RCEA THE RIMINI CENTRE FOR ECONOMIC ANALYSIS

WP 14-08

Dimitrios D. Thomakos
University of Peloponnese, Greece
and
The Rimini Centre for Economic Analysis

"OPTIMAL LINEAR FILTERING, SMOOTHING AND TREND EXTRACTION FOR PROCESSES WITH UNIT ROOTS AND COINTEGRATION"

Optimal Linear Filtering, Smoothing and Trend Extraction

for Processes with Unit Roots and Cointegration[*]

Dimitrios D. Thomakos[†]

This version: March 26, 2008

Abstract

In this paper I propose a novel optimal linear filter for smoothing, trend and signal extraction for time series with a unit root. The filter is based on the Singular Spectrum Analysis (SSA) methodology, takes the form of a particular moving average and is different from other linear filters that have been used in the existing literature. To best of my knowledge this is the first time that moving average smoothing is given an optimality justification for use with unit root processes. The frequency response function of the filter is examined and a new method for selecting the degree of smoothing is suggested. I also show that the filter can be used for successfully extracting a unit root signal from stationary noise. The proposed methodology can be extended to also deal with two cointegrated series and I show how to estimate the cointegrating coefficient using SSA and how to extract the common stochastic trend component. A simulation study explores some of the characteristics of the filter for signal extraction, trend prediction and cointegration estimation for univariate and bivariate series. The practical usefulness of the method is illustrated using data for the US real GDP and two financial time series.

Keywords: cointegration, forecasting, linear filtering, singular spectrum analysis, smoothing, trend extraction and prediction, unit root.

Contents

1 Introduction

In a series of papers Phillips (1996), (1998) and (2005) gave an alternative formulation and modeling approach to stochastic processes with unit roots. Phillips' original aim appears to have been an attempt in formally showing that what he, in previous research, called 'spurious regression' was nothing more than a manifestation of a specific underlying structure for the asymptotic limit of the unit root process itself. His idea was ingenious because he showed us how a unit root process can be expressed via deterministic functions of time. Using a known result about the orthogonal decomposition of stochastic processes, the Karhunen-Loève (KL) decomposition, Phillips formally analyzed the properties of a particularly simple regression where the realization of a unit root process was regressed on a set of orthogonal trigonometric functions of time and showed how to interpret the regression results, what the implications were for the notion of 'spurious regression' and what the implications were for modeling and predicting stochastic trends.

This pioneering work has not found widespread use in spite of a very important implication: within that framework of Phillips one can meaningfully smooth a unit root process, extract the underlying smoothed series and predict the smoothed series itself, as well as the residual deviations of the original series from its smooth component. For economics and finance, where most would admit that the majority of available data have unit root-type, non-stationary characteristics, this is of practical significance: it allows one to perform standard time series operations (smoothing and trend extraction) without having to face any theoretical problem. For example, extracting stochastic trend components can be used in defining potential output from a real GDP series or for defining a 'fair price' path for an asset and, in addition, allows for an analysis of the resulting residual series using standard methods for stationary processes.

There is, of course, substantial previous literature that dealt with filtering and smoothing of non-stationary (including unit root) processes in economics (and of course other fields) but its focus was that of trend ("signal") extraction and smoothing based on mainly cyclical (e.g. business cycles) considerations and was related to the extraction of components of certain frequencies. In addition, that line of research was not really focused on the unit root model (it was dealt only as a special case). The work of Hodrick and Prescott (1997) and the filter named after them is probably the best example of this line of research. Related work has been done by King and Rebelo (1993), Baxter and King (1999) and Christiano and Fitzgerald (2003). It is interesting to note that in the first three out of these four papers the words "business cycles" appears in their titles! Pollock (2000) also relates to this line of research. Schleicher (2003)

summarizes some of this past work. It is important to note here that these papers dealt with optimal (in a mean squared error - MSE - sense) filters as well but from a different starting point and with a different aim in mind that what is done in the present work. Earlier work on the topic of smoothing non-stationary time series includes the seminal paper of Bell (1984) and the subsequent paper of Kohn and Ansley (1987). A convenient, matrix-based representation of the optimal MSE filter for the separation of a non-stationary signal from (stationary or not) noise is given by McElroy (2005). Book summaries of smoothing and filtering using the state space approach, which includes models for non-stationary time series, can be found in Harvey (1989) and Durbin and Koopman (2001).

In this paper I expand on the ideas of Phillips and provide a number of new results on smoothing stochastic processes with unit roots. Using the method of Singular Spectrum Analysis (SSA), and a number of already existing results, I derive an asymptotically optimal linear filter for smoothing and trend extraction for unit root processes. It turns out that this filter takes the form of a particular moving average and I derive explicit expressions for the filtering weights. This new result is important because it provides a theoretical justification for moving average smoothing in the context of unit root processes and has large potential for empirical applications. As in Phillips (2005) I also derive an h-step ahead, out-of-sample predictor for the smoothed series, which turns out to be recursively defined as a simple average of the past (actual and predicted) values of the smoothed series itself. In addition, I propose a new, data-based method for selecting the degree of smoothing which is applicable in both the current approach and the approach of Phillips. The proposed filter can also be used successfully in the context of non-stationary signal extraction type problems. Finally, I show how the methodology of this paper can handle common stochastic trend extraction in the context of a system of two cointegrating series, with the cointegrating coefficient being estimated using SSA.

In developing the results that follow I use material that is now readily accessible through books and monographs. Any results that are well known are not repeated; exceptions are: (a) an outline of the SSA method, which is reviewed in the next section and (b) a few other necessary items that are replicated in the appendix. The interested reader can consult the following sources for additional information about the mathematical background: Rao (1973) for matrix algebra results, including the spectral and singular value decompositions; Priestley (1981) and Fuller (1995) for results on linear filters, moving averages and and orthogonal decompositions of stationary processes (including diagonalization of autocovariance matrices) - Fuller (1995) also has the necessary results on limits of sample moments of unit root processes; Golyandina *et al.* (2001) is the only complete reference that deals with the core mathematics of SSA and should be

consulted for detailed results about the application of SSA in a more general context as well as for the specific results on SSA forecasting which are not provided later in the discussion. See also Elsner and Tsonis (1996) for an earlier but much less complete book reference on SSA. Optimal filtering in SSA for stationary series is discussed in Allen and Smith (1996).

The outline of the paper is as follows. In section 2 there is a brief review of the SSA method. Section 3 contains the paper's main results, where the SSA method is applied in the context of a stochastic process with a unit root and the form of the asymptotically optimal linear filter is derived, followed by with a discussion on its properties and related methodology. In section 4 I provide results from a simulation analysis while in section 5 there are empirical illustrations using quarterly data on the US real GDP and weekly prices of Brent Oil and the Euro/US Dollar exchange rate. Section 6 offers some concluding remarks. A summary of the notation used in the paper and some necessary results are given in the appendix.

2   The SSA Method

In this section I provide a brief outline of the SSA method. SSA is the empirical implementation of the KL orthogonal decomposition to sample data and equivalent to principal components (up to a point) in multivariate analysis. Orthogonal decompositions similar to the one applied in SSA have been known in time series for many years but were mainly used as theoretical tools rather as inference methods. SSA has been used heavily in atmospheric sciences, where it was essentially developed with this name, see Broomhead and King (1986), and where most of its applications can be found. Two references that use SSA in the context of economic and financial data are Lisi and Medio (1997) and Thomakos, Wang and Wille (2002).

2.1   The Trajectory Matrix

Consider a univariate stochastic processes $\{X_t\}_{t\in\mathbb{Z}}$ and suppose that a realization of size n from this process is available $\boldsymbol{X}_n \stackrel{\text{def}}{=} [x_1, x_2, \ldots, x_n]$. Denote by $k \geq 2$ the lead parameter, possibly allowing it to be $k = o(n)$, and define the $(k \times 1)$ lead column vectors $\boldsymbol{x}_t$ as:

$$\boldsymbol{x}_t \stackrel{\text{def}}{=} [x_t, x_{t+1}, x_{t+2}, \ldots, x_{t+k-1}]^\top \tag{1}$$

for $t = 1, 2, \ldots, N$ where $N \stackrel{\text{def}}{=} n - k + 1$. These vectors group together k time-adjacent observations and are supposed to describe the local state of the underlying process. Using the lead

vectors form the $(N \times k)$ trajectory matrix $\boldsymbol{T}$ by stacking them as follows:

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_N^\top \end{bmatrix} \tag{2}$$

Alternatively, $\boldsymbol{T}$ can defined through a set of k lead vectors of different dimension, namely the $(N \times 1)$ lead vectors $\mathbf{x}_t \overset{\text{def}}{=} [x_t, x_{t+1}, x_{t+2}, \ldots, x_{t+N-1}]^\top$ that form the columns of $\boldsymbol{T}$. We can, therefore, have the equivalent representation:

$$\boldsymbol{T} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k] \tag{3}$$

It will be convenient to keep both equations (2) and (3) for the discussion that follows.[1]

Besides the application in SSA, the trajectory matrix can be used to unify a number of common time series procedures, such as filtering and autoregressive modeling. For example, let $\boldsymbol{\beta}$ denote any known, fixed $(k \times 1)$ vector and consider the following:

- For $k = 2$ and $\boldsymbol{\beta} \overset{\text{def}}{=} [-1, 1]$ we can obtain the first differences of the realization as $\boldsymbol{T}\boldsymbol{\beta}$.

- For any $k \geq 2$ and $\boldsymbol{\beta} \overset{\text{def}}{=} [1/k, 1/k, \ldots, 1/k]$ we can obtain a k-order moving average for the realization as $\boldsymbol{T}\boldsymbol{\beta}$.

- For autoregressive modeling let $\boldsymbol{\beta}^*$ denote the parameter vector and $\boldsymbol{u}$ denote the vector of innovations. Write $\boldsymbol{T}\boldsymbol{\beta}^* = \boldsymbol{u}$ and define the $(k \times 1)$ and $(k \times k)$ restriction matrices:

$$\boldsymbol{q} \overset{\text{def}}{=} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}, \quad \boldsymbol{Q} \overset{\text{def}}{=} \begin{bmatrix} \boldsymbol{I}_{k-1} \\ \boldsymbol{0}_{k-1}^\top \end{bmatrix} \tag{4}$$

so that the restricted parameter vector $\boldsymbol{\beta}$ is written as $\boldsymbol{\beta}^* \overset{\text{def}}{=} \boldsymbol{q} - \boldsymbol{Q}\boldsymbol{\beta}$. Then, the least-squares problem for estimating $\boldsymbol{\beta}$ is given by:

$$\min_{\boldsymbol{\beta}} \boldsymbol{u}^\top \boldsymbol{u} = (\boldsymbol{q} - \boldsymbol{Q}\boldsymbol{\beta})^\top \boldsymbol{T}^\top \boldsymbol{T} (\boldsymbol{q} - \boldsymbol{Q}\boldsymbol{\beta}) \tag{5}$$

with the usual solution $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{Q}^\top \boldsymbol{T}^\top \boldsymbol{T} \boldsymbol{Q}\right)^{-1} \boldsymbol{Q}^\top \boldsymbol{T}^\top \boldsymbol{T} \boldsymbol{q}$.

---

[1]Since the analysis here is confined to univariate stochastic processes I will avoid using terms relating to principal components analysis although the trajectory matrix can be seen as a way to obtain multivariate observations from a univariate process.

2.2 Diagonal Averaging

The trajectory matrix is a Hankel matrix, having constant, positive-sloping skew diagonal elements. As a result the underlying time series can be obtained back from the trajectory matrix by a process called diagonal averaging. More formally, the trajectory matrix is obtained as a result of an operation, say $\mathcal{H}(\cdot)$, to the realization $\boldsymbol{X}_n$, as:

$$\mathcal{H}(\boldsymbol{X}_n) \mapsto \boldsymbol{T} \tag{6}$$

and the $\mathcal{H}(\cdot)$ operator has to be invertible since we should be able to recover the original series from the trajectory matrix. Therefore we can write:

$$\mathcal{H}^{-1}(\boldsymbol{T}) = \mathcal{D}(\boldsymbol{T}) \mapsto \boldsymbol{X}_n \tag{7}$$

where $\mathcal{D}(\cdot)$ is the operator for diagonal averaging which we describe next.[2]

Transfer the lead vectors of the trajectory matrix from equation (2) into the k-block diagonal of the $(N \times n)$ band matrix $\boldsymbol{B}$ as:

$$\boldsymbol{B}(\boldsymbol{T}) \equiv \boldsymbol{B} \stackrel{\text{def}}{=} \begin{bmatrix} \boldsymbol{x}_1^\top & 0 & \ldots & 0 \\ 0 & \boldsymbol{x}_2^\top & \ldots & 0 \\ 0 & \ldots & \ldots & 0 \\ 0 & \ldots & \ldots & \boldsymbol{x}_N^\top \end{bmatrix} \tag{8}$$

and note that the column arithmetic averages (of the non-zero elements) are equal to the original elements of the realization. We can formalize this operation as follows. Let $\boldsymbol{J}_N$ denote the $(N \times 1)$-dimensional unit vector and let $\boldsymbol{s}$ denote a $(n \times 1)$ vector with elements that correspond to inverse of the number of the non-zero elements in the rows of $\boldsymbol{B}$, that is:

$$\boldsymbol{s}^\top \stackrel{\text{def}}{=} [1, 1/2, \ldots, 1/k-1, \overbrace{1/k, \ldots, 1/k}^{n-2(k-1)}, 1/k-1, \ldots, 1/2, 1] \tag{9}$$

Using $\boldsymbol{J}_N$ and s the diagonal averaging operator can be defined as:

$$\mathcal{D}(\boldsymbol{T}) \stackrel{\text{def}}{=} \left(\boldsymbol{J}_N^\top \boldsymbol{B} \right) \odot \boldsymbol{s}^\top \tag{10}$$

where $\odot$ is the Hadamard (element-by-element) product between two matrices.

The diagonal averaging operation, which is essential in what follows, is an optimal operation in the following sense. For any matrix $\boldsymbol{X}$, not necessarily a trajectory matrix, the Hankel matrix, say $\boldsymbol{X}_o$, that can best approximate $\boldsymbol{X}$ using the Frobenius matrix norm $\|\cdot\|_M^2$ (see the appendix

---

[2]To the best of my knowledge this representation of diagonal averaging has not yet appeared in the SSA literature.

for the definition) is obtained through diagonal averaging in the sense that (see Buchstaber, 1994, and Golyandina *et al.*, 2001):

$$\|\boldsymbol{X} - \boldsymbol{X}_o\|_M^2 = \min_{\boldsymbol{Z} \in \mathcal{M}_H} \|\boldsymbol{X} - \boldsymbol{Z}\|_M^2 \tag{11}$$

where $\mathcal{M}_H$ is the set of conformable Hankel matrices and where the elements of $\boldsymbol{X}_o = \mathcal{H}\left[\mathcal{D}(\boldsymbol{X})\right]$.

2.3 Decomposition & Minimum Norm Approximation

Using known results about the singular value decomposition (SVD) of an $(N \times k)$ matrix such as $\boldsymbol{T}$ we can decompose it as:

$$\boldsymbol{T} = \sum_{i=1}^{k} \sqrt{\lambda_j} \mathbf{v}_j \mathbf{u}_j^\top = \boldsymbol{V} \boldsymbol{\Lambda}^{1/2} \boldsymbol{U}^\top \tag{12}$$

where $\sqrt{\lambda_j}$ denotes singular values and $\mathbf{v}_j, \mathbf{u}_j$ denotes the left and right singular vectors respectively. The SVD decomposition has, however, an interpretation based on the cross-moment matrix of $\boldsymbol{T}$, that is $\boldsymbol{T}^\top \boldsymbol{T}$. For a zero-mean, stationary process this would be the symmetric matrix of sample autocovariances of all orders from $j = 1, 2, \ldots, k$. Denoting by $\widehat{\gamma}(s) \stackrel{\text{def}}{=} n^{-1}\mathbf{x}_j\mathbf{x}_{j+s}$, for $j = 1, 2, \ldots, k$ and $s = 0, 1, \ldots, k - j$ the sample cross-moments and placing them into the symmetric matrix $\widehat{\boldsymbol{\Gamma}}(k)$ we have that:

$$\widehat{\boldsymbol{\Gamma}}(k) \stackrel{\text{def}}{=} \frac{1}{n}\left(\boldsymbol{T}^\top \boldsymbol{T}\right) \tag{13}$$

The singular values of $\boldsymbol{T}$ are the (scaled square roots of the) eigenvalues from the spectral decomposition of $\widehat{\boldsymbol{\Gamma}}(k)$ and the right singular vectors $\mathbf{u}_j$ are the corresponding eigenvectors of $\widehat{\boldsymbol{\Gamma}}(k)$, that is:

$$\widehat{\boldsymbol{\Gamma}}(k) = \sum_{j=1}^{k} \lambda_j \mathbf{u}_j \mathbf{u}_j^\top \tag{14}$$

For a stationary process with finite second moments the matrix of the autocovariances contains essentially all the information that we need for modeling and forecasting the realization $\boldsymbol{X}_n$. It is therefore appropriate to work with the decompositions of equations (12) and (14). The relative magnitude of the eigenvalues of $\widehat{\boldsymbol{\Gamma}}(k)$ can tell us how much information is contained within the cross-moments of the process. Consider, for example, the cumulative proportion of the first r eigenvalues, i.e. $\ell_r \stackrel{\text{def}}{=} \sum_{j=1}^{r} \lambda_j / \sum_{j=1}^{k} \lambda_j$. If this proportion is very high, say over 90%, then we should be able to approximate $\widehat{\boldsymbol{\Gamma}}(k)$, and therefore $\boldsymbol{T}$, with only r of the components in (12) or (14). It can be shown, see the appendix, that this approximation is an operation of minimum norm and is thus an optimal approximation.

Using the properties of the SVD matrices and the first $r < k$ components we obtain that the minimum norm approximation of $\boldsymbol{T}$ is given by:

$$\boldsymbol{T}_r \overset{\text{def}}{=} \boldsymbol{T} \sum_{j=1}^{r} \mathbf{u}_j \mathbf{u}_j^\top = \boldsymbol{T} \boldsymbol{Q}_r, \quad \text{for } \boldsymbol{Q}_r \overset{\text{def}}{=} \sum_{j=1}^{r} \mathbf{u}_j \mathbf{u}_j^\top \tag{15}$$

and we can see that this approximation is a linear operation on $\boldsymbol{T}$. Moreover, $\boldsymbol{T}_r$ has the property that is is a sum of r orthogonal (uncorrelated) elementary rank-one matrices since we can also write:

$$\boldsymbol{T}_r = \sum_{j=1}^{r} \boldsymbol{T} \mathbf{u}_j \mathbf{u}_j^\top = \sum_{j=1}^{r} \boldsymbol{T}_j \tag{16}$$

where $\boldsymbol{T}_j \overset{\text{def}}{=} \boldsymbol{T} \mathbf{u}_j \mathbf{u}_j^\top$ and $\boldsymbol{T}_i \boldsymbol{T}_j^\top = \boldsymbol{0}$, for $i \neq j$.

Using the diagonal averaging operator $\mathcal{D}(\boldsymbol{T}_r)$ we can recover an optimal approximation to the original realization.[3] This is clearly a smoothing operation and therefore we obtain an optimal (minimum norm) filter for the original series which is denoted as:

$$\mathcal{D}(\boldsymbol{T}_r) \mapsto \boldsymbol{X}_{n,r} \tag{17}$$

or, equivalently, as $\{x_{t,r}\}_{t=1}^{n}$. The residual series is denoted by $u_{t,r} = x_t - x_{t,r}$.

3 Optimal Smoothing for Unit Root Processes

3.1 Application of SSA

The previous section provided a brief outline of the SSA method. While SSA has been successfully applied in time series analysis across different fields, there is no formal work about the properties of SSA in the context of stochastic processes that contain a unit root. In this section I bridge this gap with existing literature and show that SSA can be applied in the context of unit root processes. In particular, I derive the asymptotically optimal (minimum norm) approximation based on SSA and thus derive an optimal filter for smoothing unit root processes.

Consider a stochastic process $\{X_t\}_{t \in \mathbb{N}_+}$ with a unit root, that is:

$$X_t = X_{t-1} + \eta_t \quad \rightarrow \quad X_t = \sum_{j=1}^{t} \eta_j \tag{18}$$

with $X_0 = 0$ and η_t a sequence of i.i.d. random variables with mean zero and variance σ_η^2. The i.i.d. assumption about η_t is innocuous and used for simplicity only as a number of other

---

[3]Note the double optimality involved here: both the SVD approximation is a minimum norm operation *and* the diagonal averaging is a minimum norm operation.

assumptions, such as mixing, do not alter the results that I obtain below. Suppose that a realization $\boldsymbol{X}_n = [x_1, x_2, \ldots, x_n]$ is available for this process and that you wish to apply SSA to it. What are the implications of the unit root assumption for the resulting decomposition and minimum norm approximation? The answer can be readily obtained using known results from the unit root literature. I then obtain new, explicit expressions for the asymptotic optimal approximation/linear filter. These expressions can be used for smoothing and trend extraction, as well as for trend prediction.

As before we need the matrix of autocovariances $\widehat{\boldsymbol{\Gamma}}(k)$. Under the unit root hypothesis this matrix diverges as $n \to \infty$ but it converges to a stochastic matrix if scaled by n. Using standard results, see the appendix, it can be shown that:

$$\frac{1}{n}\widehat{\boldsymbol{\Gamma}}(k) = \sum_{j=1}^{k} \lambda_j \mathbf{u}_j \mathbf{u}_j^\top \Rightarrow \sigma_\eta^2 w \boldsymbol{J}_{k,k} \tag{19}$$

where $\Rightarrow$ denotes weak convergence, w is a stochastic integral and $\boldsymbol{J}_{k,k}$ is a matrix of ones. This limit matrix has a very simple spectral decomposition with one positive stochastic eigenvalue given by $\sigma_\eta^2 w k$ and associated orthonormal eigenvector $\boldsymbol{J}_k/\sqrt{k}$. Note that the stochastic nature of the limit matrix is confined to the eigenvalue and not to the eigenvector. In fact we have:

Proposition 1. Under the assumptions of equation (18) we have that the eigenvalues and eigenvectors of the decomposition of the matrix of autocovariances $n^{-1}\widehat{\boldsymbol{\Gamma}}(k)$ obey the following:

1. $\lambda_1 \Rightarrow \sigma_\eta^2 w k$, $\lambda_j \Rightarrow 0$, for $2 \leq j \leq k$, and $\ell_1 \Rightarrow 1$

2. $\mathbf{u}_1 \Rightarrow \boldsymbol{J}/\sqrt{k}$

It follows that, in large samples, it will not make any difference whether one uses the empirical eigenvector $\mathbf{u}_1$ or the asymptotic eigenvector $\boldsymbol{J}_k/\sqrt{k}$.

Proceeding for simplicity with the asymptotic eigenvector we have that the minimum norm approximation of equation (15) now becomes:

$$\boldsymbol{T}_{1,\infty} = \boldsymbol{T}\boldsymbol{Q}_{1,\infty}, \quad \text{for } \boldsymbol{Q}_{1,\infty} \overset{\text{def}}{=} \frac{1}{k}\boldsymbol{T}\boldsymbol{J}_k\boldsymbol{J}_k^\top \tag{20}$$

The matrix $\boldsymbol{T}_{1,\infty}$ has a special structure that reveals the nature of the underlying smoothing operation that takes place. In particular, note that:

$$\boldsymbol{T}_{1,\infty} = \begin{bmatrix} k^{-1}\sum_{t=1}^{k} x_t \\ k^{-1}\sum_{t=2}^{k+1} x_t \\ \vdots \\ k^{-1}\sum_{t=N}^{n} x_t \end{bmatrix} \boldsymbol{J}_k^\top \tag{21}$$

10

so that it has identical columns and its rows are k-period rolling averages of the realization. Therefore, application of the diagonal averaging operator $\mathcal{D}(\boldsymbol{T}_{1,\infty})$ will produce a smoothed series $\mathcal{D}(\boldsymbol{T}_{1,\infty}) \mapsto \boldsymbol{X}_{n,1} \stackrel{\text{def}}{=} \{x_{s,1}\}_{s=1}^{n}$ that will be based on these averages. In particular, after some algebra, we obtain an explicit expression for the smoothed series as:

$$x_{s,1} \stackrel{\text{def}}{=} \begin{cases} \frac{1}{sk} \sum_{j=1}^{s} \sum_{t=j}^{k+(j-1)} x_t, & s \le k-1 \\ \frac{1}{k^2} \sum_{j=1}^{k} \sum_{t=s-k+j}^{s+(j-1)} x_t, & k \le s \le n-k+1 \\ \frac{1}{(n-s+1)k} \sum_{j=s}^{n} \sum_{t=s-k+1}^{s} x_t, & s > n-k+1 \end{cases} \tag{22}$$

The above representation has a very interesting structure since (a) it is composed from local cumulative k-period averages and (b) these cumulative averages do not have the same number of terms (that is, the same degree of smoothness) at the beginning and end of the series. These properties show that the asymptotically optimal filter automatically preserves the original structure of the series (cumulation of η's maps into cumulation of averages) and takes into account end effects. An example will clarify the structure of the smoothed series $x_{s,1}$. Take $k = 4$ and note that we have:

$$x_{1,1} = \frac{x_1+x_2+x_3+x_4}{4},$$
$$x_{2,1} = \frac{x_1+2x_2+2x_3+2x_4+x_5}{8},$$
$$x_{3,1} = \frac{x_1+2x_2+3x_3+3x_4+2x_5+x_6}{12},$$
$$x_{4,1} = \frac{x_1+2x_2+3x_3+4x_4+3x_5+2x_6+x_7}{16},$$
$$\vdots$$

The smoothed series takes the form of a symmetric moving average with weights that decline (increase) linearly from the center value of the average, the weights summing-up to one. In addition, it automatically takes care of the end of the series so that the first smoothed value is a forward moving average and the last smoothed value is a backward moving average.

3.2 Properties of the Smoothed Series

Note that when $k \le s \le n-k+1$, i.e. when excluding the end-points of the smoothed series, we can express the moving average in a standard linear filter format as:

$$x_{s,1} = \psi(L)x_s = \frac{1}{k^2} \sum_{j=-k+1}^{k-1} (k-|j|)x_{s+j}, \quad \text{for } k \le s \le n-k+1 \tag{23}$$

and, furthermore, note that this average can be thought of as the solution to the local optimization problem:

$$x_{s,1} = \underset{\mu_s}{\operatorname{argmin}} \sum_{j=-k+1}^{k+1} f_j(x_{s+j}-\mu_s)^2 \tag{24}$$

11

where f_j is the frequency of occurrence of each x_{s+j} in the average. The associated polynomial of the filter $\psi(L)$ can be factored as:

$$\psi(L) = \frac{1}{k^2} L^{k-1} \psi^*(L^{-1}), \quad \text{for } \psi^*(z) \stackrel{\text{def}}{=} 1 + \sum_{j=1}^{k-1} \left[(j+1)z^j + (k-j)z^{(k-1)+j} \right] \quad (25)$$

The roots of the filter polynomial are determined by the roots of $\psi^*(z) = 0$ and it can be shown that they are all equal to unity in absolute value and, in particular: if k is even then it has two repeated real roots at $z = -1$ and the rest are repeated complex roots of the same conjugate pair; if k is odd it has only repeated complex roots from different conjugate pairs.

It is also straightforward to compute the frequency response function of the filter polynomial $\psi(L)$ so as to examine the effects of different degrees of smoothing in the original series. The frequency response function is defined as the Fourier transform of the filter polynomial and we have:

$$\mathcal{R}(\omega) \stackrel{\text{def}}{=} \frac{1}{k^2} \sum_{j=-k+1}^{k-1} (k - |j|) \exp(-2\pi i \omega j) = \frac{1}{k} \left[1 + \frac{2}{k} \sum_{j=1}^{k-1} j \cos(2\pi\omega j) \right] \quad (26)$$

where ω is the frequency. It is now easy to visually see the effects of smoothing as a function of the frequency ω.

In Figure 1 I plot $|\mathcal{R}(\omega)|^2$ against frequency for various values of $k = 8, 16, 20, 40$. Assuming that k is selected as $k = \sqrt{n}$ these values correspond to the following sample sizes: 16 years of quarterly observations, 64 years of quarterly or 21 years of monthly observations, and over 5 years of daily observations. As expected, we see that the smoother enhances the low frequency component of the original series but for low values of k allows for some power to pass from higher frequencies. It is possible to select the value of k so as to significantly reduce the power at certain frequencies. For example, taking k equal to the cycle period (e.g. 4 for quarterly or 12 for monthly data) we can smooth the original series retaining the low frequency component and erasing most of the cyclical component, if one exists. In Figure 2 I plot $|\mathcal{R}(\omega)|^2$ against frequency for $k = 4, 12$ and mark the frequency that corresponds to the period noted before.

We can also examine the properties of the residual series $u_{s,1} \stackrel{\text{def}}{=} x_s - x_{s,1}$. First note that, using the factorization of the filter polynomial in equation (25), we have that, for $k \leq s \leq n-k+1$:

$$u_{s,1} = [1 - \psi(L)] x_s = \left[1 - k^{-2} L^{k-1} \psi^*(L^{-1}) \right] x_s \quad (27)$$

so that the residuals can be thought of as $k - 1$-period 'quasi-differences'. Note that these residuals contain forward looking information about the series, they are computed using terms after x_s. In this sense one can possibly call them 'predictive residuals'. In Figure 3 I plot the sample realization from a unit root process with $\eta_t \sim \mathcal{N}(0, 0.2^2)$, $n = 500$ and $k = [\sqrt{n}] = 22$, and

12

the sample autocorrelation function of the original series and the residual series. The differences in the behavior between the two series are apparent. Moreover, note that after the removal of the smooth component there is a substantial amount of serial correlation that remains in the residual series. However, this autocorrelation can be adequately addressed using a standard autoregressive model to account for the cyclical patterns in the residual autocorrelation function.

All the above present an entirely new result that justifies the use of moving averages as optimal smoothers in the context of unit root processes. The implications for the use of moving averages in smoothing and trend extraction of economic and financial time series are evident: the proposed method allows for a precise extraction of the main k components of a stochastic trend using an optimization-based approach, with explicit expressions for the resulting smoothing weights and specific properties attached to the smoothed and residual series.

3.3 Connection with Phillips' Approximation and Signal Extraction Problems

There are certain similarities and differences between the approach taken in the work of Phillips, say e.g. (2005), the approach of "non-stationary signal extraction", e.g. Bell (1984), and the current approach. The main similarities between the current approach and Phillips' approach is that the smoothing of the unit root process is based on asymptotic considerations and that both approaches use versions of the Karhunen-Loève (KL) decomposition. In particular, here I use the asymptotic minimum norm approximation (the sample KL version) and the resulting eigenvector to construct the smoothed series $x_{s,1}$. Phillips motivates his smoother by the use of the KL decomposition to the limit process of the appropriately scaled x_t given by:

$$\frac{1}{\sqrt{n}} x_{[n\cdot]} \Rightarrow B(r) \stackrel{\text{def}}{=} \sqrt{2} \sum_{j=1}^{\infty} \frac{\sin\left[(j-1/2)\pi r\right]}{(j-1/2)\pi} \xi_j \tag{28}$$

for $r \in [0,1]$ and the ξ_j being i.i.d. normal random variables $\xi_j \sim \mathcal{N}(0, \sigma^2)$. The above theoretical relationship can be empirically fitted using observations from a realization and can also be used for trend prediction. Letting $\left\{\phi_j(r) \stackrel{\text{def}}{=} \sqrt{2} \sin\left[(k-1/2)\pi r\right]\right\}_{j=1}^{\infty}$ denote the system of orthogonal deterministic functions of time we have the linear regression:

$$x_t \stackrel{\text{def}}{=} \widehat{x}_{t,k} + \widehat{u}_{t,k} = \sum_{j=1}^{k} \widehat{b}_j \phi_j(t/n) + \widehat{u}_{t,k} \tag{29}$$

where k denotes the number of 'trend coordinates' that are to be used in reconstructing the series. For $k \to \infty$ and $k = o(n)$ we have that the above regression can reproduce the entire realization. For low values of k one can extract the underlying trend components. Note that the regression coefficient estimators $\widehat{b}_j$ converge to random variables in this context, not constants (see Phillips (2005) and references therein for details).

While in essence both approaches are trying to do the same thing, i.e. smooth the realization of a unit root process, there are also two differences between them. First, the approach of Phillips (2005) is a global approach, as it is based on the application of global least squares to the entire realization for extracting the trend components. The smoothing is a by-product of this global fitting, essentially through the use of the trigonometric basis functions. In contrast, the current approach is a local approach as it is based on the application of local smoothing via the use of moving averages. Of course both approaches can achieve the same degree of smoothing by appropriate choices of the smoothing parameter k. A second difference is that the method in Phillips (2005) does not come out of an optimization framework and does not come with a 'structural' interpretation. With 'structural' interpretation I simply mean that the approach that is proposed here is related to a well-understood notion of smoothing, that of a moving average. It is obvious that both methods have well-defined interpretations in the context of the unit root assumption. Finally, note that the choice of the smoothing parameter is opposite in the two methods: in the SSA-based method of the previous section more smoothing is performed by allowing k to increase; in Phillips' method more smoothing is performed by allowing k to decrease.

There are more differences than similarities with other methods that have as their basic idea that of "signal extraction" or the isolation of a particular, well-defined component of the underlying stochastic process X_t using the realization $\boldsymbol{X}_n$. An important, theoretical and practical difference, is that all such methods require the a priori specification of a parametric model for the "signal" and the "noise" (whose precise definition varies by discipline). Without postulating such a model it is not possible to apply any of the optimal filters that appear in the relevant literature. Such a model comes with along with parameter estimation, estimation uncertainty, the possibility of structural breaks, misspecification, etc. This is not to claim that problems like structural breaks cannot occur within the unit root framework; they do. However, the simplicity of the unit root model, and the proposed smoothing method of this paper, do have a certain sense of robustness to such problems.

To illustrate the differences between smoothing in the signal extraction framework and the current framework consider a simple example: the random walk plus noise model (also known as local level model). Let Y_t denote the observable stochastic process and let X_t denote the unobservable signal, which has a unit root. They are assumed to be related by the following state space model:

$$
\begin{aligned}
Y_t &= X_t + \epsilon_t \\
X_t &= X_{t-1} + \eta_t
\end{aligned}
\tag{30}
$$

where $\epsilon_t \sim$ i.i.d.$(0, \sigma_\epsilon^2)$ is the observational noise and where $\eta_t \sim$ i.i.d.$(0, \sigma_\eta^2)$ is the signal noise, with ϵ_t independent of η_s for all (t, s). The properties of this model depend on the "signal-to-noise" ratio $q \stackrel{\text{def}}{=} \sigma_\eta^2/\sigma_\epsilon^2$: as $q \to 0$ the signal is buried in noise and is difficult to recover; as $q \to \infty$ the model collapses to the standard unit root model of equation (18). Accurate MSE extraction of the signal component requires estimation (via the Kalman filter) of the two variance parameters and then fixed point smoothing. Is the method proposed in this paper capable of separating the signal from the noise in this set-up? To examine this let us construct the relevant trajectory matrices and the corresponding autocovariance matrices. Denoting the trajectory matrices in standard fashion as $\boldsymbol{T}_Y$, $\boldsymbol{T}_X$ and $\boldsymbol{T}_\epsilon$ and the corresponding matrices of sample autocovariances as $\widehat{\boldsymbol{\Gamma}}_j(k)$ for $j = Y, X, \epsilon$ we immediately obtain:

$$
\begin{aligned}
\boldsymbol{T}_Y &= \boldsymbol{T}_X + \boldsymbol{T}_\epsilon \\
\widehat{\boldsymbol{\Gamma}}_Y(k) &= \widehat{\boldsymbol{\Gamma}}_X(k) + \widehat{\boldsymbol{\Gamma}}_\epsilon(k) + \widehat{\boldsymbol{\Gamma}}_{X,\epsilon}(k) + \widehat{\boldsymbol{\Gamma}}_{\epsilon,X}(k)
\end{aligned}
\tag{31}
$$

where $\widehat{\boldsymbol{\Gamma}}_{X,\epsilon}(k), \widehat{\boldsymbol{\Gamma}}_{\epsilon,X}(k)$ are the cross-covariances between the signal and the noise trajectories. Under the assumptions of equation (30), and using standard results, we have that asymptotically $\widehat{\boldsymbol{\Gamma}}_\epsilon(k) \Rightarrow \sigma_\epsilon^2 \boldsymbol{I}_k$ and $\widehat{\boldsymbol{\Gamma}}_{\epsilon,X}(k) \Rightarrow \chi$, where χ is a stochastic matrix. However, since $\widehat{\boldsymbol{\Gamma}}_X$ does not converge unless scaled by n we end up having $n^{-1}\widehat{\boldsymbol{\Gamma}}_\epsilon(k) \Rightarrow \boldsymbol{0}_{k,k}$, $n^{-1}\widehat{\boldsymbol{\Gamma}}_{\epsilon,X}(k) \Rightarrow \boldsymbol{0}_{k,k}$ and therefore:

$$
\frac{1}{n}\widehat{\boldsymbol{\Gamma}}_Y(k) \approx \frac{1}{n}\widehat{\boldsymbol{\Gamma}}_X(k) \Rightarrow \sigma_\eta^2 w \boldsymbol{J}_{k,k}
\tag{32}
$$

exactly as in equation (19). This result has not appeared in the SSA or filtering literature and is of practical significance: using SSA for a unit root process contaminated with noise we can extract the underlying non-stationary signal directly, at least asymptotically. Combining previous results on stationary SSA with the results from the previous sections we can also select k, the degree of smoothing appropriately: as $q \to 0$ then $k \to n/2$ with $n \to \infty$; as $q \to \infty$ then $k = o(n)$, e.g. $k = \sqrt{n}$.

For comparison with the SSA approach I reproduce below McElroy's (2005) matrix-based formulas for Kalman fixed point smoothing for the local level model. Letting $\boldsymbol{\Delta}$ denote the $(n \times n - 1)$ matrix with -1 on its principal diagonal and 1 in its first lower diagonal and $\boldsymbol{Y}_n \stackrel{\text{def}}{=} [y_1, y_2, \ldots, y_n]$ denote the $(1 \times n)$ vector of observations we have that the $(n \times n)$ matrix of optimal smoothing coefficients is given by:

$$
\boldsymbol{F}_n(\sigma_\epsilon^2, \sigma_\eta^2) \equiv \boldsymbol{F}_n \stackrel{\text{def}}{=} \sigma_\eta^{-2} \left(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top \sigma_\epsilon^{-2} + \boldsymbol{I}_n \sigma_\eta^{-2} \right)^{-1} = \left(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top q + \boldsymbol{I}_n \right)^{-1}
\tag{33}
$$

so that the optimal MSE signal estimate is given by $\widehat{\boldsymbol{X}}_n \stackrel{\text{def}}{=} \mathsf{E}[\boldsymbol{X}|\boldsymbol{Y}] = \boldsymbol{Y}_n \boldsymbol{F}_n$.

Figure 4 illustrates the above using a sample realization from equation (30) with both ϵ_t and η_t being normally distributed and $q = 1\%$, the noise variance being 100 times greater that

the signal variance, and with $k = \sqrt{n}$ as before. The lower panel of the figure shows the true signal and the two smoothed series, the one based on SSA and the other on the application of fixed point smoothing (with the parameters estimated). It is clear from the figure that the non-parametric SSA smoother performs on par with the parametric fixed point smoother. We further explore the performance of the proposed methodology in the context of signal extraction in the simulation section.

Remark 1. In the signal extraction framework one can accommodate a comparison between the proposed method and the Hodrick-Prescott (1997) HP-filter that is used frequently in trend extraction and smoothing in economics. It can be shown, see for example Schlicht (2005), Dermoune *et al.* (2007) and earlier references therein, that the HP filter can be derived from a signal extraction model similar to equation (30) but where the signal process X_t has two instead of one unit roots, i.e. $(1 - B)^2 X_t = \eta_t$, with B being the backshift operator. The filtered values can be computed using exactly the same formula as in equation (33) before with $\boldsymbol{\Delta}$ being defined as $(n \times n - 2)$ with -2 on its principal diagonal and 1's on the first upper and lower diagonals. Schlicht (2005) and Dermoune *et al.* (2007) also proposed methods to consistently estimate σ_η^2, σ_ϵ^2 and q from the data. The method proposed in this paper can easily be used in the HP model context and we illustrate this in the empirical applications' section.

3.4   Trend Prediction

Both the signal extraction approach and Phillips' (2005) approach can be used to extrapolate the smoothed series and thus make signal/trend predictions. The signal extraction formulas are well known and thus omitted. Phillips (2005) also provides an explicit expression for the h-step ahead predictor for the fitted k trend components, say $\widehat{x}_{t+h,k}$.

Below I give the h-step ahead trend predictor based on the SSA method and the resulting smoothed series. Using known results from SSA about the continuation of reconstructed components, see Golyandina *et al.* (2001), the formula the defines the extrapolation coefficients is given by:

$$\boldsymbol{\alpha} = \frac{1}{1 - \nu^2} \sum_{j=1}^{r} u_{k,j} \mathbf{u}_j^{k-1} \tag{34}$$

where $\nu \stackrel{\text{def}}{=} u_{k,1}^2 + u_{k,2}^2 + \cdots + u_{k,r}^2$ is the sum of squares of the last element of the eigenvectors $j = 1, 2, \ldots, r$ (called the verticality coefficient) and where $\mathbf{u}_j^{k-1}$ denotes the $(k-1 \times 1)$ vector with the first $k-1$ elements of eigenvectors $j = 1, 2, \ldots, r$. Successful continuation of a reconstructed component, i.e. the smoothed series, requires that $\nu^2 < 1$.

In the current context we have that $r = 1$ and $\mathbf{u}_1$ has the particularly simple form $\mathbf{u}_1 = \mathbf{J}_k/\sqrt{k}$. Note that the verticality coefficient becomes $\nu^2 = 1/k$ which is always less than one and thus the predictor is well defined. Doing some algebra we finally get that the prediction parameter vector $\boldsymbol{\alpha}$ simplifies to $\boldsymbol{\alpha} = [1/(k-1), 1/(k-1), \ldots, 1/(k-1)]^\top$ and the the smoothed series h-step ahead prediction is defined recursively as a simple average of its past $k-1$ values, that is:

$$\widehat{x}_{s+h,1} = \frac{1}{k-1} \sum_{j=1}^{k-1} \widehat{x}_{s+h-j,1} \tag{35}$$

3.5 Selecting the degree of smoothing

A practical problem is how to select the degree of smoothing. Both the current methodology and Phillips' methodology cannot use an MSE-like criterion for selecting the degree of smoothing since there is no underlying model that defines a signal. Is there any other way in which we can select the degree of smoothing based on an optimization criterion? I propose such a criterion below, applicable both to the current methodology and to Phillips' methodology, and structure the problem as follows. To maintain the idea of smoothing and trend extraction k should be kept relatively high (low) and an appropriate value should be selected based on a different objective function than the residual error variance, i.e. the MSE. A suitable alternative could be the following. Assume that you would like to fit the "correct trend on average". You would then expect to obtain about an equal amount of positive and negative residuals: sometimes the actual series would be higher than the smoothed trend and sometimes it would be lower. An objective function that fits this idea is the (absolute value of the) average sign of the residuals from the selected smoothed trend. Minimizing this function we can not only obtain a meaningful "optimal" value of k but we can also compute appropriate residual error bands.

More formally, for any value $k \in \{k_{\min}, \ldots, k_{\max}\}$ define the average sign of the fitted residuals[4] as:

$$\widehat{m}(k) \overset{\text{def}}{=} \frac{1}{n} \sum_{t=1}^{n} sign(u_{t,1}) \tag{36}$$

Since the sign function can be decomposed as $sign(x) = I(x > 0) - I(x < 0)$, for $I(\cdot)$ being the indicator function, we can also decompose $\widehat{m}(k)$ as $\widehat{m}(k) = \widehat{m}_+(k) - m_-(k)$, the respective averages of the positive and negative residuals. A plausible 'optimal' choice for k is that value that minimizes the absolute value of $\widehat{m}(k)$, that is the value that produces, on average, an equal

---

[4]I use the notation for the residuals of the proposed method in the paper, $u_{t,1}$ but the same can be done with the residuals from Phillips' method $\widehat{u}_{t,k}$.

17

number of positive and negative residuals, that is $\widehat{m}_+(k) \approx \widehat{m}_-(k)$. This value, say k^*, would give the trend component of the series that will be an approximately "unbiased" estimate of the series' main trend component. We have then that k^* would be given as solution to the following optimization problem:

$$k^* : \underset{k \in \{k_{\min}, \ldots, k_{\max}\}}{\operatorname{argmin}} |\widehat{m}(k)| \tag{37}$$

There is no rule for selecting $k_{\min}$ and $k_{\max}$ but one can follow the suggestions in the end of section 3.3 after equation (32). For Phillips' approach $k_{\min}$ can be taken to be 1 but this will not work for the SSA approach where $k_{\min} \geq 2$.[5]

If one wants to use predicted trend values, say $\widehat{x}_{t+h,1}$ in selecting k^* then a suitable approach is historical simulation, which can be done as follows. Split the available realization in two parts of $n_0 + n_1 = n$ observations, an estimation and an evaluation sample respectively. Using a rolling window of n_0 observations compute $n_1 - h$ h-step ahead out-of-sample trend predictions. Then, for each value of k, compute:

$$\widehat{m}_1(k) \overset{\text{def}}{=} (n_1 - h)^{-1} \sum_{t=n_0+1}^{n-h} sign(\widehat{u}_{t+h,1}) \tag{38}$$

where $\widehat{u}_{t+h,1} \overset{\text{def}}{=} x_{t+h} - \widehat{x}_{t+h,1}$ and find k^* from minimizing $|\widehat{m}_1(k)|$.

Based on the predictions from k^* one can form the corresponding residual deviations $\widehat{u}^*_{t+h,1} \overset{\text{def}}{=} x_{t+h} - \widehat{x}^*_{t+h,1}$ and compute their standard deviation, say s_{k^*}. Using the standard deviation one can construct trend prediction bands $[L(\tau), U(\tau)]$ given by:

$$[L(\tau), U(\tau)] \overset{\text{def}}{=} \left[\widehat{x}^*_{t+h,1} - \tau \cdot s_{k^*}, \widehat{x}^*_{t+h,1} + \tau \cdot s_{k^*} \right] \tag{39}$$

for any $\tau \in \mathbb{R}_+$. An interesting empirical question that relates to these prediction bands is their coverage ratio, i.e. the fraction of out-of-sample observations that fall within the bands during the evaluation sample. This ratio is defined as:

$$CR(\tau) \overset{\text{def}}{=} (n_1 - h)^{-1} \sum_{t=n_0+1}^{n-h} I\left\{ L(\tau) \leq x_{t+h} \leq U(\tau) \right\} \tag{40}$$

I further explore this method for selecting k and its properties in the simulation section that follows.

3.6 Cointegration and common stochastic trend extraction

The methodology proposed in the earlier sections of the paper can be extended to handle common stochastic trend extraction in the context of a simple bivariate cointegrating system - systems

---

[5]In fact in both approaches $k_{\min}$ can be selected to achieve any minimum degree of smoothing, e.g. 5.

of larger dimension should, in principle, be handled in a similar fashion but I do not pursue this here. Phillips (2005) has, in the context of his methodology, a much more comprehensive treatment of cointegration in the context of stochastic trend extraction.

Consider a unit root process Z_t that is the (unobservable) common stochastic trend component of two other non-stationary processes Y_t and X_t that are related by the following system of equations:

$$
\begin{aligned}
Y_t &= \alpha Z_t + \epsilon_t \\
X_t &= Z_t + u_t \\
Z_t &= Z_{t-1} + \eta_t
\end{aligned}
\tag{41}
$$

with $\epsilon_t \sim$ i.i.d.$(0, \sigma_\epsilon^2)$, $u_t \sim$ i.i.d.$(0, \sigma_u^2)$ and $\eta_t \sim$ i.i.d.$(0, \sigma_\eta^2)$ where I assume for simplicity that $\epsilon_t, u_s, \eta_\ell$ are independent for all (t, s, ℓ).

It is easy to see that Y_t and X_t are cointegrated, i.e. a linear combination of them forms a stationary process, with the scalar parameter α being the cointegrating coefficient. To see this note that:

$$
\begin{aligned}
Y_t - \alpha X_t &= \alpha Z_t + \epsilon_t - \alpha Z_t - \alpha u_t \rightarrow \\
Y_t &= \alpha X_t + \zeta_t
\end{aligned}
\tag{42}
$$

where $\zeta_t \overset{\text{def}}{=} \epsilon_t - \alpha u_t$ is a stationary process. The inference problems here are (a) estimation of the α parameter and (b) extraction of a smoothed version of the common stochastic trend Z_t.

It should be immediately clear that (b) above can easily be handled by a straightforward application of the signal extraction approach of equations (30) to (32) to the X_t process. However, there is potential loss of information in doing this since Z_t enters into both Y_t and X_t and the two series are connected — not to mention the case where $\epsilon_t, u_s, \eta_\ell$ are not i.i.d. and dependent between them. Therefore a more efficient approach would be to use information from both processes in extracting Z_t and in what follows I propose an SSA-based way to handle both (a) and (b) above.

The estimation of α can be accomplished immediately if we use the properties of the matrix of the autocovariances for each process. Using a similar notation to equations (30) to (32), and the corresponding properties of the series, note that we have:

$$
\begin{aligned}
n^{-1}\widehat{\boldsymbol{\Gamma}}_Y(k) &\approx \alpha^2 n^{-1}\widehat{\boldsymbol{\Gamma}}_Z(k) \Rightarrow \alpha^2 \sigma_\eta^2 w \boldsymbol{J}_{k,k} \\
n^{-1}\widehat{\boldsymbol{\Gamma}}_X(k) &\approx n^{-1}\widehat{\boldsymbol{\Gamma}}_Z(k) \Rightarrow \sigma_\eta^2 w \boldsymbol{J}_{k,k}
\end{aligned}
\tag{43}
$$

The above relationships suggest a simple estimator for α based on the empirical (estimated) eigenvalues of the two autocovariance matrices. We have:

Proposition 2. Under the assumptions of equation (41) and the properties of Proposition 1 we have that a consistent, SSA-based, estimator of the cointegrating cofficient α is given by the square root of the ratio of the two leading estimated eigenvalues of $n^{-1}\widehat{\mathbf{\Gamma}}_Y(k)$ and $n^{-1}\widehat{\mathbf{\Gamma}}_X(k)$ as:

$$\widehat{\alpha} \stackrel{\text{def}}{=} \sqrt{\widehat{\lambda}_{1,Y}/\widehat{\lambda}_{1,X}} \Rightarrow \alpha \qquad (44)$$

The finite sample properties of the above estimator are briefly examined in the simulation analysis of the next section.

Once an estimator of α is available we can combine information from both Y_t and X_t in extracting a smoothed version of the unobserved common stochastic trend Z_t. This is accomplished by averaging as before:

$$\begin{aligned} Y_t + \alpha X_t &= 2\alpha Z_t + \epsilon_t + u_t \rightarrow \\ \Psi_t(\alpha) \stackrel{\text{def}}{=} (2\alpha)^{-1}(Y_t + \alpha X_t) &= Z_t + \xi_t \end{aligned} \qquad (45)$$

where $\xi_t \stackrel{\text{def}}{=} (2\alpha)^{-1}(\epsilon_t + u_t)$. We can now apply SSA to the empirical counterpart of $\Psi_t(\alpha)$, i.e. $\psi_t(\widehat{\alpha})$ or combine the individually SSA-smoothed Y_t and X_t series directly as:

$$\psi_{s,1}(\widehat{\alpha}) \stackrel{\text{def}}{=} (2\widehat{\alpha})^{-1}(y_{s,1} + \widehat{\alpha} x_{s,1}) \qquad (46)$$

The resulting $\psi_{s,1}(\widehat{\alpha})$ series would be the smooth approximation to the common stochastic trend component Z_t and will utilize information from both Y_t and X_t. If no smoothing is desired then the common stochastic trend component is directly approximated by $\psi_t(\widehat{\alpha})$. The performance of the above smoother is also briefly examined in the simulation analysis that follows.

4   Simulation Analysis

In this section I evaluate the proposed methodology using three simulations. For the first simulation I use as the data generating process the local level model of equation (30) with different values for the signal-to-noise ratio q and different assumptions on the distribution of η_t. For a sample size of $n = 250$ observations I perform signal extraction using: (a) the Kalman fixed point smoother of equation (33); (b) the SSA approach of this paper with fixed k and k^*; (c) the approach of Phillips with fixed k and k^*; and (d) exponential smoothing. For (a) and (d) the relevant parameters are estimated from the data. I perform 500 replications and report average mean-squared and mean-absolute errors for the residuals from the true signal. Specifically, if $\widehat{x}_{t,ij}$ denotes the signal estimate for the ith replication for the jth method I compute:

$$AMSE_j \stackrel{\text{def}}{=} \frac{1}{500}\sum_{i=1}^{500}\left[\frac{1}{n}\sum_{t=1}^{n}(x_{t,i} - \widehat{x}_{t,ij})^2\right], \quad AMAE_j \stackrel{\text{def}}{=} \frac{1}{500}\sum_{i=1}^{500}\left[\frac{1}{n}\sum_{t=1}^{n}|x_{t,i} - \widehat{x}_{t,ij}|\right] \quad (47)$$

For the same simulations I also report the average value of the selected k^* using the methodology of the previous section. These results are summarized in Tables 1, 2 and 3.

The results in Table 1 are extremely encouraging. The fixed $k = \sqrt{n}$ filter is very competitive to the Kalman fixed point filter, which is optimal for the data generating process. For the AMSE on the top panel of the table we can observe that as q increases the performance of the optimal k^* filter also becomes very competitive to the Kalman filter, eventually matching it. For the AMAE on the bottom panel of the table we see a similar situation as with the AMSE results. Overall we can say that (a) the fixed k SSA-based filter of this paper can be used reliably for this type of data generating process; and (b) the optimal k^* SSA-based filter is to be preferred when the signal-to-noise ratio is relatively large. Phillip's smoother is trailing after the SSA smoother, its performance being slightly worse. The exponential smoother performs well only for $q = 1$, as expected. Finally, the results in Table 2 summarize the average k^* that was selected across replications.

In Table 3 I present a variation of the results in Table 1. Here I take the signal noise to have been generated by a Student $t_{(2)}$ distribution scaled by a factor σ_η. This distribution does not have finite second moments so the meaning of q is not the same as before. Nevertheless it is instructive to look at the performance of the various filters. We can see that the performance of almost all filters deviates more substantially than before from the performance of the Kalman fixed point filter and their differences do not significantly diminish as q increases (this was expected since for any value of q the signal has infinite variance.) It is only the optimal k^* SSA-based filter that can come close to the Kalman fixed point filter and its performance improves as q increases.

For the second simulation I examine the out-of-sample prediction approach for selecting the degree of smoothing given in equations (38) to (40). I generate 500 replications from a unit root model of size $n = 250$ observations. I then use a rolling window of $n_0 = 200$ observations and an evaluation period of $n_1 = 50$ observations to construct trend prediction bands as in equation (39) and to examine the coverage ratio of equation (40). I take $\tau = 1$, i.e. to consider prediction bands one standard deviation away from the trend. The results in Table 4 below present: the average coverage ratio (ACR), its standard deviation (SD), the median coverage ratio ($Q_{0.50}$), the 10% and 90% quantiles of the coverage ratio ($Q_{0.10}$ and $Q_{0.90}$), and the average selected value of k^*.

The results between the SSA approach and Phillip's approach are similar. The average coverage ratio is about 66% and its distribution is symmetric (the median coverage ratios are practically the same as the mean coverage ratios). This is quite suggestive, given the underlying

normality of η_t, since the coverage probability of a standard normal distribution between ± 1 (standard deviations) is about 68%. In addition we can observe that, on average across replications, the standard deviation of the coverage ratio is proportional to the difference between the two quantiles, i.e. $(Q_{0.90} - Q_{0.10})/SD \approx 1.22$ (1.17 for SSA and 1.28 for Phillips); the corresponding ratio in a standard normal distribution is 1.65. These descriptive statistics, which are based only on 50 observations, are very supporting the proposed methodology for selecting the degree of smoothing in an out-of-sample context: they indicate that the trend selected in this way is approximately "unbiased".

I conclude this section with results from a third simulation, about the finite sample properties of the estimator of the cointegrating coefficient given by Proposition 2 in equation (44) and the properties of the smoother $\psi_{s,1}(\widehat{\alpha})$ in equation (46). I use the system in equation (41) as the data generating process with different values for the parameters and the sample size. I consider three sample sizes $n = \{100, 200, 400\}$ and five values of $\alpha = \{0.10, 0.25, 0.75, 1.00, 2.00\}$. I perform 500 replications, as before, and I compute (a) the average mean absolute deviation of the estimates from their true parameter value and (b) the relative average mean-squared error between the actual common stochastic trend and the corresponding smoother of equation (46). In both cases I take $k = \sqrt{n}$. Specifically, for the average mean absolute deviation I compute:

$$AMAD \stackrel{\text{def}}{=} \frac{1}{500} \sum_{i=1}^{500} |\widehat{\alpha}_i - \alpha| \tag{48}$$

The results given in Table 5 indicate that the proposed estimator does a reasonably good job in estimating the true cointegrating parameter and is indeed consistent. It is possible that its performance in small samples is better in an intermediate range of values for α: for $n = 100$ the largest $AMAD$ is found for the two extremes, i.e. for $\alpha = 0.10$ and $\alpha = 2.00$. This disappears for the larger samples.

For the relative average mean-squared error I do the following. In each replication run I compute the smoother of equation (46) as well as the smoothers based on X_t alone and on $Y_t/\widehat{\alpha}$ alone, that can also approximate Z_t. For each of the smoothers, say $\widehat{z}_{t,1}$, I compute:

$$AMSE(\widehat{z}_{t,1}) \stackrel{\text{def}}{=} \frac{1}{500} \sum_{i=1}^{500} \left[\frac{1}{n} \sum_{t=1}^{n} (Z_t - \widehat{z}_{t,1})^2 \right] \tag{49}$$

and then report the relative average mean-squared errors as:

$$RAMSE_Y \stackrel{\text{def}}{=} \frac{AMSE(y_{t,1}/\widehat{\alpha})}{AMSE(\psi_{t,1}(\widehat{\alpha}))}, \text{ and } RAMSE_X \stackrel{\text{def}}{=} \frac{AMSE(x_{t,1})}{AMSE(\psi_{t,1}(\widehat{\alpha}))} \tag{50}$$

For the results in Table 6 I use the following parameter combination[6]: $\sigma_\epsilon^2 = \sigma_u^2 = 0.2^2$, $\sigma_\eta^2 = 0.1^2$

---

[6]Other parameter combinations produced similar results and are available on request.

and the three noise series were drawn from three independent normal distributions. Two results about the performance of $\psi_{t,1}(\widehat{\alpha})$ are immediately evident:

- if α takes small values, here $\alpha = 0.10$ and $\alpha = 0.25$, and we are close to the no cointegration case then using $\psi_{t,1}(\widehat{\alpha})$ will improve the results compared to using $y_{t,1}/\widehat{\alpha}$ alone but not compared to using $x_{t,1}$ alone — this is of course expected since in that case its essentially X_t that contains the useful information about Z_t.

- if, on the other hand, α takes larger values then using $\psi_{t,1}(\widehat{\alpha})$ is better than using either of the individual smoothers.

All in all the results in Tables 5 and 6 are supportive of the proposed methodology for bivariate cointegrated systems, both for estimation of the cointegrating coefficient and for extracting a smoothed version of the common stochastic trend component.

5   Empirical Illustrations

In this section I present two empirical illustrations of the proposed methodology. First I compare the SSA-based filter of this paper with the Hodrick-Prescott (HP) filter for the series of quarterly observations of the U.S. real GDP (series GDPC96 from the Federal Reserve Bank of St. Louis Database, in billions of chained 2000 dollars). The series has $n = 244$ quarterly observations from 1947:Q1 to 2007:Q4. Then I use the SSA-based filter for smoothing and trend extraction for the weekly prices of Brent crude Oil and of the Euro/US Dollar exchange rate.[7] The series have $n = 477$ weekly observarions from 01/04/1999 to 02/19/2008.

5.1   U.S. Real GDP

The HP filter has been used widely in smoothing trending economic time series and the original 1997 paper was followed by a large literature on optimal filtering. Here we show that the SSA-based filter of this paper performs on par with the HP filter, either when a naive value is used for the smoothing parameter q or when the value of the smoothing parameter is optimized. Consider first the case where $q = 1600$, the value suggested by Hodrick and Prescott in their 1997 paper for quarterly data, and compare it with the SSA-based filter with k fixed to $k = \sqrt{n}$. Figure 5 contains the results. There are several things to notice: first, see that both filters

---

[7]These series are used in the EurOil Index project found at http://econ.uop.gr/ ∼thomakos/EurOil_Index.html where more information can be found.

achieve practically the same degree of smoothing over the entire length of the series and produce practically identical residuals, with the exception of the end of the series. This is due to the construction of the two filters: in the SSA case the last value of the smoothed trend is an unweighted backward moving average, see equation (22). This can be seen as a shortcoming of the SSA-based filter which brings us to a question: are the two filters, as applied, comparable? The answer is clearly no!

As noted in Remark 1 at the end of section 3.3, the underlying stochastic model on which the HP filter is based is for a time series with two not one unit root. Therefore the results of Figure 5 are OK but are not directly comparable. To make a meaningful comparison we need to apply the SSA filter in the first differences of the GDP series and then cumulate the extracted signal. The results from this approach are presented in Figure 6.

We can now make a meaningful comparison between the two filters and we can see that they now match everywhere, both in the smooth trends and the corresponding residuals; see the differences in the scatterplots between Figure 5 and Figure 6. We therefore see that the proposed SSA-based filter achieves the same degree of smoothness as the standard HP filter, after taking into account the stochastic model under which the HP filter operates.

An interesting byproduct of this approach, of applying the SSA filter in the first differences of the GDP series, is the extracted smooth trend. Figure 7 has this trend for both $k = \sqrt{n}$ and for $k = 4$, the latter corresponding to the quarterly frequency of the data - see the squared frequency response for $k = 4$ in Figure 2. We also present for comparison the smooth trend obtained from applying Kalman fixed point smoothing based on the local level model of equation (30) - the estimated signal-to-noise ratio is $\widehat{q} = 0.068$ which makes signal extraction in first differences a more interesting exercise than in levels. Both short and long-term cycles in U.S. output are evident in the three panels of the figure that show the smooth trends. Note the similarity of the SSA-based smooth trend for $k = \sqrt{n}$ and the KFP-based smooth trend - remember that the close performance of these two filtering methods in the context of a similar signal-to-noise ratio was seen in the simulations of the previous section.

In concluding this section I note that all the above analysis was repeated with both k and q (in HP filter) selected in an optimal fashion but the results were qualitatively similar to the ones already presented and are available on request.

5.2 Oil Prices and the Euro/US Dollar Exchange Rate

The two financial time series analyzed in this section are different in nature from the U.S. Real GDP series of the previous section. Unlike the GDP series which was quite smooth and a model with two unit roots was more appropriate, the Oil and Euro/US Dollar series exhibit characteristics consistent with one unit root. In this section I present some graphical results from SSA-based smoothing of the two series and also for the construction of prediction error bands for the extracted trends using the methodology at the end of section 3.5.

Figures 8 and 9 contain the results from smoothing the weekly Oil series. Two values for k are used the fixed value $k = \sqrt{n} = 21$ and the optimal value $k^* = 9$. The data with the smooth trends and the corresponding residual series are in Figure 8 while the autocorrelation functions of the residuals are given in Figure 9. Note the similarity between the results in these two figures and the results from a sample realization from a unit root process in Figure 3. In particular note that there is significant residual autocorrelation that exhibits a cyclical pattern. Figures 10 and 11 contain the related results from smoothing the Euro/US Dollar exchange rate series. The optimal value of k in this case was $k^* = 8$. The underlying patterns are very similar to that of the Oil series, i.e. are consistent with an underlying unit root process. Whether the cyclicality of the residuals in both series can be (partly) attributed to market conditions or its an artifact of the unit root behavior is of course an open question.

In Figure 12 I present the results from a rolling, one week ahead, out-of-sample trend prediction with corresponding one ($\tau = 1$) standard deviation bands. I use a rolling window of $n_0 = 277$ weeks and an evaluation window of $n_1 = 200$ weeks in the figure. The s_{k^*}'s, the prediction error standard deviations are \$5.32 and \$0.03, over the 200 weeks evaluation period. The coverage ratio $CR(1)$ of equation (40) was equal to 65% for the Oil and 69% for the Euro/US Dollar series, a proportion quite consistent with the simulation results presented in Table 4. It is interesting to note that for both series their last actual values exceed the predicted trend values; Oil, in particular, is at the upper prediction band. Both series have (as of 02/27/2008) exceeded their trend values.

Finally, I use the methodology of section 3.6 to examine the possibility of extracting a smooth, common stochastic trend component between the two series. After standardizing the data, to express them in a comparable numerical scale, I apply the smoother of equation (46) and present the result in Figure 13. The averaging operation can clearly be seen in the figure as the smooth trend component runs between the two series. It captures quite well the common evolution of these two assets that have moved closely together over time.

While I make no formal claims about the presence of cointegration, and the possible effects that standardization has on the data, it is very interesting to note that the estimator $\widehat{\alpha}$ that was proposed in equation (44) is almost identical for the estimator that one obtains using the methodology in Phillips (2005). Using the same value for k in both methods I find that, for the standardized data, $\widehat{\alpha} = 1.0592$ while Phillips' estimator is 1.0878. The closeness of the estimates is quite suggestive about the properties of the two methods.

6   Concluding Remarks

Applying the methodology of Singular Spectrum Analysis (SSA) I derive an optimal linear filter for stochastic processes with a unit root. The filter takes the form of a particular moving average and is different from other linear filters that have been used in the existing literature for smoothing time series that have non-stationary characteristics, including unit roots. To best of my knowledge this is the first time that moving average smoothing is given an optimality justification for use with unit root processes.

The frequency response function of the filter is examined, a new method for selecting the degree of smoothing is suggested and I also derive the trend prediction function, which takes a particularly simple form. I compare the proposed filter with Phillips' (2005) methodology for extracting trends from time series unit roots and also with the well known Hodrick-Prescott (1997) (HP) filter used widely for smoothing economic time series. I then show that it can be used for extracting a non-stationary signal from stationary noise with the same degree of success as the filter that corresponds to the data generating process itself.

I also provide an extension of the proposed methodology to the case where we have a system with two series that share a common, non-stationary component and are cointegrated. An SSA-based, consistent estimator of the cointegrating coefficient is suggested along with the corresponding smoother of the common trend.

A simulation study explores some of the characteristics of the filter under a controlled environment and I illustrate its practical usefulness using data for the US real GDP and two financial time series, Oil prices and the Euro/US Dollar exchange rate. For the GDP data I show how to use the SSA-based filter of this paper to get the same results as the HP filter and indicate an alternative method for extracting the cyclical component of output.

Further research is needed for examining the full potential of the proposed method. Interesting extensions of the current work include: finding the appropriate SSA-based filter for processes

with two rather than one unit root, so as to match the data generating process that underlies the application of the HP filter; examining the spectral shape and autocorrelation function of the residual series obtained after smoothing to see whether their apparent periodicities are genuine or an artifact of the underlying unit root process; comparing the performance of the proposed SSA-based filter with other existing filters in the literature; further working on the potential of SSA for application in larger systems of cointegrated series.

References

1. Allen, M.R. and L. A. Smith, 1996. "Optimal Filtering in Singular Spectrum Analysis", *Physics Letters A*, vol. 234, pp. 419-428.

2. Baxter, M., and R. G. King, 1999. "Measuring Business-cycles: Approximate Band-pass Filters for Economic Time Series", *Review of Economics and Statistics*, vol. 81, pp. 575-593.

3. Bell, W., 1984. "Signal Extraction for Non-stationary Time Series", *The Annals of Statistics*, vol. 12, pp. 646-664.

4. Broomhead, D. S. and G. P. King, 1986. "Extracting qualitative dynamics from experimental data", *Physica D*, vol. 20, pp. 217-236.

5. Buchstaber, V. M., 1994. "Time Series Analysis and Grassmanians", in S. Gindikin (Ed.), *Applied Problems of Radon Transforms*, AMS TransactionsSeries 2, Vol. 162, AMS, Providence, RI.

6. Christiano, L. and T. Fitzgerald, 2003. "The Band Pass Filter", *International Economic Review*, vol. 44, pp. 435-465.

7. Dermoune, A., Djehiche, B. and N. Rahmania, 2007. "A Consistent Estimator of the Smoothing Parameter in the Hodrick-Prescott Filter", *Journal of the Japan Statistical Society*, forthcoming.

8. Durbin, J. and S. Koopman, 2001. *Time Series Analysis by State Space Methods*, Oxford: Oxford University Press.

9. Elsner, J. B. and A. A. Tsonis, 1996. *Singular Spectrum Analysis: A New Tool in Time Series Analysis*, Plenum Press: New York.

10. Fuller, W., 1995. *Time Series Analysis*, 2nd edition, New York: John Wiley.

11. Golyandina, N., Nekrutkin, V. and A. Zhigljavsky, 2001. *Analysis of Time Series Structure: SSA and Related Techniques*, Monographs on Statistics and Applied Probability, Chapman and Hall.

12. Harvey, A., 1989. *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge: Cambridge University Press.

13. Hodrick, R. and E. C. Prescott, 1997. "Postwar U.S. Business Cycles: An Empirical Investigation", *Journal of Money, Credit and Banking*, vol. 29, pp. 1-16.

14. King, R.G. and S. T. Rebelo, 1993. "Low Frequency Filtering and Business Cycles", *Journal of Economic Dynamics and Control*, vol. 17, pp. 207-232.

15. Kohn, R. and Ansley, C., 1987. "Signal Extraction for Finite Non-stationary Time Series", *Biometrika*, vol. 74, pp. 411-421.

16. Lisi, F. and A. Medio, 1997, "Is a Random Walk the Best Exchange Rate Predictor", *International Journal of Forecasting*, vol. 13, pp. 255-267.

17. McElroy, T., 2005. "Matrix Formulas for Nonstationary Signal Extraction", SRD Research Report No. RRS2005-04, U.S. Census Bureau.

18. Phillips, P. C. B., 1996, "Spurious Regression Unmasked", CFDP 1135.

19. Phillips, P. C. B., 1998, "Econometric Analysis of Fisher's Equation", CFDP 1180.

20. Phillips, P. C. B., 2005, "Challenges of Trending Time Series Econometrics", *Mathematics and Computers in Simulation*, vol. 68, pp. 401-416.

21. Pollock, D. S. G., 2000. "Trend Estimation and De-trending via Rational Square-wave Filters", *Journal of Econometrics*, vol. 99, pp. 317-334.

22. Priestley, M. B., , *Spectral Analysis and Time Series*, San Diego: Academic Press.

23. Rao, C. R., 1973, *Linear Statistical Inference and its Applications*, New York: John Wiley.

24. Schleicher, C., 2003. "Kolmogorov-Wiener Filters for Finite Time-Series", mimeo.

25. Schlicht, E., 2005. "Estimating the Smoothing Parameter in the So-called Hodrick-Prescott Filter", *Journal of the Japan Statistical Society*, vol. 35, pp. 99-119.

26. Thomakos, D. D., Wang, T. and L. Wille, 2002, "Modeling and Forecasting Realized Futures Volatility using SSA", *Physica A*, vol 312, pp. 505-519.

Appendix

A1. Matrix and vector notation, special matrices

Matrices are denoted by uppercase, italic bold letters such as $\boldsymbol{X}, \boldsymbol{Y}$ etc. and the transpose of any such matrices by $\boldsymbol{X}^\top, \boldsymbol{X}^\top$ etc. Generic vectors as well as the vectors making the rows of any $(N \times k)$ matrix are denoted by lowercase, italic bold letters such as $\boldsymbol{x}_i^\top$ for $i = 1, 2, \ldots, N$ and I write $\boldsymbol{X} \stackrel{\text{def}}{=} [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]^\top$. The vectors making the columns of any such matrix are denoted by lowercase, non-italic bold letters such as $\mathbf{x}_j$, for $j = 1, 2, \ldots, k$ and I write $\boldsymbol{X} \stackrel{\text{def}}{=} [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k]$.

The identity matrix of order k is denoted by $\boldsymbol{I}_k$ and its columns are denoted by $\boldsymbol{i}_j$ for $j = 1, 2, \ldots, k$. An $(N \times k)$ matrix of unities is denoted by $\boldsymbol{J}_{N,k}$ and any of its columns (rows) by $\boldsymbol{J}_N$ ($\boldsymbol{J}_k^\top$). Evidently $\boldsymbol{J}_{N,k} = \boldsymbol{J}_N \boldsymbol{J}_k^\top$.

A2. Vector and matrix norms

The inner product of any two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ is defined as usual $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \stackrel{\text{def}}{=} \boldsymbol{x}^\top \boldsymbol{y}$ and by $\|\boldsymbol{x}\|^2$ I denote the corresponding Euclidean norm $\|\boldsymbol{x}\|^2 \stackrel{\text{def}}{=} \boldsymbol{x}^\top \boldsymbol{x}$. A matrix norm can be defined via the vector norm as follows: any $(N \times k)$ matrix $\boldsymbol{X} \stackrel{\text{def}}{=} [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k]$ can be expressed as a $(Nk \times 1)$ vector $\boldsymbol{x}$ by stacking its columns using the $\mathsf{vec}(\cdot)$ operator as $\boldsymbol{x} = \mathsf{vec}(\boldsymbol{X}) \stackrel{\text{def}}{=} \left[\mathbf{x}_1^\top, \mathbf{x}_2^\top, \ldots, \mathbf{x}_k^\top\right]^\top$. The norm of $\boldsymbol{x}$ is then given by $\|\boldsymbol{x}\|^2 = \sum_{j=1}^{k} \|\mathbf{x}_j\|^2$. We take the norm of $\boldsymbol{x}$ to be the appropriate matrix norm for $\boldsymbol{X}$ and write:

$$\|\boldsymbol{X}\|_M^2 \stackrel{\text{def}}{=} \sum_{j=1}^{k} \|\mathbf{x}_j\|^2 = \sum_{i=1}^{N} \sum_{j=1}^{k} x_{ij}^2 \tag{51}$$

A3. Spectral and singular value decompositions of a matrix

For any $(N \times k)$, $N > k$, matrix $\boldsymbol{X}$ of rank $r \leq k$ define the rank r, $(k \times k)$ symmetric matrix $\boldsymbol{C}_k$ as $\boldsymbol{C}_k \stackrel{\text{def}}{=} \boldsymbol{X}^\top \boldsymbol{X}$ and denote its spectral decomposition as follows:

$$\boldsymbol{C}_k = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top = \sum_{j=1}^{r} \lambda_j \mathbf{u}_j \mathbf{u}_j^\top \tag{52}$$

where $\boldsymbol{U} \stackrel{\text{def}}{=} [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k]$ denotes the $(k \times k)$ matrix of orthonormal eigenvectors, $\|\mathbf{u}_j\| = 1$ and $\mathbf{u}_i^\top \mathbf{u}_j = 0$, and $\boldsymbol{\Lambda}$ denotes the diagonal matrix of its eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ and $\lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_k = 0$.

Next, define the $(N \times N)$ matrix $\boldsymbol{V} \stackrel{\text{def}}{=} [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N]$ to be the matrix of orthonormal eigenvectors of the rank r, $(N \times N)$ matrix $\boldsymbol{X} \boldsymbol{X}^\top$ and note that only the first r is these eigenvectors

correspond to the non-zero eigenvalues of $\boldsymbol{X}\boldsymbol{X}^\top$. It can be shown that for $j = 1, 2, \ldots, r$ we have that $\mathbf{v}_j \stackrel{\text{def}}{=} \boldsymbol{X}\mathbf{u}_j/\sqrt{\lambda_j}$ and that $\boldsymbol{X}$ has a representation as a sum of r rank-one elementary matrices of the form:

$$\boldsymbol{X} = \sum_{j=1}^{r} \sqrt{\lambda_j}\mathbf{v}_j\mathbf{u}_j^\top \tag{53}$$

The above corresponds to the singular value decomposition (SVD) of $\boldsymbol{X}$ and has the following optimality property. Let $\mathcal{M}_q$, $q < r$, denotes the set of $(N \times k)$ matrices of the form $\boldsymbol{Y} = \sum_{j=1}^{q} \mathbf{a}_j\mathbf{b}_j^\top$, for $\{\mathbf{b}_j\}_{j=1}^q$ an orthonormal collection of vectors. Then, the optimal approximation with respect to $\|\cdot\|_M^2$ of $\boldsymbol{X}$ by a matrix in $\mathcal{M}_q$ is given by the SVD of $\boldsymbol{X}$ as follows:

$$\|\boldsymbol{X} - \boldsymbol{Y}_o\|_M^2 = \min_{\boldsymbol{Y}\,\in\,\mathcal{M}_r} \|\boldsymbol{X} - \boldsymbol{Y}\|_M^2, \quad \text{for } \boldsymbol{Y}_o \stackrel{\text{def}}{=} \sum_{j=1}^{q} \sqrt{\lambda_j}\mathbf{v}_j\mathbf{u}_j^\top \tag{54}$$

See Rao (1973) and Golyandina *et al.* (2001) for more detailed results and explanations about the spectral and the SVD decompositions.

A4. Limits of sample cross-moments for a unit root process

Let X_t denote a unit root process so that $X_t = X_{t-1} + \eta_t$ with $X_0 = 0$ and η_t a stationary process. For simplicity assume that η_t forms a sequence of i.i.d. random variables with mean zero and variance σ_η^2, although all results continue to hold for η_t obeying mixing conditions. Let n denote the length of a realization from x_t and let k be either fixed or $k = o(n)$. If we denote by $\widehat{\gamma}(k)$ the sample cross-moment of order k, $\widehat{\gamma}(k) \stackrel{\text{def}}{=} n^{-1}\sum_{t=k+1}^{n} x_t x_{t-k}$ we seek the asymptotic limit of $\widehat{\gamma}(k)$. The required limit is obtained from known results, see for example Fuller (1995), as follows. First, rewrite $\widehat{\gamma}(k)$ as:

$$\widehat{\gamma}(k) = \frac{1}{2n}\sum_{t=k+1}^{n} x_t^2 + \frac{1}{2n}\sum_{t=k+1}^{n} x_{t-k}^2 - \frac{1}{2n}\sum_{t=k+1}^{n}(x_t - x_{t-k})^2 \tag{55}$$

and then note that both the first and second term on the right-hand side of the above equation converge when scaled by n to the following stochastic integral:

$$\frac{1}{2n^2}\sum_{t=k+1}^{n} x_t^2 + \frac{1}{2n^2}\sum_{t=k+1}^{n} x_{t-k}^2 \Rightarrow \sigma_\eta^2 w \stackrel{\text{def}}{=} \sigma_\eta^2 \int_0^1 W(r)^2 \tag{56}$$

where $W(r)$ denotes standard Brownian motion and $\Rightarrow$ signifies weak convergence in the appropriate space. Since the last term on the right-hand side of equation (55) converges in probability to a constant we have that, for all k, $n^{-1}\widehat{\gamma}(k) \Rightarrow \sigma_\eta^2 w$. Therefore, the $(k \times k)$ variance-covariance matrix $\widehat{\boldsymbol{\Gamma}}(k)$ converges to the following stochastic matrix:

$$\frac{1}{n}\widehat{\boldsymbol{\Gamma}}(k) \Rightarrow \sigma_\eta^2 w \boldsymbol{J}_{k,k} \tag{57}$$

The $\boldsymbol{J}_{k,k}$ matrix has one positive eigenvalue equal to $\lambda_1 = k$ and all other eigenvalues equal to zero, $\lambda_j = 0$ for $2 \leq j \leq k$. The eigenvector corresponding to λ_1 is given by $\mathbf{u}_1 = \boldsymbol{J}_k/\sqrt{k}$.

Tables

Table 1. Dispersion of residuals from true signal, $n = 250$,
$\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and $\eta_t \sim \mathcal{N}(0, \sigma_\eta^2)$

	Average Mean Squared Error					
q	KFP	SSA-$\sqrt{n}$	PP-$\sqrt{n}$	ES	SSA-O	PP-O
0.01	0.22	0.25	0.31	0.50	0.80	0.32
0.10	0.06	0.08	0.09	0.15	0.09	0.18
1.00	0.02	0.07	0.07	0.06	0.03	0.17
	Average Mean Absolute Error					
q	KFP	SSA-$\sqrt{n}$	PP-$\sqrt{n}$	ES	SSA-O	PP-O
0.01	0.38	0.40	0.44	0.55	0.68	0.45
0.10	0.20	0.22	0.24	0.30	0.24	0.33
1.00	0.11	0.20	0.21	0.20	0.14	0.31

Notes to Table 1. The column nomenclature is as follows: KFP denotes Kalman fixed point smoothing; SSA-$\sqrt{n}$ and PP-$\sqrt{n}$ denote SSA and Phillips' smoothing with k fixed to $\sqrt{n}$; SSA-O and PP-O denote SSA and Phillips' smoothing with k^* selected via equation (37).

Table 2. Selected Degree of Smoothing k^*

	$n=250$		$n=500$	
q	SSA-O	PP-O	SSA-O	PP-O
0.01	6	7	7	9
0.10	6	7	8	10
1.00	6	8	8	11

Note to Table 2. For column nomenclature see note of Table 1.

Table 3. Dispersion of residuals from true signal, $n = 250$,

$\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$ **and** $\eta_t \sim t_{(2)}\sigma_\eta$

q	KFP	SSA-$\sqrt{n}$	PP-$\sqrt{n}$	ES	SSA-O	PP-O
	Mean Squared Error					
0.01	0.58	0.93	1.13	1.50	1.03	2.18
0.10	0.16	0.63	0.68	0.61	0.31	1.40
1.00	0.03	0.68	0.75	0.47	0.24	1.59
	Mean Absolute Error					
0.01	0.57	0.61	0.68	0.83	0.74	0.95
0.10	0.30	0.49	0.53	0.49	0.35	0.78
1.00	0.14	0.49	0.53	0.33	0.26	0.81

Note to Table 3. Column nomenclature and table entries as in Table 1.

Table 4. Coverage Ratio Statistics &

Out-of-sample Selection of Degree of Smoothing

$X_t = X_{t-1} + \eta_t \ , \ \eta_t \sim \mathcal{N}(0, \sigma_\eta^2)$

Method	ACR	SD	$Q_{0.10}$	$Q_{0.50}$	$Q_{0.90}$	k^*
SSA-O	0.66	0.06	0.58	0.66	0.72	4
PP-O	0.63	0.07	0.54	0.64	0.72	7

Notes for Table 4. Column nomenclature is as follows: ACR is the average coverage ratio; SD is its corresponding standard deviation; Q_α is the $\alpha\%$ quantile of the coverage ratio; k^* is the average selected optimal k. Row nomenclature as in Table 1. Table entries are averages across replications.

Table 5. Average Mean Absolute Deviation of estimator $\widehat{\alpha}$ from true value

n	$\alpha = 0.10$	$\alpha = 0.25$	$\alpha = 0.75$	$\alpha = 1.00$	$\alpha = 2.00$
100	0.095	0.052	0.043	0.051	0.075
200	0.037	0.023	0.021	0.023	0.041
400	0.014	0.010	0.011	0.012	0.020

Notes for Table 5. Table entries report the AMAD of equation (48) on the performance of the estimator $\widehat{\alpha}$ of the cointegrating parameter given in equation (44).

Table 6. Relative Average Mean Squared Error for the smoother $\psi_t(\widehat{\alpha})$

n	$\alpha = 0.10$	$\alpha = 0.25$	$\alpha = 0.75$	$\alpha = 1.00$	$\alpha = 2.00$
			$RAMSE_Y$		
100	1.716	1.454	1.091	1.047	0.992
200	1.694	1.351	1.054	1.025	0.997
400	1.613	1.240	1.029	1.013	0.999
			$RAMSE_X$		
100	0.579	0.826	1.029	1.046	1.066
200	0.599	0.871	1.013	1.028	1.036
400	0.676	0.920	1.009	1.014	1.019

Notes for Table 6. Table entries report the RAMSE of equation (50) on the relative performance of the smoother for the common stochastic trend component given equation (46). Values greater than 1 indicate superior performance of $\psi_t(\widehat{\alpha})$ relative to the smoothers based on the individual series alone.
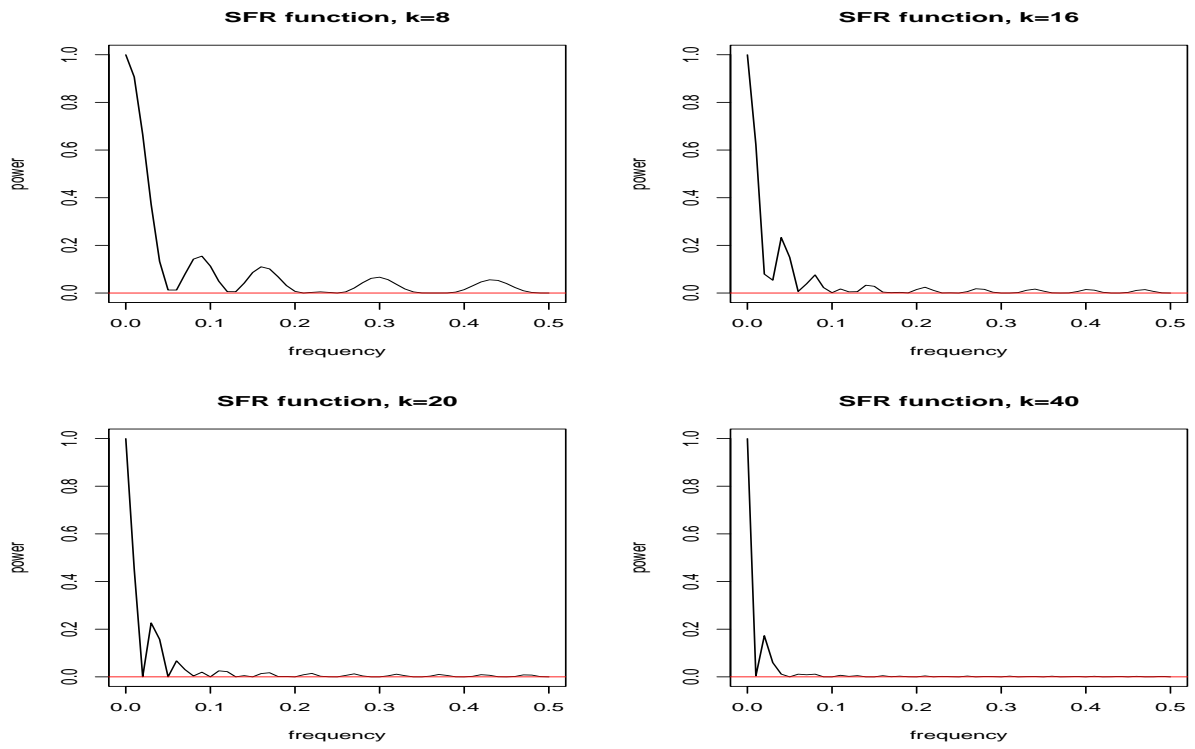
Figure 1: Squared Frequency Response $|\mathcal{R}(\omega)|^2$ for various values of k
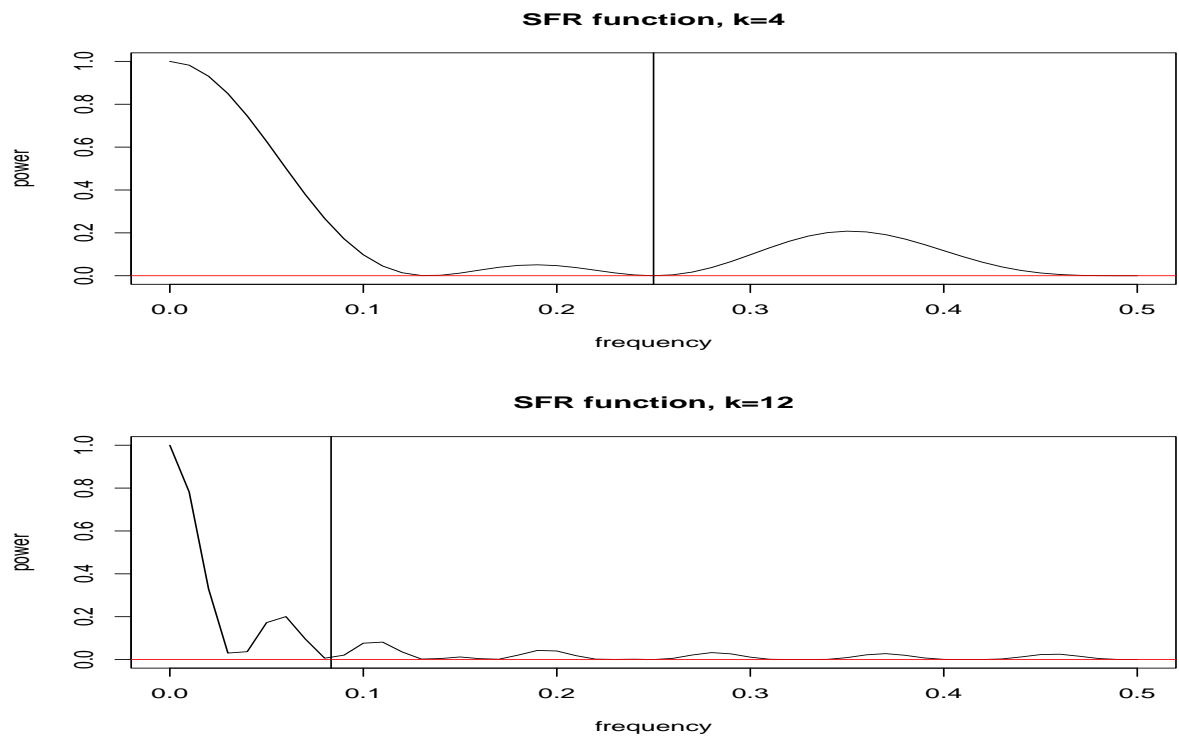
Figure 2: Squared Frequency Response $|\mathcal{R}(\omega)|^2$ - erasing specific frequencies

36
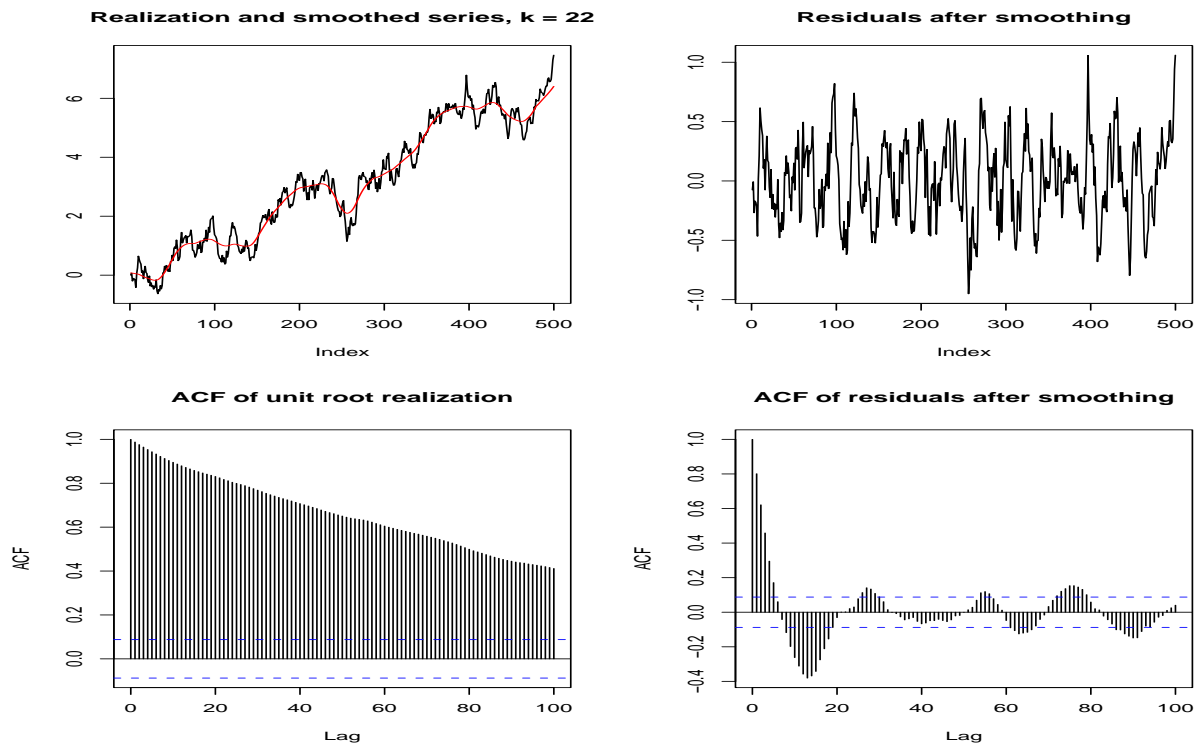
Figure 3: Sample realization from a unit root process and its smooth and residual components
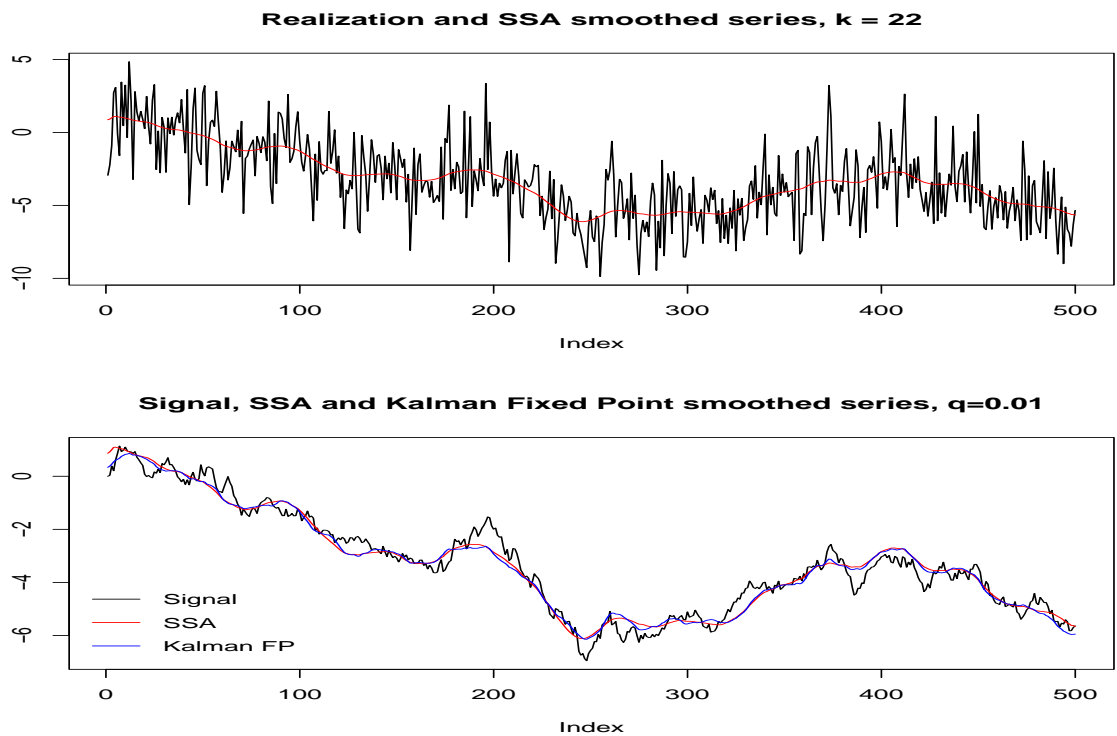
Figure 4: Sample realization from a local level model, signal and smooth series
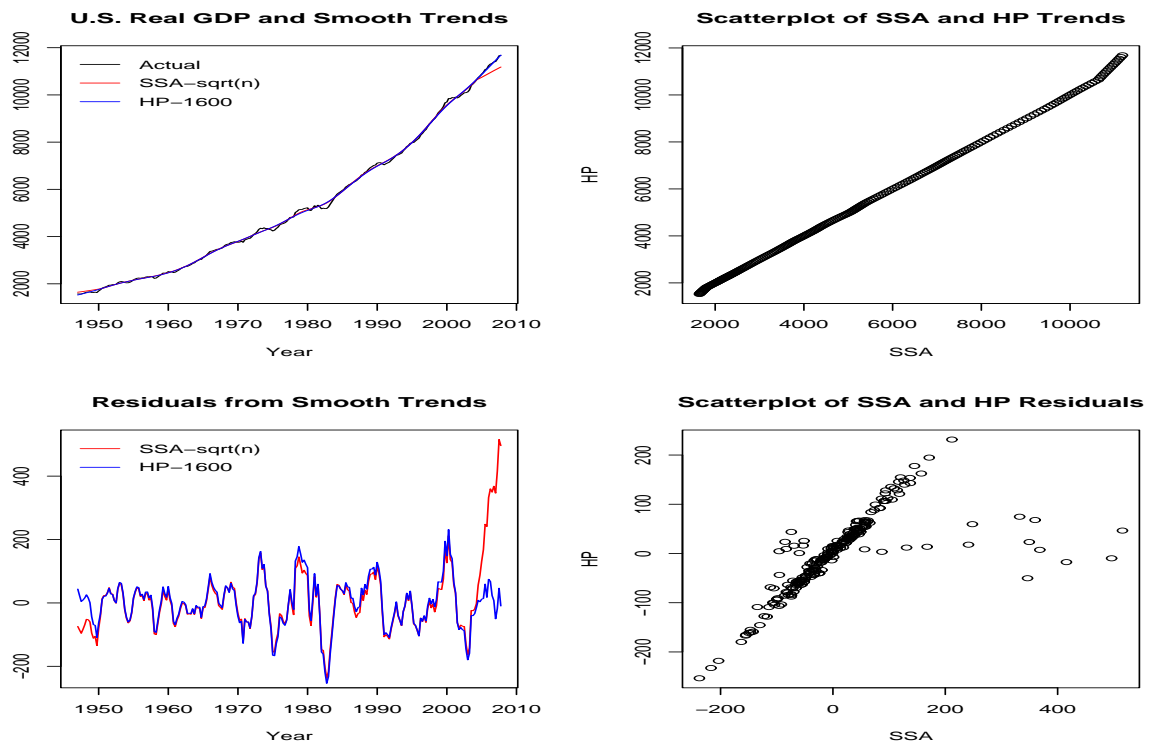
Figure 5: SSA and HP Smoothing of the U.S. Real GDP - smoothing parameters are fixed to $k = \left\lceil \sqrt{244} \right\rceil = 15$ and $q = 1600$ respectively
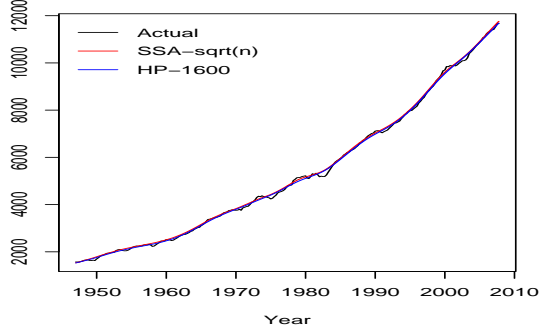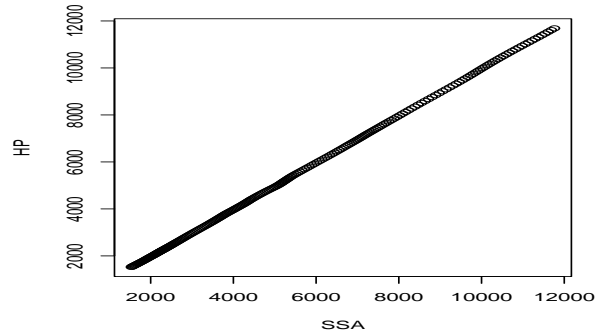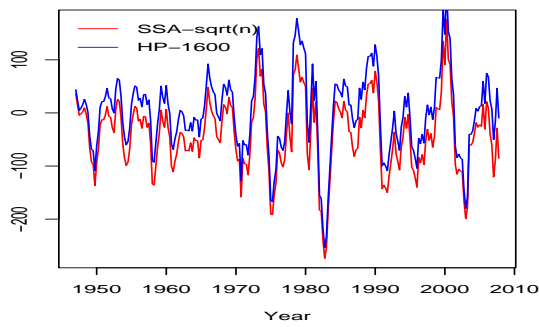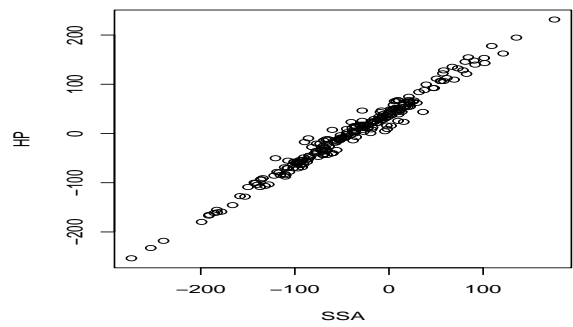
Figure 6: SSA and HP Smoothing of the U.S. Real GDP, SSA filter applied in differences - smoothing parameters are fixed to $k = \left[\sqrt{244}\right] = 15$ and $q = 1600$ respectively
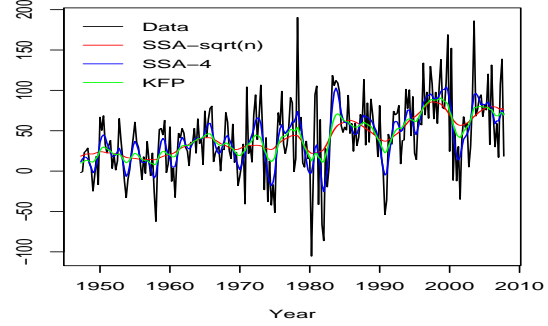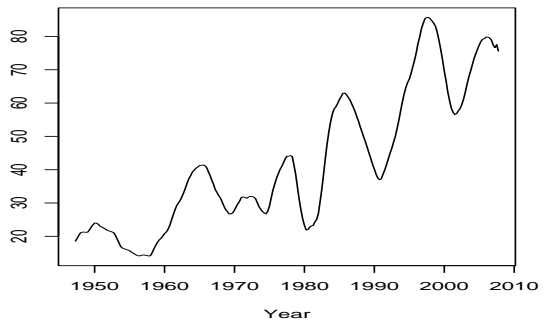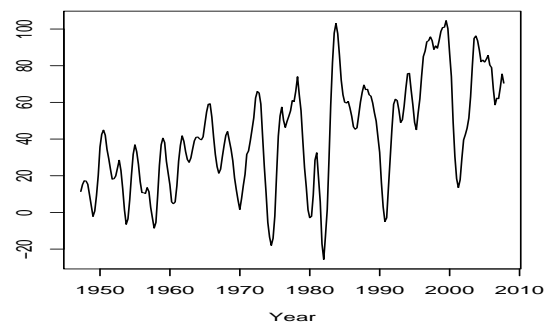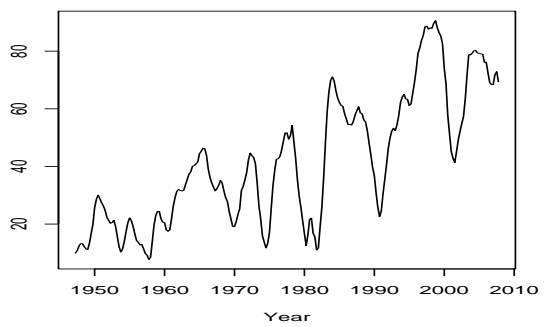
Figure 7: SSA and KFP-based smooth trend extracted from first difference of U.S. Real GDP - smoothing parameter is fixed to $k = \left[\sqrt{244}\right] = 15$ and $k = 4$ respectively
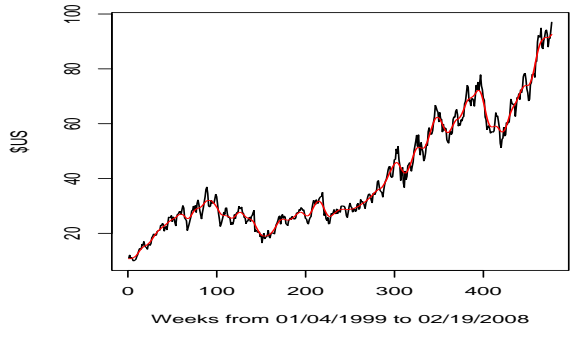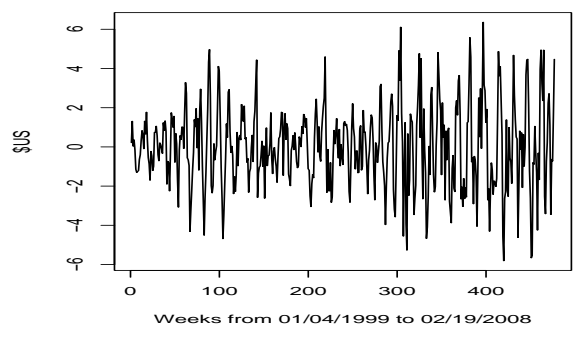
Figure 8: Data for weekly prices of Brent Oil, SSA-based smooth trends and residuals

Oil − ACF for Residuals from Smooth Trend, k=sqrt(n)


Oil− ACF for Residuals from Smooth Trend, optimal k
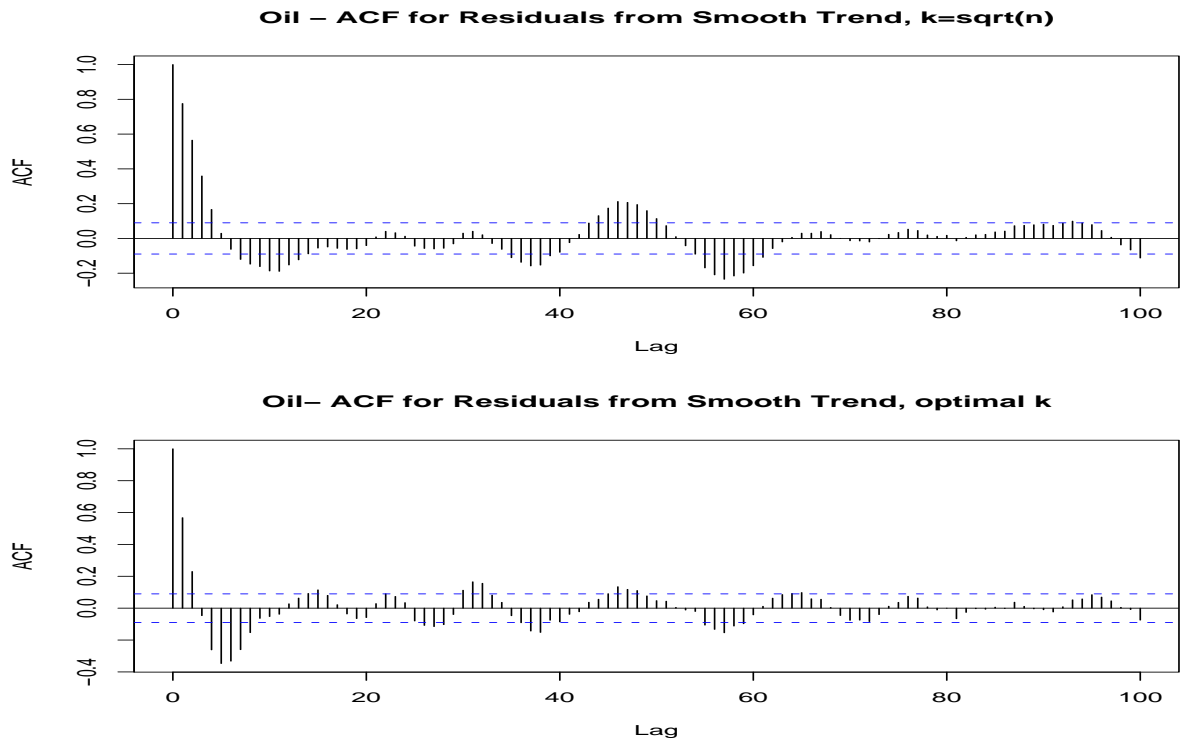
Figure 9: Autocorrelation functions for residual series from Figure 8

kly Euro/$US Exchange Rate and Smooth Trend, k

Residuals from Smooth Trend, k=sqrt(n)

kly Euro/$US Exchange Rate and Smooth Trend, optimal k

Residuals from Smooth Trend, optimal k
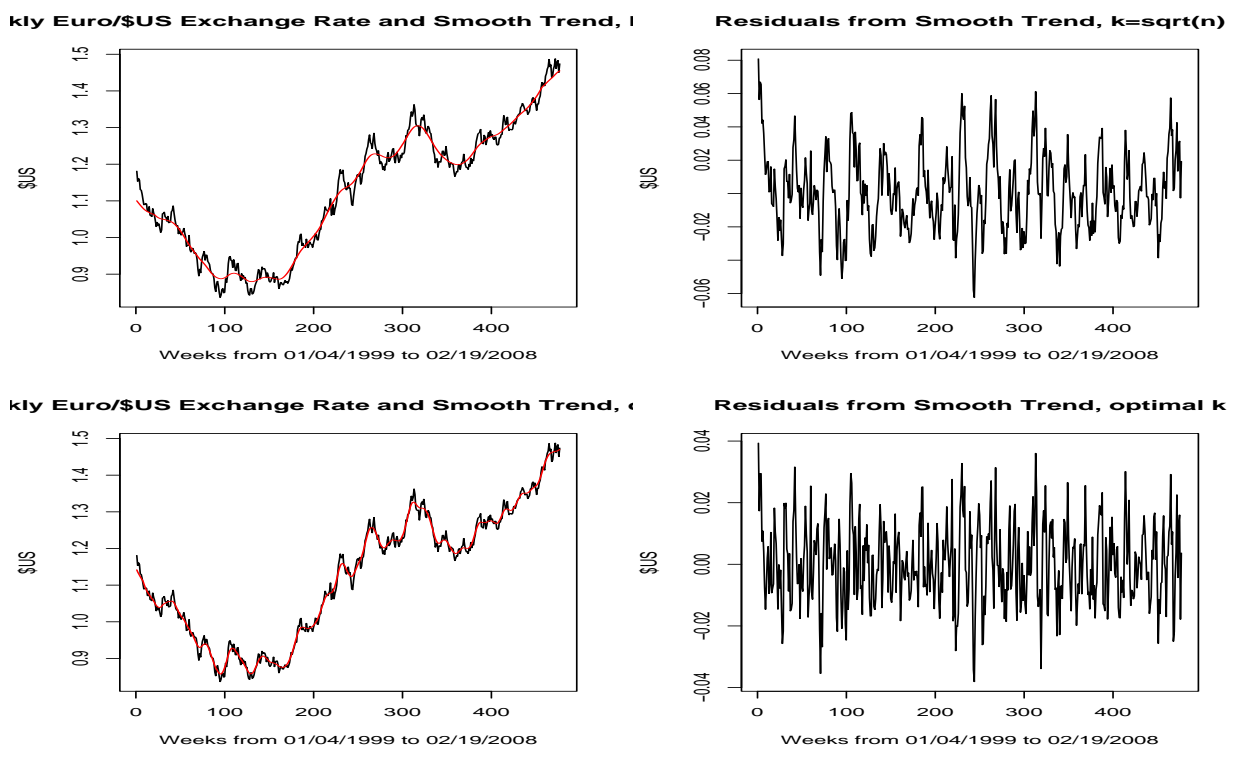
Figure 10: Data for weekly prices of Euro/US Dollar exchange rate, SSA-based smooth trends and residuals
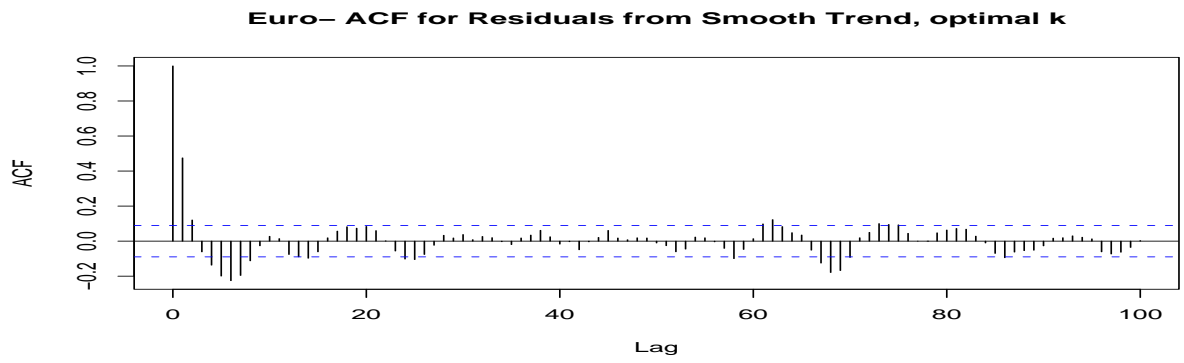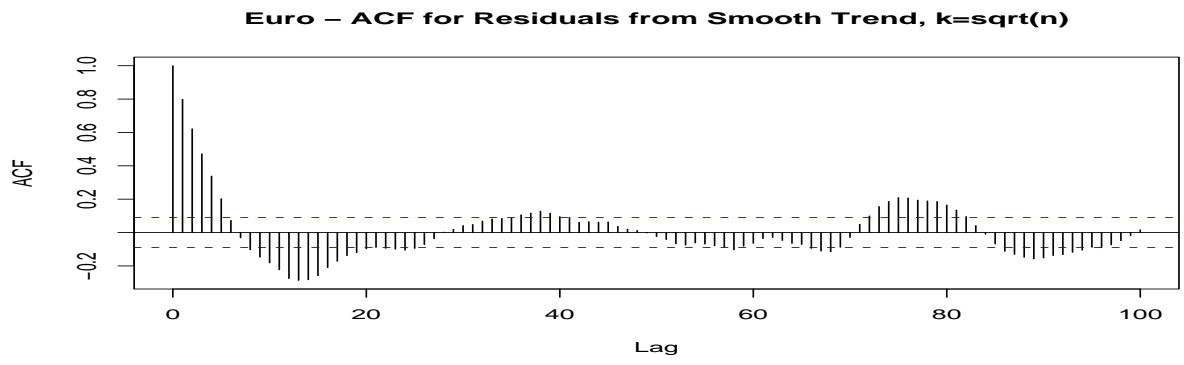
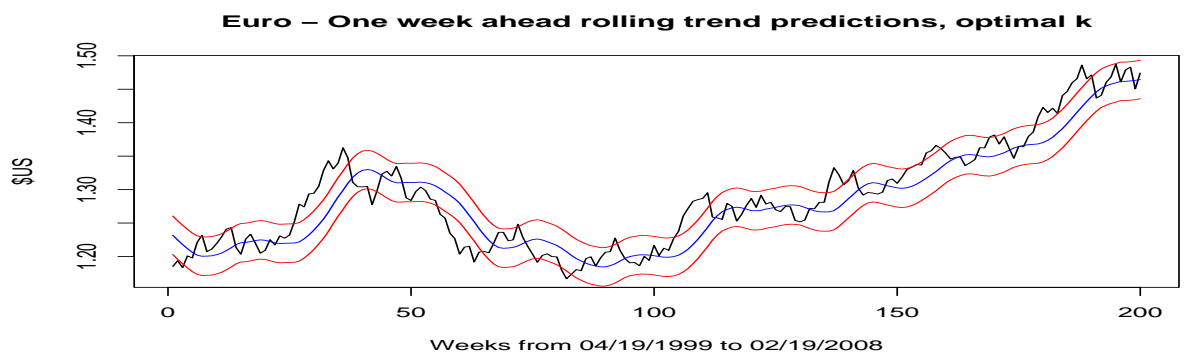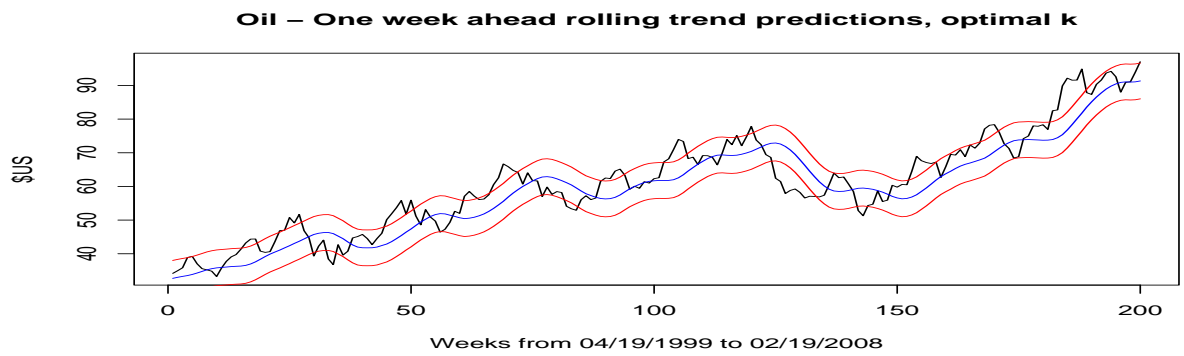Figure 11: Autocorrelation functions for residual series from Figure 10

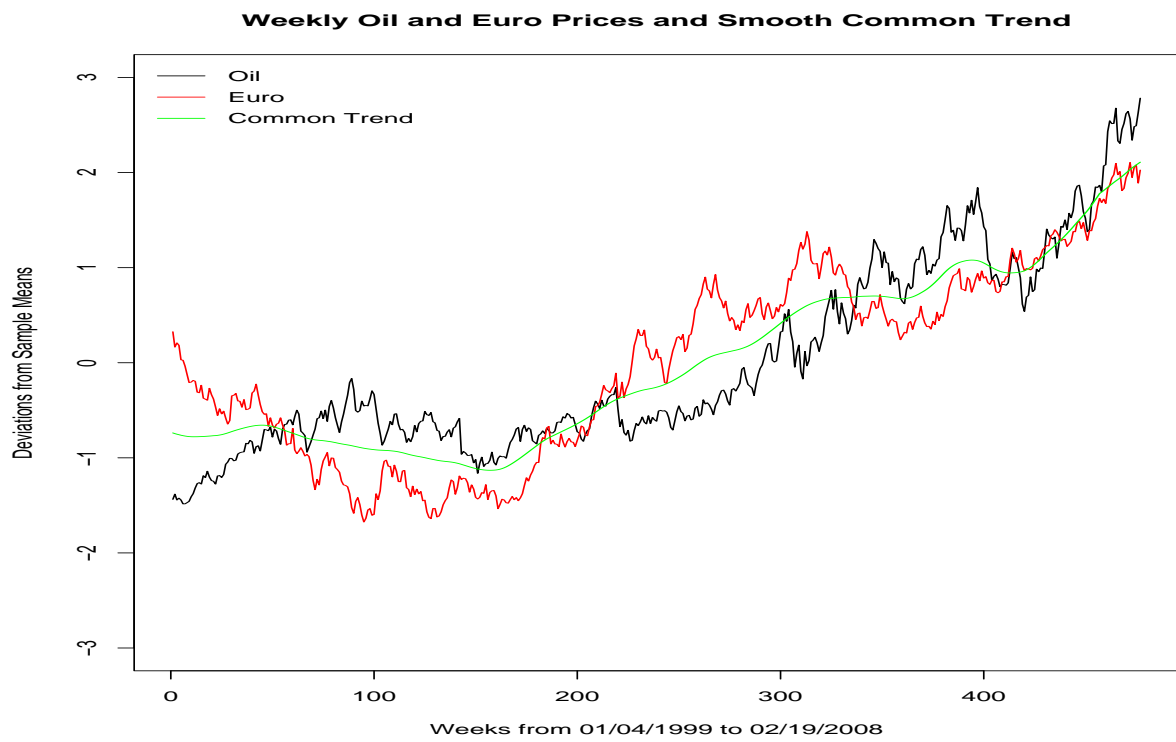Figure 12: Rolling out-of-sample trend predictions and one standard deviation prediction bands

Figure 13: Oil and Euro weekly series, standardized, and common trend component