# A simple state space model of house prices

Rainer Schulz and Axel Werwatz

## 1 Introduction

For most people, purchasing a house is a major decision. Once purchased, the house will by far be the most important asset in the buyer's portfolio. The development of its price will have a major impact on the buyer's wealth over the life cycle. It will, for instance, affect her ability to obtain credit from commercial banks and therefore influence her consumption and savings decisions and opportunities. The behavior of house prices is therefore of central interest for (potential) house buyers, sellers, developers of new houses, banks, policy makers or, in short, the general public.

An important property of houses is that they are different from each other. Hence, while houses in the same market (i.e., the same city, district or neighborhood) will share some common movements in their price there will at all times be idiosyncratic differences due to differences in maintenance, design or furnishing. Thus, the average or median price will depend not only on the general tendency of the market, but also on the composition of the sample. To calculate a price index for real estate, one has to control explicitly for idiosyncratic differences. The hedonic approach is a popular method for estimating the impact of the characteristics of heterogenous goods on their prices.

The statistical model used in this chapter tries to infer the common component in the movement of prices of 1502 single-family homes sold in a district of Berlin, Germany, between January 1980 and December 1999. It combines hedonic regression with Kalman filtering. The Kalman filter is the standard statistical tool for filtering out an unobservable, common component from idiosyncratic, noisy observations. We will interpret the common price component as an index of house prices in the respective district of Berlin. We assume that the index follows an autoregressive process. Given this assumption, the model is writable in state space form.

The remainder of this chapter is organized as follows. In the next section we propose a statistical model of house prices and discuss its interpretation and estimation. Section 4 introduces the data, while Section 5 describes the quantlets used to estimate the statistical model. In this section we present also the estimation results for our data. The final section gives a summary.

## 2 A Statistical Model of House Prices

### 2.1 The Price Function

The standard approach for constructing a model of the prices of heterogeneous assets is hedonic regression (Bailey, Muth and Nourse, 1963; Hill, Knight and Sirmans, 1997; Shiller, 1993). A hedonic model starts with the assumption that on the average the observed price is given by some function $f(I_t, X_{n,t}, \beta)$. Here, $I_t$ is a common price component that "drives" the prices of all houses, the vector $X_{n,t}$ comprises the characteristics of house $n$ and the vector $\beta$ contains all coefficients of the functional form.

Most studies assume a log-log functional form and that $I_t$ is just the constant of the regression for every period (Clapp and Giaccotto, 1998; Cho, 1996). In that case

$$p_{n,t} = I_t + x_{n,t}^\top \beta + \varepsilon_{n,t} \ . \tag{1}$$

Here, $p_{n,t}$ denotes the log of the transaction price. The vector $x_{n,t}$ contains the transformed characteristics of house $n$ that is sold in period $t$. The idiosyncratic influences $\varepsilon_{n,t}$ are white noise with variance $\sigma_\varepsilon^2$.

Following Schwann (1998), we put some structure on the behavior of the common price component over time by assuming that the common price component follows an autoregressive moving average (ARMA) process. For our data it turns out that the following AR(2) process

$$I_t = \phi_1 I_{t-1} + \phi_2 I_{t-2} + \nu_t \tag{2}$$

with $I_0 = 0$ suffices. This autoregressive specification reflects that the market for owner-occupied houses reacts sluggish to changing conditions and that any price index will thus exhibit some autocorrelation. This time-series-based way of modelling the behavior of $I_t$ is more parsimonious than the conventional hedonic regressions (which need to include a seperate dummy variable for each time period) and makes forecasting straightforward.

## 2.2 State Space Form

We can rewrite our model (1) and (2) in *State Space Form* (SSF) (Gourieroux and Monfort, 1997). In general, the SSF is given as:

$$\alpha_t = c_t + T_t \alpha_{t-1} + \varepsilon_t^s \tag{3a}$$

$$y_t = d_t + Z_t \alpha_t + \varepsilon_t^m \tag{3b}$$

$$\varepsilon_t^s \sim (0, R_t) \ , \ \varepsilon_t^m \sim (0, H_t) \ . \tag{3c}$$

The notation partially follows Harvey (1989; 1993). The first equation is the *state equation* and the second is the *measurement equation.* The characteristic structure of state space models relates a series of unobserved values $\alpha_t$ to a set of observations $y_t$. The unobserved values $\alpha_t$ represent the behavior of the system over time (Durbin and Koopman, 2001).

The unobservable state vector $\alpha_t$ has the dimension $K \geqslant 1$, $T_t$ is a square matrix with dimension $K \times K$, the vector of the observable variables $y_t$ has the dimension $N_t \times 1$. Here, $N_t$ denotes the number of observations $y_{t,n}$ in period $t \leqslant T$. If the number of observations varies through periods, we denote

$$N \stackrel{\text{def}}{=} \max_{t=1,\cdots,T} N_t \ .$$

The matrix $Z_t$ contains constant parameters and other exogenous observable variables. Finally, the vectors $c_t$ and $d_t$ contain some constants. The system matrices $c_t$, $T_t$, $R_t$, $d_t$, $Z_t$, and $H_t$ may contain unknown parameters that have to be estimated from the data.

In our model—that is (1) and (2)—, the common price component $I_t$ and the quality coefficients $\beta$ are unobservable. However, whereas these coefficients

are constant through time, the price component evolves according to (2). The parameters $\phi_1$, $\phi_2$, and $\sigma_\nu^2$ of this process are unknown.

The observed log prices are the entries in $y_t$ of the measurement equation and the characteristics are entries in $Z_t$. In our data base we observe three characteristics per object. Furthermore, we include the constant $\beta_0$. We can put (1) and (2) into SSF by setting

$$\alpha_t = \begin{bmatrix} I_t \\ \phi_2 I_{t-1} \\ \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \; T_t = \begin{bmatrix} \phi_1 & 1 & 0 & 0 & 0 & 0 \\ \phi_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \; \varepsilon_t^s = \begin{bmatrix} \nu_t \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{4a}$$

$$y_t = \begin{bmatrix} p_{1,t} \\ \dots \\ p_{N_t,t} \end{bmatrix}, \; Z_t = \begin{bmatrix} 1 & 0 & x_{1,t}^\top \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_{N_t,t}^\top \end{bmatrix}, \; \varepsilon_t^m = \begin{bmatrix} \varepsilon_{1,t} \\ \vdots \\ \varepsilon_{N_t,t} \end{bmatrix} \tag{4b}$$

For our model, both $c_t$ and $d_t$ are zero vectors. The transition matrices $T_t$ are non time-varying. The variance matrices of the state equation $R_t$ are identical for all $t$ and equal to a $6 \times 6$ matrix, where the first element is $\sigma_\nu^2$ and all other elements are zeros. $H_t$ is a $N_t \times N_t$ diagonal matrix with $\sigma_\varepsilon^2$ on the diagonal. The variance $\sigma_\varepsilon^2$ is also an unknown parameter.

The first two elements of the state equation just resemble the process of the common price component given in (2). However, we should mention that there are other ways to put an AR(2) process into a SSF (see Harvey, 1993, p. 84). The remaining elements of the state equation are the implicit prices $\beta$ of the hedonic price equation (1). Multiplying the state vector $\alpha_t$ with row $n$ of the matrix $Z_t$ gives $I_t + x_{t,n}^\top \beta$. This is just the functional relation (1) for the log price without noise. The noise terms of (1) are collected in the SSF in the vector $\varepsilon_t^m$. We assume that $\varepsilon_t^m$ and $\varepsilon_t^s$ are uncorrelated. This is required for identification (Schwann, 1998, p. 274).

# 3 Estimation with Kalman Filter Techniques

## 3.1 Kalman Filtering given all parameters

Given the above SSF and all unknown parameters $\psi \stackrel{\text{def}}{=} (\phi_1, \phi_2, \sigma_\nu^2, \sigma_\varepsilon^2)$, we can use Kalman filter techniques to estimate the unknown coefficients $\beta$ and the process of $I_t$. The Kalman filter technique is an algorithm for estimating the unobservable state vectors by calculating its expectation conditional on information up to $s \leqslant T$. In the ongoing, we use the following general notation:

$$a_{t|s} \stackrel{\text{def}}{=} \mathrm{E}[\alpha_t | \mathcal{F}_s] \tag{5a}$$

denotes the filtered state vector and

$$P_{t|s} \stackrel{\text{def}}{=} \mathrm{E}[(\alpha_t - a_{t|s})(\alpha_t - a_{t|s})^\top | \mathcal{F}_s] \tag{5b}$$

denotes the covariance matrix of the estimation error and $\mathcal{F}_s$ is a shorthand for the information available at time $s$.

Generally, the estimators delivered by Kalman filtering techniques have minimum mean-squared error among all linear estimators (Shumway and Stoffer, 2000, Chapter 4.2). If the initial state vector, the noise $\varepsilon^m$ and $\varepsilon^s$ are multivariate Gaussian, then the Kalman filter delivers the optimal estimator among all estimators, linear and nonlinear (Hamilton, 1994, Chapter 13).

The Kalman filter techniques can handle missing observations in the measurement equation (3b). For periods with less than $N$ observations, one has to adjust the measurement equations. One can do this by just deleting all elements of the measurement matrices $d_t$, $Z_t$, $H_t$ for which the corresponding entry in $y_t$ is a missing value. The quantlets in XploRe use this procedure. Another way to take missing values into account is proposed by Shumway and Stoffer (1982; 2000): replace all missing values with zeros and adjust the other measurement matrices accordingly. We show in Appendix 6.1 that both methods deliver the same results. For periods with no observations the Kalman filter techniques recursively calculate an estimate given recent information (Durbin and Koopman, 2001).

## 3.2 Filtering and state smoothing

The Kalman *filter* is an algorithm for sequently updating our knowledge of the system given a new observation $y_t$. It calculates one step predictions conditional on $s = t$. Using our general expressions, we have

$$a_t = \mathrm{E}[\alpha_t | \mathcal{F}_t]$$

and

$$P_t = \mathrm{E}[(\alpha_t - a_t)(\alpha_t - a_t)^\top | \mathcal{F}_t] \ .$$

Here we use the standard simplified notation $a_t$ and $P_t$ for $a_{t|t}$ and $P_{t|t}$. As a by-product of the filter, the recursions calculate also

$$a_{t|t-1} = \mathrm{E}[\alpha_t | \mathcal{F}_{t-1}]$$

and

$$P_{t|t-1} = \mathrm{E}[(\alpha_t - a_{t|t-1})(\alpha_t - a_{t|t-1})^\top | \mathcal{F}_{t-1}] \ .$$

We give the filter recursions in detail in Subsection 5.3.

The Kalman *smoother* is an algorithm to predict the state vector $\alpha_t$ given the whole information up to $T$. Thus we have with our general notation $s = T$ and

$$a_{t|T} = \mathrm{E}[\alpha_t | \mathcal{F}_T]$$

the corresponding covariance matrix

$$P_{t|T} = \mathrm{E}[(\alpha_t - a_{t|T})(\alpha_t - a_{t|T})^\top | \mathcal{F}_T] \ .$$

We see that the filter makes one step predictions given the information up to $t \in \{1, \ldots, T\}$ whereas the smoother is backward looking. We give the smoother recursions in detail in Subsection 5.5.

## 3.3 Maximum likelihood estimation of the parameters

Given the system matrices $c_t$, $T_t$, $R_t$, $d_t$, $Z_t$, and $H_t$, Kalman filtering techniques are the right tool to estimate the elements of the state vector. However, in our model some of these system matrices contain unknown parameters $\psi$. These parameters have to be estimated by maximum likelihood.

Given a multivariate Gaussian error distribution, the value of the log likelihood function $l(\psi)$ for a general SSF is up to an additive constant equal to:

$$-\frac{1}{2}\sum_{t=1}^{T}\ln|F_t| - \frac{1}{2}\sum_{t=1}^{T} v_t^\top F_t^{-1} v_t \ . \tag{9}$$

Here,

$$v_t \stackrel{\text{def}}{=} y_t - d_t - Z_t a_{t|t-1} \tag{10}$$

are the *innovations* of the filtering procedure and $a_{t|t-1}$ is the conditional expectation of $\alpha_t$ given information up to $t-1$. As we have already mentioned, these expressions are a by-product of the filter recursions. The matrix $F_t$ is the covariance matrix of the innovations at time $t$ and also a by-product of the Kalman filter. The above log likelihood is known as the *prediction error decomposition form* (Harvey, 1989). Periods with no observations do not contribute to the log likelihood function.

Starting with some initial value, one can use numerical maximization methods to obtain an estimate of the parameter vector $\psi$. Under certain regularity conditions, the maximum likelihood estimator $\tilde{\psi}$ is consistent and asymptotically normal. One can use the information matrix to calculate standard errors of $\tilde{\psi}$ (Hamilton, 1994).

## 3.4 Diagnostic checking

After fitting a SSF, one should check the appropriateness of the results by looking at the *standardized residuals*

$$v_t^{st} = F_t^{-1/2} v_t \ . \tag{11}$$

If all parameters of the SSF were known, $v_t^{st}$ would follow a multivariate standardized normal distribution (Harvey, 1989, see also (9)). We know that $F_t$ is a symmetric matrix and that it should be positive definite (recall that it is just the covariance matrix of the innovations $v_t$). So

$$F_t^{-1/2} = C_t \Lambda_t^{-1/2} C_t^\top \ , \tag{12}$$

where the diagonal matrix $\Lambda_t$ contains all eigenvalues of $F_t$ and $C_t$ is the matrix of corresponding normalized eigenvectors (Greene, 2000, p.43). The standardized residuals should be distributed normally with constant variance, and should show no serial correlation. It is a signal for a misspecified model when the residuals do not possess these properties. To check the properties, one can use standard test procedures. For example, a Q-Q plot indicates if the quantiles of the residuals deviate from the corresponding theoretical quantiles of a normal distribution. This plot can be used to detect non-normality. The Jarque-Bera test for normality can also be used for testing non-normality of the residuals (Bera and Jarque, 1982). This test is implemented in XploRe as `jarber`.

In the empirical part, we combine Kalman filter techniques and maximum likelihood to estimate the unknown parameters and coefficients of the SSF for the house prices in a district of Berlin.

# 4 The Data

The data set is provided by the *Gutachterausschuß für Grundstückswerte in Berlin*, an expert commission for Berlin's real estate market. The commission collects information on all real estate transactions in Berlin in a data base called *Automatisierte Kaufpreissammlung*.

Here, we use data for 1502 sales of detached single-family houses in a district of Berlin for the years 1980 to 1999, stored in MD*BASE. Besides the price, we observe the size of the lot, the floor space, and the age of the house. The data set `XFGhouseprice` contains the log price observations for all 80 quarters. There are at most $N = 43$ observations in any quarter. The following lines of XploRe code

```
Y = read("XFGhouseprice.dat")
Y[1:20,41:44]
```

can be used to take a look at the entries of `XFGhouseprice`. Every column gives the observations for one quarter. Thus, in columns 41 to 44 we find the observations for all quarters of 1990. If less than 43 transactions are observed in a quarter the remaining entries are filled with the missing value code `NaN`. Only in the first quarter of the year 1983 we observe 43 transactions.

The corresponding data set `XFGhousequality` contains the observed characteristics of all houses sold. They are ordered in the following way: each column contains all observations for a given quarter. Remember that for every house we observe log size of the lot, log size of the floor space and age. The first three rows of a column refer to the first house in $t$, the next three to the second house and so on.

Let us look at the characteristics of the first two observations in 1990:1. Just type the following lines in the XploRe input window

```
X = read("XFGhousequality.dat")
X[1:6,41]'
```

After compiling, you get the output

```
[1,]   6.1048   4.7707  53  6.5596  5.1475  13
```

The size of the lot for the second house is about 706 square meters (just take the antilog). The size of the floor space is 172 square meters and the age is 13 years.

The following table shows summary statistics of our Berlin house price data.

```
"=========================================================="
" Summary statistics for the Berlin house price data      "
"=========================================================="
" Sample for 80 quarters with 1502 observations           "
"                                                         "
"     Observations per period                             "
```

```
"     ------------------------------------------------"
"        Minimum =  4    Average = 18.77    Maximum = 43  "
"                                                        "
"     Transaction prices (in thousand DM)                "
"     ------------------------------------------------"
"        Minimum =  100.00      Average   = 508.46        "
"        Maximum = 1750.01      Std. Dev. = 197.92        "
"                                                        "
"     Size of the lot (in square meters)                 "
"     ------------------------------------------------"
"        Minimum =  168.00      Average   = 626.18        "
"        Maximum = 2940.00      Std. Dev. = 241.64        "
"                                                        "
"     Size of the floor space (in square meters)         "
"     ------------------------------------------------"
"        Minimum =  46.00       Average   = 144.76        "
"        Maximum = 635.00       Std. Dev. = 48.72         "
"                                                        "
"     Age of the building (in years)                     "
"     ------------------------------------------------"
"        Minimum =   0          Average   = 28.59         "
"        Maximum = 193          Std. Dev. = 21.58         "
"================================================"
                                          XFGsssm1.xpl
```

Not surprisingly for detached houses there are large differences in the size of the lot. Some houses were new in the period of the sale while one was 193 years old. That is a good example for the potential bias of the average price per quarter as a price index. If we do not control explicitly for depreciation we might obtain a low price level simply because the houses sold in a quarter were old.

Nevertheless, the average price per quarter can give an indication of the price level. Figure 1 shows the average price per quarter along with confidence intervals at the 90% level. Instead of the average price, we could also calculate an average adjusted price, where the most important characteristic is used for the adjustment. Such adjustment is attained by dividing the price of every house by—for example—the respective size of the lot. However, even in that case we would control only for one of the observed characteristics. In our model we will control for all of the observed characteristics.

# 5 Estimating and filtering in XploRe

## 5.1 Overview

The procedure for Kalman filtering in XploRe is as follows: first, one has to set up the system matrices using `gkalarray`. The quantlet adjusts the measurement matrices for missing observations.

After the set up of the system matrices, we calculate the Kalman filter with `gkalfilter`. This quantlet also calculates the value of the log likelihood

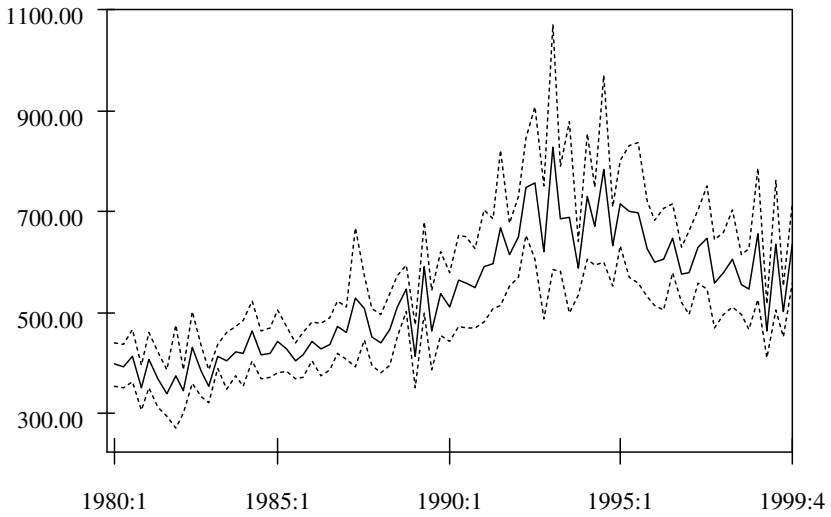


Figure 1: Average price per quarter, units are Deutsche Mark (1 DM $\approx$ 0.511 EURO). Confidence intervals are calculated for the 90% level.

XFGsssm2.xpl

function given in equation (9). That value will be used to estimate the unknown parameters of the system matrices with numerical maximization (Hamilton, 1994, Chapter 5). The first and second derivatives of the log likelihood function will also be calculated numerically. To estimate the unknown state vectors—given the estimated parameters—we use the Kalman smoother `gkalsmoother`. For diagnostic checking, we use the standardized residuals (11). The quantlet `gkalresiduals` calculates these residuals.

## 5.2 Setting the system matrices

```
gkalarrayOut = gkalarray(Y,M,IM,XM)
      sets the system matrices for a time varying SSF
```

The Kalman filter quantlets need as arguments arrays consisting of the system matrices. The quantlet `gkalarray` sets these arrays in a user-friendly way. The routine is especially convenient if one works with time varying system matrices. In our SSF (4), only the system matrix $Z_t$ is time varying. As one can see immediately from the general SSF (3), possibly every system matrix can be time varying.

The quantlet uses a three step procedure to set up the system matrices.

1. To define a system matrix all constant entries must be set to their respective values and all time varying entries must be set to an arbitrary number (for example to 0).

2. One must define an index matrix for every system matrix. An entry is set to 0 when its corresponding element in the system matrix is constant and to some positive integer when it is not constant.

3. In addition, for every time varying system matrix, one also has to specify a data matrix that contains the time varying entries.

`gkalarray` uses the following notation: `Y` denotes the matrix of all observations $[y_1, \ldots, y_T]$, `M` denotes the system matrix, `IM` denotes the corresponding index matrix and `XM` the data matrix.

If all entries of a system matrix are constant over time, then the parameters have already been put directly into the system matrix. In this case, one should set the index and the data matrix to 0.

For every time varying system matrix, only constant parameters—if there are any—have already been specified with the system matrix. The time-varying coefficients have to be specified in the index and the data matrix.

In our example, only the matrices $Z_t$ are time varying. We have

$$\texttt{Z} \stackrel{\text{def}}{=} \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

$$\texttt{IZ} \stackrel{\text{def}}{=} \begin{bmatrix} 0 & 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 4 & 5 & 6 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & (3N+1) & (3N+2) & (3N+3) \end{bmatrix}$$

$$\texttt{XZ} \stackrel{\text{def}}{=} \texttt{XFGhousequality}$$

The system matrix $Z_t$ has the dimension $(N \times 6)$. The non-zero entries in the index matrix `IZ` prescribe the rows of `XFGhousequality`, which contain the time varying elements.

The output of the quantlet is an array that stacks the system matrices one after the other. For example, the first two rows of the system matrix $Z_{41}$ are

```
[1,]        1        0        1   6.1048   4.7707       53
[2,]        1        0        1   6.5596   5.1475       13
```

XFGsssm3.xpl

It is easy to check that the entries in the last three columns are just the characteristics of the first two houses that were sold in 1990:1 (see p. 6).

## 5.3 Kalman filter and maximized log likelihood

```
{gkalfilOut,loglike} = gkalfilter(Y,mu,Sig,ca,Ta,Ra,
                                  da,Za,Ha,l)
```

Kalman filters a time-varying SSF

We assume that the initial state vector at $t = 0$ has mean $\mu$ and covariance matrix $\Sigma$. Recall, that $R_t$ and $H_t$ denote the covariance matrix of the state noise and—respectively—of the measurement noise. The general filter recursions are as follows:

Start at $t = 1$: use the initial guess for $\mu$ and $\Sigma$ to calculate

$$\begin{aligned} a_{1|0} &= c_1 + T_1\mu \\ P_{1|0} &= T_1\Sigma T_1^\top + R_1 \\ F_1 &= Z_1 P_{1|0} Z_1^\top + H_1 \end{aligned}$$

and

$$\begin{aligned} a_1 &= a_{1|0} + P_{1|0}Z_1^\top F_1^{-1}(y_1 - Z_1 a_{1|0} - d_1) \\ P_1 &= P_{1|0} - P_{1|0}Z_1^\top F_1^{-1} Z_1 P_{1|0} \end{aligned}$$

**Step at** $t \leqslant T$: using $a_{t-1}$ and $P_{t-1}$ from the previous step, calculate

$$\begin{aligned} a_{t|t-1} &= c_t + T_t a_{t-1} \\ P_{t|t-1} &= T_t P_{t-1} T_t^\top + R_t \\ F_t &= Z_t P_{t|t-1} Z_t^\top + H_t \end{aligned}$$

and

$$\begin{aligned} a_t &= a_{t|t-1} + P_{t|t-1}Z_t^\top F_t^{-1}(y_t - Z_t a_{t|t-1} - d_t) \\ P_t &= P_{t|t-1} - P_{t|t-1}Z_t^\top F_t^{-1} Z_t P_{t|t-1} \end{aligned}$$

The implementation for our model is as follows: The arguments of `gkalfilter` are the data matrix `Y`, the starting values `mu` ($\mu$), `Sig` ($\Sigma$) and the array for every system matrix (see section 5.2). The output is a $T + 1$ dimensional array of $[a_t \quad P_t]$ matrices. If one chooses $l = 1$ the value of the log likelihood function (9) is calculated.

Once again, the $T + 1$ matrices are stacked "behind each other", with the $t = 0$ matrix at the front and the $t = T$ matrix at the end of the array. The first entry is $[\mu \quad \Sigma]$.

How can we provide initial values for the filtering procedure? If the state matrices are non time-varying and the transition matrix $T$ satisfies some stability condition, we should set the initial values to the unconditional mean and variance of the state vector. $\Sigma$ is given implicitly by

$$\mathrm{vec}(\Sigma) = (I - T \otimes T)^{-1}\mathrm{vec}(R) \; .$$

Here, vec denotes the vec-operator that places the columns of a matrix below each other and $\otimes$ denotes the Kronecker product. Our model is time-invariant. But does our transition matrix fulfill the stability condition? The necessary and sufficient condition for stability is that the characteristic roots of the transition matrix $T$ should have modulus less than one (Harvey, 1989, p. 114). It is easy to check that the characteristic roots $\lambda_j$ of our transition matrix (4a) are given as

$$\lambda_{1,2} = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2} \; .$$

For example, if $\phi_1$ and $\phi_2$ are both positive, then $\phi_1 + \phi_2 < 1$ guarantees real characteristic roots that are smaller than one (Baumol, 1959, p. 221). However, when the AR(2) process of the common price component $I_t$ has a unit root, the stability conditions are not fulfilled. If we inspect Figure 1, a unit root seems quite plausible. Thus we can not use this method to derive the initial values.

If we have some preliminary estimates of $\mu$, along with preliminary measures of uncertainty—that is a estimate of $\Sigma$—we can use these preliminary estimates as initial values. A standard way to derive such preliminary estimates is to use OLS. If we have no information at all, we must take diffuse priors about the initial conditions. A method adopted by Koopman, Shephard and Doornik (1999) is setting $\mu = 0$ and $\Sigma = \kappa I$ where $\kappa$ is an large number. The large variances on the diagonal of $\Sigma$ reflect our uncertainty about the true $\mu$.

| | coefficient | | $t$-statistic | $p$-value |
|---|---|---|---|---|
| log lot size | 0.2675 | | 15.10 | 0.0000 |
| log floor space | 0.4671 | | 23.94 | 0.0000 |
| age | -0.0061 | | -20.84 | 0.0000 |
| | Regression diagnostics | | | |
| $R^2$ | 0.9997 | Number of observations | | 1502 |
| $\overline{R}^2$ | 0.9997 | F-statistic | | 64021.67 |
| $\hat{\sigma}_\varepsilon^2$ | 0.4688 | Prob(F-statistic) | | 0.0000 |

Table 1: Results for hedonic regression

We will use the second approach for providing some preliminary estimates as initial values. Given the hedonic equation (1), we use OLS to estimate $I_t$, $\beta$, and $\sigma_m^2$ by regressing log prices on lot size, floor space, age and quarterly time dummies. The estimated coefficients of lot size, floor space and age are reported in Table 1. They are highly significant and reasonable in sign and magnitude. Whereas lot size and floor space increase the price on average, age has the opposite effect. According to (1), the common price component $I_t$ is a time-varying constant term and is therefore estimated by the coefficients of the quarterly time dummies, denoted by $\{\hat{I}_t\}_{t=1}^{80}$. As suggested by (2), these estimates are regressed on their lagged values to obtain estimates of the unknown parameters $\phi_1$, $\phi_2$, and $\sigma_s^2$. Table 2 presents the results for an AR(2) for the $\hat{I}_t$ series. The residuals of this regression behave like white noise. We should remark that

$$\hat{\phi}_1 + \hat{\phi}_2 \approx 1$$

and thus the process of the common price component seems to have a unit root.

Given our initial values we maximize the log likelihood (9) numerically with respect to the elements of $\psi^* \stackrel{\text{def}}{=} (\phi_1, \phi_2, \log(\sigma_\nu^2), \log(\sigma_\varepsilon^2))$. Note that $\psi^*$ differs from $\psi$ by using the logarithm of the variances $\sigma_\nu^2$ and $\sigma_\varepsilon^2$. This transformation is known to improve the numerical stability of the maximization algorithm, which employs `nmBFGS` of XploRe's `nummath` library. Standard errors are computed from inverting the Hessian matrix provided by `nmhessian`. The output of the maximum likelihood estimation procedure is summarized in Table 3, where we

| | coefficient | | $t$-statistic | $p$-value |
|---|---|---|---|---|
| constant | 0.5056 | | 1.3350 | 0.1859 |
| $\hat{I}_{t-1}$ | 0.4643 | | 4.4548 | 0.0000 |
| $\hat{I}_{t-2}$ | 0.4823 | | 4.6813 | 0.0000 |
| | Regression diagnostics | | | |
| $R^2$ | 0.8780 | Number of observations | | 78 |
| $\overline{R}^2$ | 0.8747 | F-statistic | | 269.81 |
| $\hat{\sigma}_\nu^2$ | 0.0063 | Prob(F-statistic) | | 0.0000 |

Table 2: Time series regression for the quarterly dummies

report the estimates of $\sigma_\nu^2$ and $\sigma_\varepsilon^2$ obtained by retransforming the estimates of $\log(\sigma_\nu^2)$ and $\log(\sigma_\varepsilon^2)$).

| | estimate | std error | $t$-value | $p$-value |
|---|---|---|---|---|
| $\hat{\psi}_1 = \hat{\phi}_1$ | 0.783 | 0.501 | 1.56 | 0.12 |
| $\hat{\psi}_2 = \hat{\phi}_2$ | 0.223 | 0.504 | 0.44 | 0.66 |
| $\hat{\psi}_1 = \hat{\sigma}_\nu^2$ | 0.0016 | 0.012 | 1.36 | 0.17 |
| $\hat{\psi}_2 = \hat{\sigma}_\varepsilon^2$ | 0.048 | 0.002 | 26.7 | 0 |
| average log likelihood | | 0.9965 | | |

Table 3: Maximum likelihood estimates of the elements of $\psi$ `XFGsssm4.xpl`

Note that the maximum likelihood estimates of the AR coefficients $\phi_1$ and $\phi_2$ approximately sum to 1, again pointing towards a unit root process for the common price component.

## 5.4 Diagnostic checking with standardized residuals

```
{V,Vs} = gkalresiduals(Y,Ta,Ra,da,Za,Ha,gkalfilOut)
    calculates innovations and standardized residuals
```

The quantlet `gkalresiduals` checks internally for the positive definiteness of $F_t$. An error message will be displayed when $F_t$ is not positive definite. In such a case, the standardized residuals are not calculated.

The output of the quantlet are two $N \times T$ matrices `V` and `Vs`. `V` contains the innovations (10) and `Vs` contains the standardized residuals (11).

The Q-Q plot of the standardized residuals in Figure 2 shows deviations from normality at both tails of the distribution.

This is evidence, that the true error distribution might be a unimodal distribution with heavier tails than the normal, such as the $t$-distribution. In this case the projections calculated by the Kalman filter no longer provide the conditional expectations of the state vector but rather its best linear prediction.

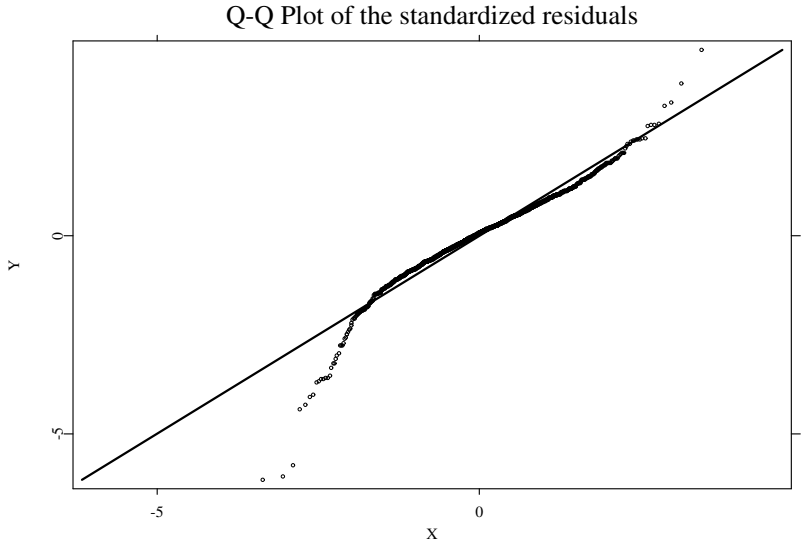


Figure 2: Deviations of the dotted line from the straight line are evidence for a nonnormal error distribution

XFGsssm5.xpl

Moreover the estimates of $\psi$ calculated from the likelihood (9) can be interpreted as pseudo-likelihood estimates.

## 5.5 Calculating the Kalman smoother

```
gkalsmoothOut = gkalsmoother(Y,Ta,Ra,gkalfilOut)
    provides Kalman smoothing of a time-varying SSF
```

The Kalman filter is a convenient tool for calculating the conditional expectations and covariances of our SSF (4). We have used the innovations of this filtering technique and its covariance matrix for calculating the log likelihood. However, for estimating the unknown state vectors, we should use in every step the whole sample information up to period $T$. For this task, we use the Kalman smoother.

The quantlet `gkalsmoother` needs as argument the output of `gkalfilter`. The output of the smoother is an array with $[a_{t|T} \quad P_{t|T}]$ matrices. This array of dimension $T+1$ starts with the $t=0$ matrix and ends with the matrix for $t=T$. For the smoother recursions, one needs $a_t$, $P_t$ and $P_{t|t-1}$ for $t=1\ldots T$. Then the calculation procedure is as follows:

**Start at** $t = T$:

$$\begin{aligned} a_{T|T} &= a_T \\ P_{T|T} &= P_T \end{aligned}$$

**Step at** $t < T$:

$$\begin{aligned} P_t^* &= P_t T_{t+1}^\top P_{t+1|t}^{-1} \\ a_{t|T} &= a_t + P_t^*(a_{t+1|T} - T_{t+1} a_t) \\ P_{t|T} &= P_t + P_t^*(P_{t+1|T} - P_{t+1|t}) P_t^{*\top} \end{aligned}$$

The next program calculates the smoothed state vectors for our SSF form, given the estimated parameters $\tilde{\psi}$. The smoothed series of the common price component is given in Figure 3. The confidence intervals are calculated using the variance of the first element of the state vector.

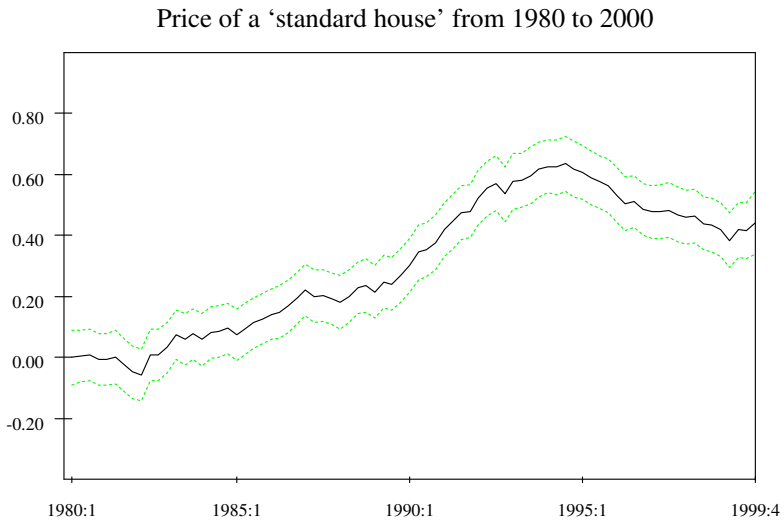


Figure 3: Smoothed common price component. Confidence intervals are calculated for the 90% level.

XFGsssm6.xpl

Comparison with the average prices given in Figure 1 reveals that the common price component is less volatile than the simple average. Furthermore, a table for the estimated hedonic coefficients—that is $\beta$—is generated, Table 4.

Recall that these coefficients are just the last three entries in the state vector $\alpha_t$. According to our state space model, the variances for these state variables are zero. Thus, it is not surprising that the Kalman smoother produces constant estimates through time for these coefficients. In the Appendix 6.2 we give a formal proof of this intuitive result.

```
[1,] "========================================="
[2,] " Estimated hedonic coefficients          "
[3,] "========================================="
[4,] " Variable        coeff.  t-Stat.  p-value  "
[5,] " ----------------------------------------- "
[6,] " log lot size    0.2664   21.59    0.0000   "
[7,] " log floor area  0.4690   34.33    0.0000   "
[8,] " age            -0.0061  -29.43    0.0000   "
[9,] "========================================="
```

Table 4: Estimated hedonic coefficients $\beta$. XFGsssm6.xpl

The estimated coefficient of log lot size implies that, as expected, the size of the lot has an positive influence on the price. The estimated relative price increase for an one percent increase in the lot size is about 0.27%. The estimated effect of an increase in the floor space is even larger. Here, a one percent increase in the floor space lets the price soar by about 0.48%. Finally, note that the price of a houses is estimated to decrease with age.

# 6 Appendix

## 6.1 Procedure equivalence

We show that our treatment of missing values delivers the same results as the procedure proposed by Shumway and Stoffer (1982; 2000). For this task, let us assume that the $(N \times 1)$ vector of observations $t$

$$y_t^\top = \begin{bmatrix} y_{1,t} & \cdot & y_{3,t} & \cdot & y_{5,t} & \cdots & y_{N,t} \end{bmatrix}$$

has missing values. Here, observations 2 and 4 are missing. Thus, we have only $N_t < N$ observations. For Kalman filtering in XploRe, all missing values in $y_t$ and the corresponding rows and columns in the measurement matrices $d_t$, $Z_t$, and $H_t$, are deleted. Thus, the adjusted vector of observations is

$$y_{t,1} = \begin{bmatrix} y_{1,t} & y_{3,t} & y_{5,t} & \cdots & y_{N,t} \end{bmatrix}$$

where the subscript 1 indicates that this is the vector of observations used in the XploRe routines. The procedure of Shumway and Stoffer instead rearranges the vectors in such a way that the first $N_t$ entries are the observations—and thus given by $y_{t,1}$—and the last $(N - N_t)$ entries are the missing values. However, all missing values must be replaced with zeros.

For our proof, we use the following generalized formulation of the measurement equation

$$\begin{bmatrix} y_{t,1} \\ y_{t,2} \end{bmatrix} = \begin{bmatrix} d_{t,1} \\ d_{t,2} \end{bmatrix} + \begin{bmatrix} Z_{t,1} \\ Z_{t,2} \end{bmatrix} \alpha_t + \begin{bmatrix} \varepsilon^m_{t,1} \\ \varepsilon^m_{t,2} \end{bmatrix}$$

and

$$\operatorname{cov} \begin{pmatrix} \varepsilon^m_{t,1} \\ \varepsilon^m_{t,2} \end{pmatrix} = \begin{bmatrix} H_{t,11} & H_{t,12} \\ H_{t,12} & H_{t,22} \end{bmatrix} .$$

$y_{t,1}$ contains the observations and $y_{t,2}$ the missing values. The procedure of Shumway and Stoffer employs the generalized formulation given above and sets $y_{t,2} = 0$, $d_{t,2} = 0$, $Z_{t,2} = 0$, and $H_{t,12} = 0$ (Shumway and Stoffer, 2000, p. 330). We should remark that the dimensions of these matrices also depend on $t$ via $(N - N_t)$. However, keep notation simple we do not make this time dependency explicit. It is important to mention that matrices with subscript 1 and 11 are equivalent to the adjusted matrices of XploRe's filtering routines.

First, we show by induction that both procedures deliver the same results for the Kalman filter. Once this equivalence is established, we can conclude that the smoother also delivers identical results.

**PROOF:**

Given $\mu$ and $\Sigma$, the terms $a_{1|0}$ and $P_{1|0}$ are the same for both procedures. This follows from the simple fact that the first two steps of the Kalman filter do not depend on the vector of observations (see Subsection 5.3).

Now, given $a_{t|t-1}$ and $P_{t|t-1}$, we have to show that also the filter recursions

$$a_t = a_{t|t-1} + P_{t|t-1} Z_t^\top F_t^{-1} v_t \ , \quad P_t = P_{t|t-1} - P_{t|t-1} Z_t^\top F_t^{-1} Z_t P_{t|t-1} \tag{13}$$

deliver the same results. Using *ss* to label the results of the Shumway and Stoffer procedure, we obtain by using

$$Z_{t,ss} \stackrel{\text{def}}{=} \begin{bmatrix} Z_{t,1} \\ 0 \end{bmatrix}$$

that

$$F_{t,ss} = \begin{bmatrix} Z_{t,1} P_{t|t-1} Z_{t,1}^\top & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} H_{t,11} & 0 \\ 0 & H_{t,22} \end{bmatrix} \ .$$

The inverse is given by (Sydsæter, Strøm and Berck, 2000, 19.49)

$$F_{t,ss}^{-1} = \begin{bmatrix} F_{t,1}^{-1} & 0 \\ 0 & H_{t,22}^{-1} \end{bmatrix} \tag{14}$$

where $F_{t,1}$ is just the covariance matrix of the innovations of XploRe's procedure. With (14) we obtain that

$$Z_{t,ss}^\top F_{t,ss}^{-1} = \begin{bmatrix} Z_{t,1}^\top F_{t,1}^{-1} & 0 \end{bmatrix}$$

and accordingly for the innovations

$$v_{t,ss} = \begin{bmatrix} v_{t,1} \\ 0 \end{bmatrix} \ .$$

We obtain immediately

$$Z_{t,ss}^\top F_{t,ss}^{-1} v_{t,ss} = Z_{t,1}^\top F_{t,1}^{-1} v_{t,1} \ .$$

Plugging this expression into (13)—taking into account that $a_{t|t-1}$ and $P_{t|t-1}$ are identical—delivers

$$a_{t,ss} = a_{t,1} \quad \text{and} \quad P_{t,ss} = P_{t,1} \ .$$

This completes the first part of our proof.

The Kalman smoother recursions use only system matrices that are the same for both procedures. In addition to the system matrices, the output of the filter is used as an input, see Subsection 5.5. But we have already shown that the filter output is identical. Thus the results of the smoother are the same for both procedures as well. □

## 6.2 Smoothed constant state variables

We want to show that the Kalman smoother produces constant estimates through time for all state variables that are constant by definition. To proof this result, we use some of the smoother recursions given in Subsection 5.5. First of all, we rearrange the state vector such that the last $k \leqslant K$ variables are constant. This allows the following partition of the transition matrix

$$T_{t+1} = \begin{bmatrix} T_{11,t+1} & T_{12,t+1} \\ 0 & I \end{bmatrix} \tag{15}$$

with the $k \times k$ identity matrix $I$. Furthermore, we define with the same partition

$$\tilde{P}_t \stackrel{\text{def}}{=} T_{t+1} P_t T_{t+1}^\top = \begin{bmatrix} \tilde{P}_{11,t} & \tilde{P}_{12,t} \\ \tilde{P}_{12,t} & \tilde{P}_{22,t} \end{bmatrix}$$

The filter recursion for the covariance matrix are given as

$$P_{t+1|t} = T_{t+1} P_t T_{t+1}^\top + R_{t+1}$$

where the upper left part of $R_{t+1}$ contains the covariance matrix of the disturbances for the stochastic state variables. We see immediately that only the upper left part of $P_{t+1|T}$ is different from $\tilde{P}_t$.

Our goal is to show that for the recursions of the smoother holds

$$P_t^* = \begin{bmatrix} M_{11,t} & M_{12,t} \\ 0 & I \end{bmatrix} , \tag{16}$$

where both $M$s stand for some complicated matrices. With this result at hand, we obtain immediately

$$a_{t|T}^k = a_{t+1|T}^k = a_T^k \tag{17}$$

for all $t$, where $a_{t|T}^k$ contains the last $k$ elements of the smoothed state $a_{t|T}$.

Furthermore, it is possible to show with the same result that the lower right partition of $P_{t|T}$ is equal to the lower right partition of $P_T$ for all $t$. This lower right partition is just the covariance matrix of $a_{t|T}^k$. Just write the smoother recursion

$$P_{t|T} = P_t(I - T_{t+1}^\top P_t^{*\top}) + P_t^* P_{t+1|T} P_t^{*\top} .$$

Then check with (15) and (16) that the lower-right partition of the first matrix on the right hand side is a $k \times k$ matrix of zeros. The lower-right partition of the second matrix is given by the the lower-right partition of $P_{t+1|T}$.

**PROOF:**

Now we derive (16): We assume that the inverse of $T_{t+1}$ and $T_{11,t+1}$ exist. The inverses for our model exist because we assume that $\phi_2 \neq 0$. For the partitioned transition matrix (Sydsæter, Strøm and Berck, 2000, 19.48) we derive

$$T_{t+1}^{-1} = \begin{bmatrix} T_{11,t+1}^{-1} & -T_{11,t+1}^{-1} T_{12,t+1} \\ 0 & I \end{bmatrix} . \tag{18}$$

Now, it is easy to see that

$$P_t^* = T_{t+1}^{-1} \tilde{P}_t P_{t+1|t}^{-1} . \tag{19}$$

We have (Sydsæter, Strøm and Berck, 2000, 19.49)

$$P_{t+1|t}^{-1} = \begin{bmatrix} \Delta_t & -\Delta_t \tilde{P}_{12,t} \tilde{P}_{22,t}^{-1} \\ -\tilde{P}_{22,t}^{-1} \tilde{P}_{12,t} \Delta_t & \tilde{P}_{22,t}^{-1} + \tilde{P}_{22,t}^{-1} \tilde{P}_{12,t} \Delta_t \tilde{P}_{12,t} \tilde{P}_{22,t}^{-1} \end{bmatrix} \tag{20}$$

with $\Delta_t$ as a known function of the partial matrices. If we multiply this matrix with the lower partition of $\tilde{P}_t$ we obtain immediately $[0\ I]$. With this result and (18) we derive (16). □

**Acknowledgement**

The authors acknowledge support by the Deutsche Forschungsgemeinschaft via Sonderforschungsbereich 373 "Quantifikation and Simulation ökonomischer Prozesse" at Humboldt-Universität zu Berlin.

# References

Bailey, M. J., Muth, R. F. and Nourse, H.O. (1963). A regression method for real estate price index construction, *Journal of the American Statistical Association* **58**: 933–942.

Baumol, W. (1959). *Economic Dynamics*, 2nd ed., Macmillan, New York.

Bera, A. K. and Jarque, C. M. (1982). Model Specification Tests: a Simultaneous Approach, *Journal of Econometrics* **20**: 59–82.

Cho, M. (1996). House price dynamics: a survey of theoretical and empirical issues, *Journal of Housing Research* **7:2**: 145–172.

Clapp, J. M. and Giaccotto, C. (1998). Price indices based on the hedonic repeat-sales method: application to the housing market, *Journal of Real Estate Finance and Economics* **16:1**: 5–26.

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

Engle, R. F. and M. W. Watson (1981). A One-Factor Multivariate Time Series Model of Metropolitan Wage Rates, *Journal of the American Statistical Association* **76**: 774–781.

Gourieroux, C. and Monfort, A. (1997). *Time Series and Dynamic Models*, Cambridge University Press, Cambridge.

Greene, W. H. (2000). *Econometric Analysis. Fourth Edition*, Prentice Hall, Upper Saddle River, New Jersey.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton University Press, Princeton, New Jersey.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

Harvey, A. C. (1993). *Time Series Models*, 2. edn, Harvester Wheatsheaf, New York.

Hill, R. C., Knight, J. R. and Sirmans, C. F. (1997). Estimating Capital Asset Price Indexes, *Review of Economics and Statistics* **79**: 226–233.

Koopman, S. J., Shepard, N. and Doornik, J. A. (1999). Statistical Algorithms for Models in State Space Using SsfPack 2.2, *Econometrics Journal* **2**: 107–160.

Peña, D., Tiao, G. C. and Tsay, R. S. (2001). *A Course in Time Series Analysis*, Wiley, New York.

Schwann, G. M. (1998). A real estate price index for thin markets, *Journal of Real Estate Finance and Economics* **16:3**: 269–287.

Shiller, R. J. (1993). *Macro Markets. Creating Institutions for Managing Society's Largest Economic Risks*, Clarendon Press, Oxford.

Shumway, R. H. and Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm, *Journal of Time Series Analysis* **3**: 253–264.

Shumway, R. H. and Stoffer, D. S. (2000). *Time Series Analysis and Its Applications*, Springer, New York, Berlin.

Sydsæter, K., Strøm, A. and Berck, P. (2000). *Economists' Mathematical Manual*, 3. edn, Springer, New York, Berlin.