An implementation of the Differential Filter for Computing Gradient and Hessian of the Log-likelihood of Nonstationary Time Series Models

Genshiro Kitagawa
Mathematics and Informatics Center, The University of Tokyo

September 27, 2022

Abstract
The state-space model and the Kalman filter provide us with unified and computationally efficient procedure for computing the log-likelihood of the diverse type of time series models. This paper presents an algorithm for computing the gradient and the Hessian matrix of the log-likelihood by extending the Kalman filter without resorting to the numerical difference. Different from the previous paper [11], it is assumed that the observation noise variance $R = 1$. It is known that for univariate time series, by maximizing the log-likelihood of this restricted model, we can obtain the same estimates as the ones for the original state-space model. By this modification, the algorithm for computing the gradient and the Hessian becomes somewhat complicated. However, the dimension of the parameter vector is reduce by one and thus has a significant merit in estimating the parameter of the state-space model especially for relatively low dimensional parameter vector. Three examples of nonstationary time series models, i.e., trend model, standard seasonal adjustment model and the seasonal adjustment model with AR component are presented to exemplified the specification of structural matrices.

Key words State-space model, Kalman filter, log-likelihood, gradient, Hessian matrix, seasonal adjustment model, autoregressive model.

1 Introduction: The Maximum Likelihood Estimation of a State-Space Model

We consider a linear Gaussian state-space model

$$x_n = F_n(\theta)x_{n-1} + G_n(\theta)v_n$$  \hspace{1cm} (1)

$$y_n = H_n(\theta)x_n + w_n,$$  \hspace{1cm} (2)

where $y_n$ is a one-dimensional time series, $x_n$ is an $m$-dimensional state vector, $v_n$ is a $k$-dimensional Gaussian white noise, $v_n \sim N(0, Q_n(\theta))$, and $w_n$ is a one-dimensional white noise, $w_n \sim N(0, 1)$. $F_n(\theta)$, $G_n(\theta)$ and $H_n(\theta)$ are $m \times m$ matrix, $m \times k$ matrix and $m$ vector, respectively. $\theta$ is the $p$-dimensional parameter vector of the state-space model such as the variances of the noise inputs and unknown coefficients in the matrices $F_n(\theta)$, $G_n(\theta)$, $H_n(\theta)$ and $Q_n(\theta)$. For simplicity of the notation, hereafter, the parameter $\theta$ and the suffix $n$ will be omitted. It is noted that for the state-space model of the univariate time series, the assumption that $R = 1$ does not lose any generality, since it is known that even with this assumption we can obtain the same estimates of the parameters of the model by a proper transformation (Kitagawa (2020)).

Various models used in time series analysis can be treated uniformly within the state-space model framework. Further, many problems of time series analysis, such as prediction, signal extraction, decomposition, parameter estimation and interpolation, can be formulated as the estimation of the state of a state-space model.

Given the time series $Y_N \equiv \{y_1, \ldots, y_N\}$ and the state-space model (1) and (2), the one-step-ahead predictor $x_{n|n-1}$ and the filter $x_{n|n}$ and their variance covariance matrices $V_{n|n-1}$ and $V_{n|n}$ are obtained by the following Kalman filter (Anderson and Moore (2012) and Kitagawa (2020)): 

\[
\begin{align*}
    x_{n|n-1} &= F_n(\theta)x_{n-1|n-1} + G_n(\theta)v_n \\
    x_{n|n} &= x_{n|n-1} + K_n(\theta)(y_n - H_n(\theta)x_{n|n-1}) \\
    V_{n|n-1} &= F_n(\theta)V_{n-1|n-1}F_n(\theta)^T + G_n(\theta)Q_n(\theta)G_n(\theta)^T \\
    V_{n|n} &= V_{n|n-1} - K_n(\theta)H_n(\theta)V_{n|n-1} \\
    K_n(\theta) &= V_{n|n-1}H_n(\theta)^T(V_{n|n} + W_n)^{-1} \\
    W_n &= H_n(\theta)V_{n|n-1}H_n(\theta)^T + W_n
\end{align*}
\]
One-step-ahead prediction

\[ \begin{align*}
x_{n|n-1} &= Fx_{n-1|n-1} \\
V_{n|n-1} &= FV_{n-1|n-1}F^T + GQ_nG^T.
\end{align*} \tag{3} \]

Filter

\[ \begin{align*}
K_n &= V_{n|n-1}H^T(HV_{n|n-1}H^T + R)^{-1} \\
x_n &= x_{n|n-1} + K_n(y_n - Hx_{n|n-1}) \\
V_n &= (I - K_nH)V_{n|n-1}.
\end{align*} \tag{4} \]

Given the data \( Y_N \), the likelihood of the time series model is defined by

\[ L(\theta) = p(Y_N|\theta) = \prod_{n=1}^{N} g_n(y_n|Y_{n-1}, \theta), \tag{5} \]

where \( g_n(y_n|Y_{n-1}, \theta) \) is the conditional distribution of \( y_n \) given the observation \( Y_{n-1} \) and is a normal distribution given by

\[ g_n(y_n|Y_{n-1}, \theta) = \frac{1}{\sqrt{2\pi r_n}} \exp \left\{ -\frac{\varepsilon_n^2}{2r_n} \right\}, \tag{6} \]

where \( \varepsilon_n \) and \( r_n \) are the one-step-ahead prediction error and its variance defined by

\[ \begin{align*}
\varepsilon_n &= y_n - Hx_{n|n-1} \\
r_n &= H_nV_{n|n-1}H_n^T + 1 \tag{7} \]

Therefore, the log-likelihood of the state-space model is obtained as

\[ \ell(\theta) = \log L(\theta) = \sum_{n=1}^{N} \log g_n(y_n|Y_{n-1}, \theta) \]

\[ = -\frac{1}{2} \left( N \log 2\pi \hat{\sigma}^2 + \sum_{n=1}^{N} \log r_n + N \right), \tag{8} \]

where the maximum likelihood estimate of the variance is given by

\[ \hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} \frac{\varepsilon_n^2}{r_n}. \tag{9} \]

The maximum likelihood estimates of the parameters of the state-space model can be obtained by maximizing the log-likelihood function \( \ell(\theta) \). In general, since the log-likelihood function is mostly nonlinear, the maximum likelihood estimates are obtained by using a numerical optimization algorithm based on the quasi-Newton method. According to this method, using the value \( \ell(\theta) \) of the log-likelihood and the first derivative (gradient) \( \partial \ell/\partial \theta \) for a given parameter \( \theta \), the maximizer of \( \ell(\theta) \) is automatically estimated by repeating

\[ \theta_k = \theta_{k-1} + \lambda_k B_{k-1}^{-1} \frac{\partial \ell}{\partial \theta}, \tag{10} \]

where \( \theta_0 \) is an initial estimate of the parameter. The step width \( \lambda_k \) is automatically determined and the inverse matrix \( H_{k-1}^{-1} \) of the Hessian matrix is obtained recursively by the DFP or BFGS algorithms (Fletcher (2013)).
Here, the gradient of the log-likelihood function is usually approximated by numerical difference, such as
\[
\frac{\partial \ell(\theta)}{\partial \theta_j} \approx \frac{\ell(\theta_j + \Delta \theta_j) - \ell(\theta_j - \Delta \theta_j)}{2\Delta \theta_j},
\] (11)
where \(\Delta \theta_j\) is defined by \(C|\theta_j|\), for some small \(C\) such as 0.0001. The numerical difference usually yields reasonable approximation to the gradient of the log-likelihood. However, since it requires \(2p\) times of log-likelihood evaluations, the amount of computation becomes considerably large if the dimension of the parameters is large. Further, if the the maximum likelihood estimates lie very close to the boundary of admissible domain, which sometimes occur in regularization problems, it becomes difficult to obtain the approximation to the gradient of the log-likelihood by the numerical difference.

Analytic derivative of the log-likelihood of time series models were considered by many authors. For example, Kohn and Ansley (1985) gave method for computing likelihood and its derivatives for an ARMA model. Zadrozny (1989) derived analytic derivatives for estimation of linear dynamic models. Kulikova (2009) presented square-root algorithm for the evaluation of the likelihood gradient to avoid numerical instability of the recursive algorithm for log-likelihood computation.

In this paper, the gradient and Hessian of the log-likelihood of linear state-space model are given under the assumption that the observation noise variance is 1. By this method, the dimension of the unknown parameter is reduced by one, but instead the derivative of the observation noise variance must be computed simultaneously. Details of the implementation of the algorithm for the trend model, the standard seasonal adjustment model and the seasonal adjustment model with stationary AR component are given. For each implementation, comparison with a numerical difference method is shown. In section 2, algorithm for obtaining the gradient and the Hessian of the log-likelihood is presented. Application of the method is exemplified with the three models, i.e., the trend model, the standard seasonal adjustment model and the seasonal adjustment model with autoregressive component are shown in section 3.

2 The Gradient and the Hessian of the log-likelihood

The recursive algorithm for computing the gradient and the Hessian of the log-likelihood is essentially the same as the one shown in Kitagawa(2020b). However, we assume that the observation noise variance is 1, the expression for the gradient and the Hessian matrix become simple. Instead, we need to evaluate the first and the second derivative of the observation noise variance \(\sigma^2\).

2.1 The gradient of the log-likelihood

From (8), the gradient of the log-likelihood is obtained by
\[
\frac{\partial \ell(\theta)}{\partial \theta} = -\frac{1}{2} \left( N \frac{\partial \hat{\sigma}^2}{\partial \theta} + \sum_{n=1}^{N} \frac{1}{r_n} \frac{\partial r_n}{\partial \theta} \right),
\] (12)
where, from (7) and (9), the derivatives of the one-step-ahead prediction \(\varepsilon_n\), the one-step-ahead prediction error variance \(r_n\) and the observation noise variance are obtained by
\[
\frac{\partial \varepsilon_n}{\partial \theta} = -H \frac{\partial x_{n|n-1}}{\partial \theta} - \frac{\partial H}{\partial \theta} x_{n|n-1},
\] (13)
\[
\frac{\partial r_n}{\partial \theta} = H \frac{\partial V_{n|n-1}}{\partial \theta} + \frac{\partial H}{\partial \theta} V_{n|n-1} H^T + H V_{n|n-1} \frac{\partial H^T}{\partial \theta},
\] (14)
\[
\frac{\partial \hat{\sigma}^2}{\partial \theta} = \frac{1}{N} \sum_{n=1}^{N} \left( \frac{2 \varepsilon_n}{r_n} \frac{\partial \varepsilon_n}{\partial \theta} - \varepsilon_n \frac{\partial r_n}{\partial \theta} \right).\] (15)

To evaluate these quantities, we need the derivatives of the one-step-ahead predictor of the state \(\frac{\partial x_{n|n-1}}{\partial \theta}\) and its variance covariance matrix \(\frac{\partial V_{n|n-1}}{\partial \theta}\) which can be obtained recursively in parallel to the Kalman filter.
algorithm:

[One-step-ahead-prediction]

\[
\frac{\partial x_{n|n-1}}{\partial \theta} = F \frac{\partial x_{n-1|n-1}}{\partial \theta} + \frac{\partial F}{\partial \theta} x_{n-1|n-1}
\]
\[
\frac{\partial V_{n|n-1}}{\partial \theta} = F \frac{\partial V_{n-1|n-1}}{\partial \theta} F^T + \frac{\partial F}{\partial \theta} V_{n-1|n-1} F^T + F V_{n-1|n-1} \frac{\partial F^T}{\partial \theta}
\]
\[
\quad + G \frac{\partial Q}{\partial \theta} G^T + \frac{\partial G}{\partial \theta} Q G^T + G Q \frac{\partial G^T}{\partial \theta}.
\]

(16)

[Filter]

\[
\frac{\partial K_n}{\partial \theta} = \left( \frac{\partial V_{n|n-1}}{\partial \theta} H^T + V_{n|n-1} \frac{\partial H^T}{\partial \theta} \right) r_n^{-1} - V_{n|n-1} H^T \frac{\partial r_n}{\partial \theta} r_n^{-2}
\]
\[
\frac{\partial x_{n|n}}{\partial \theta} = \frac{\partial x_{n|n-1}}{\partial \theta} + K_n \frac{\partial \varepsilon_n}{\partial \theta} + \frac{\partial K_n}{\partial \theta} \varepsilon_n
\]
\[
\frac{\partial V_{n|n}}{\partial \theta} = \frac{\partial V_{n|n-1}}{\partial \theta} - \frac{\partial K_n}{\partial \theta} H V_{n|n-1} - K_n H \frac{\partial V_{n|n-1}}{\partial \theta}.
\]

(17)

2.2 Hessian of the Log-likelihood of the State-space Model

The Hessian (the second derivative) of the log-likelihood can also be obtained by a recursive formula, since, from [12], it is given as

\[
\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T} = -\frac{N}{2} \left( \frac{1}{\sigma^2} \frac{\partial^2 \hat{\sigma}^2}{\partial \theta \partial \theta^T} - \frac{1}{(\hat{\sigma}^2)^2} \frac{\partial \hat{\sigma}^2}{\partial \theta} \frac{\partial \hat{\sigma}^2}{\partial \theta^T} \right) - \frac{1}{2} \sum_{n=1}^{N} \left( \frac{1}{r_n} \frac{\partial^2 x_n}{\partial \theta \partial \theta^T} - \frac{1}{r_n^2} \frac{\partial \varepsilon_n}{\partial \theta} \frac{\partial r_n}{\partial \theta^T} \right),
\]

where, from [14]–[15], \( \frac{\partial^2 \varepsilon_n}{\partial \theta \partial \theta^T} \), \( \frac{\partial^2 r_n}{\partial \theta \partial \theta^T} \) and \( \frac{\partial^2 \hat{\sigma}^2}{\partial \theta \partial \theta^T} \) are obtained by

\[
\frac{\partial^2 \varepsilon_n}{\partial \theta \partial \theta^T} = -\frac{\partial H}{\partial \theta} \frac{\partial x_{n|n-1}}{\partial \theta} - \frac{\partial H}{\partial \theta} \frac{\partial x_{n|n-1}}{\partial \theta} - H \frac{\partial^2 x_{n|n-1}}{\partial \theta \partial \theta^T} - \frac{\partial^2 H}{\partial \theta \partial \theta^T} x_{n|n-1}
\]

(18)

\[
\frac{\partial^2 r_n}{\partial \theta \partial \theta^T} = \frac{\partial H}{\partial \theta} \frac{\partial V_{n|n-1}}{\partial \theta} H^T + \frac{\partial H}{\partial \theta} \frac{\partial V_{n|n-1}}{\partial \theta} H^T + H \frac{\partial^2 V_{n|n-1}}{\partial \theta \partial \theta^T} H^T
\]
\[
\quad + H \frac{\partial V_{n|n-1}}{\partial \theta} H^T + H \frac{\partial V_{n|n-1}}{\partial \theta} H^T + H \frac{\partial^2 V_{n|n-1}}{\partial \theta \partial \theta^T} H^T
\]
\[
\quad + \frac{\partial H}{\partial \theta} \frac{\partial V_{n|n-1}}{\partial \theta} H^T + \frac{\partial H}{\partial \theta} \frac{\partial V_{n|n-1}}{\partial \theta} H^T + H \frac{\partial^2 V_{n|n-1}}{\partial \theta \partial \theta^T} H^T + \frac{\partial^2 H}{\partial \theta \partial \theta^T}.
\]

(19)

\[
\frac{\partial^2 \hat{\sigma}^2}{\partial \theta \partial \theta^T} = \frac{1}{N} \sum_{n=1}^{N} \left\{ \frac{2}{r_n} \left( \frac{\partial \varepsilon_n}{\partial \theta} \frac{\partial \varepsilon_n}{\partial \theta^T} + \varepsilon_n \frac{\partial^2 \varepsilon_n}{\partial \theta \partial \theta^T} \right) + \frac{2 \varepsilon_n}{r_n} \frac{\partial r_n}{\partial \theta} \frac{\partial r_n}{\partial \theta^T} \right\}
\]
\[
\quad - \frac{1}{r_n} \left( 2 \varepsilon_n \frac{\partial \varepsilon_n}{\partial \theta} \frac{\partial r_n}{\partial \theta^T} + 2 \varepsilon_n \frac{\partial \varepsilon_n}{\partial \theta} \frac{\partial r_n}{\partial \theta^T} + \varepsilon_n^2 \frac{\partial^2 r_n}{\partial \theta \partial \theta^T} \right) \}
\]

(20)

To evaluate the Hessian, the following computation should be performed along with the recursive formula for the log-likelihood and the gradient of the log-likelihood.

\[
\frac{\partial^2 x_{n|n-1}}{\partial \theta \partial \theta^T} = \frac{\partial F}{\partial \theta} \frac{\partial x_{n-1|n-1}}{\partial \theta} + \frac{\partial F}{\partial \theta} \frac{\partial x_{n-1|n-1}}{\partial \theta} + F \frac{\partial^2 x_{n-1|n-1}}{\partial \theta \partial \theta^T} + \frac{\partial^2 F}{\partial \theta \partial \theta^T} x_{n-1|n-1}.
\]
\[
\frac{\partial^2 V_{n|n-1}}{\partial \theta \partial \theta^T} = \frac{\partial F}{\partial \theta} \frac{\partial V_{n-1|n-1}}{\partial \theta^T} F + \frac{\partial F}{\partial \theta} \frac{\partial V_{n-1|n-1}}{\partial \theta^T} F + F \frac{\partial^2 V_{n-1|n-1}}{\partial \theta \partial \theta^T} F + F \frac{\partial V_{n-1|n-1}}{\partial \theta^T} \frac{\partial F}{\partial \theta} \\
+ F \frac{\partial V_{n-1|n-1}}{\partial \theta} \frac{\partial F}{\partial \theta^T} + \frac{\partial F}{\partial \theta} \frac{\partial V_{n|n-1}}{\partial \theta^T} V_{n|n-1} F + \frac{\partial F}{\partial \theta} \frac{\partial V_{n|n-1}}{\partial \theta^T} V_{n|n-1} F + \frac{\partial F}{\partial \theta} \frac{\partial V_{n|n-1}}{\partial \theta^T} \frac{\partial F}{\partial \theta} \\
+ F V_{n|n-1} \frac{\partial^2 F}{\partial \theta^T} + \frac{\partial G}{\partial \theta} \frac{\partial Q}{\partial \theta^T} G^T + \frac{\partial G}{\partial \theta} \frac{\partial Q}{\partial \theta^T} G^T + G \frac{\partial^2 Q}{\partial \theta^T} G^T + G \frac{\partial Q}{\partial \theta^T} G^T \\
+ G \frac{\partial Q}{\partial \theta} G^T + \frac{\partial^2 G}{\partial \theta^T} Q G^T + \frac{\partial G}{\partial \theta} \frac{\partial Q}{\partial \theta^T} G^T + \frac{\partial G}{\partial \theta} \frac{\partial Q}{\partial \theta^T} G^T + G \frac{\partial^2 Q}{\partial \theta^T} G^T \\
\]

3 Examples

In order to implement the differential filter, it is necessary to specify the first and second derivatives of \(F, G, H\) and \(Q\) along with the original state-space model. In this section, we shall consider three typical cases. The first two examples are the trend model and the standard seasonal adjustment model, for which three matrices (or vectors), \(F, G\) and \(H\) do not contain unknown parameters and thus the derivatives of these matrices become 0. This makes the algorithm for the gradient and the Hessian of the log-likelihood presented in the previous section considerably simple. The third example is the seasonal adjustment model with AR component. For this model, the matrix \(F\) depends on the unknown AR coefficients, although the derivative of \(F\) is very sparse. However, since we usually use a nonlinear transformation of the parameters and the Levinson’s formula between partial autocorrelations coefficients and AR coefficients, to ensure the stationarity condition, the expression of the non-zero elements of the derivatives of \(F\) becomes fairly complex.

3.1 Trend model

The trend model is a typical example of the case where only the noise covariance \(Q\) depends on the unknown parameter \(\theta\). Consider a trend model

\[y_n = T_n + w_n,\] (22)

where \(T_n\) is the trend component that typically follow the following model

\[(1 - B)^k T_n = v_n,\] (23)

where \(B\) is the back-shift operator satisfying \(B T_n = T_{n-1}\), \(v_n\) and \(w_n\) are assumed to be Gaussian white noise with variances \(\tau^2\) and 1, respectively (Kitagawa and Gersch (1984, 1996) and Kitagawa (2020a)). Note that for \(k = 1\) and \(k = 2\), the model \(23\) becomes \(T_n = T_{n-1} + v_n\) and \(T_n = 2T_{n-1} - T_{n-2} + v_n\), respectively.
and that in the Kalman filter, the essentially the same filtering results can be obtained by assuming that $\sigma^2 = 1$ (Kitagawa (2020a)) and thus the dimension of the unknown parameter vector is reduced by one, i.e., in the case of the trend model the dimension of the parameter becomes one.

This trend model can be expressed in the state-space model form as

$$ x_n = F x_{n-1} + G v_n $$
$$ y_n = H x_n + w_n, $$

with $v_n \sim N(0, Q)$ and $w_n \sim N(0,1)$ and the state vector $x_n$ and the matrices $F, G, H, Q$ and $R$ are defined by

$$ x_n = [T_n T_n]$$
$$ F = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} $$

for $k = 1$ and

$$ x_n = \begin{bmatrix} T_n \\ T_{n-1} \end{bmatrix}, \quad F = \begin{bmatrix} 2 & -1 \\ 1 & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} $$

for $k = 2$.

In this state-space representation, the parameter is $\tau^2$, and the $F, G, H$ and $R$ do not depend on the parameter. Therefore, we have $\frac{\partial F}{\partial \theta} = \frac{\partial G}{\partial \theta} = \frac{\partial H}{\partial \theta} = \frac{\partial R}{\partial \theta} = 0$ and $\frac{\partial^2 F}{\partial \theta \partial \theta^T} = \frac{\partial^2 G}{\partial \theta \partial \theta^T} = \frac{\partial^2 H}{\partial \theta \partial \theta^T} = \frac{\partial^2 R}{\partial \theta \partial \theta^T} = 0$.

In actual likelihood maximization, since there is a positivity constraint, $\tau^2 > 0$, it is frequently used a log-transformation,

$$ \theta = \log(\tau^2), $$

and maximize the log-likelihood with respect to this transformed parameter $\theta$. In this case,

$$ \frac{\partial Q}{\partial \theta} = \frac{\partial^2 Q}{\partial \theta \partial \theta^T} = \tau^2. $$

Since log-transformation is a monotone increasing function, we can get the same parameter by solving this modified optimization problem.

In this case, the recursive algorithm for gradient of the log-likelihood becomes significantly simple as follows:

$$ \frac{\partial \ell(\theta)}{\partial \theta} = -\frac{1}{2} \left( N \frac{\partial \hat{\sigma}^2}{\partial \theta} + \sum_{n=1}^{N} \frac{1}{r_n} \frac{\partial r_n}{\partial \theta} \right), $$

where, from (7), the derivatives of the one-step-ahead predtion error $\varepsilon_n$, the one-step-ahead prediction error variance $r_n$ and the observation noise variance $\hat{\sigma}^2$ are obtained by

$$ \frac{\partial \varepsilon_n}{\partial \theta_i} = -H \frac{\partial x_{n|n-1}}{\partial \theta_i} $$
$$ \frac{\partial r_n}{\partial \theta_i} = H \frac{\partial V_{n|n-1}}{\partial \theta_i} - H^T $$

$$ \frac{\partial \hat{\sigma}^2}{\partial \theta} = \frac{1}{N} \sum_{n=1}^{N} \left( \frac{2 \varepsilon_n}{r_n} \frac{\partial \varepsilon_n}{\partial \theta} - \frac{\varepsilon_n^2}{r_n} \frac{\partial r_n}{\partial \theta} \right), $$

(32)
The formula for obtaining the derivatives of the state and its variance covariance matrix become

\[
\begin{align*}
\frac{\partial x_{n|n-1}}{\partial \theta_i} &= F \frac{\partial x_{n-1|n-1}}{\partial \theta_i}, \\
\frac{\partial V_{n|n-1}}{\partial \theta_i} &= F \frac{\partial V_{n-1|n-1}}{\partial \theta_i} F^T + G \frac{\partial Q}{\partial \theta_i} G^T, \\
\frac{\partial K_n}{\partial \theta_i} &= \frac{\partial V_{n|n-1}}{\partial \theta_i} H^T r_n - V_{n|n-1} H^T r_n^2 \frac{\partial r_n}{\partial \theta_i}, \\
\frac{\partial x_{n|n}}{\partial \theta_i} &= \frac{\partial x_{n|n-1}}{\partial \theta_i} + \frac{\partial K_n}{\partial \theta_i} \varepsilon_n + K_n \frac{\partial x_{n|n-1}}{\partial \theta_i}, \\
\frac{\partial V_{n|n}}{\partial \theta_i} &= (I - K_n H) \frac{\partial V_{n|n-1}}{\partial \theta_i} - \frac{\partial K_n}{\partial \theta_i} H V_{n|n-1}.
\end{align*}
\]

The Hessian (the second derivative) of the log-likelihood is also obtained by a recursive formula, since, from \([12]\), it is given as

\[
\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T} = -\frac{N}{2} \left( \frac{1}{\sigma^2} \frac{\partial^2 \hat{\sigma}^2}{\partial \theta \partial \theta^T} - \frac{1}{(\sigma^2)^2} \frac{\partial \hat{\sigma}^2}{\partial \theta} \frac{\partial \hat{\sigma}^2}{\partial \theta^T} \right) + \frac{1}{2} \sum_{n=1}^{N} \left( \frac{1}{r_n} \frac{\partial^2 r_n}{\partial \theta \partial \theta^T} - \frac{1}{r_n^2} \frac{\partial r_n}{\partial \theta} \frac{\partial r_n}{\partial \theta^T} \right),
\]

where, from \([14]\), \(\frac{\partial^2 \varepsilon_n}{\partial \theta_i \partial \theta_j}\) and \(\frac{\partial^2 \hat{\sigma}^2}{\partial \theta_i \partial \theta_j}\) are obtained by

\[
\begin{align*}
\frac{\partial^2 \varepsilon_n}{\partial \theta_i \partial \theta_j} &= -H \frac{\partial^2 x_{n|n-1}}{\partial \theta_i \partial \theta_j}, \\
\frac{\partial^2 r_n}{\partial \theta_i \partial \theta_j} &= H \frac{\partial^2 V_{n|n-1}}{\partial \theta_i \partial \theta_j} H^T, \\
\frac{\partial^2 \hat{\sigma}^2}{\partial \theta_i \partial \theta_j} &= \frac{N}{2} \sum_{n=1}^{N} \left( \frac{2}{r_n} \left( \frac{\partial \varepsilon_n}{\partial \theta} \frac{\partial \varepsilon_n}{\partial \theta^T} + \varepsilon_n \frac{\partial^2 \varepsilon_n}{\partial \theta \partial \theta^T} \right) + \frac{2 \sigma_n^2}{r_n^3} \frac{\partial \varepsilon_n}{\partial \theta} \frac{\partial r_n}{\partial \theta} + \frac{2 \sigma_n^2}{r_n^2} \frac{\partial \varepsilon_n}{\partial \theta} \frac{\partial \varepsilon_n}{\partial \theta^T} + \frac{\sigma_n^2}{r_n^2} \frac{\partial^2 \varepsilon_n}{\partial \theta \partial \theta^T} \right).
\end{align*}
\]

Therefore, to evaluate the Hessian, the following computation should be performed along with the recursive formula for the log-likelihood and the gradient and Hessian of the log-likelihood. Note that since in this example, the parameter is one dimensional, we should read \(i = j\) in the following formula.

\[
\begin{align*}
\frac{\partial^2 x_{n|n-1}}{\partial \theta_i \partial \theta_j} &= F \frac{\partial^2 x_{n-1|n-1}}{\partial \theta_i \partial \theta_j}, \\
\frac{\partial^2 V_{n|n-1}}{\partial \theta_i \partial \theta_j} &= F \frac{\partial^2 V_{n-1|n-1}}{\partial \theta_i \partial \theta_j} F^T + G \frac{\partial^2 Q}{\partial \theta_i \partial \theta_j} G^T, \\
\frac{\partial^2 K_n}{\partial \theta_i \partial \theta_j} &= r_n^{-1} \frac{\partial^2 V_{n|n-1}}{\partial \theta_i \partial \theta_j} H^T + 2 r_n^{-3} V_{n|n-1} H^T \frac{\partial r_n}{\partial \theta_i} \frac{\partial r_n}{\partial \theta_j} \\
&\quad - r_n^{-2} \left( \frac{\partial V_{n|n-1}}{\partial \theta_i} H^T \frac{\partial r_n}{\partial \theta_j} + \frac{\partial V_{n|n-1}}{\partial \theta_j} H^T \frac{\partial r_n}{\partial \theta_i} + V_{n|n-1} H^T \frac{\partial^2 r_n}{\partial \theta_i \partial \theta_j} \right), \\
\frac{\partial^2 x_{n|n}}{\partial \theta_i \partial \theta_j} &= \frac{\partial^2 x_{n|n-1}}{\partial \theta_i \partial \theta_j} + K_n \frac{\partial^2 \varepsilon_n}{\partial \theta_i \partial \theta_j} + \frac{\partial K_n}{\partial \theta_i} \frac{\partial \varepsilon_n}{\partial \theta_j} + \frac{\partial K_n}{\partial \theta_j} \frac{\partial \varepsilon_n}{\partial \theta_i} + \frac{\partial^2 K_n}{\partial \theta_i \partial \theta_j} \varepsilon_n, \\
\frac{\partial^2 V_{n|n}}{\partial \theta_i \partial \theta_j} &= (I - K_n H) \frac{\partial^2 V_{n|n-1}}{\partial \theta_i \partial \theta_j} - \frac{\partial^2 K_n}{\partial \theta_i \partial \theta_j} H V_{n|n-1} - \frac{\partial K_n}{\partial \theta_i} H \frac{\partial V_{n|n-1}}{\partial \theta_j} - \frac{\partial K_n}{\partial \theta_j} H \frac{\partial V_{n|n-1}}{\partial \theta_i}.
\end{align*}
\]
For Whard (whole sale hardware) data (Kitagawa (2020a)), $N = 155$, the system noise variance parameter $\theta = \tau^2$ of the trend model with $m_1 = 1$ was estimated using the initial value $\tau^2 = 0.5$, i.e., $\theta_0 = \log(0.5) = -0.69315$. By a numerical optimization procedure the maximum likelihood estimate of the parameter is obtained as $\hat{\theta} = 1.50393$, i.e., $\hat{\tau}^2 = e^{\hat{\theta}} = 4.49933$.

Table [1] shows the log-likelihoods, the gradients, the Hessians and the observation noise variances of the initial and the final estimates. For comparison, the values obtained by the numerical difference are also shown in the table. The log-likelihood of the model with these initial and final estimates are $\ell(\theta) = -307.6616$ and $-317.9243$, respectively. It can be seen that the analytic gradient and the Hessian coincide with ones obtained by the numerical differentiation at least up to 4th digit.

Figure 1 shows the change of the log-likelihood, the gradient, Hessian and the observation noise variance for various values of the system noise variance $\tau^2$. In this case, the log-likelihood has only one local maximum and the gradient is unimodal. On the other hand, the Hessian has two peaks and two troughs.

Table [2] shows the results for the second order trend model. In this case, the final estimate obtained by the numerical optimization procedure depends on the initial estimate and two cases are shown in the table. If the initial estimate is set to $\theta_0 = -13.8155$, i.e., $\tau^2_0 = 10^{-4}$, the final estimate is $\hat{\theta} = -0.53318$, i.e., $\hat{\tau}^2 = 0.58674$ with the log-likelihood value $\ell(\hat{\theta}) = 296.179$. On the other hand, if we set the initial estimate as $\theta_0 = -6.90776$, i.e., $\tau^2_0 = 10^{-1}$, the final estimate becomes $\hat{\theta} = -7.0399$, i.e., $\hat{\tau}^2 = 0.87619 \times 10^{-3}$ with $\ell(\hat{\theta}) = 283.739$. Comparing the log-likelihood values, $\hat{\tau}^2 = 0.58674$ is the maximum likelihood estimate of the second order trend model.

Figure 2 shows the change of the log-likelihood, the gradient, the Hessian and the observation noise variance for various values of the system noise variance $\tau^2$ for the second order trend model. In this case, the log-likelihood is bimodal and the gradient attains zero at three points, two local maxima and one local minimum. The Hessian of the two local maximum likelihood estimate, $0.58674$ and $0.87619 \times 10^{-3}$ are $6.15072$ and $1.66045$, respectively. This indicates that the estimate $0.58674$ has sharper peak in the log-likelihood function.

Table 1: Comparison of the results by the numerical difference and the proposed analytic method for the first order trend model ($m_1 = 1$). Left: initial values, Right: final estimate.

| Parameter | Difference | Analytic | Difference | Analytic |
|-----------|------------|----------|------------|----------|
| $\tau^2$  | 0.50000    | 0.50000  | 4.49933    | 4.49933  |
| $\theta$  | -0.69315   | -0.69315 | 1.50393    | 1.50393  |
| $\ell(\theta)$ | -307.6616 | -307.6616 | 317.9243   | 317.9243 |
| $\frac{\partial \ell(\theta)}{\partial \theta}$ | 9.87647 | 9.87647  | $-2.7365 \times 10^{-7}$ | $-2.86304 \times 10^{-7}$ |
| $\frac{\partial^2 \ell(\theta)}{\partial \theta^2}$ | -3.54544 | -3.54545 | $-2.34221$ | $-2.34219$ |
| $\hat{\sigma}^2$ | $5.50118 \times 10^{-4}$ | $5.50152 \times 10^{-4}$ | $1.52463 \times 10^{-4}$ | $1.52497 \times 10^{-4}$ |
Figure 1: The log-likelihood, the gradient, the Hessian and the variance of the observation noise of the trend models with order 1. The horizontal axes are the value of the system noise variance in log-scale.

Table 2: Comparison of the results by the numerical difference and the proposed analytic method for the second order trend model ($m_1 = 2$)

|                      | Difference | Analytic   | Difference | Analytic   |
|----------------------|------------|------------|------------|------------|
| $\theta_0$           | -13.8155   | -13.8155   | -0.53318   | -0.53318   |
| $\tau_0^2$           | 0.100 × 10^{-5} | 0.100 × 10^{-5} | 0.58674    | 0.58674    |
| $\ell(\theta)$       | -270.02065 | -270.02065 | 296.17899  | 296.17899  |
| $\frac{\partial \ell(\theta)}{\partial \theta}$ | 1.68866 | 1.68866 | 5.6238 × 10^{-7} | 5.5944 × 10^{-7} |
| $\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta}$ | 0.010951 | 0.010953 | -6.150639 | -6.15072  |
| $\hat{\sigma}^2$     | 1.57779 × 10^{-3} | 1.57791 × 10^{-3} | 3.55274 × 10^{-4} | 3.55286 × 10^{-4} |
| $\theta$             | -6.90776   | -6.90776   | -7.0399    | -7.0399    |
| $\hat{\tau}^2$       | 0.1000 × 10^{-2} | 0.1000 × 10^{-2} | 0.87619 × 10^{-3} | 0.87619 × 10^{-3} |
| $\ell(\theta)$       | -283.72502 | -283.72502 | 283.73930  | 283.73930  |
| $\frac{\partial \ell(\theta)}{\partial \theta}$ | -0.214378 | -0.21437 | 0.24466 × 10^{-8} | -0.6156 × 10^{-8} |
| $\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta}$ | -1.58184 | -1.58183 | -1.66045   | -1.66045   |
| $\hat{\sigma}^2$     | 1.1221 × 10^{-3} | 1.12215 × 10^{-3} | 1.13024 × 10^{-3} | 1.13029 × 10^{-3} |
3.2 The standard seasonal adjustment model

As the second example, we consider a standard seasonal adjustment model

\[ y_n = T_n + S_n + w_n, \]  

where \( T_n \) and \( S_n \) are the trend component and the seasonal component that typically follow the following model

\[
\begin{align*}
T_n &= 2T_{n-1} - T_{n-2} + u_n, \\
S_n &= -(S_{n-1} + \cdots + S_{n-p+1}) + v_n.
\end{align*}
\]

The noise terms \( u_n, v_n \) and \( w_n \) are assumed to be Gaussian white noise with variances \( \tau_1^2, \tau_2^2 \) and \( \sigma^2 \), respectively (Kitagawa and Gersch (1984, 1996) and Kitagawa (2020a)).

This seasonal adjustment model with two component models can be expressed in state-space model form as

\[
\begin{align*}
x_n &= Fx_{n-1} + Gv_n \\
y_n &= Hx_n + w_n,
\end{align*}
\]

with \( v_n \sim N(0, Q) \) and \( w_n \sim N(0, 1) \) and the state vector \( x_n \) and the matrices \( F, G, H, Q \) and \( R \) are defined by

\[
\begin{align*}
x_n &= \begin{bmatrix} T_n \\ T_{n-1} \\ S_n \\ S_{n-1} \\ \vdots \\ S_{n-p+2} \end{bmatrix}, \\
F &= \begin{bmatrix} 2 & -1 & 1 & \cdots & -1 \\ 1 & 1 & -1 & \cdots & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & 1 \end{bmatrix}, \\
G &= \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}
\end{align*}
\]
\[ H = \begin{bmatrix} 1 & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} \]
\[ Q = \begin{bmatrix} \tau_1^2 & 0 \\ 0 & \tau_2^2 \end{bmatrix}, \quad R = 1. \]  

(40)

It is noted that, similar to the trend model, it is possible to assume that \( R = 1 \). In this case, the parameter is \((\tau_1^2, \tau_2^2)^T\), and the \( F, G, H \) and \( R \) do not depend on the parameter. In actual likelihood maximization, since there are positivity constrains, \( \tau_1^2 > 0 \) and \( \tau_2^2 > 0 \), we use the log-transformation,

\[ \theta_1 = \log(\tau_1^2), \quad \theta_2 = \log(\tau_2^2). \]  

(41)

In this case,

\[ \frac{\partial Q}{\partial \theta_1} = \frac{\partial^2 Q}{\partial \theta_1^2} = \begin{bmatrix} \tau_1^2 & 0 \\ 0 & 0 \end{bmatrix}, \quad \frac{\partial Q}{\partial \theta_2} = \frac{\partial^2 Q}{\partial \theta_2^2} = \begin{bmatrix} 0 & 0 \\ 0 & \tau_2^2 \end{bmatrix}, \quad \frac{\partial^2 Q}{\partial \theta_1 \partial \theta_2} = \frac{\partial^2 Q}{\partial \theta_2 \partial \theta_1} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \]

\[ \frac{\partial R}{\partial \theta_i} = 0, \quad \frac{\partial^2 R}{\partial \theta_i \partial \theta_j} = 0 \quad (i, j = 1, 2). \]  

(42)

Since \( F, G \) and \( H \) do not depend on \( \theta \) and \( \frac{\partial F}{\partial \theta} = 0, \frac{\partial G}{\partial \theta} = 0 \) and \( \frac{\partial H}{\partial \theta} = 0 \) hold, we can use the same recursive algorithm for gradient and the Hessian of the log-likelihood as the one for the trend model.

For Whard data, the standard seasonal adjustment model with \( m_1 = 2, m_2 = 1 \) is estimated using the initial estimates of parameters, \( \theta = (\log \tau_1^2, \log \tau_2^2)^T = (-5.29831, -14.98848)^T \). The log-likelihood of the model with these initial parameters is \( \ell(\theta) = -377.0849 \) and the gradient and the Hessian obtained by the numerical difference and the proposed method are shown in the Table 3. It can be seen that the numerical differentiation coincides with the analytic derivative up to 5th digit.

The maximum likelihood estimates of the system noises are \( \hat{\tau}_1^2 = 0.021185, \hat{\tau}_2^2 = 0.0068434 \) with \( \ell(\hat{\theta}) = 380.6569 \).

| Initial \( \hat{\tau}_1^2 \) | Proposed method | Numerical Difference |
|--------------------------|-----------------|---------------------|
| \[ 0.005000 \quad 0.0068160 \] | \[ 0.005000 \quad 0.0068160 \] | \[ 0.005000 \quad 0.0068160 \] |
| \[ -5.29831 \quad -4.98848 \] | \[ -5.29831 \quad -4.98848 \] | \[ -5.29831 \quad -4.98848 \] |
| \( \ell(\theta) \) | \(-377.0852\) | \(-377.0849\) |
| \( \frac{\partial \ell(\theta)}{\partial \theta} \) | \[4.82627, -0.31490\] | \[4.82627, -0.31490\] |
| \( \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T} \) | \[3.17294, -0.23066\] | \[3.17296, -0.23066\] |

| Optimized \( \hat{\tau}_1^2 \) | Proposed method | Numerical Difference |
|--------------------------|-----------------|---------------------|
| \[ 0.021185 \quad 0.0068434 \] | \[ 0.021185 \quad 0.0068434 \] | \[ 0.021185 \quad 0.0068434 \] |
| \( \hat{\theta} \) | \[ -3.8545 \quad -4.9845 \] | \[ -3.8545 \quad -4.9845 \] |
| \( \ell(\hat{\theta}) \) | \[380.6569\] | \[380.6569\] |
| \( \frac{\partial \ell(\hat{\theta})}{\partial \theta} \) | \[ -2.3897 \times 10^{-5} \quad -0.3747 \times 10^{-5} \] | \[ -1.46922 \times 10^{-6} \quad -0.38865 \times 10^{-6} \] |
| \( \frac{\partial^2 \ell(\hat{\theta})}{\partial \theta \partial \theta^T} \) | \[3.71968 \quad -0.20363 \] | \[3.71968 \quad -0.20363 \] |

|                  | \[ -0.20363 \quad 0.18135 \] | \[ -0.20363 \quad 0.18135 \] |
Figure 3 shows the contour of the log-likelihood of the seasonal adjustment model. The horizontal and the vertical axes indicate the common logarithms of $\tau_1^2$ and $\tau_2^2$, respectively. The log-likelihood has a similar values for smaller $\tau_2^2$ and form a plateau. The small value of $\frac{\partial^2 l(\theta)}{\partial \theta_2 \partial \theta_2}$, 0.18135, corresponds to this phenomenon.

### 3.3 Seasonal adjustment model with stationary AR component

The third example is a seasonal adjustment model with stationary AR component

$$y_n = T_n + S_n + p_n + w_n,$$

where $T_n$ and $S_n$ are the trend component and the seasonal component introduced in the previous subsection and $p_n$ is an AR component with AR order $m_3$ defined by

$$p_n = \sum_{j=1}^{m_3} a_j p_{n-j} + v^{(p)}_n.$$

Here $v^{(p)}_n$ is a Gaussian white noise with variance $\tau_3^2$. The model contains $3+m_3$ parameters and the parameter vector is given by $\theta = (\theta_1, \ldots, \theta_{3+m_3})^T \equiv (\log \tau_1^2, \log \tau_2^2, \log \tau_3^2, \theta_4, \ldots, \theta_{m_3+3})^T$. 
The matrices $F, G, H, Q$ and $R$ are defined by

$$
\begin{bmatrix}
T_n \\
T_{n-1}
\end{bmatrix},
F = \begin{bmatrix}
2 & -1 \\
1 & 1
\end{bmatrix},
G = \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 0
\end{bmatrix},
H = \begin{bmatrix}
1 & 0 & 1 & 0 & \cdots & 1 & 0 & \cdots & 0
\end{bmatrix}
$$

(45)

$$
H = \begin{bmatrix}
1 & 0 & 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0
\end{bmatrix}
$$

(46)

The relation between the parameter $\theta_j$ and the variances and AR coefficients are as follows.

$$
\tau_j^2 = e^{\theta_j}, (j = 1, \ldots, 3), \quad \beta_j = C \frac{e^{\theta_j} - 1}{e^{\theta_j} + 1}
$$

(47)

$$
\{a_j^{(m)} = a_j^{(m-1)} - \beta_m a_{m-j}, j = 1, \ldots, m\}, \quad \text{for } m = 1, \ldots, 3.
$$

(48)

Note that the equation (48) is the relation between the AR coefficients of order $m-1$ and those of the order $m$ used in the Levinson’s algorithm (Kitagawa (2020a)).

In this case,

$$
\frac{\partial Q}{\partial \theta_1} = \frac{\partial^2 Q}{\partial \theta_1 \partial \theta_1} = \begin{bmatrix}
\tau_1^2 & 0 & 0 \\
0 & \tau_2^2 & 0 \\
0 & 0 & \tau_3^2
\end{bmatrix}, \quad \frac{\partial Q}{\partial \theta_2} = \frac{\partial^2 Q}{\partial \theta_2 \partial \theta_2} = \begin{bmatrix}
0 & \tau_1^2 & 0 \\
0 & 0 & \tau_2^2 \\
0 & 0 & 0
\end{bmatrix}
$$

$$
\frac{\partial Q}{\partial \theta_3} = \frac{\partial^2 Q}{\partial \theta_3 \partial \theta_3} = \begin{bmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}, \quad \frac{\partial^2 Q}{\partial \theta_i \partial \theta_j} = \begin{bmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0
\end{bmatrix}, \quad \text{for } i \neq j
$$

(49)

$$
\begin{align*}
\left( \frac{\partial F}{\partial \theta_k} \right)_{pq} &= \frac{\partial a_q^{(m)}}{\partial \theta_k} \\
\left( \frac{\partial^2 F}{\partial \theta_q \partial \theta_k} \right)_{pq} &= \begin{cases}
\frac{\partial a_q^{(m)}}{\partial \theta_q} & \text{if } \begin{cases}
k = 4, \ldots, m_3 + 3, p = m_1 + M_2(p-1) + 1, \text{ and} \\
q = m_1 + M_2(p-1) + j, (j = 1, \ldots, m_3)
\end{cases} \\
0 & \text{otherwise}
\end{cases}
\end{align*}
$$

where $\left( \frac{\partial F}{\partial \theta_k} \right)_{pq}$ and $\left( \frac{\partial^2 F}{\partial \theta_q \partial \theta_k} \right)_{pq}$ denote the $(p, q)$ components of the matrices $\frac{\partial F}{\partial \theta_k}$ and $\frac{\partial^2 F}{\partial \theta_q \partial \theta_k}$, respectively.

and $\frac{\partial a_q^{(m)}}{\partial \theta_k}$ and $\frac{\partial^2 a_q^{(m)}}{\partial \theta_q \partial \theta_k}$ are obtained by

$$
\frac{\partial a_q^{(m)}}{\partial \theta_k} = \frac{\partial a_q^{(m)}}{\partial \beta_i} \frac{\partial \beta_i}{\partial \theta_k}, \quad i = 1, \ldots, m
$$

(50)

$$
\frac{\partial^2 a_q^{(m)}}{\partial \theta_q \partial \theta_k} = \frac{\partial^2 a_q^{(m)}}{\partial \beta_i \partial \beta_j} \frac{\partial \beta_i}{\partial \theta_q} \frac{\partial \beta_j}{\partial \theta_k} + \frac{\partial a_q^{(m)}}{\partial \beta_i} \frac{\partial^2 \beta_i}{\partial \theta_q \partial \theta_j}, \quad i, j = 1, \ldots, m
$$

(51)
\[
\frac{\partial \beta_k}{\partial \theta_i} = \begin{cases} 
C_k & \text{for } k = i \\
0 & \text{for } k \neq i
\end{cases}
\] (52)

\[
\frac{\partial^2 \beta_k}{\partial \theta_i \partial \theta_j} = \begin{cases} 
D_k & \text{for } k = i = j \\
0 & \text{otherwise}
\end{cases}
\] (53)

\[
\frac{\partial a_{k}^{(m)}}{\partial \beta_i} = \begin{cases} 
0 & \text{for } k = m \text{ and } i < m \\
1 & \text{for } k = m \\
\frac{\partial a_{k}^{(m-1)}}{\partial \beta_i} - \beta_m \frac{\partial a_{m-k}^{(m-1)}}{\partial \beta_i} & \text{for } k < m \text{ and } i < m \\
-a_{m-k}^{(m-1)} & \text{for } k < m \text{ and } i = m
\end{cases}
\] (54)

\[
\frac{\partial^2 a_{k}^{(m)}}{\partial \beta_i \partial \beta_j} = \begin{cases} 
0 & \text{for } k = m \\
\frac{\partial^2 a_{k}^{(m-1)}}{\partial \beta_i \partial \beta_j} - \beta_m \frac{\partial a_{m-k}^{(m-1)}}{\partial \beta_i \partial \beta_j} - \beta_m \frac{\partial^2 a_{m-k}^{(m-1)}}{\partial \beta_i \partial \beta_j} & \text{for } k < m \text{ and } i < m \\
\frac{\partial a_{m-k}^{(m-1)}}{\partial \beta_j} & \text{for } k < m \text{ and } i = m
\end{cases}
\] (55)

In the above equations, \( C_i \) and \( D_i \) are the first and the second derivatives of the nonlinear transformation \( C_i \frac{e^{\theta_i}}{e^{\theta_i} + 1} \) and are given by

\[ C_i = 2C \frac{e^{\theta_i}}{(e^{\theta_i} + 1)^2}, \quad D_i = 2C \frac{e^{\theta_i} (1 - e^{\theta_i})}{(e^{\theta_i} + 1)^3}. \] (56)

For \( m_1 = 2, m_2 = 1 \), the matrices \( \frac{\partial F}{\partial \theta}, \frac{\partial G}{\partial \theta}, \frac{\partial H}{\partial \theta}, \frac{\partial^2 F}{\partial \theta \partial \theta^T}, \frac{\partial^2 G}{\partial \theta \partial \theta^T} \) and \( \frac{\partial^2 H}{\partial \theta \partial \theta^T} \) are given by

\[
\frac{\partial F}{\partial \theta} = \begin{bmatrix} 0 & \cdots & 0 & 0 \\
\vdots & \ddots & \vdots & \vdots \\
0 & \cdots & 0 & 0 \\
0 & \cdots & 0 & \frac{\partial F_k}{\partial \theta_i} \end{bmatrix}, \quad \frac{\partial G}{\partial \theta} = \begin{bmatrix} 0 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & 0 \\
0 \\
0 \end{bmatrix}, \quad \frac{\partial H}{\partial \theta} = \begin{bmatrix} 0 \\
\vdots \\
0 \\
0 \end{bmatrix}
\] (57)

\[
\frac{\partial^2 F}{\partial \theta \partial \theta^T} = \begin{bmatrix} 0 & \cdots & 0 & 0 \\
\vdots & \ddots & \vdots & \vdots \\
0 & \cdots & 0 & 0 \\
0 & \cdots & 0 & \frac{\partial^2 F_k}{\partial \theta \partial \theta^T} \end{bmatrix}, \quad \frac{\partial^2 G}{\partial \theta \partial \theta^T} = \begin{bmatrix} 0 & \cdots & 0 \\
\vdots & \ddots & \vdots \\
0 & \cdots & 0 \\
0 \end{bmatrix}, \quad \frac{\partial^2 H}{\partial \theta \partial \theta^T} = \begin{bmatrix} 0 \\
\vdots \\
0 \\
0 \end{bmatrix}
\] (58)

where for \( m_3 = 1, 2 \) and \( 3, \frac{\partial F_3}{\partial \theta} \) and \( \frac{\partial^2 F_3}{\partial \theta \partial \theta^T} \) are respectively give by:

For \( m_3 = 1 \)

\[ F = [a_1], \quad \frac{\partial F}{\partial \theta_1} = [C_1], \quad \frac{\partial^2 F}{\partial \theta_1 \partial \theta_1} = [D_1]. \]

For \( m_3 = 2 \)

\[ F = \begin{bmatrix} a_1 & a_2 \\
1 & 0 \end{bmatrix} \]
The derivative of the one-step-ahead predictor and the filter differential filter shown in (12)-(21) become considerably simple as follows.

\[
\frac{\partial F}{\partial \theta_1} = C_1 \begin{bmatrix} 1 - \beta_2 & 0 \\ 0 & 0 \end{bmatrix}, \quad \frac{\partial^2 F}{\partial \theta_1^2} = D_1 \begin{bmatrix} 1 - \beta_2 & 0 \\ 0 & 0 \end{bmatrix}, \quad \frac{\partial^2 F}{\partial \theta_2 \partial \theta_1} = -C_1 C_2 \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \frac{\partial^2 F}{\partial \theta_2^2} = D_2 \begin{bmatrix} -\beta_1 & 1 \\ 0 & 0 \end{bmatrix}
\]

For \( m_3 = 3 \),

\[
F = \begin{bmatrix} a_1^{(3)} & a_2^{(3)} & a_3^{(3)} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad a_1^{(3)} = a_1^{(2)} - a_3^{(3)} a_2^{(2)} = \beta_1 - \beta_2 \beta_1 - \beta_3 \beta_2,
\]

\[
\frac{\partial F}{\partial \theta_1} = C_1 \begin{bmatrix} 1 - \beta_2 & \beta_3 (\beta_2 - 1) & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \frac{\partial^2 F}{\partial \theta_1^2} = D_1 \begin{bmatrix} 1 - \beta_2 & \beta_3 (\beta_2 - 1) & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \frac{\partial^2 F}{\partial \theta_1 \partial \theta_2} = C_1 \begin{bmatrix} -1 & \beta_3 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \frac{\partial^2 F}{\partial \theta_2 \partial \theta_3} = D_2 \begin{bmatrix} -1 & \beta_1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \frac{\partial^2 F}{\partial \theta_3 \partial \theta_3} = C_3 \begin{bmatrix} -1 & \beta_1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}
\]

Since we have \( \frac{\partial \ell}{\partial \theta} = \frac{\partial H}{\partial \theta} = \frac{\partial R}{\partial \theta} = 0 \) and \( \frac{\partial^2 \ell}{\partial \theta \partial \theta^T} = \frac{\partial^2 H}{\partial \theta \partial \theta^T} = \frac{\partial^2 R}{\partial \theta \partial \theta^T} = 0 \) for the current model, the differential filter shown in [12]-[21] become considerably simple as follows. [The gradient of the log-likelihood]

\[
\frac{\partial \ell(\theta)}{\partial \theta} = -\frac{1}{2} \left( \frac{N}{\delta^2} \frac{\partial^2 \ell}{\partial \theta \partial \theta} + \sum_{n=1}^{N} \frac{\partial r_n}{\partial \theta} \right), \quad (59)
\]

where

\[
\frac{\partial \varepsilon_n}{\partial \theta} = -H \frac{\partial x_n|n-1}{\partial \theta}, \quad \frac{\partial r_n}{\partial \theta} = H \frac{\partial V_n|n-1}{\partial \theta} H^T, \quad (60)
\]

\[
\frac{\partial^2 \varepsilon_n}{\partial \theta^2} = \frac{1}{N} \sum_{n=1}^{N} \left( \frac{\partial^2 \varepsilon_n}{\partial \theta \partial \theta^T} - \frac{\partial r_n}{\partial \theta} \right) , \quad (61)
\]

[The derivative of the one-step-ahead predictor and the filter]

\[
\frac{\partial x_n|n-1}{\partial \theta} = F \frac{\partial x_{n-1}|n-1}{\partial \theta} + F \frac{\partial F}{\partial \theta} \frac{\partial x_{n-1}|n-1}{\partial \theta} + F \frac{\partial F}{\partial \theta} \frac{\partial x_{n-1}|n-1}{\partial \theta} + \frac{\partial F}{\partial \theta} + G \frac{\partial Q}{\partial \theta} G^T \]

\[
\frac{\partial V_n|n-1}{\partial \theta} = F \frac{\partial V_{n-1|n-1}}{\partial \theta} F^T + \frac{\partial F}{\partial \theta} \frac{\partial V_{n-1|n-1}}{\partial \theta} F^T + F \frac{\partial F}{\partial \theta} + G \frac{\partial Q}{\partial \theta} G^T \]

\[
\frac{\partial K_n}{\partial \theta} = F \frac{\partial V_{n|n-1}}{\partial \theta} H^T r_n^{-1} - V_{n|n-1} H^T \frac{\partial r_n}{\partial \theta} r_n^{-2} \]

15
\[
\frac{\partial x_{n|n}}{\partial \theta} = \frac{\partial x_{n|n-1}}{\partial \theta} + K_n \frac{\partial \hat{\varepsilon}_n}{\partial \theta} + \frac{\partial K_n}{\partial \theta} \varepsilon_n
\]
\[
\frac{\partial V_{n|n}}{\partial \theta} = \frac{\partial V_{n|n-1}}{\partial \theta} - \frac{\partial K_n}{\partial \theta} H V_{n|n-1} - K_n H \frac{\partial V_{n|n-1}}{\partial \theta}.
\]

(63)

[The Hessian of the log-likelihood]
\[
\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T} = -\frac{N}{2} \left( \frac{1}{\sigma^2} \frac{\partial^2 \hat{\sigma}^2}{\partial \theta \partial \theta^T} - \frac{1}{(\hat{\sigma}^2)^2} \frac{\partial \hat{\sigma}^2}{\partial \theta} \frac{\partial \hat{\sigma}^2}{\partial \theta^T} \right) - \frac{1}{2} \sum_{n=1}^{N} \left( \frac{1}{r_n} \frac{\partial^2 r_n}{\partial \theta \partial \theta^T} - \frac{1}{r_n^2} \frac{\partial r_n \partial r_n}{\partial \theta} \right),
\]

where \(\frac{\partial^2 \hat{\varepsilon}_n}{\partial \theta \partial \theta^T}\), \(\frac{\partial^2 r_n}{\partial \theta \partial \theta^T}\), and \(\frac{\partial^2 \hat{\sigma}^2}{\partial \theta \partial \theta^T}\) are obtained by
\[
\frac{\partial^2 \varepsilon_n}{\partial \theta \partial \theta^T} = -H \frac{\partial^2 x_{n|n-1}}{\partial \theta \partial \theta^T}
\]
\[
\frac{\partial^2 r_n}{\partial \theta \partial \theta^T} = H \frac{\partial^2 V_{n|n-1}}{\partial \theta \partial \theta^T} H^T
\]
\[
\frac{\partial^2 \hat{\sigma}^2}{\partial \theta \partial \theta^T} = \frac{1}{N} \sum_{n=1}^{N} \left\{ 2 \left( \frac{\partial \hat{\varepsilon}_n \partial \varepsilon_n}{\partial \theta \partial \theta^T} + \varepsilon_n \frac{\partial^2 \hat{\varepsilon}_n}{\partial \theta \partial \theta^T} \right) + \frac{2 \hat{\varepsilon}_n^2}{r_n^3} \frac{\partial r_n \partial r_n}{\partial \theta} \right\}
\]

(64)

[The second derivatives of the one-step-ahead predictor and the filter]
\[
\frac{\partial^2 x_{n|n-1}}{\partial \theta \partial \theta^T} = \frac{\partial F \frac{\partial x_{n-1|n-1}}{\partial \theta}}{\partial \theta} + \frac{\partial F \frac{\partial x_{n-1|n-1}}{\partial \theta}}{\partial \theta} + \frac{\partial F \frac{\partial x_{n-1|n-1}}{\partial \theta}}{\partial \theta} + \frac{\partial^2 F}{\partial \theta \partial \theta^T} x_{n-1|n-1}
\]
\[
\frac{\partial^2 V_{n|n-1}}{\partial \theta \partial \theta^T} = \frac{\partial F \frac{\partial V_{n-1|n-1}}{\partial \theta}}{\partial \theta} + \frac{\partial F \frac{\partial V_{n-1|n-1}}{\partial \theta}}{\partial \theta} + \frac{\partial F \frac{\partial V_{n-1|n-1}}{\partial \theta}}{\partial \theta} + \frac{\partial^2 F}{\partial \theta \partial \theta^T} V_{n-1|n-1}
\]
\[
+ \frac{\partial F \frac{\partial V_{n-1|n-1}}{\partial \theta}}{\partial \theta} + \frac{\partial F \frac{\partial V_{n-1|n-1}}{\partial \theta}}{\partial \theta} + \frac{\partial F \frac{\partial V_{n-1|n-1}}{\partial \theta}}{\partial \theta} + \frac{\partial^2 F}{\partial \theta \partial \theta^T} V_{n-1|n-1}
\]
\[
+ \frac{\partial F \frac{\partial V_{n-1|n-1}}{\partial \theta}}{\partial \theta} + \frac{\partial F \frac{\partial V_{n-1|n-1}}{\partial \theta}}{\partial \theta} + \frac{\partial F \frac{\partial V_{n-1|n-1}}{\partial \theta}}{\partial \theta} + \frac{\partial^2 F}{\partial \theta \partial \theta^T} V_{n-1|n-1}
\]
\[
+ \frac{\partial^2 V_{n|n-1}}{\partial \theta \partial \theta^T} = \frac{\partial^2 V_{n|n-1}}{\partial \theta \partial \theta^T} \frac{H^T r_n}{r_n} \left( \frac{\partial V_{n-1|n-1}}{\partial \theta} H^T \frac{\partial r_n}{\partial \theta} + \frac{\partial V_{n-1|n-1}}{\partial \theta} H^T \frac{\partial r_n}{\partial \theta} + V_{n-1|n-1} \frac{H^T \frac{\partial r_n}{\partial \theta}}{r_n} \right) r_n^{-2}
\]
\[
+ 2 V_{n|n-1} H^T \frac{\partial r_n}{\partial \theta} \frac{\partial r_n}{\partial \theta} r_n^{-3}
\]

(65)

Table 4 shows the gradients and the Hessian matrix obtained by the differential filter and the numerical differencing method. The initial estimates of the parameters are set to be \(\theta = (\log(0.25682 \times 10^{-3}), \log(1.0), \log(0.52499), 1.7099, -0.89985)\) and the log-likelihood of the model is \(\ell(\theta) = -348.9595\).

The maximum likelihood estimates of the model are \(\tau^2_1 = 1.8824 \times 10^{-4}, \tau^2_2 = 1.1348 \times 10^{-2}, \tau^2_3 = 6.2550 \times 10^{-2}, a_1^{(2)} = 1.6546\) and \(a_2^{(2)} = -0.6884\) with \(\ell(\hat{\theta}) = 387.9554\). The gradients and the Hessian matrix for this maximum likelihood estimates are shown in Table 5. In this case as well, the analytic derivative matches the numerical differentiation up to the fifth digit.
Table 4: Comparison of the gradient vectors and the Hessian matrix obtained by the differential filter and the numerical differencing.

| By differential filter: |  |  |  |  |  |
|-------------------------|------------------|------------------|------------------|------------------|------------------|
| $\frac{\partial \ell(\theta_0)}{\partial \theta}$ | 3.472775 | −12.340431 | 3.244577 | 21.235746 | 4.505459 |
| | 0.320932 | 1.233847 | −2.574948 | −4.783933 | −0.262825 |
| | 1.233847 | −11.299148 | 5.956742 | 7.917642 | 0.399997 |
| $\frac{\partial^2 \ell(\theta_0)}{\partial \theta \partial \theta^T}$ | −2.574948 | 5.956742 | −10.176196 | −11.233390 | −0.372234 |
| | −4.783933 | 7.917642 | 11.233390 | −24.934335 | 0.822854 |
| | −0.262825 | 0.399997 | −0.372234 | 0.822854 | 3.126537 |

| By numerical differencing: |  |  |  |  |  |
|---------------------------|------------------|------------------|------------------|------------------|------------------|
| $\frac{\partial \ell(\theta_0)}{\partial \theta}$ | 3.472775 | −12.340431 | 3.244577 | 21.235747 | 4.505459 |
| | 0.320913 | 1.233881 | −2.574897 | −4.783931 | −0.262818 |
| | 1.233881 | −11.300011 | 5.956570 | 7.917585 | 0.400087 |
| $\frac{\partial^2 \ell(\theta_0)}{\partial \theta \partial \theta^T}$ | −2.574897 | 5.956570 | −10.176824 | −11.233393 | −0.372323 |
| | −4.783931 | 7.917585 | 11.233393 | −24.934451 | 0.822888 |
| | −0.262818 | 0.400087 | −0.372323 | 0.822888 | 3.126465 |

4 Summary

The gradient and the Hessian matrix of the log-likelihood of the reduced order linear state-space model are given. Details of the implementation of the algorithm for trend model, the standard seasonal adjustment model, and the seasonal adjustment model with stationary AR component are given. For each implementation, comparison with a numerical difference method is shown.

Aknowledgements

This work was supported in part by JSPS KAKENHI Grant Number 18H03210. The author is grateful to the project members, Prof. Kunitomo, Prof Nakano, Prof. Kyo, Prof. Sato, Prof. Tanokura and Prof. Nagao for their stimulating discussions.
Table 5: Comparison of the gradient vectors and the Hessian matrix obtained by the differential filter and the numerical differencing.

By differential filter:

|          | \(\frac{\partial\ell}{\partial\hat{\theta}}\) | \(\frac{\partial^2\ell}{\partial\theta^2}\) |
|----------|-------------------------------|---------------------------------|
|          | 0.000001                      | -0.000000                      |
|          | -0.512604                     | 0.036523                       |
|          | 0.036523                      | -0.305611                      |
|          | -0.186673                     | 0.371657                       |
|          | -0.000000                     | 0.000000                       |
|          | 0.029211                      | -0.349766                      |
|          | 0.036523                      | -0.305611                      |
|          | -8.066285                     | -0.000000                      |
|          | -8.066400                     | -0.000003                      |
|          | 7.186088                      | 7.186088                       |
|          | -10.588254                    | -10.588417                     |

By numerical differencing:

|          | \(\frac{\partial\ell}{\partial\hat{\theta}}\) | \(\frac{\partial^2\ell}{\partial\theta^2}\) |
|----------|-------------------------------|---------------------------------|
|          | 0.000002                      | -0.000001                      |
|          | -0.512613                     | 0.036524                       |
|          | 0.036524                      | -0.305652                      |
|          | -0.186672                     | 0.371642                       |
|          | -0.000001                     | 0.000003                       |
|          | 0.029217                      | -0.349770                      |
|          | 0.036524                      | -0.305652                      |
|          | -8.066400                     | -0.000003                      |
|          | -8.066305                     | -0.000005                      |
|          | 7.186086                      | 7.186086                       |
|          | 0.000002                      | -10.588254                     |

References

[1] Akaike, H. (1980b), “Seasonal adjustment by a Bayesian modeling”, J. Time Series Anal., 1, 1–13.

[2] Akaike, H. and Ishiguro, M. (1983), “Comparative study of X-11 and Bayesian procedure of seasonal adjustment,” Applied Time Series Analysis of Economic Data, U.S. Census Bureau.

[3] Anderson, B. D. O, and Moore, J. B. (2012). Optimal filtering. DOver Publications, New York.

[4] Box, G.E.P., Hillmer, S.C. and Tiao, G.C. (1978) “Analysis and modeling of seasonal time series”, in Seasonal Analysis of Time Seres, ed.Zelher, A., US Bureau of the Census, Economic Research Report ER-1, 309–334.

[5] Fletcher, R. (2013). Practical methods of optimization. John Wiley & Sons.

[6] Kitagawa, G. (1987). Non-Gaussian state-space modeling of nonstationary time series. Journal of the American Statistical Association, 82(400), 1032–1041.

[7] Kitagawa, G. (1989). Non-Gaussian seasonal adjustment, Computers & Mathematics with Applications, Vol.18, No.6/7, pp. 503−514.

[8] Kitagawa, G. (1994). The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother, Annals of the Institute of Statistical Mathematics, Vol. 46, No.4, pp. 605–623.

[9] Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models, Journal of Computational and Graphical Statistics, Vol.5, no.1, pp. 1–25.

[10] Kitagawa, G. (2020a). Introduction to Time Series Modeling with Applications in R, Monographs on Statistics and Applied Probability 166, CRD Press, Chapman & Hall, New York.

[11] Kitagawa, G. (2020b). Computation of the Gradient and the Hessian of the Log-likelihood of the State-space Model by the Kalman Filter, arXiv preprint arXiv:2011.09638

[12] Kitagawa, G. and Gersch, W. (1984), “A smoothness priors-state space modeling of trend and seasonality”, J. Amer. Statist. Assoc., 79, 378−389.
[13] Kitagawa, G. and Gersch, W. (1996), *Smoothness Priors Analysis of Time Series*, Lecture Notes in Statistics, **116**, Springer, New York.

[14] Kohn, R., and Ansley, C. F. (1985). “Computing the likelihood and its derivatives for a gaussian ARMA model”. *Journal of Statistical Computation and Simulation*, 22(3-4), 229–263.

[15] Konishi, S. and Kitagawa, G. (2008), *Information Criteria and Statistical Modeling*, Springer Series in Statistics, pp-273, Springer, New York.

[16] Kulikova, M. V. (2009). “Likelihood Gradient Evaluation Using Square-Root Covariance Filters”, *IEEE Transactions on Automatic Control*, Vol. 54, Issue 3, 646-651.

[17] Takeuchi, K. (1976). “Distributions of information statistics ans criteria for adequacy of models”, *Mathematical Science*, **1553**, 12–18 (in Japanese).

[18] Zadrozny, P. A. (1989). “Analytic derivatives for estimation of linear dynamic models”, *Computers Math. Applic.*, Vol. 18, No. 6/7, 539-553.