 1, \ldots, n.$$

Then, the autocovariance function of $x(t)$ can be expressed as

$$\Sigma_x(u) = \sum_{v=0}^{s}\sum_{w=0}^{s} \mathscr{B}(v)\Xi(u+w-v)\mathscr{B}(w)^T + D(u), \quad u = 0, \pm 1, \ldots. \tag{6.14}$$

Ordinary factor analysis deals with only the case of $u = 0$.

Giving vector-matrix notation to the model (6.14), it can be reduced to a simultaneous equation system, which shows that (6.14) is a natural extension of ordinary factor analysis. Let

$$\begin{aligned} x' &= (x(t)', \ldots, x(t-a)') \\ z' &= (z(t)', \ldots, z(t-a-s)') \\ \varepsilon' &= (\varepsilon(t)', \ldots, \varepsilon(t-a)'), \end{aligned}$$

where $a \geq s$. Furthermore, let

$$\mathscr{B} = \begin{pmatrix} \mathscr{B}(0) & \mathscr{B}(1) & \cdots & \mathscr{B}(s) & 0 & \cdots & 0 & 0 \\ 0 & \mathscr{B}(0) & \cdots & \mathscr{B}(s-1) & \mathscr{B}(s) & \cdots & 0 & 0 \\ \vdots & & \ddots & & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \mathscr{B}(0) & \cdots & \cdots & \mathscr{B}(s-1) & \mathscr{B}(s) \end{pmatrix}.$$

Then the dynamic factor model can be concisely expressed as

$$x = \mathscr{B}z + \varepsilon.$$

From this equation we can extract the covariance structure

$$\Sigma_x = \mathscr{B}\Xi\mathscr{B}' + D,$$

where

$$\Sigma_x = \begin{pmatrix} \Sigma_x(0) & & & \\ \Sigma_x(1) & \Sigma_x(0) & & \\ \vdots & & \ddots & \\ \Sigma_x(a) & \Sigma_x(a-1) & \cdots & \Sigma_x(0) \end{pmatrix},$$

and

$$\begin{aligned} \Xi &= \{\Xi(i-j); i,j = 1,\ldots,a+s+1\} \\ D &= \{D(i-j); i,j = 1,\ldots,a+1.\} \end{aligned}$$

Σ_x, to be calculated from multivariate time series data, corresponds to the sample covariance matrix in the usual multivariate analysis. (If $s = 0$, it reduces to the ordinary factor analysis.) The method proposed by Molenaar (1985) is to reformulate a dynamic factor model to a structural equation model (Jöreskog (1978), Bollen (1989), Jöreskog and Sörbom (1998)), and to estimate the parameters by MLE under the assumption that Σ_x follows Wishart distribution.

Hence what is done in Molenaar (1985) is not the maximum likelihood estimation but the pseudo maximum likelihood estimation, because the covariance matrix in dynamic factor model does not exactly follow the Wishart distribution. Here the MLE is performed as if the Wishart distribution were a true data generating distribution. In that sense, MLE here is the pseudo MLE. This reminds us of the term 'Gaussian MLE' in time series analysis which assumes Gaussian innovation though it may not be a correct assumption. For pseudo maximum likelihood estimators, see Gourieroux et al. (1984) for example.

So far the theoretical framework is constructed for stationary time series. Molenaar et al. (1992) proposed a method to cope with nonstationary case where the deterministic function of time is introduced in the mean structure. Precisely, let $\bar{t}$, $S(t)$ be the mean and variance of $t \in \{1,2,\ldots,n\}$, and prepare the normalized trend function $\tau(t) = (t-\bar{t})/S(t)$. Then the latent factor can be expressed as the sum of linear trend and stationary process $z(t)$ as $f(t) = \gamma\tau(t) + z(t)$. Suppose q-latent vectors are assumed in the analysis. Then a $q \times 1$ vector γ can yield factor-wise different slope. This requires only minor modifications to the procedure of

stationary case because we only have to replace (6.14) by

$$\Sigma_x(u) = \sum_{v=0}^{s}\sum_{w=0}^{s} \mathscr{B}(v)(\gamma\gamma' + \delta(u+w-v)I_q)\mathscr{B}(w)^T + D(u), \quad u = 0, \pm 1, \ldots,$$

where δ stands for Kronecker's delta.

6.3.2 Dynamic Factor and Structural Time Series Model

The way of modeling trend largely depends on the perspective inherent to a specific field of science. For example, in psychological experimental data, such simple structures that can be described by polynomial is assumed. Together with it is the restriction of data length, more complex modeling than polynomials are rarely employed in modeling growth curve or learning curve. On the contrary in economic time series, deterministic functions of time are not enough to describe the actual trend because it seems to repeat up and down, or even abrupt change within several hundreds of data points. In such cases, global trend described by deterministic function is not appropriate but *locally constant* or *locally linear* models are sometimes relevant. A trend model represented by such local models is often referred to stochastic trend. This section explains how the common stochastic trend is treated in the framework of structural time series model.

Structural Time Series Model

let x_t $(t = 1, \ldots, n)$ be a observed p-variate time series of which elements are possibly non-stationary. Let $q \times 1$ vector z_t be the unobserved common trend factor. Suppose the data is generated by the following equations,

$$\begin{aligned} x_t &= \mathscr{B}z_t + z_0 + \varepsilon_t, \quad \mathrm{var}(\varepsilon_t) = \Sigma_\varepsilon \\ z_t &= z_{t-1} + \beta + \eta_t, \quad \mathrm{var}(\eta_t) = \Sigma_\eta, \end{aligned} \tag{6.15}$$

where $\mathscr{B}$ is factor loading matrix $(p \times q)$ and the drift term β is time invariant. From the first row to the q-row of z_0 are restricted to zero and the rest of the elements are freely estimated. We call this specification $\bar{z}$. Same as ordinary factor analysis, the above model lacks identifiability with respect to the rotation of factor. Therefore the model is usually estimated under the assumptions, for example, that $\Sigma_\eta = I$ and the upper triangular elements of $\mathscr{B}$ are zero. Afterwards the factors will be rotated so that the interpretations of the factors should be easy. As long as the identifiability problem is cleared, likelihood calculation, parameter estimation, state estimation are routinely processed by the standard arguments of structural time series model. We do not go further about the state space representation and the Kalman filter, which have been already explained in Chapter 3.

Connection to Cointegration

Suppose all the elements of observed time series x_t become weakly stationary after the first difference, or for each elements $x_{i,t}(i = 1, \ldots, p)$ of x_t, $\Delta x_{i,t} = x_{i,t} - x_{i,t-1}$ follows some weakly stationary process. Alternatively, as x_t is defined as the accumulation of a stationary process, it is a nonstationary process of which mean is stochastically evolves. If a linear combination $\gamma' x_t$ is stationary, the multivariate time series x_t is said to be cointegrated, and γ is usually referred to as the cointegration vector. Intuitively speaking, because all the elements of the multivariate process x_t share the common trend component, their linear combination results in stationary residuals.

In the structural time series model defined right before, there exists $p-q$ linear combinations of level time series x_t that reduce to stationary processes, or $p-q$ cointegration relationships exist. Here we note that the cointegration vectors are $p-q$ vectors in the null space of factor loading matrix, or the rows of A such that $A\mathscr{B} = 0$. Consequently the observational equation (6.15) becomes

$$Ax_t = Az_0 + A\varepsilon_t, \tag{6.16}$$

hence it is seen that Ax_t is $(p-q) \times 1$ vector stationary process. In this example, it has collapsed to be a simpler structure, the multivariate white noise process with its mean Az_0 and the covariance $A\Sigma_\varepsilon A'$.

Let us take a look at a simple case ($p = 2$, $q = 1$). Then the model can be written as

$$\begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = \begin{pmatrix} 1 \\ \theta \end{pmatrix} z_t + \begin{pmatrix} 0 \\ \bar{z} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} \tag{6.17}$$

$$z_t = z_{t-1} + \beta + \eta_t,$$

where $\mathrm{var}(\eta_t) = \sigma_\eta^2$. Cointegration vectors are normalized so as to satisfy $A = (1, \alpha)$ otherwise only the directions are estimable. Because we must have $1 + \alpha\theta = 0$, we obtain $\alpha = -1/\theta$. Multiply (6.17) by the cointegration vector, we obtain

$$x_{1t} = (1/\theta)x_{2t} + (-1/\theta)\bar{z} + \varepsilon_t,$$

where $\varepsilon_t = \varepsilon_{1t} - \varepsilon_{2t}/\theta$. As mentioned above, a dynamic factor model for nonstationary time series has restrictions among the level variables. Due to this fact, a cointegration model can be derived from a dynamic factor model.

A merit of employing the structural time series model to estimate the dynamic factor analysis may be found on the ease of modeling even if the factors are nonstationary, or on the convenience that the trajectories of latent factors and their distributions can be automatically

brought as the by-products of model fitting. At the same time, the structural time series approach is not very much suited for the construction of rigorous procedure for the determination of number of factors. Of course, it may be possible to define a practical rule by the use of information criteria, but the models described in section 6.3.4 will offer more convenient framework to determine the number of latent factors.

6.3.3 Dynamic Factor and Multivariate ARMA Model

Instead of the structural time series model used in section 6.3.2, the latent factor process can be described by ARMA process. If the dimension of the factor process is more than two, it is naturally described by multivariate ARMA process. This section reviews Pe a and Box (1987) and related works.

Canonical Component Model

Let X_t be $p \times 1$ vector-valued stationary time series, and the demeaned process $x_t = X_t - \mu_X$ (where $\mu_X = \mathrm{E}(X_t)$) will be regarded as observation afterwards. Then suppose the following model,

$$x_t = \mathscr{B} z_t + \varepsilon_t,$$

where z_t is unobservable factor of dimension $q \times 1$ ($q \leq p$), $\mathscr{B}$ is a $p \times q$ matrix of factor loadings with its rank$\mathscr{B} = r$, and ε_t is a p dimension white noise process with its covariance matrix Σ_ε of full rank. We restrict $\mathscr{B}$ to be an orthogonal matrix such that $\mathscr{B}'\mathscr{B} = I_q$. Furthermore, we assume that the unobserved factor z_t follows an multivariate ARMA process

$$\phi_z(L) z_t = \theta_z(L) a_t,$$

where $\phi_z(L)$ and $\theta_z(L)$ are the polynomial of lag operator L of order p_z and q_z respectively,

$$\begin{aligned} \phi_z(L) &= I_q - \phi_z(1)L - \cdots - \phi_z(p_z)L^{p_z} \\ \theta_z(L) &= I_q - \theta_z(1)L - \cdots - \theta_z(q_z)L^{q_z}. \end{aligned}$$

Alternatively it can be written as $z_t \sim \mathrm{VARMA}(p_z, q_z)$. here it is assumed that all the roots of $|\phi_z(L)| = 0$ and $|\theta_z(L)| = 0$ should lie outside the unit circle, which ensures the stationarity and invertibility of z_t. Also assumed is that a_t is independent normal white noise sequence with its mean zero and the positive definite covariance matrix Σ_a. To simplify the argument, let us begin with the case where the $r \times r$ matrices $\phi(\cdot)$ and $\theta(\cdot)$ are both diagonal. This is called the uncoupled-factors model.

The results of Pe a and Box (1987) can be summarized as follows. Although the model for x_t apparently has too much flexibility, the parameters to be estimated have to meet a lot of restrictions. As a result, the identification of the model is easy. Especially, (a) all the AR coefficient matrices ($\phi_x(h)$) shares q-eigenvectors that span the subspace $S(\mathscr{B})$ generated by $\mathscr{B}$, (b) the rank of all the coefficient matrices $\psi_x(h)$ of pure MA representation is q, and their column vector belongs to $S(\mathscr{B})$. In other words, (a) tells that the AR coefficient matrices of x_t are tightly constrained, and (b) shows that the MA coefficient matrices are anchored to the AR coefficient matrices.

The above results are just to characterize the uncoupled-factor models. In actual model identification procedure, it is recommended to use covariance matrix according to the following argument. Let $\Gamma_x(h) = \mathrm{E}(x_{t-h}x_t')$ be the covariance matrix of the observation x_t, and let $\Gamma_z(h) = \mathrm{E}(z_{t-h}z_t')$ be the covariance matrix of the unobservable factor z_t. Then we have

$$\begin{aligned} \Gamma_x(0) &= \mathscr{B}\Gamma_z(0)\mathscr{B}' + \Sigma_\varepsilon \\ \Gamma_x(h) &= \mathscr{B}\Gamma_z(h)\mathscr{B}', \quad h \geq 1 \end{aligned}$$

and if $h \geq 1$ the rank of $\Gamma_x(h)$ is equal to the number of common factor, q. If all the factors are mutually independent at all lags and Σ_a is diagonal, then $\Gamma_z(h)$ is also diagonal. Subsequently $\Gamma_x(h)$ is symmetric for $h \geq 1$, and the column vectors of $\mathscr{B}$ are the eigenvectors of $\Gamma_x(h)$ corresponding to the eigenvalues $\gamma_i(h)$. (At the same time, $\{\gamma_i(h)\}$ are the diagonal elements of $\Gamma_z(h)$.) Therefore the actual data analysis starts with estimating $\Gamma_x(1)$, $\Gamma_x(2)$, and so on from the observation, and calculating their eigenvalues and eigenvectors afterwards.

After guessing the number of latent factors, the next thing to do is to rotate the factors so that they can be easily interpreted. For this purpose, the following transformation will be performed to isolate the influence of factors. Let $\mathscr{B}^-$ be a generalized inverse of the loading matrix $\mathscr{B}$, then we have $z_t = \mathscr{B}^- x_t - \mathscr{B}^- \varepsilon_t$. Let B be a $(p-q) \times p$ matrix by which a matrix M to transform data is defined by

$$M = \begin{pmatrix} \mathscr{B}^- \\ B \end{pmatrix}.$$

Then the transformed observation vector $\tilde{x}_t$ can be expressed as

$$\tilde{x}_t = Mx_t = \begin{pmatrix} \mathscr{B}^- x_t \\ Bx_t \end{pmatrix} = \begin{pmatrix} z_t + \mathscr{B}^- \varepsilon_t \\ B\mathscr{B}z_t + B\varepsilon_t \end{pmatrix} = \begin{pmatrix} \tilde{x}_{1t} \\ \tilde{x}_{2t} \end{pmatrix}.$$

To isolate the influence of the factors, we only have to choose so as to satisfy $B\mathscr{B} = O$. (As the columns of B, choose $p-q$ eigenvectors that correspond to zero eigenvalues of $\mathscr{B}\mathscr{B}'$.) There still remains some arbitrariness on the choice of $\mathscr{B}^-$. However, if it is justified to assume $\Sigma_\varepsilon = I$, we can make $\tilde{x}_{1t}$ and $\tilde{x}_{2t}$ mutually uncorrelated. That is why Moore-Penrose generalized inverse

($\mathscr{B}^- = (\mathscr{B}'\mathscr{B})^{-1}\mathscr{B}'$) is preferred. This transformation is called the canonical transformation, see Box and Tiao (1977).

The assumption that the factor processes are mutually uncorrelated is too unrealistic to be accepted in real data analysis. When the correlations of the latent factor processes are taken into account, such model is called a coupled-factors model in Pe a and Box (1987). In coupled-factors models, the eigenvectors of $\Gamma_x(h)$ and $\phi_x(h)$ are no longer the column vector of $\mathscr{B}$. Diagonalization of $\Gamma_z(h)$ and $\phi_z(h)$ is required to reduce all the arguments to the uncoupled-factors model. However, it can be shown that the canonical transformation can be constructed without the knowledge of the matrices to diagonalize $\Gamma_z(h)$ and $\phi_z(h)$. The determination of the number of factors can be conducted in a similar manner as uncoupled-factors model making the ranks of $\Gamma_x(h)$ ($h \geq 1$) and $\phi_x(1) - \theta_x(1)$ of clue to go on. If it works or not will finally depend on the correlation of factors. The example found in Pe a and Box (1987) (analysis of monthly wheat prices observed in various states in Spain) gives an impression that there is no difficulty in the model identification and extraction of the latent factor processes.

6.3.4 Latent Factor without Explicit Model

So far we reviewed the covariance structure analysis approach, the structural time series approach, and the canonical transformation approach. These procedures share one commonality that they give more or less explicit model to the unobservable latent factors. This section overviews the different approaches for dynamic factor models where explicit models are not necessarily considered. One of them is reduced rank autoregressive models, and the other is error correction models. For the presentation purpose, we start our argument with a vector autoregressive model. Note that a reduced regression model can be used in a usual regression model, and an error correction model is often used in a single equation approach though the concept of cointegration is essentially a multivariate one.

Reduced Rank Autoregressive Model

Let x_t be a $p \times 1$ dimensional vector of observation at time t. The following p-variate autoregressive model

$$x_t = \sum_{j=1}^{s} \Phi_j x_{t-j} + \varepsilon_t \tag{6.18}$$

to which x_t obeys is called the reduced rank (auto)regression model if the lag operator $I - \Phi(L) = \sum_j \Phi_j L^j$ is expressed as

$$x_t = A(L)B(L)x_t + \varepsilon_t = \sum_{u=1}^{s_1}\sum_{v=1}^{s_2} A_u B_v x_{t-u-v} + \varepsilon_t$$

by a $p \times q$ ($q \leq p$) dimensional operator $A(L) = A_1 L + \cdots + A_{s_1} L^{s_1}$ and a $q \times p$ dimensional operator $B(L) = B_1 L + \cdots + B_{s_2} L^{s_2}$, where $s_1 + s_2 = s$. Reinsel (1983) discussed the parameter estimation and model selection for the case of $s_2 = 0$, namely for

$$x_t = \sum_{j=1}^{s} A_j B_0 x_{t-j} + \varepsilon_t.$$

The above model is often referred to as *index model*, probably because $B_0 x_{t-j}$ is interpreted as an index which summarizes many variables. On the other hand, Velu et al. (1986) discussed the parameter estimation and its asymptotics for the case of $s_1 = 0$,

$$x_t = A_0 \sum_{j=1}^{s} B_j x_{t-j} + \varepsilon_t.$$

They also mention the relation between the canonical transformations in Box and Tiao (1977) and the reduced rank regression models. As Velu et al. (1986) mentions, these two approaches are almost same. The canonical transformation dig the larger eigenvalues of a covariance matrix to find predictable q-variables, while the reduced rank regression models pay attention to the smaller eigenvalues to drop unnecessary $p - q$ components. It can be said that Box and Tiao (1977) proposed a reasonable procedure, but it is until Velu et al. (1986) that the statistical inference of Box-Tiao's scheme is rigorously established. Furthermore, Ahn and Reinsel (1988) developed the estimation theory for the nested reduced rank regression models. The reduced rank (auto)regression model is said to be nested if Φ_j in (6.18) satisfies

$$\text{rank}(\Phi_j) = q_j \geq \text{rank}(\Phi_{j+1}) = q_{j+1}, \quad j = 1, \ldots, s-1.$$

It is possible to give various interpretations for the nested models. For example, if time lag is large enough, past observation from some channel does not have significant information on present state. Similar physical interpretation would be possible even if the time index is replaced by the spatial index.

Error Correction Model

In multivariate vector autoregression (6.18), we assume that all the elements of observation vector x_t are stationary after the first difference, or for each elements $x_{i,t} (i = 1, \ldots, p)$ of x_t, $\Delta x_{i,t} = x_{i,t} - x_{i,t-1}$ follows some weakly stationary process. On the other hand, regardless of the stationarity of x_t, it is possible to give (6.18) an error correction representation as follows (Engle and Granger, 1987),

$$\Delta x_t \quad = \quad \Pi x_{t-1} + \sum_{j=1}^{s-1} \Phi_j^* \Delta x_{t-j} + \varepsilon_t$$

$$\Phi_j^* = -\sum_{i=j+1}^{s} \Phi_j, \quad \Pi = -\left(I - \sum_{j=1}^{p} \Phi_j\right).$$

This model is commonly interpreted as follows. Changes in observation vector is accounted for not only by past changes but by the linear combination of the level time series Πx_{t-1}. If $\Pi = O$, this multivariate process has p independent nonstationary components. Else if Π is of full rank, x_t must be stationary which contradicts to the assumption we made. Therefore we are not interested in these extreme cases. The important cases are when the rank of Π is greater than zero and less than the dimension of observation vector. Then Πx_{t-1} reduces to a stationary process, which is interpreted as a long-run relationship held among economic variables. In fact, Πx_{t-1} cannot be identically zero (or constant) but randomly fluctuate. These errors are the deviations from the long-run relationship, and these errors affect the present changes in the feedback loop. That is why this model is referred to an error correction model.

As is explained, in an error correction model, we are not interested in the latent factors but in the dimensions of factors and/or the existence of linear restrictions. The most popular method in econometrics to determine the number of latent common trends (or the rank of cointegration) is the likelihood ratio procedure proposed by Johansen (1988). There are a lot of empirical works, review papers and text books on Johansen's LR procedure. See Johansen (1995) and Chapter 30 of Hamilton (1994) among others. There are other works to determine the number of common trends, but we only mention Stock and Watson (1988) here.

From the frequentist perspective, if the observed vector time series x_t is nonstationary, it is an usual practice to use an error correction representation. From the Bayesian perspective, however, it is possible to discuss the reduced rank regression models directly for the level (nonstationary) time series. As such a work, we mention Geweke (1996).

6.4 PCA in Time Domain and Two Step Procedure

This chapter is so to say a two dimensional plane, which is divided into four orthants by 'principal component analysis – factor analysis' coordinate and by 'frequency domain – time domain' coordinate. In this section we discuss the final fourth orthant, principal component analysis in time domain.

In the standard settings of multivariate analysis, the data is assumed to come from random sampling from a certain multivariate distribution. In this situation, the statistical inference does not depend on the order in which the observations come, because the sample covariance matrix is essentially invariant with respect to the permutation about the subjects or items. However,

this does not necessarily hold for time series. In time series and spatial data, the correlation structure caused by time space adjacency is essential information. If the order of time series were changed, statistical analysis yields completely different results.

Conversely, if we are just interested in the contemporaneous correlation among the elements of vector time series at a fixed time, the information contained in serial correlations would be meaningless. If we accept this attitude, it means that we neglect the time series property of multivariate time series and pretend as if a couple of correlated data happened to be observed as time evolution.

There are three important applications done in this spirit. The first one is the construction of index like a business cycle. Theil (1960), Kloek and de Wit (1961), Kariya (1986) discussed this problem. The second application is nonparametric smoothing such as trend estimating via principal component analysis like Ahamad (1967). Ahamad (1967) gives us a good practical lesson. Simple minded application of principal component can yield smooth trend, but sometimes we find it an obvious factor that should have been considered prior to the analysis. Ahamad (1967) applied principal component analysis on the annual occurrence of 18 crimes in England and Wales observed from 1950 to 1963. Unless we offset the number of crime occurrence by the population prior to the analysis, the first principal component is almost the same curve as the growth rate of the teenager population there. Thirdly, like stock return time series, if the contemporaneous correlations at a time are very tight, and the auto- and cross-correlations are relatively weak so that they decay (or at least are supposed to decay) quite rapidly, there will be some reasons to focus on the contemporaneous correlations only. See Kariya (1993).

As is mentioned, the use of principal component analysis may not be very problematic if it is used as an conventional tool for nonparametric smoothing and/or data reduction. What is really problematic is the two step procedure where the outcome of principal component analysis is treated as if a new data set. In the first stage, he applies principal component analysis to a set of multivariate time series as if i.i.d. multivariate samples are successively obtained. But once the principal component analysis has been done, he regards a series of principal component as a time series, and fits time series models to each principal component. This approach is, whether useful or not, logically inconsistent. When he formally applies principal component analysis to multiple time series, he pretend as if he is not interested in time dependency in the data at all but interested in the synchronicity in a multivariate data for fixed time point. They *happened to* come in some order of time. Nevertheless, as soon as principal components are obtained, he changes his mind to assume dynamics for the obtained principal components.

The same story can be found in regression analysis, too. For example, suppose we estimate

a regression model at a fixed time point. Running the time index yields the series of estimated coefficients. In empirical research, time series models are fitted to the coefficients as if they were new observations. This approach is also inconsistent; in the first stage, the parameters in the regression model are unknown but fixed constant, while in the second stage randomness is introduced in the coefficients of the model. There might be some reasons for this approach if it is motivated by the need for interpretation of the results, which might be improved by such an ex-post smoothing. But it is questionable if this approach is used for prediction, because the goodness of the time series model fitted to the estimated regression coefficient only guarantees the goodness of prediction in the space of coefficients, which has nothing to do with the goodness of prediction in the original data space.

A similar kind of problem has been pointed out in psychometrics in the context of multilevel analysis. Suppose we have various sampling levels like district, school, children. Under such a circumstance, it is convenient but problematic if the results (or the estimated parameters) obtained by some statistical analysis on students in a school are used as data in the analysis of upper level, such as analysis on schools. Of course, a hierarchical model can give a direct formulation for this kind of problem, see Hox and Kreft (1994) and the special issue it is included in.

6.4.1 Numerical Example: Modeling Output Gap

As is explained in the previous section, time domain PCA on multiple time series is only valid for a few cases like nonparametric trend estimation or data reduction like index construction. This section, noting that one factor model without any consideration for lag corresponds to time domain PCA, considers bivariate econometric model to put stress on the importance of lagged dynamic factors. Let p_t ($t = 1, \ldots, T$) be consumer's price index and let y_t be another economic time series. (It will be specified later.) It is assumed that each of p_t and y_t has its own stochastic trend component ($\mu_{1,t}$, $\mu_{2,t}$) while shares the stationary component around trend (s_t) and the loading for each series can be different. Furthermore we assume a first order random walk for trend component, and second order stationary AR for s_t. That is, the model considered here is explicitly given as

$$\begin{aligned}
\mu_{1,t} &= \mu_{1,t-1} + v_{1,t}, \quad v_{1,t} \sim \mathrm{N}(0, \tau_1^2) \\
\mu_{2,t} &= \mu_{2,t-1} + v_{2,t}, \quad v_{2,t} \sim \mathrm{N}(0, \tau_2^2) \\
s_t &= \phi_1 s_{t-1} + \phi_2 s_{t-2} + v_{3,t}, \quad v_{3,t} \sim \mathrm{N}(0, \tau_3^2) \\
p_t &= \mu_{1,t} + c_1 s_t + w_{1,t}, \quad w_{1,t} \sim \mathrm{N}(0, \sigma_1^2)
\end{aligned} \tag{6.19}$$

$$y_t = \mu_{2,t} + c_2 s_t + w_{2,t}, \quad w_{2,t} \sim \mathrm{N}(0, \sigma_2^2).$$

Prior to the model estimation, means of $\{p_t\}$ and $\{y_t\}$ are demeaned. In addition, by a suitable normalization of differenced series we may set $\tau_1^2 = \tau_2^2$. Without loss of generality, we also assume $\tau_3^2 = 1$.

If we choose unemployment rate for y_t, the trend of y_t can be interpreted as natural unemployment rate, and deviations from natural unemployment rate may indicate the output gap. According to the theory of Phillips curve, output gap can account for the movement of price indices. The aim of numerical example here is not the empirical test of such economic theory but to accentuate the importance of lagged common factor. Now we replace 6.19 by

$$p_t = \mu_{1,t} + c_1 s_{t-\ell} + w_{1,t}, \quad (\ell = 0, 1, 2, \cdots).$$

It means that s_t concurrently accounts for y_t but may need some time delay until affecting CPI, p_t. In short, we expect y_t be a leading economic indicator. Here we adopt construction order in private sectors. Original data is plotted in the upper two panels in Figure 6.1. The sample period covers from the first quarter of 1982 to fourth quarter of 1999. Original data was monthly observed but it is converted to quarterly data here. As for CPI series, the effects of the introduction and raise of consumption tax are eliminated from the original monthly data, and are converted to quarterly data afterwards.

Estimation results are reported in Table 6.1. As the model employed here has three unobserved components for bivariate time series, it is difficult to compare it to the ordinary principal component analysis. Nevertheless, the case of $\ell = 0$ might be close to the simple-minded time domain PCA. The model with $\ell = 5$ improves the log-likelihood by almost 2. It is fair improvement if we consider the trends are dominant in both time series and there is no significant differences in the estimated τ^2s.

In the model without lag, though the estimated trend looks alike that of the model with lag 5, the variance of common AR component ten times as large, and the estimated AR coefficients (ϕ_1, ϕ_2) suggest nonstationarity ($\phi_1 + \phi_2 \approx 1$). In fact, the lower two panels in Figure 6.1 shows that the common AR component estimated from the model with $\ell = 0$ seems to have local trends. It may indicate that neglecting lag structure in the dynamic factor model interfered the extraction of appropriate common factor.

6.5 Summary and Conclusion

We surveyed the theory and methods for principal component and factor analysis for multivariate time series. This chapter is so to say a two dimensional plane, which is divided into

Table 6.1: Parameter Estimates

Parameter	Lag 0	Lag 5
τ^2	0.83×10^{-2}	0.90×10^{-2}
σ_1^2	0.14×10^{-1}	0.15×10^{-2}
σ_2^2	0.25×10^{-2}	0.18×10^{-2}
ϕ_1	1.47	0.84
ϕ_2	-0.55	-0.52
c_1	-0.08	0.10
c_2	-0.04	0.56
log-likelihood	-135.47	-133.29

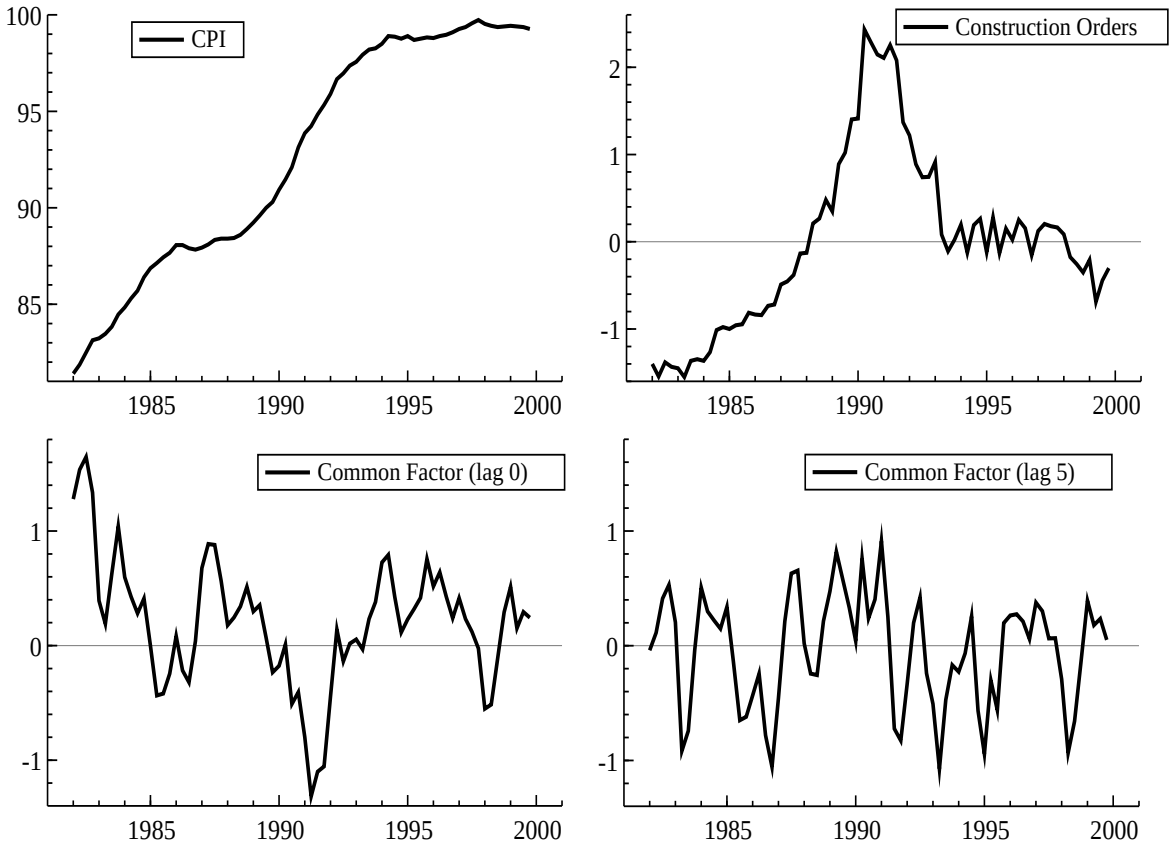


Figure 6.1: (Upper-left) CPI, (upper-right) construction order, (lower-left) common factor estimated without lag, (lower-right) common factor estimated with lag 5.

four orthants by ‘principal component analysis – factor analysis’ coordinate and by ‘frequency domain – time domain’ coordinate. Principal component/factor analysis in frequency domain utilize the discrete Fourier transformation (DFT) of time series. The central idea is that the DFT of time series ensures the asymptotic independence of the periodograms, which subsequently makes it possible to apply the classical framework that results in the eigenvalue decomposition of spectral density matrix. Factor analysis in time domain generally gives explicit models to the latent factor process. Dynamic factor analysis extends the ordinary factor analysis by incorporating lagged factors and estimate the model in terms of covariance structure analysis. The canonical transformation models describe the latent factors by multivariate ARMA process. In the methods where explicit models are not given to the factor processes, the arguments are concentrated on the rank condition of the matrix that yields factors. As such methods, reduced rank regression models and error correction models are available. On the other hand, a formal application of principal component analysis in time domain should be handled with care because such an approach discards the information contained in time series dependency, and is just focusing on the correlation of a multivariate data given at a fixed time. Its validity seems to be limited to such problems as nonparametric trend estimation and data reduction typically found in index construction.

Chapter 7

Smoothness Prior Approach to Estimate Large Scale Multifactor Models

7.1 Introduction

There are many empirical researches that suggest relative stock returns can be predicted by factors that are not consistent with the accepted paradigms of modern finance. In a sense there is a dichotomy on factor models or multifactor models. From the viewpoint of modern portfolio theory, factor models are the theoretical framework that explain the relation between stock returns and their risk. But for the most of practical analysts or institutional investors, multifactor models are tools to predict excess stock returns in the next period. Although there have been a lot of reaction to the assertion that stock returns can be predicted, this paper basically regard multifactor models as prediction models.

The aim of this paper is neither to discuss the appropriateness of the way of using multifactor models nor to present convincing evidence that stock returns can not be predicted. When the multifactor models are used for prediction, the most common method is constructed with OLS (cross-sectional regressions) and smoothing procedure such as moving average method or exponential smoothing. The merit of this method is its easiness in implementation. On the other hand, the greatest disadvantage of this method is we are obliged to determine the degree of smoothness quite empirically. The aim of this paper is to remove this ad hocery by smoothness prior approach.

However, it is shown in section 7.2 that in a large scale multifactor model, a simple application of Kalman filter produces almost the same result as the OLS without smoothing. To obtain the smooth payoff enough to be used in prediction, an extension called 'temporal effect model' is introduced in section 7.3. In section 7.4 various versions of multifactor models are compared in terms of the portfolio performance in trading simulations. In section 7.5, other possibilities

to bring predictive payoffs are sought together with the issues such as the initialization in a recursive filter and computational costs. Section 7.6 concludes.

7.2 Cross-Sectional Regression

At first we introduce conventional prediction scheme based on cross-sectional regression. Details are seen in Haugen and Baker (1996) for example. As they point out, factor models are basically risk models, or models which explain how returns are related with risks. On the other hand, at least for the most of institutional investors, they are used to predict the rate of return to particular stocks in the next period.

For a given month, we simultaneously estimate the monthly payoffs (cross-sectional regression coefficients) to a set of factor characteristics. Factor characteristics used throughout this paper are

- rate of return in previous one month
- cash flow to price
- dividend yield.

There seems to be no unanimous set of factors, and it can be different by the practical analysts. Possible ones are risk factors, liquidity factors, price level factors, factors indicating growth potentials and technical factors. Although searching predictive factors can be a very interesting issue, the problem is not discussed in this paper to focus on the recursive filter modeling of the conventional approach. Using an ordinary least squares (OLS), we estimate a cross-sectional multiple regression model.

$$r_{j,t} = \sum_{i=1}^{K} \beta_{i,t} F_{j,i,t-1} + u_{j,t} \tag{7.1}$$

where $r_{j,t}$ is rate of return to stock j in month t, $\beta_{i,t}$ is regression coefficient or payoff to factor i in month t, $F_{j,i,t-1}$ stands for exposure to factor i for stock j at the end of month $t-1$ and u_j is the unexplained component of return for stock j in month t. Let N be the number of stocks and K the number of factors considered here. We assume $(u_{1,t}, \ldots, u_{N,t})'$ are identically and independently distributed with multivariate normal with mean zero and covariance matrix $\sigma^2 I_N$. Past empirical analysis often find that the OLS estimated factor payoff series (which can be seen as a time series) are very noisy and have very poor predictive power. To get a reasonable solution as an expected payoff to a factor, the arithmetic mean of the estimated payoff over the

trailing m months are used to predict the next period rate of return. Expected rate of return are calculated as

$$\hat{r}_{j,t} = \sum_i \tilde{\beta}_{i,t} F_{j,i,t-1} \tag{7.2}$$

$$\tilde{\beta}_{i,t} = \frac{1}{m} \sum_{k=0}^{m-1} \hat{\beta}_{i,t-k} \tag{7.3}$$

where $\hat{\beta}_{i,t-k}$ denotes OLS estimates of payoff to factor i in month $t-k$. The number of one-sided moving average terms m differs by practical analysts, and is usually set to 12, 24 or 60 when the data is monthly observed. Poor performance of the prediction without smoothing, namely replacing $\tilde{\beta}_{i,t}$ in (7.2) by $\hat{\beta}_{i,t}$, will be seen in section 7.4.

More or less, the factor payoffs are supposed to have time-varying properties. On the other hand, the fluctuation of OLS estimates are so large that we cannot use it for prediction without smoothing *a posteriori*. Another route to smooth payoff is modeling time transition of factor payoffs *a priori*. For example, it is possible to assume that each payoff to factor i gradually changes according to the first order random walk,

$$\beta_{i,t} = \beta_{i,t-1} + v_{i,t} \tag{7.4}$$

where $v_{i,t}$ is identically independently distributed Gaussian random variable with zero mean and variance τ_i^2. Then equations (7.1) and (7.4) form state-space representation with suitable vector notation, and under the assumption of Gaussianity we can exploit recursive algorithm like Kalman filter. In other words, the problem here is to estimate a multiple regression model with time varying coefficients, and the time varying structure is given by a so-called smoothness prior. (Shiller (1973). See also Kitagawa and Gersch (1996) for the state-of-art techniques in this field.)

However, whether or not these smoothness prior models work depends on the dimension of observation vector (N) relative to that of state vector (K). Especially when $N \gg K$, simple application of Kalman filter does not produce smooth payoff series as we expect. As an example, we analyze the returns of stocks in TOPIX. The data is monthly observed and its sample period is from January 1985 to October 1997. The number of observations (or stocks to be considered) varies from 1,000 to 1300 during this period. As for three factors mentioned in the beginning of this section, each factors are normalized to zero mean and unit variance. Outliers, more than three standard deviations from the mean, are set to three standard deviations (times their sign).

Payoff series estimated by Kalman filter are shown in Figure 7.1. In each panel, only the filter mean is drawn. Three panels from the top respectively correspond to the three factors

chosen in the analysis, rate of return in previous one month, cash flow to price and dividend yield. The panel at the bottom shows the filter mean of the intercept term.

Figure 7.2 shows the result of OLS estimation of cross-sectional regression. Plot of payoff series in Figure 7.1 and Figure 7.2 are surprisingly alike. The reason is simple. The dimension of observation (almost 1300) is much greater than the dimension of smoothness prior (4 here), the information from sample variation always dominates the prior information.

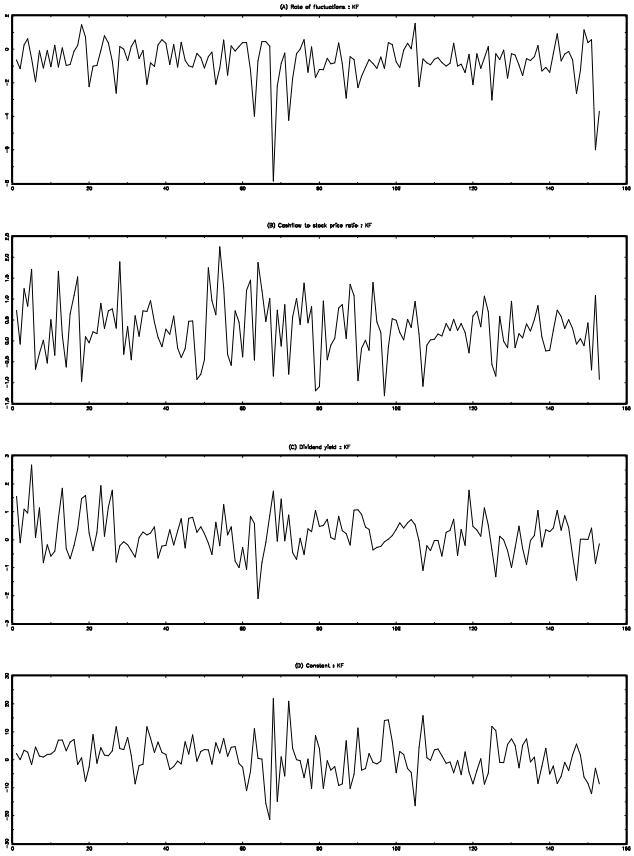


Figure 7.1: Results of Kalman filter: filter mean

The Kalman filter setup given by equations (7.1) and (7.4) is not exactly equal to the OLS based scheme by the treatment of observational noise. In the OLS based method, the variance of observational noise can differ at each time (at least *a posteriori*), whereas we assume the time-invariant σ^2 in the Kalman filter setup. Hence we rerun the Kalman filter after estimating σ_t^2 from the residual in the first run of Kalman filter. In spite of this, Figure 7.3 shows that

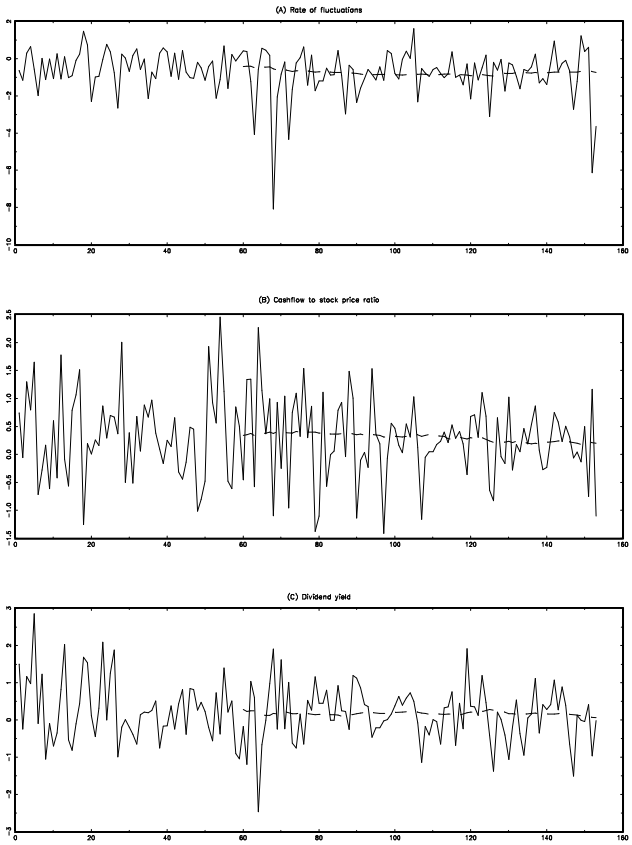


Figure 7.2: Payoff from OLS (bold) and moving averaged payoff (dashed)

estimated payoffs are almost unchanged. Hence we may conclude that in multifactor model the time-varying property of σ^2 does not affect the smoothness of payoff very much, and another modeling is needed to have smooth transition of payoff series.

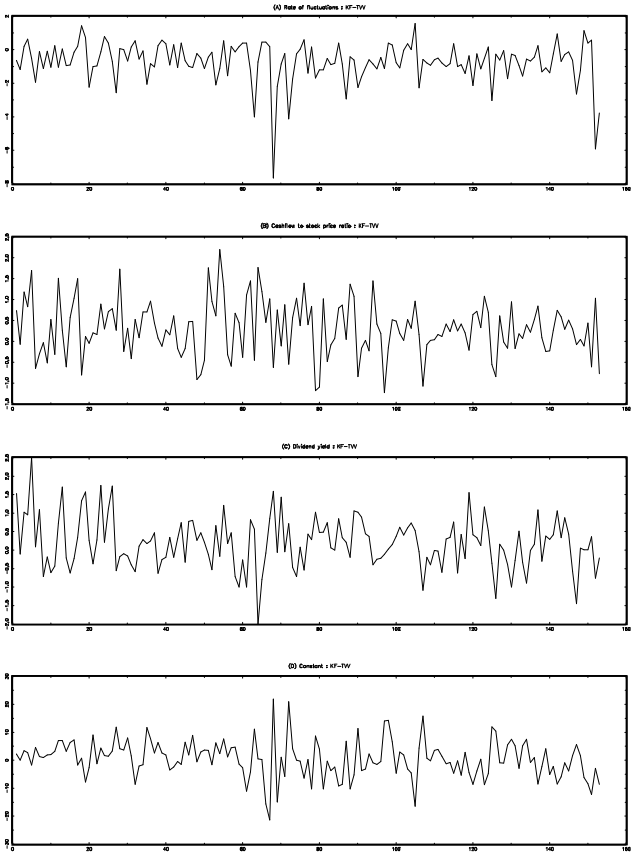


Figure 7.3: Kalman filter with time varying variance

7.3 Smoothness-Prior Approach

So far it has become clear that in the estimation of state-space model, when the observational dimension is much larger than that of state dimension, the use of the Kalman filter demands an extraordinary computational burden. When there exists such a dimension gap, a variant of the Kalman filter called the information filter gives an efficient solution, and the exact initial diffuse distribution can be provided for non-stationary state vector. Moreover, the square root decom-

position of the information matrix ensures the positive definite information matrix (therefore the covariance matrix). In this section, we give an efficient solution to the problem of estimating the vector-valued multiple regression models with time-varying coefficients.

7.3.1 Preparation

Let x_t be the $p \times 1$ state vector, and $\Sigma_{t|t-1}$ be the covariance matrix ($p \times p$) of one step ahead predictor of x_t. Let $N \times 1$ vector $z_t (t = 1, \ldots, T)$ be the vector of observations, and suppose a state-space representation is given to the model of our interest as follows,

$$x_{t+1} = F_t x_t + G_t w_t, \; w_t \sim \mathrm{N}(0, Q_t) \tag{7.5}$$

$$z_t = H_t x_t + v_t, \; v_t \sim \mathrm{N}(0, R_t) \tag{7.6}$$

where w_t and v_t are independently and normally distributed with their mean 0 and the covariance matrices Q_t, R_t respectively. Under these assumptions, the log-likelihood for this model can be given as follows in a so-called 'variance decomposition form',

$$\ell(\theta) = -\frac{1}{2}\left\{ NT \log 2\pi + \sum_{t=1}^{T} \log |\Lambda_{t|t-1}| + \sum_{t=1}^{T} (z_t - H_t x_{t|t-1})' \Lambda_{t|t-1}^{-1} (z_t - H_t x_{t|t-1}) \right\} \tag{7.7}$$

where

$$\Lambda_{t|t-1} = H_t \Sigma_{t|t-1} H_t' + R_t. \tag{7.8}$$

A $N \times p$ matrix H_t is the design matrix in the observation equation, and the argument θ of the log-likelihood ℓ signifies the stuck of the unknown (hyper-)parameters to be estimated from the data of which dimension essentially depends on the model specification. (Some authors use the transpose H_t' in the definition of the observation equation (7.6). If the reader is accustomed to that notation, the equations and formulas in the subsequent sections should be translated.)

Although the final aim of this chapter is to offer an efficient algorithm for the vector-valued multiple regression models with time-varying coefficients, for the moment we establish our arguments for the general linear Gaussian state space models. We are interested in the trade-off between the complexity of R_t and the feasibility of estimation, and two different modeling are proposed in section 2 and 3 respectively. In both cases (and even in OLS case) it is necessary to evaluate efficiently the determinant and the quadratic form in the right hand side of (7.7). The following lemma will be frequently used in this chapter.

Lemma (Anderson and Moore (1979), § 6.3)

Let Σ, R and Hbe a $p \times p$, $n \times n$ and a $p \times n$ matrix respectively. Then the following equality holds.

$$(I + \Sigma H R^{-1} H')^{-1} \Sigma = (\Sigma^{-1} + H R^{-1} H')^{-1} = \Sigma - \Sigma H (H' \Sigma H + R)^{-1} H' \Sigma \tag{7.9}$$

7.3.2 Elton-Grüber Model

Applying the matrix inversion formula to (7.8) yields

$$\Lambda_{t|t-1}^{-1} = R_t^{-1} - R_t^{-1} H_t \left[\Sigma_{t|t-1}^{-1} + H_t' R_t^{-1} H_t \right]^{-1} H_t' R_t^{-1}. \tag{7.10}$$

Even when the dimension of observation is quite large, (7.10) ensures the fast calculation of $\Lambda_{t|t-1}^{-1}$ as long as the calculation of R_t^{-1} is easy. In this case, the computation of the quadratic form is not heavy. In this subsection we consider the covariance matrix R_t of special form as follows,

$$R_t = (1-\rho)\omega^2 I_N + \rho\omega^2 \iota_N \iota_N' \tag{7.11}$$

where ι_N is the $N \times 1$ vector of which elements are all unity. As is easily seen, the off-diagonal elements are all set to ρ. This is one of the well known patterned matrix of which inverse can be easily given as

$$R_t^{-1} = \frac{1}{(1-\rho)\omega^2} \left[I_N - \frac{\rho}{1+(N-1)\rho} \iota_N \iota_N' \right], \tag{7.12}$$

see Graybill (1969) for example. Substituting (7.12) into (7.10), the remaining computational task is the inversion of $p \times p$ square matrix. In the context of financial econometrics, the use of this patterned matrix can be traced back to Elton and Grüber (1973), so for convenience sake we call this specification the Elton and Grüber type modeling.

In the Elton and Grüber type modeling, it is not so easy as in the quadratic form to give a clear short cut formula for the computation of the determinant. Nevertheless, it is still possible to device the reduction of computational task (Kawasaki et al. 2000). For the moment, we suppress the time suffix t of all the vectors and matrices to state the algorithm for a fixed time t. At first, the next relationship holds in general. Perform Cholesky decomposition as $\Sigma = SS'$, $R = PP'$, and define $C = P^{-1}HS$, then from

$$\Lambda = H\Sigma H' + R = HSS'H' + PP'$$

we obtain

$$P^{-1}\Lambda(P')^{-1} = I + CC'.$$

Because $|P^{-1}| = |P|^{-1}$ and the matrix P is lower triangular, it follows that $|P| = |P'|$. Using this, we obtain

$$|\Lambda| = |I_p + C'C| \cdot |P|^2.$$

This reduction from the observational dimension to the factor dimension is a often used technique in factor analysis. (See §8.4 of Anderson (1984) for example.)

The problem is whether or not the Cholesky decomposition of a $N \times N$ matrix $R(= PP')$, and the answer is yes for the Elton-Grüber type modeling. In the triangular matrix obtained as the Cholesky decomposition of (7.11), if it is seen row-wise, all the elements under the diagonal element take the same value. Hence we do not have to compute all the elements $(N^2/2)$ but only $2N-1$ values to complete the square root decomposition of R_t. We summarize the steps;

1. Do Cholesky decomposition $R = PP'$ where P is lower triangular.
2. Compute $|P^{-1}| = |P|^{-1}$.
3. Do Cholesky decomposition $\Sigma = DD'$.
4. Calculate $A_1 = P^{-1}H$.
5. Calculate $A_2 = A_1 D$.
6. Calculate $A_3 = A_2' A_2 + I$.
7. Compute $|A_3|$ and divide it by $|P^{-1}|^2$.

It should be noted here that in the step 4 we do not have to compute P^{-1} explicitly. Namely, to have A_1, we only have to solve a system of linear equations $PA_1 = H$ which consists of at most p equations, whereas we need to solve another system of N equation to explicitly have P^{-1}.

7.3.3 Temporal Effect Model

In the preceding section we observed that a simple-minded introduction of Kalman filter (or smoothness prior) does not produce smooth payoff in case the dimension of observation vector is very large. In this section we introduce a slightly modified Kalman filter setup which follows.

$$\begin{aligned} \beta_t &= \beta_{t-1} + v_t, \qquad v_t \sim \mathrm{NID}(0, D_1) \\ r_t &= F_{t-1}(\beta_t + w_t) + u_t \\ w_t &\sim \mathrm{NID}(0, D_2) \\ u_t &\sim \mathrm{NID}(0, \sigma^2 I_N) \end{aligned} \tag{7.13}$$

where $D_1 = \mathrm{diag}(\tau_1^2, \ldots, \tau_K^2)$, and $D_2 = \mathrm{diag}(\lambda_1^2, \ldots, \lambda_K^2)$. Variables β_t, r_t, v_t, F_{t-1}, u_t are vector or matrix notation which is obvious from equations (7.1) and (7.4).

Motivation of this setup is rather simple. As we expect more or less smoother factor payoff than the result of OLS or simple Kalman filter, here we regard that each payoff series are contaminated with one more additional noise, w_t. In other words, we expect $\beta_{i,t}$ in (7.4) can

be decomposed into 'signal' and 'temporal additive noise'. Note that β_t depends only on the present and past value of v_t, but not on w_t at all. So it seems to make sense we hereafter refer this setup to 'temporal effect model'.

Temporal effect model can be viewed as a modeling of variance-covariance matrix in the observational equation. It is easily seen that the observational equation in (7.13) is equivalent to

$$r_t = F_{t-1}\beta_t + u_t^*, \quad u_t^* \sim \mathrm{N}(0, F_{t-1}D_2F_{t-1}' + \sigma^2 I_N). \tag{7.14}$$

Hence the temporal effect model introduces off-diagonal part in the covariance matrix of the observational noise. Or we implicitly assume that the elements in $\mathrm{Var}(u_t^*)$ are expressed as the rescaled covariances of factors in previous month plus some diagonal matrix.

Kawasaki et al. (1998) gave the following specification for R_t utilizing the design matrix H_t of regression.

$$R_t = H_tD_2H_t' + \sigma^2 I \tag{7.15}$$

where $D_2 = \mathrm{diag}(\lambda_1^2, \ldots, \lambda_p^2)$. Originally, R_t expresses the magnitude of the idiosyncratic errors that cannot be explained by the common factors. The fluctuation of stock prices is often influenced by the temporal forecasts of market tendency, so the basic idea is weight the temporal shock by the observations at present. Substituting (7.15) into (7.8) yields

$$\begin{aligned} \Lambda_{t|t-1} &= H_t\Sigma_{t|t-1}H_t' + H_tD_2H_t' + \sigma^2 I \\ &= H_t(\Sigma_{t|t-1} + D_2)H_t' + \sigma^2 I \\ &\equiv H_t\Phi_{t|t-1}H_t' + \sigma^2 I. \end{aligned} \tag{7.16}$$

where $\Phi_{t|t-1}$ is defined in the last expression. Applying the matrix inversion formula to above, we obtain

$$\begin{aligned} \Lambda_{t|t-1}^{-1} &= (\sigma^2 I)^{-1} - (\sigma^2 I)^{-1}H_t\left[\Phi_{t|t-1}^{-1} + H_t'(\sigma^2 I)^{-1}H_t\right]^{-1}H_t'(\sigma^2 I)^{-1} \\ &= \sigma^{-2}I - \sigma^{-4}H_t\left[\Phi_{t|t-1}^{-1} + \sigma^{-2}H_t'H_t\right]^{-1}H_t'. \end{aligned} \tag{7.17}$$

If nothing is done, we have to compute $(\Phi_{t|t-1}^{-1} + \sigma^{-2}H_t'H_t)^{-1}$ after computing $\Phi_{t|t-1}^{-1}$. Though the required calculation is manipulation of the matrices of at most state dimension (p), it is reasonable to employ the following algorithm. At first, compute the $p \times p$ lower triangular (square) matrix $\Gamma_{t|t-1}$ that satisfies $\Phi_{t|t-1} = \Gamma_{t|t-1}\Gamma_{t|t-1}'$. Using $\Gamma_{t|t-1}$, compute $\Delta_{t|t-1} = H_t\Gamma_{t|t-1}$. Then we obtain $\Lambda_{t|t-1} = \sigma^2 I + \Delta_{t|t-1}\Delta_{t|t-1}'$, hence applying the matrix inversion formula yields

$$\Lambda_{t|t-1}^{-1} = \sigma^{-2}I - \sigma^{-4}\Delta_{t|t-1}\left[I + \sigma^{-2}\Delta_{t|t-1}'\Delta_{t|t-1}\right]^{-1}\Delta_{t|t-1}'. \tag{7.18}$$

On the other hand, we use the following lemma to compute the determinant.

Lemma

$$\left|\sigma^2 I_N + \Delta_{t|t-1}\Delta'_{t|t-1}\right| = \sigma^{2(N-p)}\left|\sigma^2 I_p + \Delta'_{t|t-1}\Delta_{t|t-1}\right| \tag{7.19}$$

(*Proof.*) We simply write Δ for $\Delta_{t|t-1}$ because there seems to be no danger of confusion. For convenience sake, we put $\tilde{\Delta} = \sigma^{-1}\Delta$. Then the following simple algebra establishes the above lemma.

$$\begin{aligned}
\left|\sigma^2 I_N + \Delta\Delta'\right| &= \sigma^{2N}\left|I_N + \sigma^{-2}\Delta\Delta'\right| \\
&= \sigma^{2N}\left|I_N + \tilde{\Delta}\tilde{\Delta}'\right| \\
&= \sigma^{2N}\left|I_p + \tilde{\Delta}'\tilde{\Delta}\right| \\
&= \sigma^{2N}\left|I_p + \sigma^{-2}\Delta'\Delta\right| \\
&= \sigma^{2N}\sigma^{-2p}\left|\sigma^2 I_p + \Delta'\Delta\right| \\
&= \sigma^{2(N-p)}\left|\sigma^2 I_p + \Delta'\Delta\right|
\end{aligned}$$

7.3.4 Application: Large-Scale Multifactor Model

In this subsection , we consider a model to explain the stock returns (often relative to the return of the index like Nikkei 225) by a set of explanatory variables that consists of financial statements data and the technical indices. Among financial institutions, the explanatory variables are often referred to as 'factors' to determine the stock returns. Though it sounds strange for statisticians, we call the regression model considered in this chapter as the multifactor model in accordance with customary practice in this field. Though many procedures have ben proposed in practice (see Haugen and baker (1996) for example), we put the model into the state space framework in this chapter as follows.

$$x_t = x_{t-1} + w_t,\ w_t \sim \mathrm{N}(0, Q) \tag{7.20}$$

$$z_t = H_t x_t + v_t,\ v_t \sim \mathrm{N}(0, R_t) \tag{7.21}$$

The system equation (7.20) describes the fluctuation of time varying coefficients that is controlled by the hyperparameters $\tau_1^2, \ldots, \tau_p^2$, which are the elements of the covariance matrix $Q = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2)$. Namely, the covariance of the system equation does not depend on time while the coefficients do. H_t is the design matrix in which the 'factors' are filled with. Although the time subscript is t, the elements of H_t are actually the data up to time $t-1$. The specification of R_t makes the difference among models.

- $R_t = \sigma^2 I_N$ (OLS, time invariant)
- $R_t = (1-\rho)\omega^2 I_N + \rho\omega^2 \iota_N \iota_N'$ (Elton-Grüber, time invariant)
- $R_t = H_t D_2 H_t' + \sigma^2 I_N$, $D_2 = \mathrm{diag}(\lambda_1^2, \ldots, \lambda_p^2)$, (Temporal Effect, time dependent)

7.3.5 Information Square Root Filter

We go back to the state space representation (7.5) and (7.6) again. Let $S_{t|t-1}$ be the Cholesky decomposition of $\Sigma_{t|t-1}$ such that $\Sigma_{t|t-1} = S_{t|t-1} S_{t|t-1}'$. Then we can establish the recursive formula not on x_t but on the transformed state $\hat{b}_{t|t}$ and $\hat{b}_{t|t-1}$ as follows,

$$\begin{aligned} \hat{b}_{t|t} &= S_{t|t}^{-1} \hat{x}_{t|t} \\ \hat{b}_{t|t-1} &= S_{t|t-1}^{-1} \hat{x}_{t|t-1}. \end{aligned}$$

In filtering step, updating formula can be given by a certain orthogonal transformation T as

$$\begin{pmatrix} S_{t|t}^{-1} & \hat{b}_{t|t} \\ 0 & * \end{pmatrix} = T \begin{pmatrix} S_{t|t-1}^{-1} & \hat{b}_{t|t-1} \\ R_t^{-1/2} H_t' & R_t^{-1/2} z_t \end{pmatrix}. \tag{7.22}$$

Let Q_t be the covariance matrix of the system noise and G_t be the coefficient matrix by which the system noise to be multiplied. Now define A_t and B_t by

$$\begin{aligned} A_t &= (F_t^{-1})' \Sigma_{t|t}^{-1} F_t^{-1} \\ B_t &= A_t G_t (G_t' A_t G_t + Q_t^{-1})^{-1}, \end{aligned}$$

then prediction step can be defined using an appropriate orthogonal transformation $\bar{T}$ as

$$\begin{pmatrix} \{(Q_t^{-1} + G_t' A_t G_t)^{1/2}\}' & B_t' & * \\ 0 & S_{t+1|t}^{-1} & \hat{b}_{t+1|t} \end{pmatrix} = \bar{T} \begin{pmatrix} (Q_t^{1/2})^{-1} & 0 & 0 \\ S_{t|t}^{-1} F_t^{-1} G_t & S_{t|t}^{-1} F_t^{-1} & \hat{b}_{t|t} \end{pmatrix}.$$

An orthogonal transformation can be implemented by the Householder transformation or by modified Gramm-Schmidt transformation. In this chapter, the Householder transformation is employed. This formula is called the information square root filter. In the multifactor model, the above formula can be more simplified. For example, we have $A_t = \Sigma_{t|t}^{-1}$, and accordingly we have $B_t = \Sigma_{t|t}^{-1}(\Sigma_{t|t}^{-1} + Q^{-1})^{-1}$. Hence the prediction step can be rewritten as

$$\begin{pmatrix} \{(Q^{-1} + \Sigma_{t|t}^{-1})^{1/2}\}' & B_t' & * \\ 0 & S_{t+1|t}^{-1} & \hat{b}_{t+1|t} \end{pmatrix} = \bar{T} \begin{pmatrix} (Q^{1/2})^{-1} & 0 & 0 \\ S_{t|t}^{-1} & S_{t|t}^{-1} & \hat{b}_{t|t} \end{pmatrix}.$$

On the other hand, filter formula remains the same as (7.22).

As a matter of fact, each of the three models considered here has a convenient structure by which the computation of $R_t^{-1/2}$ is of little additional burden. Therefore, to estimate the multifactor models proposed here, the use of the information square root filter is best.

In the decomposition model using smoothness priors, typically the observation is univariate and the state is of multi-dimension just as in the seasonal adjustment of univariate economic time series. In such a case, the Kalman filter has an advantage while the information (square root) filter has no computational advantage except the exact treatment of the diffuse initial condition.

7.4 Comparison of Portfolio Performance

There is no statistical tool to decide which approach is better, namely 'OLS plus moving average a posteriori' or Kalman filter with temporal effect model. As moving average is just a procedure, basically we can not strictly tell whether the smoothing is appropriate or not. Hence we introduce another practical criterion, the goodness of portfolio performance during the sample.

Each model produces different factor payoffs, and therefore yields different expected stock returns. Investment strategy here is to sell dear and to buy cheap. Rebalance of the portfolio should be made at the end of every month. After estimating new payoffs and obtaining new factors, each model gets new forecasts of stock returns. According to that new information, we again sell dear and buy cheap, keeping net position zero with same amount of long and short.

In every empirical analysis reported in this paper, the intercept term, $\hat{\alpha}_t$ say, is estimated but not used in prediction. It is set to zero in each prediction step. We have two reasons for that treatment. First, as the intercept term $\hat{\alpha}_t$ is common to all the stocks in TOPIX (at fixed time t), it can be interpreted as the performance of well diversified portfolio at time t. Then, in a sense, predicting $\hat{\alpha}_t$ leads to predicting the stock market index futures one month ahead, which is extremely difficult. Secondly, including $\hat{\alpha}_t$ in prediction does not affect the investment strategy here because adding or subtracting $\hat{\alpha}_t$ just alters the baseline of expected returns of all the stocks. What we need is the information about which stock is relatively cheap (or expensive), and the absolute prediction the rate of return in the next month is not our final aim. Therefore, it can be said that $\hat{r}_t$ is the prediction of the relative excess return to the market index.

The results are reported in Table 7.1. Four models are estimated. `OLS-C-60AVG` means the multifactor model including constant term is estimated, and 60 month one-sided moving average is applied to the OLS estimated factor payoffs. `KF-TE-C-IN` stands for Kalman filter with temporal effect including constant term in the regression. `IN` means in-sample prediction,

Table 7.1: Results of Trading Simulations

	OLS-C-60AVG	KF-TE-C-IN	KF-TE-C-OUT	KF-IN-C
mean	2.9859	2.9162	2.7428	1.7209
std. dev.	3.6228	3.5943	3.2129	3.1108
t-value	7.9482	7.8242	8.2324	5.3349
Sharpe ratio	2.8551	2.8105	2.9572	1.9163

that is, a set of hyperparameters estimated with full sample was used in trading simulation. On the other hand, in `KF-TE-C-OUT`, hyperparameters are re-estimated every 10 months. With re-estimated hyperparameters, prediction and filtering are repeated for 10 months. `KF-IN-C` is a kind of control. OLS without moving average yields more or less same result.

Throughout this analysis, the first 60 payoffs are used to estimate initial factor payoff at time 60 (December 1989) because the 60 terms one-sides moving average inevitably wastes the first 60 payoffs. Hence the period of trading simulations is adjusted to the possible maximum length in `OLS-C-60AVG`, namely from January 1990 to October 1997. As for the initialization of Kalman filter, the initial distributions of all state are expressed as Normal distribution with unknown mean and variance. Hence the initial conditions are also the objects of numerical optimization. The next section discusses the various way of the initialization of Kalman filter in connection with the computational efficiency.

Means reported in Table 7.1 are the averaged rates of return per month. Risk adjusted measure of investment performance (annualized Sharpe ratio) is also reported. The results of `OLS-C-60AVG` and `KF-TE-C-OUT` show that these two models are almost equivalent in performance, which can be understood by the graphs of the estimated factor payoffs (Figure 7.4). The result of `KF-TE-C-OUT` is inferior to `OLS-C-60AVG` in the averaged rate of return, but `KF-TE-C-OUT` has smaller standard deviation, which results in the almost equivalent Sharpe ratio of `OLS-C-60AVG`.

Moreover, we performed these trading simulations with fixing the number of terms in moving average to 60, which we borrowed from the preceding research on this data set. Since the performance of `OLS-C-60AVG` is premised on the knowledge of the number of moving average terms in advance, this is a kind of *data-snooping*. Without such data-snooping, Kalman filter based method brings almost equivalent performance. However, we have to pay the price for it. Computational costs are discussed in the next section, together with the orders of smoothness priors and way of the initialization of Kalman filter.

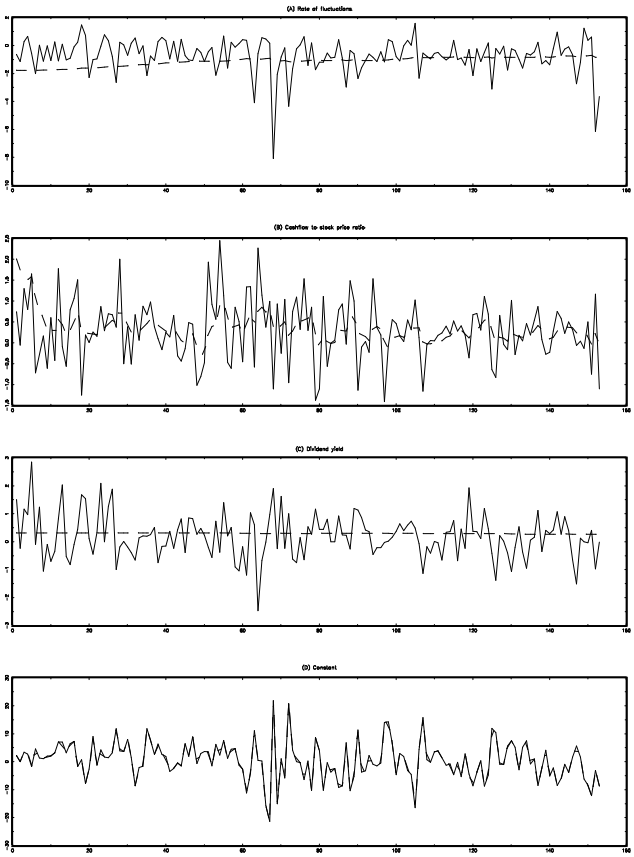


Figure 7.4: Payoff from TE model (dashed) laid on payoff from OLS (bold)

7.5 Higher Order Models, Initialization and Computational Cost

In section 7.3, we introduced the temporal effect model to obtain smooth payoff. In section 7.4, we have also seen that without temporal effect modification the performance of trading simulation based on the payoff estimated by Kalman filter shows an extremely bad result. To investigate the possibility of obtaining smoother payoff, we introduce second order smoothness prior models instead of first order ones in the following subsection 7.5.1.

In the following subsection 7.5.2, we exploratory search an appropriate treatment of initial distributions of Kalman filter. The conclusion is that, first order models are supported in terms of good portfolio performance, and that starting with diffuse prior with one-way filtering are recommended in terms of computational cost and portfolio performance.

7.5.1 Second Order Smoothness Prior

If a smoother payoff is desirable, it is quite natural to assume higher order smoothness prior models instead of (7.4). Here we adopt the second order model,

$$\beta_t = 2\beta_{t-1} - \beta_{t-2} + v_t. \tag{7.23}$$

The obvious drawback of this extension is that the dimension of state vector is twice larger than the first order model, namely from 8 to 16. (Note that the number of hyperparameters for temporal effect noise does not increase even in the higher order smoothness prior models.) In the portfolio performance evaluation done in section 7.4, we treated the unknown parameters in the initial state distribution as the objects of numerical optimization. If we adopt the same method here, a lot of computational time will be needed to optimize many unknown parameters in the second order model. For example, to estimate the model `KF-TE-C-IN` in Table 7.1, it takes two weeks and a day on a DEC Alpha 500MHz workstation with the program written in C. The results reported in Table 7.1 were calculated on the parallel computer system of the Institute of Statistical Mathematics (11 nodes of IBM RS/6000 out of 48 nodes) with the program written in FORTRAN. The computational time was reduced to nearly 30 hours to estimate `KF-TE-C-IN`.

In general, however, a parallel computer system environment is not easily accessible to the most of researchers. And the more factors we have, the more computational time it takes. In that sense it is meaningful to investigate a more efficient way of initialization of Kalman filter.

7.5.2 Initialization

We consider three different way of initialization of Kalman filter here.

1. Normal distribution with unknown mean and variance is assumed for each initial state. These parameters are among the arguments of the log-likelihood function, and searched by numerical optimization. (FF_NPR hereafter)

2. Kalman filter starts with zero mean and big variance (namely diffuse prior). To remove the effect of initial diffuse prior, first 20 or 30 prediction errors are discarded in the likelihood calculation. (FF_DFP)

3. Do the backward filtering starting with a diffuse prior. The filter distribution at the end automatically gives the estimates of initial distribution of forward filter. Then do the forward filter and compute the likelihood. Repeat this until the likelihood is numerically maximized. (BF_DFP)

For each initialization method, we apply both first and second order smoothness prior models, hence we estimate 6 models with different setups. The temporal effects are incorporated in all models, and basically .

In the empirical analysis of this section, we use the stocks in the Nikkei Stock Average (the number of stocks in it is 235 at most) and the associated factors, because TOPIX is too large to repeat the model evaluation under different settings. The same trading simulations as in section 7.4 (a risk neutral long/short strategy) are applied with the estimated payoffs from 6 models here. OLS-based method is also compared. The sample period of the performance evaluation is the same, namely from January 1990 to October 1997.

Just for the convenience, the result of OLS plus moving average is attached both in Table 7.2 and Table 7.3 although they show the same statistics. As the annualized Sharpe ratios reported in Table 7.2 and Table 7.3 are more or less same, we can hardly tell which method significantly out-performs others. In spite of the increase of computational load, the second order smoothness prior models could not find any significance improvement in the portfolio performance. At least from this experiment, it is hard to find a reason to introduce higher order difference equations for payoff series.

As for the computational load, it is not easy to compare all the six models because most of the results are computed on different computer systems and compilers. For example, it takes 37 hours 45 minutes to estimate the model FF-NPR-2 on a DEC Alpha 533MHz workstation, whereas 12 hours 29 minutes for BF-DFP-2 on the same machine. The FF-DFP-2 is supposed to be finished faster, and FF-DFP-1 much faster. Our conclusion about the way of initialization and the order of stochastic difference equation is that the forward filter starting with diffuse prior of order one is enough in terms of the risk neutral long/short investment

Table 7.2: Results of First Order Models

	FF-NPR-1	FF-DFP-1	BF-DFP-1	OLS
mean	2.3721	2.2738	2.4648	2.4165
std. dev.	3.6395	3.5624	3.5918	3.5886
t-value	6.2854	6.1553	6.6178	6.4938
Sharpe ratio	2.2578	2.2210	2.3772	2.3327

Table 7.3: Results of Second Order Models

	FF-NPR-2	FF-DFP-2	BF-DFP-2	OLS
mean	2.5226	2.1729	2.3288	2.4165
std. dev.	3.7461	3.3021	3.4976	3.5886
t-value	6.4939	6.3458	6.4210	6.4938
Sharpe ratio	2.3327	2.2795	2.3065	2.3327

strategy.

7.6 Summary and Conclusion

The multifactor models for the use of predicting the one-month ahead stock returns are discussed. As a conventional way, the cross-sectional regression models are usually estimated at first, and the estimated payoff series are averaged by a procedure like moving average or exponential smoothing. Although it is natural to give a state space representation to this procedure, we have shown that a simple-minded application of Kalman filter does not produce smooth payoff enough to be used in one-month ahead prediction, especially when we have many observations at a time, in spite of the introduction of smoothness priors. Hence we introduced a modified version of Kalman filter called the temporal effect model. This helps to get smoother payoff, and the results of the trading simulations based on the stocks in TOPIX shows that their annualized Sharpe ratios are almost same. We also investigated whether higher order stochastic difference equations bring much smoother payoff and/or much better portfolio performance, and whether different initialization of Kalman filter lead to significantly different portfolio performances. Our conclusion on that point is the forward filter starting with diffuse prior of order one is enough in terms of the risk neutral long/short investment strategy.

We briefly summarize the empirical results on Japanese stock market, the first division of the Tokyo Stock Exchange. We used monthly stock price data from january 1985 to December 1997. The coefficients derived by the OLS model are the same as those obtained by the repe-

tition of cross-sectional regression. To put it another way, the smoothness prior did not work effectively, and the model is not suitable for prediction. In Elton-Grüber type modeling, some industrial sector exhibit the smooth factor coefficients. But larger the sector is, less smooth the estimated coefficients are. This is reasonable because we have just one extra parameter (ρ) in the Elton-Grüber model, and if the size of a sector gets larger, it becomes less realistic for a single parameter to capture the correlation of many stocks at a time. Temporal effect model is flexible and has more parameters than the Elton-Grüber model, the total time required for estimation is almost comparable to that of the Elton-Grüber model. In terms of the AIC statistics, Temporal Effect model is generally the most preferred. Plotting the time varying feature of the Kalman gain reveals the differences in the specification of covariance matrix show up in the allocation of the prediction error. The profile of the time varying coefficient of reversal factor manifests the change of the sentiment and the expectations of investors in Japanese stock markets.

Acknowledgments

The authors are grateful to the Nikko Securities Co. Ltd. for providing us a set of stock data and firm characteristics data. They also wish to thank the Tokyo Stock Exchange, the Nihon Keizai Shimbun, Inc. and Toyo Keizai, Inc. who are the original data source.

Chapter 8

Estimating Term Structure Using Nonlinear Splines: A Penalized Likelihood Approach

8.1 Introduction

There have been a number of studies attempting to establish an excellent technique for estimating the term structure of interest rates from a cross-section of coupon bond prices. Under the assumption that the price of a bond is equal to the present value of its feature coupon payments and redemption, McCulloch (1971) regressed cash flows on a set of basis functions to estimate discount functions. Once the discount function is estimated, the zero-coupon yield and the forward rate can be obtained by transformations of the discount function.

Although the approach adopted by McCulloch (1971, 1975) was followed by several related studies, the approach has been criticized on a number of points. Some researchers are concerned with the choice of basis functions when defining a spline function, while others question how to place knots efficiently. Many articles on this topic have attempted to overcome the same problem that the apparently reasonable estimate of the discount function does not always lead to acceptable shapes of yield curves, especially for the forward rate curve. For example, Shea (1984) and Steely (1991) found that some spline bases, such as that chosen by McCulloch, can generate a regressor matrix with columns that are nearly perfectly collinear. As a solution, they adopted the use of B-spline bases.

The choice of basis functions and/or knot locations undoubtedly affects the estimation results. However, the present article focuses on a different point. It is considered here that instability of the estimated yield curves is caused by the ill-posed nature of the regression spline, rather than by the inappropriate choice of the basis function. By ill-posed it is meant that a model may be over-parameterized compared to the amount of sample information. Without a

addressing this ill-posed nature specifically, any modification of the choice of basis functions, approximating functional forms, or knots placement may provide only minor improvements.

Throughout this article, a penalty term is added to the original log-likelihood of a yield curve model, that is, a penalized likelihood approach is adopted for this treatment. In this sense, the work of Fisher, Nychka and Zervos (1995) is the most closely related and influential to this study. Those authors fitted smoothing splines (with B-splines bases) instead of regression splines, which is in itself an approach similar to that of Gourieroux and Scaillet (1994). Smoothing splines have a penalty for excess roughness, and a single parameter (smoothing parameter) controls the size of this penalty. An increase in the penalty should be interpreted as a reduction in the effective number of parameters in the estimation (on the effective number of parameters, see e.g., Hastie and Tibshirani, 1990, p.52). Fisher et al. recommend the use of generalized cross-validation (GCV) to choose the roughness penalty (see also Craven and Wahba ,1979; Wahba, 1990). Adaptive choice of smoothing parameter by GCV circumvents the need to exogenously supply the number and location of knots. Fisher et al. chose a number of knots equal to one third of the sample size in fitting the smoothing spline.

One important conclusion obtained by Fisher, Nychka and Zervos (1995) from their simulation studies is that smoothing splines can be used to spline an arbitrary transformation of the discount function. Their simulation results demonstrate that the best way to estimate yield curves is to place spline bases on the forward rate curve.

It is widely known that the discount function $\delta(t)$ and the instantaneous forward rate $f(t)$ are related by

$$f(t) = -\delta'(t)/\delta(t), \tag{8.1}$$

where $\delta'(t)$ is the derivative of the discount function $\delta(\cdot)$ evaluated at the point t. The term structure $\eta(t)$ is tied to the discount function $\delta(t)$ by

$$\eta(t) = -\ln(\delta(t))/t. \tag{8.2}$$

See for example Anderson et al. (1996) for the derivations of these relationships. Hence, it is not necessary to start by approximating the discount function $\delta(t)$. From equations (8.1) and (8.2), it is recognized that if splines are placed on $\eta(t)$ or $f(t)$, then $\delta(t)$ will be expressed as an exponential function with an approximating function for $\eta(t)$ or $f(t)$ as its argument. That is, splining the term structure or the forward rate is equivalent to exponential splining of the discount function. This will be revisited in detail in section 8.2.

Prior to Fisher et al., several authors proposed exponential splining of the discount function. Vasicek and Fong (1982) proposed an exponential spline, while Chambers, Carleton and

Waldman (1984) suggested the use of an exponential polynomial to model the discount function. Langetieg and Smoot (1989) fitted a cubic B-spline to the term structure for an interest rate, which is eventually equivalent to fitting an exponential spline to the discount function. Coleman, Fisher and Ibbotson (1992) approximated the instantaneous forward rate $f(t)$ using a piecewise constant function, a technique that turns out to be equivalent to assuming exponential function bases to approximate $\delta(t)$.

By fitting a smoothing spline with cubic B-spline bases, Fisher et al. compared all three options; splining $\delta(t)$, $\eta(t)$ and $f(t)$, and determined the roughness penalty using GCV in all three cases. From a theoretical viewpoint, however, the application of GCV is questionable except for the case of splining $\delta(t)$. As point out themselves (footnote No. 10 and appendix B), GCV cannot be applied unless the regressor is expressed as a linear combination of basis functions. In other words, a basis function can be nonlinear in t as is usual with many nonparametric regression schemes, but the regression functional should be linear with respect to the unknown parameters. Clearly this does not hold in splining $\eta(t)$ or $f(t)$. Supposing that the splined term structure $\eta_s(t)$ is expressed as $\eta_s(t) = \sum w_k \phi_k(t)$, where $\{\phi_k(t); k = 1, 2, ...\}$ is a set of spline bases with coefficients w_k, then (8.2) implies that the splined discount function $\delta_s(t)$ is expressed as $\delta_s(t) = \exp(-t \sum w_k \phi_k(t))$. Here, δ_s is clearly not linear in $\{w_k\}$.

Sharing the motivation of Fisher et al., the aim of the present study is to propose a theoretically valid criterion to choose smoothing parameters even when the regression functional is not always linear with respect to the unknown parameters. In this treatment, the generalized information criteria (GIC) introduced by Konishi and Kitagawa (1996) is tailored to various cases. Us of the GIC also makes it possible to choose the optimal number of basis functions. This is an important feature, as allowing excess knots can lead to an undesirable shape of the forward rate function. Selection of the appropriate number of basis by an objective criterion is therefore desirable.

In this paper, the standard framework for estimating the term structure of interest rates based on a bond equation is briefly reviewed, and after introducing specific forms of the exponential spline, the maximum penalized likelihood concept is presented. A construction scheme for information criteria allowing the size of the roughness penalty and the number of basis functions to be chosen is then introduced. Custom-made GICs for the evaluation of various yield curve models are presented in the appendix. Some Monte Carlo experiments are conducted to judge the best yield curve model or the yield curve to be splined, and the choice of the number of basis function is discussed. Finally, the scheme is applied to actually Japanese governmental bond data.

8.2 Penalized Likelihood Approach

8.2.1 Bond equation

Consider a set of n bonds traded on one day. Let p_α be the price of bond α, c_α be its coupon payment, which is paid at time t_1^α,...,$t_{L_\alpha}^\alpha$, let R_α be the redemption payment, and let L_α be the number of remaining payments. Following the theory of bond pricing (McCulloch, 1971), we assume that the price of a bond (plus accrued interest a_α) is equal to the present value of its future coupon payments and the redemption, i.e.,

$$p_\alpha + a_\alpha = c_\alpha \sum_{k=1}^{L_\alpha} \delta(t_k^\alpha) + R_\alpha \delta(t_{L_\alpha}^\alpha) + \varepsilon_\alpha, \qquad \alpha = 1, ..., n, \tag{8.3}$$

where $\delta(\cdot)$ is the discount function, ε_α are independent and normally distributed with mean of zero and variance σ^2. The discount function $\delta(t)$ gives the present value of a monetary unit, e.g., \$1.00 after t years. Most researchers follow McCulloch (1971) in explicitly constraining cash flows from different bonds due at the same time to be discounted at the same rate, and estimate the discount function $\delta(\cdot)$ from which the term structure can be derived.

We begin here with a expository comment on the most basic case where splines are placed on the discount function. In this case, $\delta(\cdot)$ is expressed as a linear combination of a set of m underlying basis functions, as follows.

$$\delta(t; w) = 1 + \sum_{k=1}^{m} w_k \phi_k(t) = 1 + w' \phi(t), \tag{8.4}$$

where $\phi(t) = (\phi_1(t), ..., \phi_m(t))'$ is an m-dimensional vector constructed from a set of basis functions $\{\phi_j(t); j = 1, \cdots, m\}$, and $w = (w_1, ..., w_m)'$ is an unknown parameter vector to be estimated. It follows from equations (8.3) and (8.4) that the bond price model based on a linear combination of basis functions is as follows.

$$f_B(y_\alpha | t_\alpha; w, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_\alpha - c_\alpha' \Phi_\alpha w)^2}{2\sigma^2} \right\}, \tag{8.5}$$

where $t_\alpha = (t_1^\alpha, ..., t_{L_\alpha}^\alpha)'$ is the vector of the points of time at which payments occur, $y_\alpha = p_\alpha + a_\alpha - L_\alpha c_\alpha - R_\alpha$, $\Phi_\alpha = (\phi(t_1^\alpha), ..., \phi(t_{L_\alpha - 1}^\alpha), \phi(t_{L_\alpha}^\alpha))'$ and $c_\alpha = (c_\alpha, ..., c_\alpha, c_\alpha + R_\alpha)'$, respectively. This specification is very convenient for parameter estimation because the functional form of (8.4) is linear with respect to the unknown parameters.

A number of functional forms have been proposed for the basis functions $\phi_j(t)$. McCulloch (1971) used a quadratic spline to estimate the discount function. To avoid a "knuckles" effect on the forward rate curve, McCulloch (1975) increased the order of the estimating functions

and used a cubic spline. Mastronikola (1991) considered a more complex cubic spline, and Schaefer (1981) proposed a set of approximating functions derived from Bernstein polynomials to estimate the discount function. Steeley (1991) concluded that by employing B-spline bases, splines can be viewed as a robust alternative to Bernstein polynomials. When it comes to introducing a penalty term for the estimation of these models, model evaluation can be performed basically in the same way as Fisher, Nychka and Zervos (1995). However, the situation is not as straight-forward for the exponential spline models detailed below.

8.2.2 Exponential spline

One of the main criticisms directed at both cubic splines and Bernstein polynomial functions as a choice of approximating function is that these approaches can lead to forward rate curves that exhibit undesirable properties for long maturities. Vasicek and Fong (1982) presented a method known as an exponential spline that can be used to produce asymptotically flat forward curves for long maturities.

One attractive feature of characterizing the discount function as essentially exponential in shape is that this view accords with modern equilibrium theories of the term structure, suggesting that for many plausible stochastic processes the discount function will have an exponential form.

As an estimation technique, the method of Vasicek and Fong (1982) appears somewhat indirect. They suggested applying the following transform to the argument t of the discount function $\delta(t)$.

$$t = -\left(\frac{1}{\alpha}\right)\ln(1-x) \quad \text{for} \quad 0 \leq x < 1.$$

This has the effect of transforming the discount function from an approximately exponential function of t to an approximately linear function of x. Those authors employed a cubic spline to estimate the transformed discount function. In terms of the original variable t, this is equivalent to estimating the discount function by a third-order exponential spline, i.e., between each pair of knot points, $\delta(t)$ takes the form:

$$\delta(t) = \sum_{j=0}^{3} b_j e^{-\alpha j t}.$$

Although Vasicek and Fong (1982) claim to have tested exponential splines successfully, they provided no evidence. Shea (1985) subsequently presented some empirical results and concluded that there is no evidence to support the claim that exponential splines produce more stable estimates of the term structure than polynomial splines. According to Shea (1985), the discount function often deviates from the expected exponential decay form.

Chambers, Carleton and Waldman (1984) advocated an exponential polynomial for modeling the discount function. This approach is equivalent to estimating the term structure using a simple polynomial, i.e.,

$$\delta(t;w) = \exp\left(-\sum_{k=1}^{m} w_k t^k\right). \tag{8.6}$$

The authors analyzed the effects of varying the polynomial degree j from one to five, and concluded that a third- or fourth-degree polynomial is sufficient to approximate the term structure.

Langetieg and Smoot (1989) fitted a cubic B-spline to the term structure of an interest rate, a technique they refer to as the exponential yields model. In this model, $\{\phi_k(t)\}_{k=1}^{m}$ are cubic B-spline bases. From equation (8.2), $\eta(t) = \sum w_k \phi_k(t)$ implies $\delta(t) = \exp(-t\sum w_k \phi_k(t))$. Hence, splining the term structure of an interest rate is equivalent to fitting an exponential spline model to the discount function:

$$\delta(t;w) = \exp\left(-t\sum_{k=1}^{m} w_k \phi_k(t)\right). \tag{8.7}$$

Langetieg et al. found that their model gave better results than that of Vasicek et al., and argued that it is not surprising since the exponential transformation model can be viewed as an approximation of the exponential yields model.

Coleman, Fisher and Ibbotson (1992) treated the forward rate curve, rather than the discount function, as the fundamental variable primarily for mathematical convenience. They approximated the instantaneous forward rate $f(t)$ by a piecewise constant function, as follows.

$$f(t) = f_j \quad \text{for} \quad t_{j-1} < t \leq t_j$$

The forward rate curve is assigned a constant over the periods (t_{j-1}, t_j). It then follows that this is equivalent to using exponential splines to estimate the discount function. For example, for $t_2 < t < t_3$, the discount function would be evaluated as

$$\delta(t) = \exp\{-[f_1 t_1 + f_2(t_2 - t_1) + f_3(t - t_2)]\}.$$

The discount function produced by this method will be continuous, whereas its first derivative will be discontinuous.

Fisher, Nychka and Zervos (1995) suggest the placement of a B-spline on the forward rate curve:

$$f(t) = \sum_{k=1}^{m} w_k \phi_k(t). \tag{8.8}$$

From the equation (8.1), (8.8) can be rewritten as

$$\delta(t;w) = \exp\left\{-\sum_{k=1}^{m} w_k \psi_k(t)\right\} \tag{8.9}$$

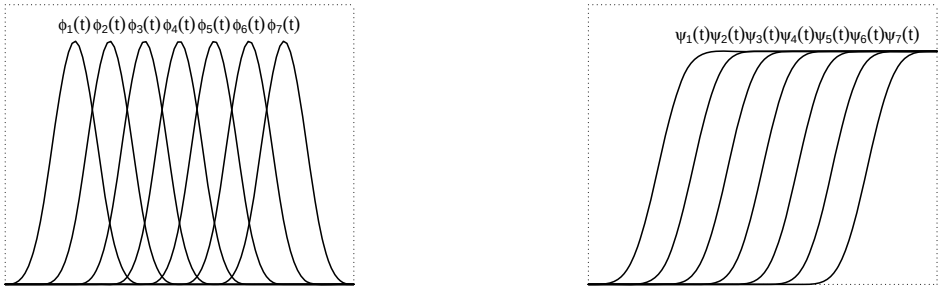


Figure 8.1: Basis functions for the B-spline and its integral

where $\psi_k(t) = \int_0^t \phi_k(s)ds$. The functional form in (8.9) resembles the exponential spline specification (8.7), but the choice of basis function is different. Figure 8.1 shows the definition of a cubic B-spline basis over six equally spaced knots and the corresponding integral.

Combining the equations (8.3) and the structures of the discount function (8.6), (8.7) or (8.9), we obtain a slightly different form of the bond pricing model, as follows.

$$f_E(y_\alpha|t_\alpha;w,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y_\alpha - c'_\alpha \delta(t_\alpha;w))^2}{2\sigma^2}\right\}, \tag{8.10}$$

where $y_\alpha = p_\alpha + a_\alpha$, $\delta(t_\alpha;w) = (\delta(t_1^\alpha;w),...,\delta(t_{L\alpha}^\alpha;w))'$ is the discount vector, and $w = (w_1,...,w_m)'$ is an unknown parameter vector to be estimated from the data.

An important point is that the discount function is not a linear combination of basis functions in fitting the B-spline for either $\eta(t)$ or $f(t)$. We will return to this point when we construct the model selection criterion.

8.2.3 Penalized likelihood

Here we present the maximum penalized likelihood method for estimating the unknown coefficients w and σ^2 in the bond price model (8.5) and (8.10). For parameter estimation in the bond price model (8.5), the maximum likelihood estimate of the weights is given explicitly by $\hat{w} = (B'B)^{-1}B'y$, where $B = (\Phi'_1 c_1,...,\Phi'_n c_n)'$, $y = (y_1,...,y_n)'$. In practice, however, the maximum likelihood method does not yield satisfactory results because the parameter estimates tend to be unstable and lead to overfitting. For example, suppose there is a hump in the estimated discount function, perhaps due to overfitting. No matter how small the hump, the derived forward rate may be negative at some maturity unless the discount function is non-increasing

everywhere. The same problem can also arise in the nonlinear model (8.10) that model only guarantees the positiveness of the discount function.

These instabilities of the estimated yield curves all originate from the ill-posed nature of the regression spline, rather than from any inappropriate choice of the basis function. To avoid overfitting, a penalty term on the smoothness of the unknown coefficients is introduced into the log-likelihood. Specifically, we maximize

$$l_\lambda(w,\sigma^2) = \sum_{\alpha=1}^{n} \log f_{(\cdot)}(y_\alpha|t_\alpha;w,\sigma^2) - \frac{n\lambda}{2}\sum_{j=2}^{m}(\Delta^2 w_j)^2, \tag{8.11}$$

where λ is the smoothing parameter controlling the smoothness of the discount function, and $\Delta w_k = w_k - w_{k-1}$ is the difference operator.

Given λ and m, the unknown parameters w and σ^2 can be obtained as the solution of $\partial l_\lambda(w,\sigma^2)/\partial w = 0$ and $\partial l_\lambda(w,\sigma^2)/\partial\sigma^2 = 0$. If these estimation equations can be solved explicitly in terms of w and σ^2, then by replacing the unknown parameters with the parameter estimates $\hat{w}$ and $\hat{\sigma}^2$, we have the bond price model constructed by the penalized likelihood: $f_{(\cdot)}(y_\alpha|x_\alpha;\hat{w},\hat{\sigma}^2)$.

If the bond model belongs to the class of f_B, it is always possible to define the maximum penalized estimates of w and σ^2 explicitly, as follows. Using an $(m-k)\times m$ difference matrix D_k defined by

$$D_k = \begin{pmatrix} (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k & 0 & \cdots & 0 \\ 0 & (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & (-1)^0{}_kC_0 & \cdots & (-1)^k{}_kC_k \end{pmatrix}.$$

with ${}_nC_k = n!/\{k!(n-k)!\}$, the penalty term in (8.11) can be represented by $\sum_{j=2}^{m}(\Delta^2 w_j)^2 = w'D_2'D_2w$. Hence, for fixed λ and m, the maximum penalized likelihood estimates of w and σ^2 in the bond price model (8.5) are explicitly provided by

$$\hat{w} = \left(B'B + n\beta D_2'D_2\right)^{-1}B'y, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{\alpha=1}^{n}\left\{y_\alpha - c_\alpha'\Phi_\alpha\hat{w}\right\}^2, \tag{8.12}$$

where $\beta = \lambda\sigma^2$.

The greatest advantage of employing an f_B-class model is that because (8.12) involves only simple linear operations, much less computation is required. It is also observed that f_B cases are reduced to a penalized least-squares estimation. The term $\beta D_2'D_2$ derives from the stochastic constraints on the unknown parameters, and can also be viewed as a kind of ridge factor.

Shea (1984) and Steely (1991), criticizing McCulloch's cubic spline specification, made reference to the multicolliearity problem on the regression spline. Although they turned to

another choice of basis function, (8.12) suggests that there exists some possibility that the McCulloch's specification could be stabilized by penalization. This is discussed in sections 8.4 and 8.5 through simulation and empirical analysis.

On the other hand, if the bond models belong to the f_E class, explicit estimators are no longer available. In such a case, a numerical maximization procedure must be invoked. In this article, the Newton-Raphson method based on the first and second derivatives of the penalized likelihood function is adopted for estimation.

An important remaining problem is the criterion by which we should choose the smoothing parameter λ and the number of basis functions m. Here, we derive a criterion for evaluating the bond price model from an information-theoretic point of view. Once the criterion is established, the optimum roughness penalty λ and number of bases m are determined by searching the grid of $(\log\lambda, m)$.

8.3 Information criteria for model evaluation

Suppose we have n bonds $\{(t_\alpha, y_\alpha); t \in R^{L_\alpha}, \alpha = 1, ..., n\}$, where $t_\alpha = (t_1^\alpha, ..., t_{L_\alpha}^\alpha)'$ are the future time points at which the payments for bond α occur. Let y_α be the bond price in which the accrued interest is included, and suppose that y_α is generated from an unknown true distribution $G(y|t)$ with probability density $g(y|t)$. In practical situations, it is difficult to obtain precise information on $g(y|t)$ from a finite number of observations. Additionally, in the context of yield curve estimation, the design matrix tends to be nearly multicollinear, and the implied forward rate curve becomes uncontrollable due to the excess number of basis functions. The maximum penalized likelihood is employed in the present treatment to estimate $f_{(\cdot)}(y|t;\hat{w},\hat{\sigma}^2)$ in (8.5) or (8.10), and then the closeness of $f_{(\cdot)}(y|t;\hat{w},\hat{\sigma}^2)$ to the true model $g(y|t)$ is assess from a predictive point of view.

Let $z_1, ..., z_n$ be another set of observations drawn from $g(y|t)$. Furthermore, let $f_{(\cdot)}(z|T;\hat{w},\hat{\sigma}^2) = \Pi_{\alpha=1}^n f_{(\cdot)}(z_\alpha|t_\alpha;\hat{w},\hat{\sigma}^2)$, and $g(z|T) = \Pi_{\alpha=1}^n g(z_\alpha|t_\alpha)$. We then use the Kullback-Leibler information number (Kullback and Leibler, 1951) as an overall measure of the divergence of $f_{(\cdot)}(z|T;\hat{w},\hat{\sigma}^2)$ from $g(z|T)$, as follows.

$$\begin{aligned} KL\{g, f_{(\cdot)}\} &= E_{G(z|T)}\left[\log\frac{g(z|T)}{f_{(\cdot)}(z|T;\hat{w},\hat{\sigma}^2)}\right] \\ &= E_{G(z|T)}\left[\log g(z|T)\right] - E_{G(z|T)}\left[\log f_{(\cdot)}(z|T;\hat{w},\hat{\sigma}^2)\right]. \end{aligned} \tag{8.13}$$

The model that minimizes the Kullback-Leibler information is then chosen from among different bond price models.

The second term on the right-hand side of (8.13) plays a very important role in model evaluation because the first term depends only on the true model and is not involved in model comparison. It is clear that the minimization of $KL(g, f_{(\cdot)})$ implies the maximization of the expected log-likelihood:

$$E_{G(z|T)}\left[\log f_{(\cdot)}(z|T;\hat{w},\hat{\sigma}^2)\right]. \tag{8.14}$$

A natural estimate of the expected log-likelihood is the log-likelihood itself, $-n\log(2\pi\hat{\sigma}^2)/2 - n/2$, which is obtained by replacing the unknown distribution $G(z|T)$ with the empirical distribution.

As the log-likelihood generally provides a positive bias as an estimator of the expected log-likelihood, we should also perform bias correction. After correction, we have the following information criterion.

$$\mathrm{IC} = -2\sum_{\alpha=1}^{n}\log f_{(\cdot)}(y_\alpha|t_\alpha;w,\sigma^2) + 2nb(G),$$

where the second term $b(G)$ is an estimate of the asymptotic bias defined by

$$b(G) = E_{G(y|T)}\left[\log f_{(\cdot)}(y|T;\hat{w},\hat{\sigma}^2) - E_{G(z|T)}[\log f_{(\cdot)}(z|T;\hat{w},\hat{\sigma}^2]\right].$$

Under the assumption that the specified family of probability distributions does not necessarily contain the true model, Konishi and Kitagawa (1996) derived the asymptotic bias as a function of the empirical influence function of the estimator and the score function of the parametric model.

The influence function of the estimator $\hat{\theta} = (\hat{w}', \hat{\sigma}^2)'$ in a yield curve model $f_{(\cdot)}(z|T;\hat{w},\hat{\sigma}^2)$ is given as follows. Let $\zeta(\cdot)$ be the functional implicitly defined by

$$\int \frac{\partial}{\partial\theta}\left\{\log f_{(\cdot)}(y|t;w,\sigma^2) - \frac{\lambda}{2}w^T D_2' D_2 w\right\}\Bigg|_{\theta=\zeta(G)} dG = 0, \tag{8.15}$$

where G is the joint distribution of (t,y). The estimator $\hat{\theta}$, the solution of the maximizing penalized log-likelihood function, can be written as $\hat{\theta} = \zeta(\hat{G})$, where $\hat{G}$ is the empirical distribution function constructed by the observations. Replacing G in (8.15) with $G_\varepsilon = (1-\varepsilon)G + \varepsilon\delta(t,y)$, where $\delta(t,y)$ is a point of mass at (t,y), and differentiating with respect to ε yields the influence function of the estimator $\hat{\theta} = \zeta(\hat{G})$ in the form

$$\zeta^{(1)}(y|t;G) = J(G)^{-1}\frac{\partial}{\partial\theta}\left\{\{\log f_{(\cdot)}(y|t;w,\sigma^2) - \frac{\lambda}{2}w' D_2' D_2 w\right\}\Bigg|_{\zeta(G)},$$

where

$$J(G) = -\int \frac{\partial^2 \{\log f_{(\cdot)}(y|t;w,\sigma^2) - \frac{\lambda}{2} w' D_2' D_2 w\}}{\partial\theta\partial\theta'} dG.$$

Then, using Theorem 2.1 given in Konishi and Kitagawa (1996), we have the following theorem.

Theorem: *Let $f_{(\cdot)}(y_\alpha|t_\alpha;w,\sigma^2)$ be the yield curve models defined by (8.5) or (8.10), and $f_{(\cdot)}(y_\alpha|t_\alpha;\hat{w},\hat{\sigma}^2)$ be the model constructed via the penalized likelihood method. Supposing that the yield curve model does not necessarily contain the true model generating the data, then the information criterion for evaluating the statistical model $f_{(\cdot)}(y_\alpha|t_\alpha;\hat{w},\hat{\sigma}^2)$ is as follows.*

$$\begin{aligned} \mathrm{GIC}(m,\lambda) &= -2\sum_{\alpha=1}^{n} \log f_{(\cdot)}(y_\alpha|t_\alpha;w,\sigma^2) + 2\mathrm{tr}\left(I_G J_G^{-1}\right) \\ &= n\log(2\pi\hat{\sigma}^2) + n + 2\mathrm{tr}\left(I_G J_G^{-1}\right), \end{aligned} \tag{8.16}$$

where I_G and J_G are $(m+2)\times(m+2)$ matrices given by

$$I_G = \frac{1}{n}\sum_{\alpha=1}^{n} \frac{\partial \psi_{(\cdot)}(y_\alpha|t_\alpha;\hat{w},\hat{\sigma}^2)}{\partial\theta} \frac{\partial \log f_{(\cdot)}(y_\alpha|t_\alpha,\hat{w},\hat{\sigma}^2)}{\partial\theta'}, \quad J_G = -\frac{1}{n}\sum_{\alpha=1}^{n} \frac{\partial^2 \psi_{(\cdot)}(y_\alpha|t_\alpha;\hat{w},\hat{\sigma}^2)}{\partial\theta\partial\theta'},$$

where $\theta = (w',\sigma^2)'$, and $\psi_{(\cdot)}(y_\alpha|t_\alpha;w,\sigma^2) = \log f_{(\cdot)}(y_\alpha|x_\alpha;w,\sigma^2) - \lambda w^T D_2' D_2 w/2$.

Using this theorem, it is possible to evaluate the penalized version of yield curve models based on various functional forms. Here we fit various yield curve models belonging to either the class of $f_B(y_\alpha|t_\alpha;w,\sigma^2)$ in (8.5) or the class of $f_E(y_\alpha|t_\alpha;w,\sigma^2)$ in (8.10). The maximum penalized likelihood estimates $\hat{w}$ and $\hat{\sigma}^2$ of $f_B(y_\alpha|t_\alpha;w,\sigma^2)$ are explicitly given by equation (8.12).

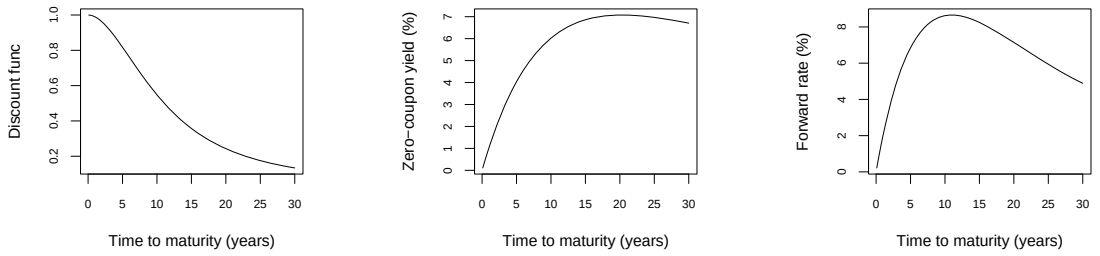
Models belonging to the f_B class can be estimated in the same way because a linear combination of basis functions are commonly used in such models to approximate the discount function. On the other hand, the parameters of models of the f_E class can be obtained only by numerically maximizing the penalized likelihood (8.11). After searching the grid $(\log\lambda, m)$ for each model, we choose a number basis functions m and smoothing parameter λ so as to minimize the information criterion $\mathrm{GIC}(m,\lambda)$ in (8.16). The appendix shows an explicit derivation of $(m+2)\times(m+2)$ matrices I_G and J_G for the two classes of bond price models, f_B and f_E, which encompass all bond price models.

Note that for fitting the smoothing spline for $\delta(t)$, the procedure suggested by Fisher, Nychka and Zervos (1995) is entirely valid, and there is no problem with the use of GCV. However, the use of GCV for fitting an exponential spline or comparing curves to be splined is no longer

theoretically justified. We inevitably resort to GIC when the regression functionals take non-linear forms such as (8.6), (8.7) and (8.9). Of course, GIC can also be constructed in the linear functional case (8.4), making it possible to compare the linear the non-linear models directly.

8.4 Monte Carlo experiments

In previous sections, a method to estimate the bond equation by the penalized likelihood method was proposed, and an information criterion to choose the most appropriate smoothing parameter and number of basis function was given. We also constructed customized GICs that allow us to compare penalized nonlinear regression models when the regression functional is non-linear with respect to the unknown parameters. Hence, we now have tools to determine which yield curve to be splined, whether the discount function $\delta(t)$, the term structure $z(t)$ or the forward rate $f(t)$. This section describes Monte Carlo experiments conducted to investigate whether there are any differences in efficiency depending on the choice of yield curve to be splined.



$\delta(t)$ $\eta(t)$ $f(t)$

Figure 8.2: Discount function, zero coupon yield, and forward rate for the Nelson-Siegel specification

We start by specifying the true functional form of the forward rate curve $f(t)$ in the experiments. The following functional form is set as the true term structure of the instantaneous forward rate.

$$f(t) = \beta_0 + \beta_1 \exp\left(-\frac{t}{\tau}\right) + \beta_2 \left[\frac{t}{\tau}\exp\left(-\frac{t}{\tau}\right)\right]. \tag{8.17}$$

This parameterization was proposed by Nelson and Siegel (1987). We then derive from $f(t)$ the discount functions $\delta(t)$ and zero coupon yield $\eta(t)$ according to the relationships (8.1) and (8.2), giving the following discount function.

$$\delta(t) = \exp\left\{-t\left[\beta_0 + \frac{(\beta_1+\beta_2)\tau}{t}\left(1-\exp\left(-\frac{t}{\tau}\right)\right) - \beta_2\exp\left(-\frac{t}{\tau}\right)\right]\right\}. \tag{8.18}$$

In the simulation, we set $\beta_0 = 0.02$, $\beta_1 = -\beta_0$, $\beta_2 = 0.2$ and $\tau = 10$. The corresponding shapes of the yield curves are illustrated in Figure 8.2.

Given $\delta(t)$, random samples were generated from the true bond price model $p_\alpha = R_\alpha \delta(t_\alpha) + \varepsilon_\alpha$ for $t_\alpha = 30 \times ((\alpha-1)/(n-1))$ and $\alpha = 1, \ldots, n$. For the error term ε_α, we consider an independent normal distribution case, $\varepsilon_\alpha \sim \mathrm{N}(0, \sigma^2)$, where $\sigma = 0.1$, and a mixture of three normal distributions cases:

$$g(\varepsilon_\alpha) = \gamma_1 \frac{1}{\sigma_1}\varphi\left(\frac{\varepsilon_\alpha}{\sigma_1}\right) + \gamma_2 \frac{1}{\sigma_2}\varphi\left(\frac{\varepsilon_\alpha}{\sigma_2}\right) + (1-\gamma_1-\gamma_2)\frac{1}{\sigma_3}\varphi\left(\frac{\varepsilon_\alpha}{\sigma_3}\right), \tag{8.19}$$

where $\varphi(x)$ denotes the density function of the standard normal distribution and the standard deviations are set as $\sigma_1 = 0.1$, $\sigma_2 = 0.2$ and $\sigma_3 = 0.5$. The mixing proportions in (8.19) are set as $\gamma_1 = 0.8$, $\gamma_2 = 0.15$ and $\gamma_3 = 0.05$. The redemption payment R_α is assumed to be 100, considering that the face value of Japanese Governmental Bonds is ¥100.

The maturity interval $[0, 30]$ is divided into equally spaced intervals, and 100 time points are chosen: $\{t_\alpha\}$ with $t_1 = 0$ and $t_n = 30$. These time points are fixed throughout the experiments. The price of artificial zero coupon bonds is then generated according to the bond equation. All yield curves ($\hat{\delta}$, $\hat{\eta}$, and $\hat{f}$) were successfully estimated, regardless of the curve fitted or basis function assumed, and the bias from the true curves (δ, η, f) were measured at the fixed time points. The squares of the biases over maturity were then averaged, and the mean-squared error (MSE) of the ith experiment is defined as

$$D_i^f = n^{-1} \sum_{\alpha=1}^{n} (\hat{f}^{(i)}(t_\alpha) - f(t_\alpha))^2.$$

The overall Monte Carlo mean $\tilde{D}^f = M^{-1}\sum_{i=1}^{M} D_i^f$ for M Monte Carlo trials and its standard deviation were then determined. $\tilde{D}^\delta$ and $\tilde{D}^\eta$ were calculated in the same way.

The results of the Monte Carlo simulations are summarized in Table 8.1. The simulation results were obtained by averaging over $M = 100$ repeated Monte Carlo trials. The standard deviations (SDs) are given in parentheses below the means. In the table, f/B indicates that B-spline bases are placed on the forward rate $f(t)$, as recommended by Fisher, Nychka and Zervos (1995). However, it should be noted that the criterion used here to choose the smoothing parameter λ differs from that employed by Fisher et al. Similarly, η/B and δ/B indicate placement of the B-spline on the zero coupon yield ($\eta(t)$) and the discount function ($\delta(t)$). The specification of the B-spline here follows that of Steeley (1991), and Eilers and Marx (1996), for example, and differs from the definition given in Fisher et al. in that extra knots are placed outside the actual maturity interval and knots are not overlapped at the ends. δ/Cubic

refers to McCulloch's natural cubic spline specification (McCulloch, 1975). Although not stated explicitly, all models are considered in the context of a penalized likelihood approach.

MSE values in the table are read as follows: for example, under independent normal error, the MSE for the estimation of $f(t)$ via forward-rate-splining (f/B) is 7.67. Hence, on average, the bound for the estimated forward rate curve is approximately $\pm$ 2.77 basis points.

The simulation results here support one of the findings in Fisher, Nychka and Zervos (1995); fitting a smoothing spline for the forward rate curve (with B-spline bases) provides the best performance, and it is not recommended to estimate the discount function first and then derive other yield curves. Although the use of GCV in Fisher et al. has no theoretical foundation, their findings were indeed correct on that point. It should be noted, however, that GIC gives a theoretically justified route to compare the various yield curve models using a roughness penalty.

The GIC is used here to choose both λ and m. The introduction of λ was aimed at resolving the ill-posed nature of the regression spline, and choosing the optimal number of basis function m originates from the experience that introducing excess basis functions often leads to an unacceptable shape of the forward rate curve, even though the discount function may be reasonably shaped. On the other hand, in terms of fitting the smoothing spline, one might suspect that choosing the number of basis functions (m) may be unnecessary because the large roughness penalty value may automatically reduce the effective number of parameters.

In light of this argument, similar experiments were performed without choosing m. Instead, a fixed number of basis functions equal to one-third of the sample size was chosen; 33 in the experiments here. The results in Table 8.2 show that all the MSE values are larger than when an appropriate number of basis functions are chosen (Table 8.1). Most notably, choosing the number of basis function results in a significant reduction of the MSE for $f(t)$, particularly when $f(t)$ is splined directly (29.7 $\rightarrow$ 7.67 for normal independent error). It therefore appears that in all cases choosing the number of basis functions improves the estimation.

When the discount function is splined directly, the regression functional is expressed as a linear combination of basis functions, for which the use of GCV is justified. Table 8.3 lists the simulation results for the δ/B and δ/Cubic cases where GCV was used instead of GIC. The upper panel of Table 8.3 corresponds to the results of Table 8.1 (λ and m selected by GCV), and the lower panel corresponds to Table 8.2 (only λ selected). If the spline bases are placed on the discount function, there is not significant difference between GCV and GIC.

An interesting new finding obtained here is that McCulloch's cubic spline specification (δ/Cubic) performs reasonably well. For example, looking at the MSE in the estimation of $f(t)$,

Obs. noise	Independent normal			Mixture normal		
Target func.	$\delta(t)$	$\eta(t)$	$f(t)$	$\delta(t)$	$\eta(t)$	$f(t)$
f/B	2.92	1.36	7.67	3.20	1.46	8.15
(std. err.)	(0.52)	(0.25)	(1.45)	(0.67)	(0.35)	(1.74)
η/B	19.22	4.87	154.96	20.54	5.13	156.32
(std. err.)	(0.53)	(0.23)	(0.92)	(0.52)	(0.25)	(0.96)
δ/B	39.81	10.20	511.10	39.99	12.25	514.17
(std. err.)	(0.56)	(0.46)	(0.82)	(0.59)	(0.66)	(0.85)
δ/Cubic	4.11	1.20	16.63	4.47	1.31	16.81
(std. err.)	(0.54)	(0.31)	(1.16)	(0.57)	(0.34)	(1.18)

Table 8.1: Simulation results for selection of smoothing parameter λ and number of basis functions m

Obs. noise	Independent normal			Mixture normal		
Target func.	$\delta(t)$	$\eta(t)$	$f(t)$	$\delta(t)$	$\eta(t)$	$f(t)$
f/B	4.60	1.80	29.7	5.41	2.04	35.31
(std. err.)	(0.62)	(0.44)	(2.73)	(0.80)	(0.63)	(7.89)
η/B	19.31	5.12	189.62	21.30	6.04	189.72
(std. err.)	(0.54)	(0.24)	(1.12)	(0.74)	(0.61)	(1.41)
δ/B	41.91	11.01	512.02	40.30	12.82	515.33
(std. err.)	(0.58)	(0.50)	(0.82)	(0.61)	(0.67)	(0.85)
δ/Cubic	4.30	1.26	16.88	4.55	1.51	16.98
(std. err.)	(0.48)	(0.42)	(1.16)	(0.57)	(0.45)	(1.41)

Table 8.2: Simulation results for selection of smoothing parameter λ only (m set to one-third the sample size)

δ/Cubic underperforms f/B but is far superior to η/B and δ/B. Considering the results shown in Tables 8.1–8.3, δ/Cubic is the second best in most cases, and sometimes beats f/B in terms of MSE. Although there has been much criticism of McCulloch's specification, most has been in the context of a regression spline. The present simulation results suggest that McCulloch's cubic spline under a penalized likelihood approach may be a better choice than the B-spline provided that the analysis is focused on estimation of the discount function. The performance of δ/Cubic is investigated again in the next section using actual bond data.

8.5 Application to real data

As an illustration of the practical application of the proposed procedure, involving appropriate selection of the smoothing parameter and number of basis functions, the model is applied to the analysis of Japanese governmental bonds observed on September 30, 2002, when 227 in-

Obs. noise	Independent normal			Mixture normal		
Target func.	$\delta(t)$	$\eta(t)$	$f(t)$	$\delta(t)$	$\eta(t)$	$f(t)$
δ/B	38.74	9.92	503.62	38.68	10.20	508.14
(std. err.)	(0.49)	(0.38)	(0.80)	(0.57)	(0.40)	(0.80)
δ/Cubic	4.04	1.18	16.25	4.29	1.25	15.89
(std. err.)	(0.50)	(0.32)	(1.04)	(0.54)	(0.26)	(1.05)
δ/B	39.83	10.57	505.59	39.05	11.19	510.73
(std. err.)	(0.50)	(0.46)	(0.84)	(0.68)	(0.63)	(0.81)
δ/Cubic	4.19	1.21	16.55	4.38	1.29	16.17
(std. err.)	(0.46)	(0.28)	(0.82)	(0.58)	(0.31)	(1.13)

Table 8.3: Simulation results for use of GCV

terest bearing bonds were traded. Data is publicly available on line from the web site of Japan Securities Dealers Association.

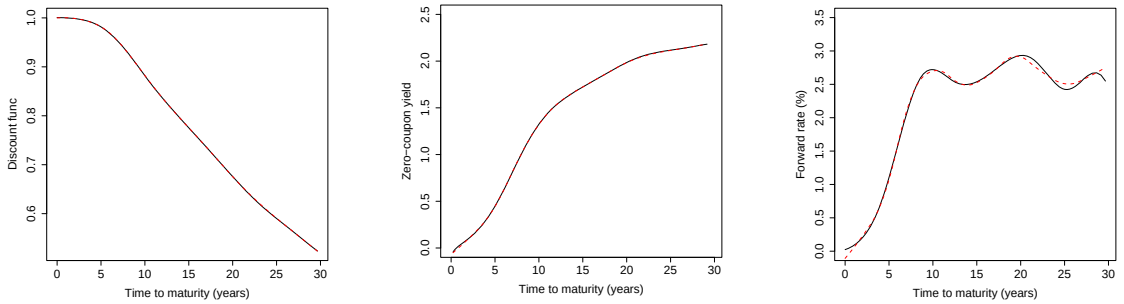
Model	GIC(m,λ)	m	λ	$\hat{\sigma}$
f/B	-534.88	8	1.89×10^{-1}	.073
δ/Cubic	-515.39	10	2.11×10^{1}	.069
δ/Quad	-345.95	8	1.10×10^{-3}	.095
η/B	-195.41	8	6.32×10^{-1}	.126
δ/B	30.66	8	6.94×10^{1}	.189
η/P	885.50	2	3.86×10^{-4}	1.610

Table 8.4: Model selection based on analysis of governmental bonds on September 30, 2002

The fitting results are summarized in Table 8.4. In addition to f/B, η/B, δ/B and δ/Cubic as defined in the previous section, δ/Quad represents estimation of the discount function using a quadratic polynomial spline, and η/P indicates an exponential polynomial model (8.6).

The best model for constructing yield curves for September 30, 2002, is estimation of the forward rate curve with B-spline bases (f/B), for which GIC is minimized by $\hat{m} = 8$ and $\hat{\lambda} = 1.89 \times 10^{-1}$. The second best is δ/Cubic with a GIC of -515.39. Estimation using the exponential polynomial model was apparently unsuccessful. According to the authors' experience, the exponential polynomial model sometimes encounters numerical difficulties and fails to converge during the nonlinear optimization procedure, even under a penalization scheme.

Estimated discount functions, zero coupon yields, and forward rate curves for the best (f/B) and second best (δ/Cubic) models are shown in Figure 8.3. In all three panels, the solid lines indicate the curves derived from f/B, and the dashed lines denote δ/Cubic. The results for the discount function and zero coupon yield curves are almost identical, whereas the results for the forward rate curves tend to differ at longer maturities. This is natural, as even if $\delta(t)$ and



discount function zero coupon yield forward rate

Figure 8.3: Estimated yield curves for September 30, 2002. Solid lines denote f/B, dashed lines denote δ/Cubic.

day	f/B	η/B	δ/B	δ/Cubic	day	f/B	η/B	δ/B	δ/Cubic
2	28.56	289.91	524.19	0	13	0	249.46	436.85	22.76
3	82.44	305.65	534.01	0	17	0	261.42	456.75	47.60
4	55.21	303.33	541.09	0	18	40.41	284.99	484.79	0
5	62.93	317.70	556.35	0	19	28.53	229.02	422.28	0
6	46.00	333.83	564.44	0	20	0	240.32	444.68	1.40
9	7.04	286.79	496.04	0	24	0	203.45	398.05	0.18
10	15.61	308.41	522.30	0	25	0	197.93	388.71	7.20
11	0	279.87	479.78	3.15	26	0	276.11	487.39	12.38
12	0	253.60	439.37	13.49	30	0	339.48	565.54	19.50
# of days attained min GIC:						9	0	0	9

Table 8.5: Results of fitting models on all business days in September 2002. Difference from minimum GIC values are reported.

$\eta(t)$ are observationally equivalent, the forward rate $f(t)$ may differ because it involves the first derivative of the discount function.

To confirm the findings of the simulation studies, the models are compared on a day-by-day basis for every business day in September 2002. In this analysis, the models are limited to f/B, η/B, δ/B and δ/Cubic for brevity. The results are tabulated in Table 8.5. The entries are GIC values minus the minimum GIC attained for the day (the table should be read row-wise). For example, for September 2, 2002, δ/Cubic is the most accurate of the four models considered, with GIC$=-500.71$. Differences from this minimum GIC are also reported in the table, and all entries are non-negative. That is, 0 indicates the best model, and larger values represent poorer performance.

For the 18 business days, f/B attains the minimum GIC for 9 days. Though this good performance of f/B looks consistent with the findings of the simulation studies, δ/Cubic occupies

the first place for the same number of days. Moreover, it is worth mentioning that δ/Cubic yields a relatively small difference from the minimum GIC even when it does not provide the best performance. η/B and δ/B consistently come in third and forth place respectively. From our numerical experience, the authors feel that δ/B habitually fits worse at longer maturities in spite of penalization. The framework considered in this article is how to fit the best curve based on the traded bond data in one day. Introduction of smoothness constraints on the adjacent business days will form the basis of future research topics.

8.6 Stability Analysis by Bootstrapping

In this section we investigate the stability of estimated curves by bootstrap method. At first, we highlight the effect of regularization. In the regularized method we employ Gaussian radial basis functions where the model is estimated by penalized likelihood and model selection is done by GIC. On the other hand, in regression spline without regularization, we just assume McCulloch's natural cubic spline, and the model is estimated by least squares. Because McCulloch (1971, 1975) does not clearly address how to choose the number of basis functions, we adopt cross-validation here.

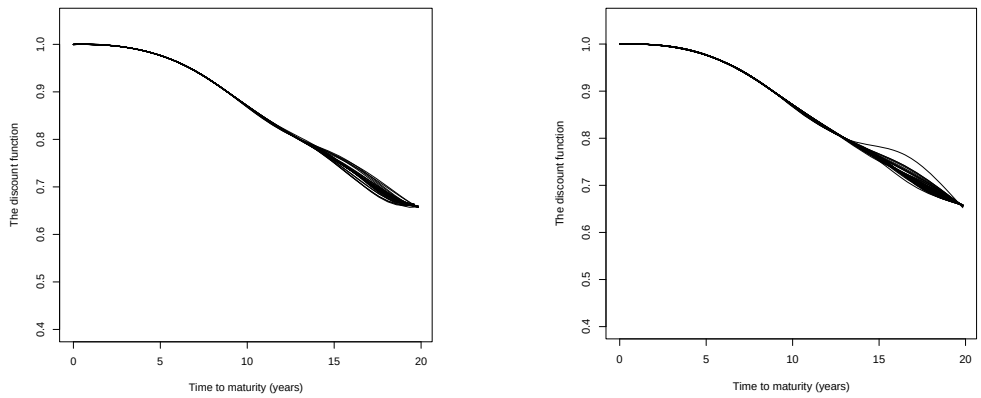
Analyzed data is 78 coupon bearing Japanese Governmental Bond traded on August 22, 2001. Figure 8.4 – 8.6 show the 100 curves estimated on 100 bootstrap sample generated from the coupon data on August 22. Here bootstrapping was done on case-based because of the possible heteroscedasticity with respect to the maturity. Take a look at Figure 8.4 might give us an impression that there seems no significant difference between the regularized results and the least square results at least in discount functions. However, it is clear from Figure 8.5 (a) that the forward rate curve is extremely stable in penalized likelihood estimation accompanied by GIC. Conversely, McCulloch's natural cubic spline without regularization sometimes yields wild curves as is apparent in Figure 8.5 (b). As we know that the McCulloch's cubic spline performs well if it is properly penalized, here we can confirm the importance of suitable penalization, and choice of basis functions is not a primary problem in yield curve estimation.

Yield curves are to be redrawn every business day. It can happen that bonds with longer maturity were not actively traded and the number of data (where the cash flow occurs) may not be sufficient. In Figures 8.5 (a) and (b), the trajectories of both curves are plumping out at longer maturities, which indicates the estimated interest rates are not very stable at longer maturities. Therefore, we should fit 'hard' model in the sparse area, otherwise there always lies a danger of overfitting, so much more if we could not choose the number of basis functions in

a reasonable way. Seemingly a good fit in the discount function does not always guarantee the goodness of fit in the forward rate,and even as small bump in the discount function may cause a big fluctuation in the forward rate.

Finally, we mention an example to show that the choice of basis function sometimes matters. In this example (using bond data on July 14, 1995) the case of McCulloch's cubic spline is also regularized, and the number of basis function and the roughness penalty parameter are also determined by GIC. Hence, only the difference between two methods is just the choice of basis functions.

Same as in the previous experiment, there is no significant difference in the estimated discount functions. But, as is clear from Figure 8.7 (a), the trajectories of forward rate based on Gaussian radial basis function are much tighter (thus superior). Figure 8.7 (b) shows that the forward rate curves estimated by cubic spline is not stable even at the maturities less than 5 years. One greatest merit of Gaussian radial basis function over cubic or B-spline is it has dispersion parameter that controls the hardness of basis functions. As is demonstrated, even when the number of data is limited, Gaussian radial basis function with penalized likelihood and GIC can automatically search the optimal dispersion of basis functions.

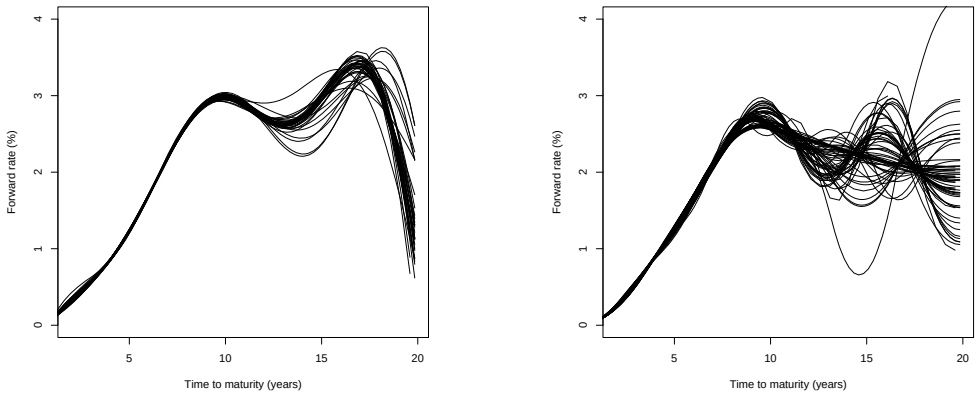


(a) (b)

Figure 8.4: Discount functions by bootstrap replication (August 22, 2001). (a): Regularized (b) : Without regularization

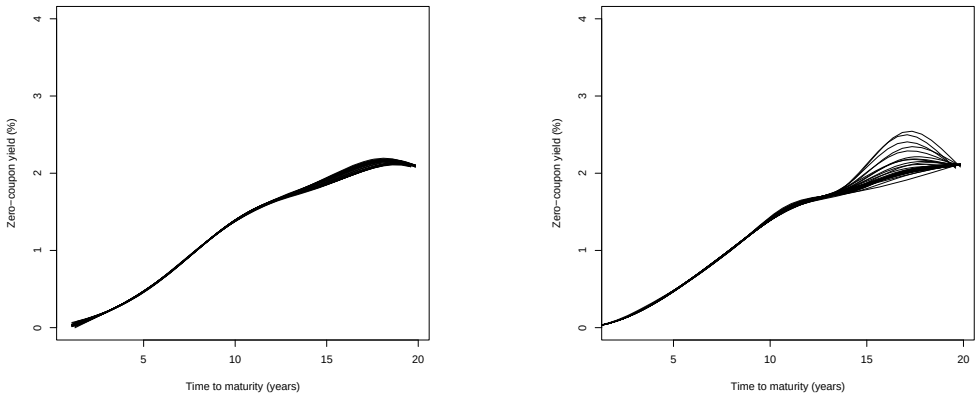
8.7 Conclusion

A penalized likelihood approach was proposed for estimation of the term structure of interest rates from a set of coupon data. In the penalized likelihood approach, the method for choosing



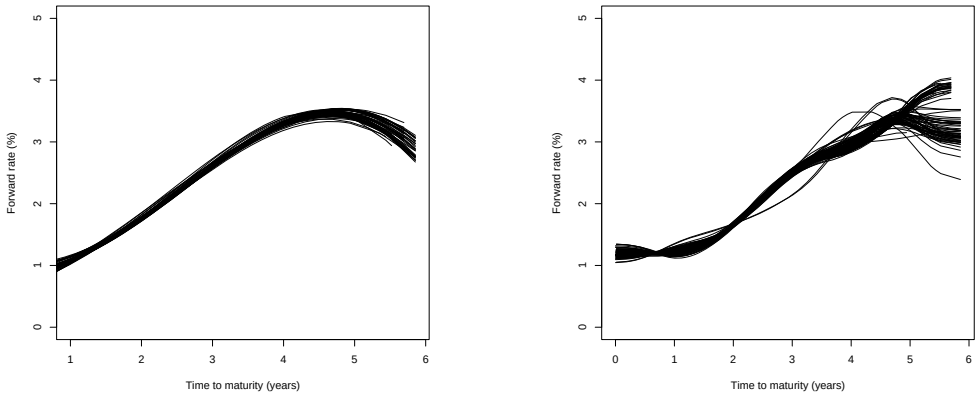
(a) (b)

Figure 8.5: Forward rate by bootstrap replication (August 22, 2001). (a): Regularized (b) : Without regularization



(a) (b)

Figure 8.6: Zero coupon yield by bootstrap replication (August 22, 2001). (a): Regularized (b) : Without regularization



(a) (b)

Figure 8.7: Forward rate by bootstrap replication (July 14, 1995). (a): Gaussian RBF (b) : Natural Cubic Spline

the smoothing parameter is important. If the nonparametric regression functional is linear in its parameters, then GCV can be used. However, if we want to spline the term structure or the forward rate, we inevitably have to estimate exponential spline models, for which GCV loses its theoretical basis. It was shown that a customized version of the generalized information criterion (GIC) can be constructed even for nonlinear spline problems. Monte Carlo studies and analysis of real data clearly showed that B-splining the forward rate with a roughness penalty provides the most accurate estimation of yield curves, confirming the findings of Fisher, Nychka and Zervos (1995) by a theoretically valid route. It was also verified that choosing the optimal number of basis functions rather than letting the single (smoothing) parameter control the number of bases reduces the estimation error. It was also shown that McCulloch's cubic spline specification works reasonably well if it is estimated with a roughness penalty.

Appendix

(a) Assuming a linear combination of a set of m underlying basis functions for the discount function, we derive the bond price model based on a linear combination of basis functions $f_B(y_\alpha|t_\alpha;w,\sigma^2)$ in (8.5). The unknown parameters w and σ^2 are estimated by the penalized likelihood method. Then we have the bond price model $f_B(y_\alpha|t_\alpha;\hat{w},\hat{\sigma}^2)$, which depends on the number of bases m and the value of the smoothing parameter. Using the theorem, we have information criteria for evaluating the statistical model $f_B(y_\alpha|t_\alpha;\hat{w},\hat{\sigma}^2)$, as follows.

$$\mathrm{GIC}(m,\lambda) = n\log(2\pi\hat{\sigma}^2) + n + 2\mathrm{tr}\left(I_G J_G^{-1}\right). \tag{8.20}$$

where I_G and J_G are $(m+2)\times(m+2)$ matrices given by

$$I_G = \frac{1}{n\hat{\sigma}^2}\begin{pmatrix} B'\Lambda/\hat{\sigma}^2 - \lambda K\hat{w}1_n' \\ p' \end{pmatrix}(\Lambda B, \hat{\sigma}^2 p),$$

$$J_G = \frac{1}{n\hat{\sigma}^2}\begin{pmatrix} B'B + n\hat{\sigma}^2\lambda K & B'\Lambda 1_n/\hat{\sigma}^2 \\ 1_n'\Lambda B/\hat{\sigma}^2 & n/2\hat{\sigma}^2 \end{pmatrix}.$$

Here, $\Lambda = \mathrm{diag}\,[y_1 - c_1'\Phi_1\hat{w}, ..., y_n - c_n'\Phi_n\hat{w}]$, $1_n = (1,1,...,1)'$, and p is an n-dimensional vector with ith element $(y_i - c_i'\Phi_i\hat{w})^2/2\hat{\sigma}^4 - 1/2\hat{\sigma}^2$. We choose m and λ as minimizers of the GIC.

(b) We next derive alternative yield curve models based on (1) exponential polynomial (Chambers, Carleton and Waldman (1984)), (2) exponential spline (Langetieg and Smoot (1989)), and (3) B-spline bases, placed on the forward rate (Fisher, Nychka and Zervos (1995)). The unknown parameters w and σ^2 are estimated by the penalized likelihood method. Then we have the bond price model $f_E(y_\alpha|t_\alpha;\hat{w},\hat{\sigma}^2)$, which depends on the number of bases m and the value of the smoothing parameter. Using the theorem, we provide information criteria for evaluating the statistical model $f_E(y_\alpha|t_\alpha;\hat{w},\hat{\sigma}^2)$ as follows.

$$\mathrm{GIC}(m,\lambda) = n\log(2\pi\hat{\sigma}^2) + n + 2\mathrm{tr}\left(I_G J_G^{-1}\right). \tag{8.21}$$

where I_G and J_G are $(m+2)\times(m+2)$ matrices given by

$$I_G = \frac{1}{n\hat{\sigma}^2}\begin{pmatrix} \Phi'\Lambda/\hat{\sigma}^2 - \lambda K\hat{w}1_n' \\ q' \end{pmatrix}(\Lambda\Phi, \hat{\sigma}^2 q),$$

$$J_G = \frac{1}{n\hat{\sigma}^2}\begin{pmatrix} \Phi'\Phi - D + n\hat{\sigma}^2\lambda K & \Phi'\Lambda 1_n/\hat{\sigma}^2 \\ 1_n'\Lambda\Phi/\hat{\sigma}^2 & n/2\hat{\sigma}^2 \end{pmatrix},$$

respectively. Here, $\Lambda = \mathrm{diag}\,[y_1 - c_1'\delta(t_1;\hat{w}), ..., y_n - c_n'\delta(t_n;\hat{w})]$, $1_n = (1,1,...,1)'$, and q is an n-dimensional vector with ith element $(y_i - c_i'\delta(t_i;\hat{w}))^2/2\hat{\sigma}^4 - 1/2\hat{\sigma}^2$, Φ and D are $n\times m$ and $m\times m$ matrices with (i,j)th elements Φ_{ij} and D_{ij}, as given by

(1) : Exponentialpolynomial

$$\Phi_{ij} = \sum_{k=1}^{L_i} c_k\delta(t_k^i;\hat{w})t_k^j,\ \ D_{ij} = \sum_{\alpha=1}^{n}[(y_\alpha - c_\alpha'\delta(t_\alpha;\hat{w}))(\sum_{k=1}^{L_\alpha} c_\alpha\delta(t_k^\alpha;\hat{w})(t_k^\alpha)^{i+j})],$$

(2) : Exponentialspline

$$\Phi_{ij} = \sum_{k=1}^{L_i} c_i\delta(t_k^i;\hat{w})f_j(t_k)t_k,\ \ D_{ij} = \sum_{\alpha=1}^{n}[(y_\alpha - c_\alpha'\delta(t_\alpha;\hat{w}))(\sum_{k=1}^{L_\alpha} c_\alpha\delta(t_k^\alpha;\hat{w})f_i(t_k^\alpha)f_j(t_k^\alpha)(t_k^\alpha)^2)],$$

(3) : Splineingforwardrate

$$\Phi_{ij} = \sum_{k=1}^{L_i} c_i\delta(t_k^i;\hat{w})f_j(t_k),\ \ D_{ij} = \sum_{\alpha=1}^{n}[(y_\alpha - c_\alpha'\delta(t_\alpha;\hat{w}))(\sum_{k=1}^{L_\alpha} c_\alpha\delta(t_k^\alpha;\hat{w})f_i(t_k^\alpha)f_j(t_k^\alpha)].$$

The values of the smoothing parameter λ and the number of basis functions m are determined as the minimizers of the GIC.

Chapter 9

Modeling Periodicity in High Frequent Financial Data

9.1 Introduction

Modeling with high frequent data in finance has become very popular since the middle of 1990's. One reason is due to the availability of such data itself. For example, a set of high frequent data was provided by a data service vendor to foster the empirical research in this field, and the conference and special issue of an academic journal were accompanied to the project. See Baillie and Dacorogna (1997), the preface of the special issue in Journal of Empirical Finance. Another reason is model development suitable for the analysis of high frequent data. Since the pioneering work of Engle and Russel (1995, 1998) which proposed a useful class of models called the Autoregressive Conditional Duration Models, it was followed by many empirical research and model ramifications such as Threshold ACD by Zhang, Russell and Tsay (1999) or Log-ACD by Bauwens and Giot (2000) among others.

The basic idea of the ACD model is to express the mean conditional duration time as a linear function of past duration time and mean conditional duration time. But we cannot fit the ACD model directly to the duration times calculated from original high frequent data because of intra-day periodic pattern. As the transactions or quotes in markets will be proportional to the actual economic activity and market time, the pattern of quote occurrence inevitably exhibits intra-day periodicity.

Section 9.2 describes commonly employed two-step procedure to remove intra-day periodicity prior to the fitting of the ACD model. The purpose of this paper is to propose simultaneous determination method for the models of intra-day periodicity and of the cluster effects of financial transactions. Section 9.3 reports the framework and worked example of conditional intensity approach to point processes developed by Ogata (1983a, b), which has been applied in

statistical modeling of seismicity analysis.

In closing this section we briefly state the data analyzed throughout this paper. The data is tick-by-tick data of yen-dollar exchange rate collected from July 10, 1997 to August 3, 1997. The data were captured from the screen of Telerate terminal, and then the obvious mis-ticks and ticks with extremely high volatility were removed prior to the analysis. The intervals corresponding to weekends (from Friday 21:00 to Sunday 21:00 GST) are removed to make the data handling easy. One feature of foreign exchange rate data is that each data record is not always an actual transaction but just a quote. Secondly, the market is basically open 24 hours a day. But the transaction will be proportional to the actual economic activity and market time, the pattern of quote occurrence inevitably exhibits intra-day periodicity.

9.2 Removing Periodicity in the ACD models

Since Engle and Russell (1998), a class of models called autoregressive conditional duration (ACD hereafter) model has received much attention, and applied to analyze the characteristics of financial markets using high frequent data. In this section we briefly review the basic framework of the ACD models, and subsequently the common recipes for removing intra-day periodicity from high frequent financial data.

Suppose t_i stands for i-th transaction. It is assumed that $0 = t_0 < t_1 < \cdots < t_n$ where n is the total number of transactions observed in the interval we consider. Then the duration time between trades is defined as $X_i = t_i - t_{i-1}$. Here we introduce the conditional mean duration time ψ_i as follows;

$$\psi_i = E[X_i | X_{i-1}, \ldots, X_1; \theta_1]. \tag{9.1}$$

θ_1 is a set of unknown parameters that describe the structure of the conditional mean duration. The basic assumption in the ACD model is, if we correctly specify ψ_i, then i-th duration time divided by ψ_i will follow some distribution of mean 1 identically and independently, namely,

$$\frac{X_i}{\psi_i} = \varepsilon_i \sim \mathscr{G}(\theta_2) \tag{9.2}$$

for some distribution $\mathscr{G}$ with mean 1. θ_2 is a set of parameters of a distribution $\mathscr{G}$. The equation (9.1) shows a general framework, and the models used actually are the one that resembles the specification of volatility in ARCH and GARCH models. If ψ_i is defined as

$$\psi_i = a_0 + \sum_{m=1}^{p} a_m X_{i-m} + \sum_{n=1}^{q} b_n \psi_{i-n}, \tag{9.3}$$

then the duration time is said to follow $ACD(p,q)$ model.

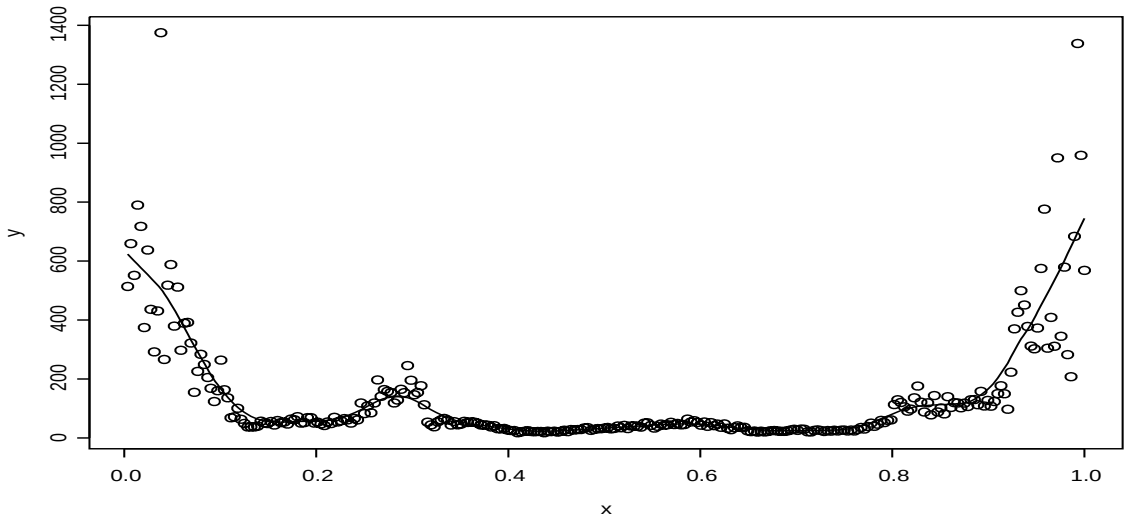


Figure 9.1: 5-minutes-bin-wise average duration time and its spline smoothing

Before the fitting of the ACD models, the removal of intra-day periodicity is usually considered. The most commonly employed recipe seems the one described as follows. Divide one day into, say, the collection of five minutes bins. Every duration time is classified into anyone of the bins. Then the mean duration time is calculated by an sample average within a bin. Now we have 288 equally spaced data points. Then perform smoothing spline for the series of mean duration. If we divide each duration time by the corresponding value implied by the fitted spline curve, then we assume that the intra-day periodicity is removed. See Engle and Russell (1995), Zhang, Russell and Tsay (1999) for example.

Figure 9.1 shows the results of applying the above stated procedure to our data. We used the function *supsmu* in S-PLUS. Both ends of the figure correspond to 21:00 GST. Figure 9.2 plots the reciprocal of duration time and the reciprocal of the estimated smoothing spline curve in Figure 9.1. Curve in Figure 9.2 clearly shows the intra-day pattern of the quote intensity of yen-dollar exchange market. Three major humps correspond to the market time of Tokyo, London and New York respectively. Engle and Russell (1995) and following works in this field generally assert this two-stage approach works fine at least for the stock market in the United States. The aim of this paper is to examine such assertion by another modeling methodology, a class of parametric models for point processes that will be estimated by the method of maximum likelihood.

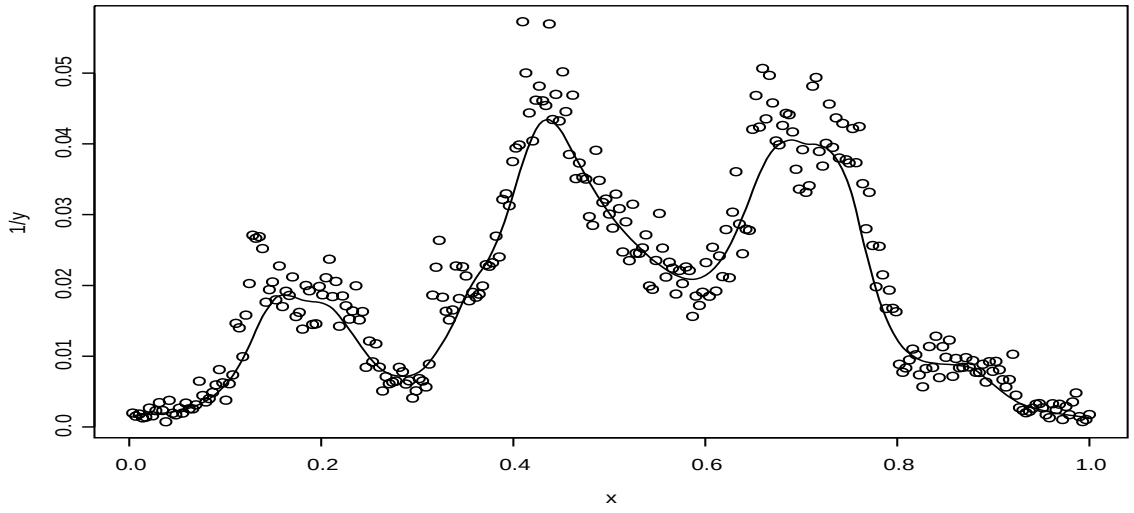


Figure 9.2: Reciprocal of 5-minutes-bin-wise average duration time and the reciprocal of smoothed curve estimated in Figure 1

9.3 Specification of Point Processes via Conditional Intensity

This section describes a general framework for analyzing various point process data by way of the conditional intensity function. An excellent review of modeling point processes mainly applied for seismicity analysis can be found in Ogata (1999).

Consider a series of events $\{t_i; 0 < t_1 < t_2 < \cdots\}$ occurring at random on the real half line $(0,\infty)$. To describe this sequence we take time differences, $X_i = t_i - t_{i-1}$, between consecutive points, and we then consider as a positive valued stochastic process. When this is distributed independently and identically, we call this a renewal process, and if its marginal is exponential distribution, then it is the stationary Poisson process.

Let $N(a,b)$ be the number of points in an interval (a,b) on the real line such that this is a nonnegative integer valued random variable. Consider a prediction of an event occurring on a small time interval. Namely, assume a point process on the real half-line $(0,\infty)$, and divide it into small intervals of length δ. Then we get a stochastic process $\{\xi_k\}$, where $\xi_k = N[(k-1)\delta, k\delta)$ is k-th random variable on the subinterval $[(k-1)\delta, k\delta)$. If δ is small enough, we may assume that $\{\xi_i\}$ is a binary process. If the point process now considered is a stationary Poisson, then $\{xi_k\}$ is identically and independently distributed Bernoulli series. But in general, the joint probability of the sequence os determined by a sequence of conditional probabilities $P\{\xi_k = 1 | \xi_1, \ldots, \xi_{k-1}\}$, $k = 1, 2, \ldots$, namely on the history of events.

Then as a derivative of the conditional probability with respect to the time, the *conditional*

intensity function $\lambda(t|\mathscr{F}_t)$ is defined by

$$P\{N(t,t+\delta)=1|\mathscr{F}_t\}=\lambda(t|\mathscr{F}_t)\delta+o(\delta),$$

or

$$\lambda(t|\mathscr{F}_t)=\lim_{\Delta\to 0}P\{\text{an event in } [t,t+\Delta]|\mathscr{F}_t\}/\Delta \tag{9.4}$$

where $\mathscr{F}_t$ is a set of observations until time t, including the history of the occurrence times of the events $\mathscr{H}_t=\{t_i;t_i<t\}$. It is known that the conditional intensity completely characterizes the corresponding point process (Liptzer and Shiryayev, 1978). Clearly a constant conditional intensity provides a stationary Poisson process. If the conditional intensity is independent of the history but dependent only on the occurrence time t, like $\lambda(t|\mathscr{F}_t)=\nu(t)$ for any nonnegative function $\nu(t)$ of t, then this means a non-stationary Poisson process which is employed in subsection 9.3.1.

There are many interesting classes of point processes which are defined by certain conditional intensity functions. One of them is Hawkes' self-exciting process described by

$$\lambda(t|\mathscr{F}_t)=\mu+\int_0^t g(t-s)dN_s=\mu+\sum_{t_i<t}g(t-t_i).$$

(See Hawkes, 1971, Hawkes and Oakes, 1974.) This functional form resembles the autoregressive model in time series analysis. In this model the expectation of an event occurring is given by a linear combination of past occurrences, where the so-called impulse response function $g(\cdot)$ measures the weights of such combinations. This type of component is considered in subsection 9.3.2.

Given a set of occurrence data $t_1,t_2,\ldots,t_n$ in an observed time interval $[0,T]$ and a parameterized conditional intensity $\lambda_\theta(t|\mathscr{F}_t)$, the likelihood is written in the form

$$L_T(\theta|t_1,t_2,\ldots,t_n;0,T)=\left\{\prod_{i=1}^n\lambda_\theta(t_i|\mathscr{F}_{t_i})\right\}\exp\left\{-\int_0^T\lambda_\theta(t|\mathscr{F}_t)dt\right\}.$$

The maximum likelihood estimate of θ is the value of the parameter vector which maximizes the logarithm of above,

$$\log L_T(\theta|t_1,t_2,\ldots,t_n;0,T)=\sum_{i=1}^n\log\lambda_\theta(t_i|\mathscr{F}_{t_i})-\int_0^T\lambda_\theta(t|\mathscr{F}_t)dt. \tag{9.5}$$

If the second term in the right hand side of (9.5) can be expressed analytically in θ, then the gradients of the log-likelihood function can be easily obtained. In such a case the maximization of the function can be carried out by using a standard nonlinear optimization technique.

Assume that we have to choose the best model among proposed competing models. The Akaike Information Criterion (Akaike, 1974),

$$\text{AIC} = -2\times(\text{maximum log-likelihood})+2\times\ (\text{number of parameters})$$

is very suitable for such model comparisons. A model with a smaller AIC is considered to be a better fit.

9.3.1 Modeling with Cyclic Part

We start with a non-stationary Poisson model, and specify the conditional intensity function as follows.

$$\lambda_\theta(t|\mathscr{F}_t) = a_0 + P_J(t) + C_K(t). \tag{9.6}$$

The second term on the right-hand side of (9.6) represents the evolutionary trend where

$$P_J(t) = \sum_{j=1}^{J} a_j \phi_j(t/T),\ 0 < t < T \tag{9.7}$$

T is the total length of the observed interval and $\phi_j(\cdot)$ is a polynomial of order j. If the data span of the high frequent data we consider is very long, then we may expect this term could capture the evolutionary change of the quote/trade intensity. The third term in (9.6) is the Fourier series

$$C_K(t) = \sum_{k=1}^{K} \{b_{2k-1}\cos(2k\pi t/T_0) + b_{2k}\sin(2k\pi t/T_0)\}, \tag{9.8}$$

which stands for cyclic effects with a given fixed cyclic length T_0. If we set T_0 equal to the length of a day, then this cyclic term represents the intra-day periodicity.

Sometimes it is useful to employ exponentiated version of trend-cycle model

$$\lambda_\theta(t|\mathscr{F}_t) = \exp\{a_0 + P_J(t) + C_K(t)\} \tag{9.9}$$

instead of (9.6) to guarantee the positivity of the conditional intensity function. This case can be viewed as the log-linear modeling of the conditional intensity function. In this case, the analytic calculation of the integral (the second part of the right-hand side in (9.5)) is not generally feasible except in special cases. See Lewis (1970) for example. However, for the slowly varying intensity model such as the exponential rate for a trend, numerical integration well approximates the integral term so that the maximum likelihood becomes feasible.

We fit the model (9.9) to the high frequent data from yen-dollar exchange market described in the section 9.2. Because the sample period covers only 16 days ($T = 16$), it is expected that any significant evolutionary movement cannot be conceived from this data. Actually, the models with polynomial term of order one, two yields bigger value of AIC, hence those are not reported here. After restricting the model class to constant plus trigonometric functions, we search the

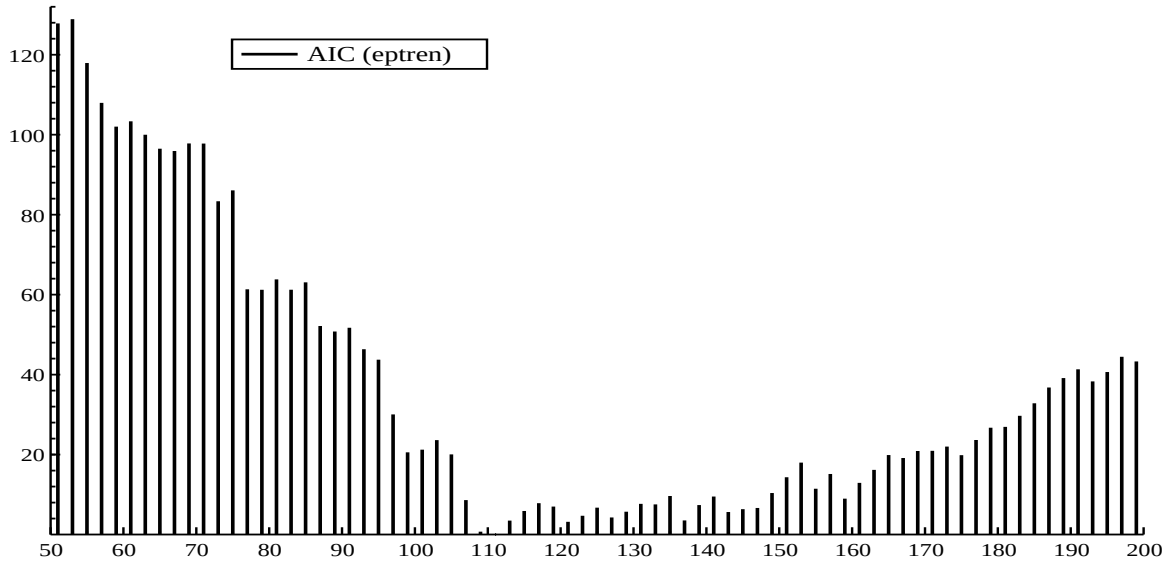


Figure 9.3: AIC difference vs. order of trigonometric function

best model up to the maximum order 1440. Figure 9.3 shows that minimum AIC is attained by the model order 111, which means a constant term and 55 pairs of trigonometric functions are required to describe the intensity of yen-dollar quote occurrence during July 1997. (The ordinates in the Figure 9.3 are the differences of AICs between each model and the minimum AIC model.) The highest frequency in the best model is approximately 26 minutes. Figure 9.4 shows the estimated cyclic intensity function. Both ends of horizontal axis are 21:00 GST, and the length of a day is set to one ($T_0 = 1$) in this analysis. Three major humps in the graph correspond to the market time of Tokyo, London and New York respectively.

9.3.2 Modeling with Cyclic and Cluster Part

The class of models we consider in this section follows;

$$\lambda_\theta(t|\mathscr{F}_t) = a_0 + P_J(t) + C_K(t) + \sum_{t_i<t} g_M(t-t_i). \tag{9.10}$$

The last term in (9.10) expresses the clustering effects such as quotes or trades invoked by precedent ones. The function $g_M(x)$ measures the increase in clustering due to a quote/trade. We may call this function a response function of an event, and parameterize it as

$$g_M(x) = \sum_{m=1}^{M} c_m x^{m-1} e^{-\alpha x}, \tag{9.11}$$

see Ogata and Akaike (1982). This model is used to examine the existence of each component by the comparison of AIC values among a possible set of configurations of (J,K,M). Different

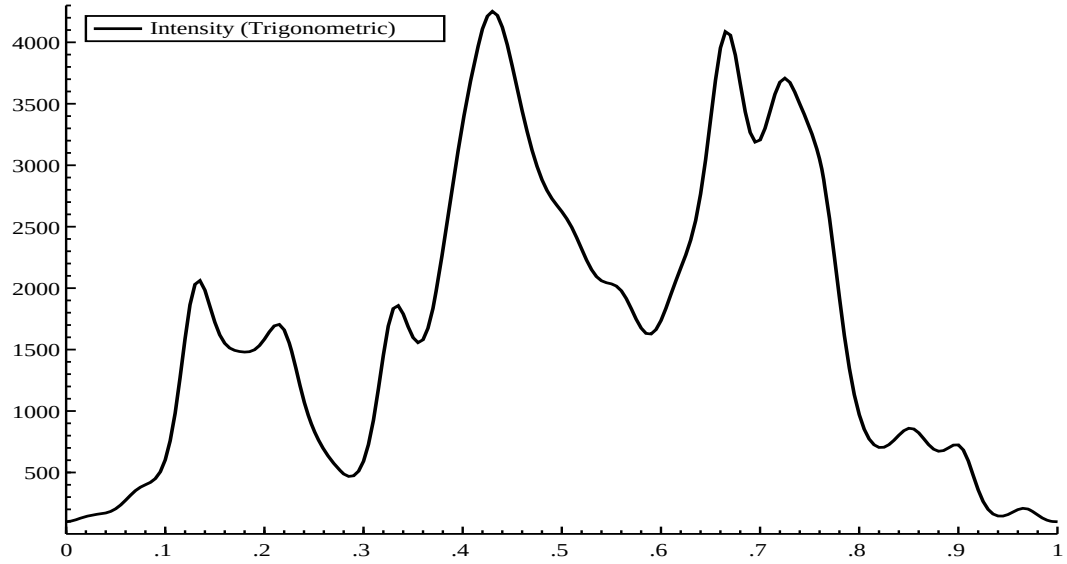


Figure 9.4: Intensity function induced by the minimum AIC model

M	K	J	AIC
2	0	0	-370461.21
2	6	0	-370667.19
2	10	0	-370857.12
2	20	0	-370910.87
2	30	0	-370921.32
2	34	0	-370924.39
2	40	0	-370918.18

Table 9.1: AIC values for several configurations

from (9.9), trend and cyclic function is not exponentiated in (9.10). It is reasonable if we expect the most of intensity will be absorbed into the cluster effect component. Then the seasonal pattern should become smoother, which lead to the ordinary trigonometric function fitting rather than exponentially transformed one.

After trying various configurations we find that $M = 2$ and $J = 0$ generally lead to smaller values of AIC. After fixing M and J to the above mentioned value, we repeat the maximum likelihood estimation changing the order of the Fourier series, K. Table 9.1 and Figure 9.5 shows the model of $K = 34$ attains minimum AIC, -370924.39. Estimated parameter values of the second order ($M = 2$) Laguerre polynomial in (9.11) are $\hat{\alpha} = 321.88$, $c_1 = 203.65$ and $c_2 = 45.36$, respectively. Because the cluster effect is almost concentrated around the origin, the intensity implied by the cluster effect part is drawn with the horizontal axis ranges only over $[0, 0.2]$, which approximately corresponds to 4 hours and 50 minutes. Numerical integration of

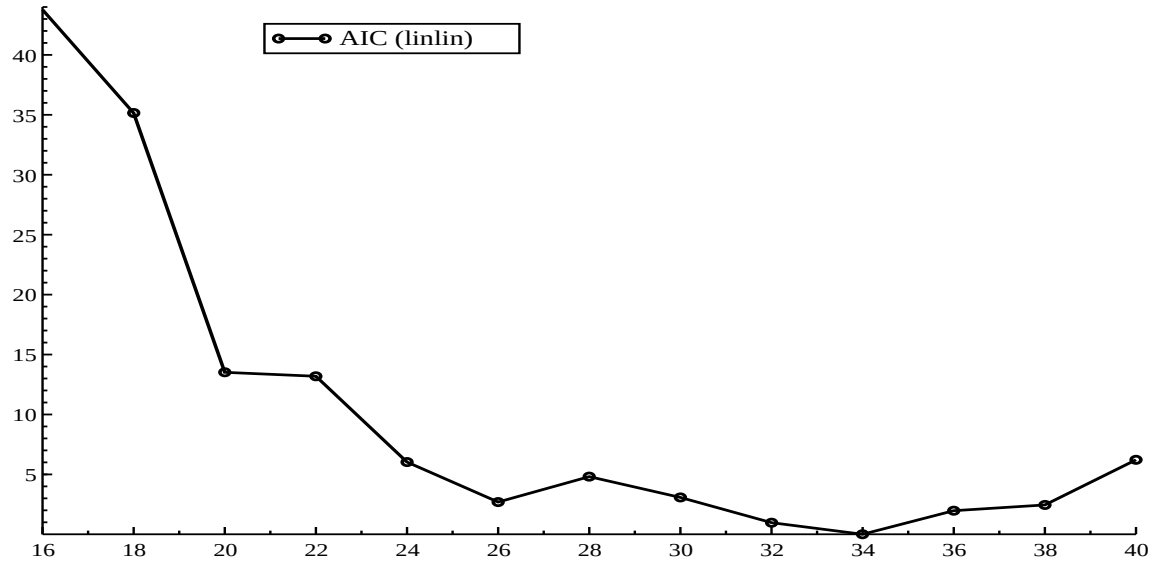


Figure 9.5: AIC difference vs. order of trigonometric function

the function with the estimated parameters over $(0, 0.00485)$ produces almost $1/2$. Hence if a quote occurred right after the previous quote, the probability of the next quote occurrence is very high.

M	K	J	AIC
0	110	0	-368454.04
2	0	0	-370461.21
2	34	0	-370924.39

Table 9.2: Comparison of AIC between the best model and other extreme models

Figure 9.7 shows the shape of the cyclic part of the intensity function if we change the order of Fourier series K from 2 to 40. Model order increases horizontally, and the most bottom-right figure corresponds to $K = 40$ case. The cases of $K = 16, 24, 34, 40$ are displayed in Figure 9.8. Compared to the Figure 9.4, we find the shape is almost alike but the scale of ordinates is very different. In Table 9.2 we report the AIC values of the best model and two extreme cases. Introducing the second order Laguerre polynomial is far better than the Fourier series modeling where the difference of AIC is more than 2000. Adding periodicity to the pure cluster effect model results in improvement of AIC by 463.18.

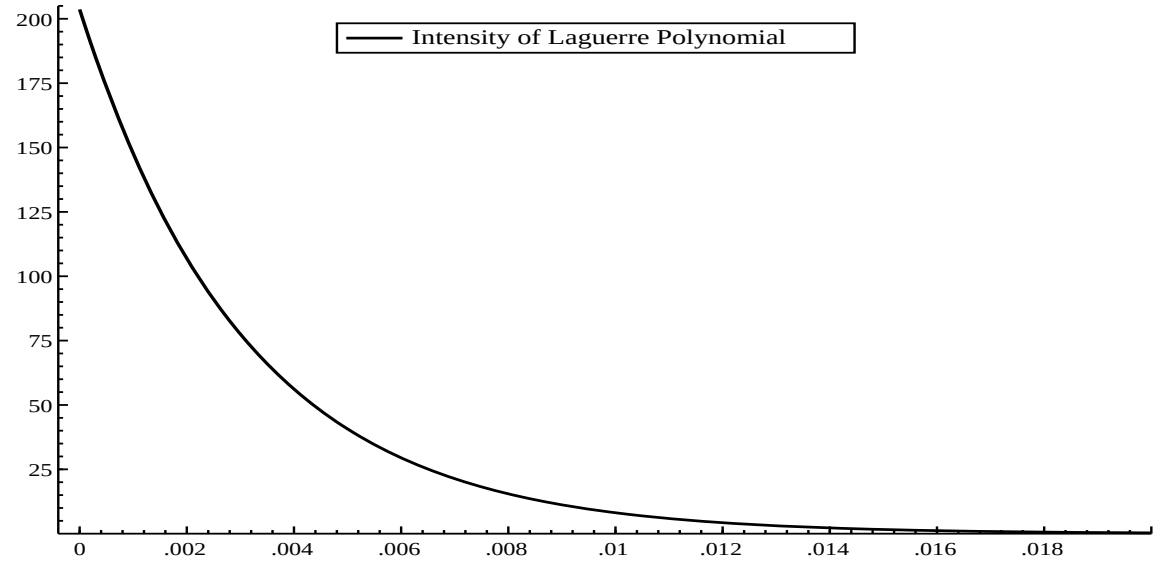


Figure 9.6: Intensity implied by the estimated Laguerre polynomial

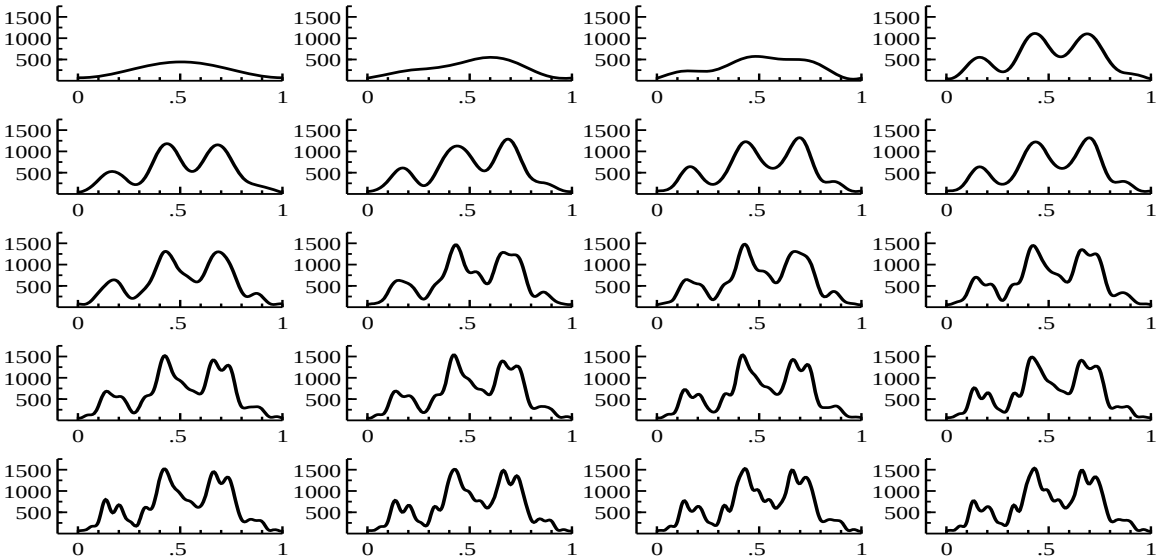


Figure 9.7: Cyclic part of the intensity function induced by the minimum AIC model. The order of trigonometric functions (K) varies from 2 to 40.

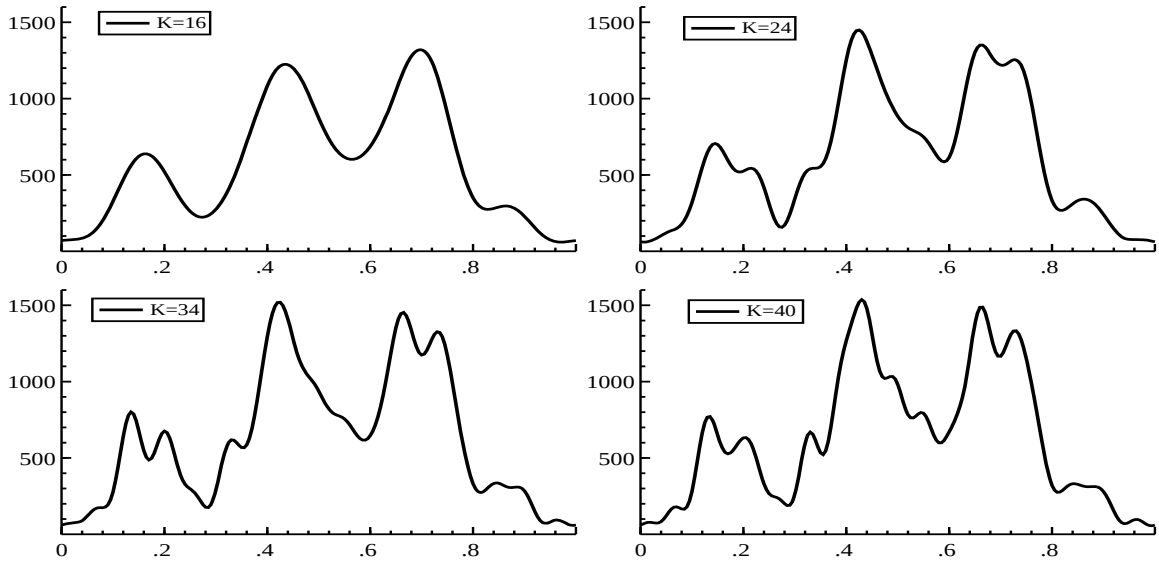


Figure 9.8: Cyclic part of the intensity function induced by the minimum AIC model.

9.4 Conclusion

A class of models for high frequent financial data is proposed. It turns out that incorporating cluster effect is essential to the modeling, which is parallel to the commonly used duration time based autoregressive models. The estimated shapes of the cyclic component of the intensity function shows the simple-minded application of built-in spline smoothing method may lead to smoother intra-day periodicity pattern.

Bibliography

[1] Ahamad, B. (1967), An analysis of crimes by the method of principal components, *Applied Statistics*, **16**, 17–35.

[2] Ahn, S. K. and G. C. Reinsel (1988) Nested Reduced Rank Autoregressive Models for Multiple Time Series, *Journal of American Statistical Association*, **83**, 849–856.

[3] Akaike, H. (1973). Information theory and extension of the maximum likelihood principle. In *Second International Symposium of Information Theory*, N. B. Petrov and F. Czaki (eds.), 267–281, Budapest: Akademiai Kiado.

[4] Akaike, H. (1974a) A new look at the statistical model identification, *IEEE Trans. Autom. Control*, **19**, 716–723.

[5] Akaike, H. (1974b). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes, *Ann. Inst. Statist. Math.*, Vol. 26, 363–387.

[6] Akaike, H. (1980a) Likelihood and the Bayes procedure, in *Bayesian Statistics*, J. M. Bernardo and M. H. De Groot and D. V. Lindley and A.F.M. Smith (eds.), University Press Valencia, 143–166.

[7] Akaike, H.(1980b). Seasonal adjustment by a Bayesian modeling, *J. Time Series Anal.*, Vol. 1, 1–13.

[8] Akaike, H. and M. Ishiguro (1980c), *BAYSEA, A Bayesian seasonal adjustment program*, Computer Science Monographs, The Institute of Statistical Mathematics.

[9] Akaike, H., Ozaki, T., Ishiguro, M., Ogata, Y., Kitagawa, G., Tamura, Y.-H., Arahata, E., Katsura, K. and Tamura, Y. (1985). TIMSAC-84 Part 1, *Computer Science Monographs No. 23*, The Institute of Statistical Mathematics, Tokyo.

[10] Akaike, H., Ozaki, T., Ishiguro, M., Ogata, Y., Kitagawa, G., Tamura, Y.-H., Arahata, E., Katsura, K. and Tamura, Y. (1985). TIMSAC-84 Part 2, *Computer Science Monographs No. 24*, The Institute of Statistical Mathematics, Tokyo.

[11] Akaike, H. and Kitagawa, G. (eds.) (1999). *The Practice of Time Series Analysis*, Springer-Verlag, New York.

[12] Anderson, B. D. O. and Moore, J. B. (1979). *Optimal filtering*, Prentice-Hall, New Jersey.

[13] Anderson, N., F. Breedon, M. Deacon, A. Derry and G. Murphy. (1996) *Estimating and Interpreting the Yield Curve*, John Wiley and Sons, Chichester.

[14] Anderson, T. W. (1963) The Use of Factor Analysis in the Statistical Analysis of Multiple Time Series, *Psychometrika*, **28**, 1–25.

[15] Anderson, T. W. (1984) *An Introduction to Multivariate Statistical Analysis*, 2nd ed., John Wiley and Sons.

[16] Ansley, C.F. and W.E. Wecker (1984) On dips in the spectrum of a seasonally adjusted time series, comment on "Issues involved with the seasonal adjustment of economic time series" by W.R. Bell and S.C. Hillmer, *Journal of Business and Economic Statistics*, 2, 323–324.

[17] Baillie, R. T. and M. M. Dacorogna (1997) High frequent data in finance, *Journal of Empirical Finance*, **4**, 69–72.

[18] Bauwens, L. and P. Giot (2000) The logarithmic ACD model: an application to the bid-ask quote process of three NYSE stocks, *Annales d'Economie et de Statistique*, **60**, 117–149.

[19] Bell, W.R. (1984) Signal extraction for nonstationary time series, *Annals of Statistics*, 13, 646–664.

[20] Bell, W. R. (1987). A note on overdifferencing and the equivalence of seasonal time series models with monthly means and models with $(0,1,1)_{12}$ seasonal parts when $\Theta = 1$, *Journal of Business and Economic Statistics*, Vol. 5, 383–387.

[21] Bollen, K. A. (1989). *Structural equations with latent variables*, John Wiley & Sons, New York.

[22] Bozdogan, H. (ed.) (1994). Proceedings of the first US/Japan conference on the frontiers of statistical modeling: an informational approach, Klewer Academic Publishers.

[23] Box, G. E. P., Hillmer, S. C. and Tiao, G. C. (1979). Analysis and modelling of seasonal time series, *NBER-Census Conference on Seasonal Analysis of Economic Time Series*, ed. Arnold Zellner, Washington D.C., 309–334.

[24] Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control, Revised ed.*, Holden-Day.

[25] Box, G. E. P. and G. C. Tiao (1977), A Canonical Analysis of Multiple Time Series, *Biometrika*, **64**, 355–365.

[26] Brillinger, D. R. (1981) *Time Series: Data Analysis and Theory*, (Expanded Edition) Holden-Day, San Francisco.

[27] Buja, A., Hastie, T. and Tibshirani, R. (1989), Linear Smoothers and Additive Models (with discussion), *Ann. Statist.*, Vol. 17, 453–555.

[28] Burman, J.P. (1980) Seasonal adjustment by signal extraction, *J. R. Statist. Soc. A*, 143, 321–337.

[29] Burridge, P. and Wallis, K. F. (1984). Unobserved-components models for seasonal adjustment filters, *J. Bus. Econ. Stat.*, Vol. 2, 350–359.

[30] Chambers, D. R., Carleton, W. T. and Waldman, D. W. (1984) A New Approach to Estimation of the Term Structure of Interest Rates, *Journal of Financial and Quantitative Analysis*, **19**, No. 3, 233–252.

[31] Cleveland, W.P. and G.C. Tiao (1976) Decomposition of seasonal time series: a model for the Census X-11 program, *Journal of the American Statistical Association*, 71, 581–587.

[32] Coleman, T. S., Fisher, L. and Ibbotson, R. G. (1992) Estimating the Term Structure of Interest Rates from Data that Include the Prices of Coupon Bonds, *The Journal of Fixed Income*, September, 85–116.

[33] Craven, P and Wahba, G. (1979) Smoothing Noisy Data with Spline Functions, *Numerishe Mathmatik*, **31**, 377 - 403.

[34] de Boor, C. and Lynch, R. E. (1966), On Splines and Their Minimum Properties, *J. Math. and Mechanics*, Vol. 15, 953–969.

[35] de Jong, P. (1991) The diffuse Kalman filter, *Annals of Statistics*, **19**, 1073–1083.

[36] Elton, E. J. and M. J. Grüber (1973) Estimating the dependence structure of share prices — Implication for portfolio selection, *Journal of Finance*, **28**, 1203–1232.

[37] Engle, R. (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation, *Econometrica*, **50**, pp. 987–1007.

[38] Engle, R. F. and C. W. J. Granger (1987) Cointegration and Error Correction: Representation, Estimation and Testing, *Econometrica*, **55**, 251–276.

[39] Engle, R. and J. Russel (1995) Forecasting the frequency of changes in quoted foreign exchange prices with the autoregressive conditional duration model, *Working Paper, University of San Diego*.

[40] Engle, R. and J. Russel (1998) Autoregressive conditional duration: a new approach for irregularly spaced transaction data *Econometrica*, **66**, 1127–1162.

[41] Findley, D.F., B.C. Monsell, W.R. Bell, M.C. Otto and B.C. Chen (1996) New capabilities and methods of the X12-ARIMA seasonal adjustment program, Unpublished manuscript.

[42] Fisher, M. E, Nychka, D and Zervos, D. (1995) Fitting the Term Structure of Interest Rates with Smoothing Splines, Federal Reserve Bank Finance and Economics Discussion Paper 95-1, January.

[43] Frankel, J., G. Galli and A. Giovannini (eds.), *The microstructure of foreign exchange markets*, University of Chicago Press, 1996.

[44] Franses, P. H. (1996a) Recent advances in modelling seasonality, *Journal of Economic Surveys*, Vol. 10, 299–345.

[45] Franses, P. H. (1996b) *Periodicity and stochastic trends in ecomnomic time series*, Oxford University Press.

[46] Gersch, W. M. (1992), Smoothness Priors, in *New Directions in Time Series Part II*, eds. D. Brillinger, P. Caines, J. Geweke, E. Parzen, M. Rosenblatt and M. S. Taqqu, The IMA Volumes in Mathematics and Its Applications Vol. 46, Springer-Verlag, 113–146.

[47] Gersch, W. M. and Kitagawa, G. (1983). The prediction of time series with trends and seasonalities, *J. Bus. Econ. Stat.*, Vol. 1, 253–264.

[48] Gersch, W. M. and Kitagawa, G. (1988), Smoothness Priors in Time Series, in *Bayesian Analysis of Time Series and Dynamic Systems*, ed. J. C. Spall, Marcel Dekker, New York, 431–476.

[49] Geweke, J. F. (1977). The Dynamic Fator Analysis of Economic Time Series Model. D. J. Aigner& A. S. Goldberger (Eds.), *Latent Variables in Socio-Economic Models*, 365–383, North-Holland, Amsterdam.

[50] Geweke, J. F. and K. Singleton (1981) Latent Variable Models for Time Series: A Frequency Domain Approach with an Application to the Permanent Income Hypothesis, *Journal of Econometrics*, **17**, 287–304.

[51] Geweke, J. F. (1996) Bayesian Reduced Rank Regression in Econometrics, *Journal of Econometrics*, **75**, 121–146.

[52] Ghysels, E., Lee, H. S. and Noh, J. (1994) Testing for unit roots in seasonal time series, *Journal of Econometrics*, Vol. 62, 415–442.

[53] Gibbons, M. R., S. A. Ross and J. Shanken (1989) A test of the efficiency of a given portfolio, *Econometrica*, **57**, 1121–1152.

[54] Gonzalez, P. and Moral, P. (1995) An analysis of the international tourism demand in Spain, *International Journal of Forecasting*, Vol. 11, 233–251.

[55] Good, I. J. and Gaskins, J. R. (1980), Density Estimation and Bump Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data, *J. Amer. Statist. Assoc.*, Vol. 75, 42–73.

[56] Gourieroux, C. , A. Monfort and A. Trognon (1984) Pseudo Maximum Likelihood Methods: Theory, *Econometrica*, **17**, 287–304.

[57] Gourieroux, C. and Scaillet, O. (1994) Estimation of the Term Structure from Bond Data, CREST Working Papers, No. 9415.

[58] Graybill, F. A. (1969) *Introduction to matrices with applications in statistics*, Belmont, California, Wadsworth.

[59] Green, P. J. and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughnes Penalty Approach*, Chapman Hall, London.

[60] Grether, D.M. and M. Nerlove (1970), Some properties of “optimal" seasonal adjustment, *Econometrica*, 38, 682–703.

[61] Gu, C. (1990), Adaptive Spline Smoothing in Non-Gaussian Regression Models, *J. Amer. Statist. Assoc.*, Vol. 85, 801–807.

[62] Gu, C. (1992), Penalized Likelihood Regression: A Bayesian Analysis, *Statistica Sinica*, Vol. 2, 255–264.

[63] Gu, C. (1993a), Smoothing Spline Density Estimation: A Dimensionless Automatic Algorithm, *J. Amer. Statist. Assoc.*, Vol. 88, 495–504.

[64] Gu, C. (1993b), Penalized Likelihood Hazard Estimation: Algorithms and Examples, in *Statistical Decision Theory and Related Topics V*, eds. S. S. Gupta and J. O. Berger, Springer-Verlag.

[65] Gu, C. and Qiu, C. (1993), Smoothing Spline Density Estimation: Theory, *Ann. Statist.*, Vol. 21, 217–234.

[66] Nerlove, M., Grether, D. M. and Carvalho, J. L. (1979), *Analysis of Economic Time Series: A Synthesis*, Academic Press, San Diego.

[67] Greville, T. N. T. (1957), On smoothing a finite table: a matrix approach, *SIAM J. Appl.*, 5, 137–154.

[68] Hamilton, J.D. (1994) *Time Series Analysis*, Princeton University Press, Princeton, New Jersey.

[69] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.

[70] Hannan, E. J., Terrell, R. D. and Tuckwell, N. (1970). The seasonal adjustment of economic time series, *International Economic Review*, Vol. 11, 24–52.

[71] Harrison, P. J. and Stevens, S. C. (1971). A Bayesian approach to short-term forecasting, *Operational Research Quarterly*, Vol. 22, 341–362.

[72] Harvey, A. C. (1984). A unified view of statistical forecasting procedures (with discussion), *Journal of Forecasting*, Vol. 3, 245–283.

[73] Harvey, A. C. (1985). Trend and cycles in macroeconomic time series, *Journal of Business and Economic Statistics*, Vol. 3, 216–227.

[74] Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*, Cambridge University Press, Victoria, Australia.

[75] Harvey, A. C. and Todd, P. H. J. (1983). Forecasting economic time series with structural and Box-Jenkins models (with discussion), *Journal of Business and Economic Statistics*, Vol. 1, 299–315.

[76] Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman and Hall, London.

[77] Hastie, T. J. and Tibshirani, R. J. (1993), varying-Coefficient Models, *J. Roy. Statist. Soc. B*, Vol. 55, 757–796.

[78] Haugen, R. A. and N. L. Baker (1996) Commonality in the determinants of expected stock returns, *Journal of Financial Economics*, **41**, 401–439.

[79] Hawkes, A. G. (1971) Point spectra of some mutually exciting point processes, *J. Roy. Statist. Soc. (B)*, **33**, 438–443.

[80] Hawkes, A. G. and D. A. Oakes (1974) A cluster process representations of self-exciting process, *J. Appl. Probab.*, **11**, 493–503.

[81] Hillmer S. C. and Tiao, G. C. (1982). An ARIMA-model-based approach to seasonal adjustment, *J. Amer. Stat. Assoc.*, Vol. 77, 63–70.

[82] Holtzman, W. H. (1962), Methodological Issues in P-technique, *Psychological Bulletin*, **59**, 243–256.

[83] Hox, J. J. and I. G. G. Kreft (1994), Multilevel Analysis Methods, *Sociological Methods and Research*, **22**, 283–299.

[84] Huber, P. J. (1981). *Robust Statistics*, Wiley, New York.

[85] Hylleberg, S. (1986) *Seasonality in regression*, Academic Press, London.

[86] Hylleberg, S., Engle, R. F., Granger, C. W. J. and Yoo, B. S. (1990) Seasonal integration and cointegration, *Journal of Econometrics*, Vol. 44, 215–238.

[87] Hylleberg, S. and Jørgensen, C. and Sørensen, N. K. (1993) Seasonality in macroeconomic time series, *Empirical Economics*, Vol. 18, 321–335.

[88] Jazwinski, A. H. (1970) *Stochastic Processes and Filtering Theory*, Academic Press, New York.

[89] Johansen, S. (1988) Statistical Analysis of Cointegration Vectors, *Journal of Economic Dynamics and Control*, **12**, 231–254.

[90] Johanasen, S. (1995) *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press, Oxford.

[91] Jöreskog, K. G. (1978), Structural Analysis of Covariance and Correlation Matrices, *Psychometrika*, **43**, 443–477.

[92] Jöreskog, K. G. and D. Sörbom (1998), *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*, Scientific Software International, Chicago.

[93] Kariya, T. (1986), *Theory and Practice of Econometric Analysis (in Japanese)*, Toyo Keizai, Tokyo.

[94] Kariya, T. (1993), *Quantitative methods for portfolio analysis; MTV approach*, Kluwer Academic, Dordrecht.

[95] Kawasaki, Y. and Franses, P. H. (1999) A model selection approach to detect seasonal unit roots, Research Memorandum No. 741, The Institute of Statistical Mathematics.

[96] Kawasaki, Y. and Franses, P. H. (1996) A model selection approach to detect seasonal unit roots, Tinbergen Institute discussion paper series TI 96–180/7, Tinbergen Institute, Erasmus University Rotterdam.

[97] Kawasaki, Y., S. Sato and S. Tachiki (1998) Smoothness prior approach to estimate large scale multifactor models, *ISM Research Memorandum No. 714*, The Institute of Statistical Mathematics, Tokyo.

[98] Kawasaki, Y., S. Sato and S. Tachiki (2000) Vector-valued multiple regression model with time varying coefficients and its application to predict excess stock returns, *Proceedings of IEEE/IAFE/INFORMS Conference on Computational Intelligence for Financial Engineering*, pp. 162–165.

[99] Kimmerdolf, G. S. and Wahba, G. (1970a), A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing Splines, *Ann. Math. Statist.*, Vol. 41, 495–502.

[100] Kimmerdolf, G. S. and Wahba, G. (1970b), Spline Functions and Stochastic Processes, *Sankhya*, Ser. A, Vol. 32, 173–180.

[101] Kimmerdolf, G. S. and Wahba, G. (1971), Some Results on Tchebycheffian Splines, *J. Math. Anal. Appl.*, Vol. 33, 82–95.

[102] Kitagawa, G. (1981). A nonstationary time series model and its fitting by a recursive filter, *J. Time Series Anal.*, Vol. 2, 103–116.

[103] Kitagawa, G. (1987). Non-Gaussian state space modeling of non stationary time series (with discussion), *J. Amer. Stat. Assoc.*, Vol. 82, 1032–1063.

[104] Kitagawa, G. (1988), Numerical Approach to Non-Gaussian Smoothing and Its Applications, *20th Interface Sympposium Computer Science and Statistics*, 379–388.

[105] Kitagawa, G. (1989). Non-Gaussian seasonal adjustment, *Computers & Mathematics with Applications*, Vol. 18, 503–514.

[106] Kitagawa, G. (1991), A Nonlinear Smoothing Method for Time Series Analysis, *Statistica Sinica*, Vol. 1, 371–388.

[107] Kitagawa, G. (1993), A Monte Carlo Filtering and Smoothing Method for Non-Gaussian Nonlinear State Space Models, *Proceedings of the 2nd U.S.-Japan Joint Seminar on Statistical Time Series Analysis*, 110–131.

[108] Kitagawa, G. (1994). The two-filter formula for smoothing and an implementation of the Gaussian sum smoother, *Ann. Inst. Statist. Math.*, Vol. 46, 605–623.

[109] Kitagawa, G. (1996). Monte carlo filter and smoother for non-Gaussian nonlinear state space models, *Journal of Computational and Graphical Statistics*, Vol. 5, 1–25.

[110] Kitagawa, G. and Gersch, W. M. (1984). A smoothness-prior state space modeling of time series with trend and seasonality, *J. Amer. Stat. Assoc.*, Vol. 79, 378–389.

[111] Kitagawa, G. and Gersch, W. M. (1985a). A Smoothness Priors Long AR Model Method for Spectral Estimation, *IEEE Trans. on Automatic Control*, AC-30, 57–65.

[112] Kitagawa, G. and Gersch, W. M. (1985a). A Smoothness Priors Time Varying AR Coefficient Modeling of Nonstationary Time Series, *IEEE Trans. on Automatic Control*, AC-30, 48–56.

[113] Kitagawa, G. and Gersch, W. M. (1996). *Smoothness priors analysis of time series, Lecture Notes in Statistics 116*, Springer-Verlag, New York.

[114] Kloek, T. and G. M. de Wit (1961), Best linear and best linear unbiased index numbers, *Econometrica*, **29**, 602–616.

[115] Konishi, S. and Kitagawa, G. (1996) Generalised Information Criteria in Model Selection, *Biometrika*., **83**, 875–890.

[116] Koopman, S. J., Harvey, A. C., Doornik, J. A. and Shephard, N. (2000). *STAMP: structural time series analyser, modeller and predictor*, Timberlake Consultants Ltd, Harrow.

[117] Kohn, R. and Ansley, C. F. (1987), A New Algorithm for Spline Smoothing Based on Smoothing A Stochastic Process, *SIAM J. Sci. Statist. Comput.*, Vol. 8, 33–48.

[118] Kohn, R. and Ansley, C. F. (1988), Smoothness Priors and Optimal Interpolation and Smoothing, in *Bayesian Analysis of Time Series and Dynamic Systems*, ed. J. C. Spall, Marcel Dekker, New York.

[119] Kullback, S. & Leibler, R. A. (1951) On Information and Sufficiency, *Ann. Math. Statist*, **22**, 79–86.

[120] Kuwana, Y., M. Susai and Y. Kawasaki (2000) How the scheduled macroeconomic announcements influence foreign exchange markets (*in Japanese*), *Proceedings of the Institute of Statistical Mathematics*, **48**, 213–227.

[121] Langetieg, T. C. and Smoot, J. S. (1989) Estimation of the Term Structure of Interest Rates, *Research in Financial Services*, **1**, 181–222.

[122] Leonard, T. (1978), Density Estimation, Stochastic Processes and Prior Information (with discussion), *J. Roy. Statist. Soc. B*, Vol. 40, 113–146.

[123] Lewis, P. A. W. (1970) Remarks on the theory, computation and application of the spectral analysis of series of events, *J. Sound. Vib.*, **12**, 353–375.

[124] Lindley, D. V. and Smith, A. F. M. (1972), Bayes estimate for the linear model, *J. R. Statist. Soc.*, B, 34, 1–41.

[125] Liptzer, R. S. and A. N. Shiryayev (1978) *Statistics of Random Processes II: Applications*, Springer-Verlag, New York.

[126] Litzenberger, R. H. and Rolfo, R. (1984) An International Study of Tax Effects on Government Bonds, *Journal of Finance*, March, 1–22.

[127] MacLean, C. J. (1974) Estimation and testing of an exponential polynomial rate function within the non-stationary Poisson process, *Biometrika*, **61**, pp. 81–86.

[128] Maekawa, K. (1994) Prewhitened unit root test, *Economics Letters*, Vol. 45, 145–153.

[129] Maravall, A. (1985) On structural time series models and the characterization of components, *Journal of Business and Economic Statistics*, Vol. 3, 350–355.

[130] Mastronikola, K. (1991) Yield Curves for Gilt-Edged Stocks: A New Model, Bank of England Discussion Paper (Technical Series), No. 49.

[131] McCulloch, J. H. (1971) Measuring the Term Structure of Interest Rates, *Journal of Business*, **44**, No. 1 (January), 19–31.

[132] McCulloch, J. H. (1975) The Tax-Adjusted Yield Curve, *Journal of Finance*, **30** (June), 811–30.

[133] Molenaar, P. C. M. (1985), A Dynamic Factor Model for the Analysis of Multivariate Time Series, *Psychometrika*, **50**, 181–202.

[134] Molenaar, P. C. M., J. G. de Gooijer and B. Schmitz (1992), Dynamic Factor Analysis of Nonstationary Multivariate Time Series, *Psychometrika*, **57**, 333–349.

[135] Nash, M. and Wahba, G. (1974), Generalized Inverses in Reproducing Kernel Spaces: An Approach to Regularization of Linear Operator Equations, *SIAM J. Math. Anal.*, Vol. 5, 974–987.

[136] Nelson, C. R. and Siegel, A. F. (1987) Parsimonious Modeling of Yield Curves, *Journal of Business*, **60**(4), 473–89.

[137] Nerlove, M. (1964) Spectral analysis of seasonal adjustment procedures, *Econometrica*, 32, 241–286.

[138] Nychka, D. (1981), Bayesian “Confidence" Intervals for Smoothing Splines, *J. Amer. Statist. Assoc.*, Vol. 83, 1134–1143.

[139] Ogata, Y. (1983a) Estimation of the parameters in the Modified Omori Formula for Aftershock Frequencies by the Maximum Likelihood Procedure, *J. Phys. Earth*, **31**, 115–124.

[140] Ogata, Y. (1983b) Likelihood Analysis of Point Processes and its Applications to Seismological Data, *Bull. Int. Statist. Inst.*, **50**, Book 2, 943–961.

[141] Ogata, Y. and K. Katsura (1985) EPTREN in TIMSAC-84 part 2 (Akaike et al. eds.), *Computer Science Monographs No. 23*, The institute of Statistical Mathematics, pp. 187–197.

[142] Ogata, Y. and K. Katsura (1985) LINLIN in TIMSAC-84 part 2 (Akaike et al. eds.), *Computer Science Monographs No. 23*, The institute of Statistical Mathematics, pp. 198–211.

[143] Ogata, Y. and H. Akaike (1982) On linear intensity models for mixed doubly stochastic Poisson and self-exciting point process, *J. Royal Statist. Soc. B*, **44**, pp. 104 – 107.

[144] Ogata, Y. (1999) Seismicity analysis through point-process modeling: a review, *Pure and Applied Geophysics*, **155**, 471–507.

[145] Okamoto, M. (1986), *Foundation of Factor Analysis (in Japanese)*, Nikka Giren, Tokyo.

[146] Osborn, D. (1990) A survey of seasonality in UK macroeconomic variables, *Journal of Forecasting*, Vol. 6, 327–336.

[147] Osborn, D., Birchenhall, C., Jensen, H. and Simpson, P. (1999) Predicting US business cycle regimes, *Journal of Business and Economic Statistics*, Vol. 17, 313–323.

[148] O'Sullivan, F., Yandell, B. S. and Raynor Jr., W. J. (1986), Automatic Smoothing of Regression Functions in Generalized Linear Models, *J. Amer. Statist. Assoc.*, Vol. 81, 96–103.

[149] Ozaki, T. (1997a). Dynamic X11 model and nonlinear seasonal adjustment I: models and computational methods (in Japanese with English abstract), *Proceedings of the Institute of Statistical Mathematics*, Vol. 45, No. 2, 265–285.

[150] Ozaki, T. (1997). Dynamic X11 model and nonlinear seasonal adjustment II: numerical examples and discussion (in Japanese with English abstract), *Proceedings of the Institute of Statistical Mathematics*, Vol. 45, No. 2, 287–300.

[151] Ozaki, T. and P. Thomson, (1994) A dynamical system approach to X-11 type seasonal adjustment, Research Memo. No. 498, The Institute of Statistical Mathematics, Tokyo.

[152] Parzen, E. (1961), An Approach to Time Series Analysis, *Ann. Math. Statist.*, Vol. 32, 951–989.

[153] Parzen, E. (1963), A New Approach to the Synthesis of Optimal Smoothing and Prediction Systems, in *Mathematical Optimization Techniques*, ed. R. Bellman, 75–108.

[154] Peña, D. and G. E. P. Box (1987), Identifying a Simplifying Structure in Time Series, *Journal of the American Statistical Association*, **82**, 836–843.

[155] Phillips, P. C. B. and S. Ouliaris (1988), Testing for Cointegration Using Principal Component Analysis, *Journal of Economic Dynamics and Control*, **12**, 205–230.

[156] Priestly, M. B., T. Subba Rao and H. Tong (1974), Application of Principal Components Analysis and Factor Analysis in the Identification of Multivariable Systems, *IEEE Trans. Automat. Contr.*, **19**, 730–734.

[157] Proietti, T. (2000) Forecasting with Structural Time Series Models, in *Economic Forecasting*, M. Clements and D. Hendry (eds.), Blackwell Publishers, Oxford.

[158] Psaradakis, Z. (1997) Testing for unit roots in time series with nearly deterministic seasonal variation, *Econometric Reviews*, Vol. 16, 422–440.

[159] Reinsel, G. C. (1983), Some Results on Multivariate Autoregressive Index Models, *Biometrika*, **70**, 145–156.

[160] Rissanen, J. (1978) Modeling by shortest data description, *Automatica*, Vol. 14, 465–471.

[161] Sato, S. (1997). Introduction to "Web-Decomp" — Seasonal Adjustment System on WWW — (in Japanese with English abstracts), *Proceedings of the Institute of Statistical Mathematics*, Vol. 45, No. 2, 233–243. See also `http://ssnt.ism.ac.jp/inets2/title.html`.

[162] Schaefer, S. M. (1981) Measuring a Tax-Specific Term Structure of Interest Rates in the Market for British Government Securities, *The Economic Journal*, **91**, 415–38.

[163] Schoenberg, I. J. (1964), Spline functions and the problems of graduation, *Proc. Natl. Acad. Sci. U. S. A.*, 52, 333–343.

[164] Schwarz, G. (1978) Estimating the dimension of a model, *Annals of Statistics*, Vol. 6, 461–464.

[165] Shea, G. S. (1984) Pitfalls in Smoothing Interest Rate Term Structure Data: Equilibrium Models and Spline Approximation, *Journal of Financial and Quantitative Analysis*, **19**, 253–269.

[166] Shea, G. S. (1985) Interest Rate Term Structure Estimation with Exponential Splines, *Journal of Finance and Quantitative Analysis*, **19**, 253–69.

[167] Shephard, N. and Pitt, M. (1997). Likelihood analysis of non-Gaussian measurement time series, *Biometrika*, Vol. 84, 653–667.

[168] Shiller, R. (1973) A Distributed Lag Estimator Derived from Smoothness Priors, *Econometrica*, **41**, 775–778.

[169] Shiskin, J. and Plewers, T.J. (1978) Seasonal adjustment of the U.S. unemployment rate, *The Statistician*, 27, 181–202

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

[171] Silverman, B. W. (1985), Some Aspects of the Spline Smoothing Approach to Nonparametric Regression Curve Fitting, *J. Roy. Statist. Soc. B*, Vol. 36, 1–52.

[172] Sims, C.A. (1978) Comments on "seasonality, causation, interpretation and implications" by C.W.J. Granger, in *Seasonal Analysis of Economic Time Series*, Proceedings of the Conference on the Seasonal Analysis of Economic Time Series, Washington D.C., 47–49, Department of Commerce, Bureau of Census, Washington D.C.

[173] Steely, J. M. (1991) Estimating the Gilt-Edged Term Structure: Basis Splines and Confidence Intervals, *Journal of Business, Finance and Accounting*, **18**, No. 4 (June), 512–29.

[174] Stock, J. H. and M. W. Watson (1988), Testing for Common Trends, *Journal of the American Statistical Association*, **83**, 1097–1107.

[175] Takeuchi, K. (1976). Distribution of information number statistics and criteria for adequacy of models (in Japanese), *Mathematical Sciences (Suri Kagaku)*, No. 153, 12–18.

[176] Theil, H. (1960), Best linear index numbers of prices and quantities, *Econometrica*, **28**, 464–480.

[177] Thury, G. and Witt, S. F. (1998) Forecasting Industrial Production Using Structural Time Series Models, *Omega Int. J. Management Sci.*, Vol. 26, 751–767.

[178] Tikhonov, A. N. (1963), Solution of Incorrectly Formulated Problems and the Regularization Method, *Soviet Math. Dokl.*, Vol. 4, 1035–1038.

[179] Titterington, D. M. (1985), Common structure of smoothing techniques in statistics, *Int. Statist. Rev.*, 53, 141–170.

[180] Tukey, J. W. (1978), Can We Predict Where "Time Series" Should Go Next?, D. R. Brillinger and G. C. Tiao (Eds.), *Directions in Time Series*, Institute of Mathematical Statistics, 1–31, Ames.

[181] Vasicek, O. A. and Fong, H. G. (1982) Term Structure Modeling Using Exponential Splines, *Journal of Finance*, **37**, No.2 (May), 339–56.

[182] Velu, R. P., G. C. Reinsel and D. W. Wichern (1986), Reduced rank models for multiple time series, *Biometrika*, **73**, 105–118.

[183] Wahba, G. (1978), Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression, *J. Roy. Statist. Soc. B*, Vol. 40, 364–372.

[184] Wahba, G. (1983), Bayesian Confidence Intervals for the Cross -Validated Smoothing Spline, *J. Roy. Statist. Soc. B*, Vol. 45, 133–150.

[185] Wahba, G. (1990), *Spline Methods for Observed Data*, SIAM, Philadelphia.

[186] Wallis, K.F. (1974) Seasonal adjustment and relation between variables, *Journal of the American Statistical Association*, 69, 18–32.

[187] Wallis, K. F. (1982). Seasonal adjustment and revision of current data: linear filters for the X-11 method, *Journal of the Royal Statistical Society, Series A*, Vol. 145, 74–85.

[188] Wecker, W. E. and Ansley, C. F. (1983), The Signal Extraction Approach to Nonlinear Regression and Spline Smoothing, *Journal of the American Statistical Association*, Vol. 78, 81–89.

[189] Weinert, H. L., Byrd, R. H. and Sidhu, G. S. (1980), A Stochastic Framework for Recursive Computation of Spline Functions: Part II Smoothing Splines, *J. Optim. Theory Appl.*, Vol. 30, 255–268.

[190] West, M. and Harrison, P. J. (1986). Monitoring and adaptation in Bayesian forecasting models, *J. Amer. Stat. Assoc.*, Vol. 81, 741–750.

[191] West, M., Harrison, P. J. and Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting (with discussion), *J. Amer. Stat. Assoc.*, Vol. 80, 73–97.

[192] Whittle, P. (1963) *Prediction and regulation by linear least square methods*, The English Universities Press Ltd.

[193] Whittaker, E. T. (1923), On a new method of graduation, *Proc. Edinborough Math. Assoc.*, 78, 81–89.

[194] Whittaker, E. T. and Robinson, G. (1924), Calculus of Observations, in *A Treasure on Numerical Calculations*, Blackie and Son Ltd., London, 303–306.

[195] Young, P. C. and D. J. Pedregal (1999), Macro-economic Relativity: Government Spending, Private Investment and Unemployment in the USA 1948–1998, *Structural Change and Economic Dynamics*, **10**, 359–380.

[196] Zhang, M. Y., J. R. Russell and R. T. Tsay (1999) A nonlinear autoregressive conditional duration model with application to financial transaction data, Mimeo, Graduate School of Business, University of Chicago.