A COMPARISON OF CONFIDENCE INTERVALS IN STATE SPACE MODELS

Jinyu Du
Southern Methodist University, jinyudu21@gmail.com

Follow this and additional works at: https://scholar.smu.edu/hum_sci_statisticalscience_etds

Part of the Applied Statistics Commons, Statistical Models Commons, and the Statistical Theory Commons

Recommended Citation
https://scholar.smu.edu/hum_sci_statisticalscience_etds/40

This Thesis is brought to you for free and open access by the Statistical Science at SMU Scholar. It has been accepted for inclusion in Statistical Science Theses and Dissertations by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.
A COMPARISON OF CONFIDENCE INTERVALS
IN STATE SPACE MODELS

Approved by:

Dr. Ronald W. Butler
C. F. Frensley Professor in Statistical Science, Southern Methodist University

Dr. Ian Richard Harris
Associate Professor in Statistical Science, Southern Methodist University

Dr. Raanju R. Sundararajan
Assistant Professor in Statistical Science, Southern Methodist University

Dr. Eric C. Larson
Associate Professor in Computer Science, Southern Methodist University
A COMPARISON OF CONFIDENCE INTERVALS
IN STATE SPACE MODELS

A Dissertation Presented to the Graduate Faculty of the
Dedman College
Southern Methodist University
in
Partial Fulfillment of the Requirements
for the degree of
Doctor of Philosophy
with a
Major in Statistical Science
by
Jinyu Du

B.S., Chemistry, ZhengZhou University
M.S., Engineering Management, Missouri S&T
M.S., Applied Mathematics and Statistics, Missouri S&T

July 31, 2023
ACKNOWLEDGMENTS

I would like to express my deep appreciation for my thesis advisor, Professor Ronald Butler for his teaching, guidance, support and encouragement. Dr. Butler was always available to help when I needed his input. His knowledge and mastery of statistics inspires me to persevere. His valuable insights greatly contributed to the completion and success of this thesis. Moreover, I would like to thank Professor Ian Harris and Professor Eric Larson for outstanding classes on mathematical statistics and machine learning, which equipped me well with the knowledge and skills necessary for my dissertation work. I also want to thank Professor Raanju Sundararajan and all the committee members for their kind, generous and valuable input and suggestions for improving the work for this thesis.

I am grateful for all the faculty members in the Statistical Science department at SMU. Their excellent teaching and mentoring helped me grow stronger in the statistics profession. Furthermore, I would like to thank our department administrator Ms. Sheila Crain for her constant loving support and coordination for our Ph.D marathon.

I would like to thank my parents and family for their support, understanding, unconditional love and unwavering faith for me. I especially thank my mother for being the anchor of my life, for her living life with perseverance, character and hope. I also appreciate the kindness and helping hands from many friends. Last but not least, I humbly bow to the divine love and intelligence that gives me and all of us the ability to think, learn, and search for a meaningful and promising life.
This thesis develops general procedures for constructing confidence intervals (CIs) of the error disturbance parameters (standard deviations) and transformations of the error disturbance parameters in time-invariant state space models (ssm). With only a set of observations, estimating individual error disturbance parameters accurately in the presence of other unknown parameters in ssm is a very challenging problem. We attempted to construct four different types of confidence intervals, Wald, likelihood ratio, score, and higher-order asymptotic intervals for both the simple local level model and the general time-invariant state space models (ssm). We show that for a simple local level model, both the likelihood ratio interval and the higher-order asymptotic interval have superior performance with underage, coverage, and overage accurate to 1% of the target values. For the general time-invariant ssm, we focus on constructing CIs for the correlation coefficient $\rho$ of the standard deviations of the accelerations in a two-dimensional object tracking example. Results show that the likelihood ratio method can achieve underage, coverage, and overage accurate to 1% of the target values, whereas the Wald method has far inferior performance. Weighing the theoretical and computational complexities of all four methods, we consider the likelihood ratio method as the most practical method for constructing confidence intervals (CIs) of the error disturbance parameters in time-invariant ssm.
# TABLE OF CONTENTS

LIST OF FIGURES ......................................................................................... ix

LIST OF TABLES ........................................................................................... xiii

CHAPTER

1 Introduction ................................................................................................. 1
  1.1. Introduction and the Objectives ............................................................. 1
  1.2. Organization of the Thesis ................................................................. 3

2 Local Level Model ....................................................................................... 4
  2.1. Constructing Confidence Intervals for the State Disturbance Standard De-
viation ........................................................................................................ 7
    2.1.1. Constructing Confidence Intervals by Inverting the Likelihood Ratio
      Test ......................................................................................................... 8
    2.1.2. Constructing Wald Confidence Intervals ........................................ 10
    2.1.3. Constructing Score Confidence Intervals ....................................... 11
    2.1.4. Saddlepoint Methodology ............................................................. 11
      2.1.4.1. Curved Exponential Families .................................................... 11
      2.1.4.2. Higher-order Asymptotics for the Signed Root of the Log-
        likelihood Ratio Test Statistic ......................................................... 13
  2.2. Simulation ................................................................................................. 16
    2.2.1. Maximum Likelihood Estimators (MLEs) for the State Space Model 16
    2.2.2. Simulation Procedure .................................................................... 22
    2.2.3. Evaluation of the Four Confidence Interval Methods ..................... 22
      2.2.3.1. Simulation Results ................................................................. 22
  2.3. Kalman Recursions ............................................................................... 35
4.3.5. Derivation for $\hat{u}_\rho(\rho)$ .................................................. 66

4.4. Simulation ................................................................. 70

4.4.1. Challenges With Constructing the Score Intervals and the Higher-order Asymptotic Intervals ........................................ 70

4.4.2. Simulation Procedure ................................................. 72

4.4.3. Evaluation of the Wald Method and the Likelihood Ratio Method .. 72

4.4.3.1. Interpretation of the Simulation Results ....................... 73

APPENDIX

A Derivation for the Local Level Model........................................ 80

B More Scatter Plots and Histograms for the Local Level Model ............ 84

C Score Test Statistic Is Invariant to the Parameterization .................. 100

BIBLIOGRAPHY ................................................................. 102
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Confidence interval determination using the likelihood ratio test</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>A scatter plot of MLEs overlaid with the 95% predictive interval, $T = 60, \beta = 0.2$</td>
<td>18</td>
</tr>
<tr>
<td>2.3</td>
<td>Histogram of standard error of $\hat{\beta}$, where $se(\hat{\beta}) &lt; 30, T = 60, \beta = 0.2$.</td>
<td>18</td>
</tr>
<tr>
<td>2.4</td>
<td>Histogram of standard error of $\hat{\delta}$, where $se(\hat{\delta}) &lt; 30, T = 60, \beta = 0.2$.</td>
<td>19</td>
</tr>
<tr>
<td>2.5</td>
<td>Histogram of standard error of $\hat{\delta}_0$, where $se(\hat{\delta}_0) &lt; 30, T = 60, \beta = 0.2$.</td>
<td>19</td>
</tr>
<tr>
<td>2.6</td>
<td>A scatter plot of MLEs overlaid with the 95% predictive interval, $T = 60, \beta = 2.8$.</td>
<td>20</td>
</tr>
<tr>
<td>2.7</td>
<td>Histogram of standard error of $\hat{\beta}$, where $se(\hat{\beta}) &lt; 30, T = 60, \beta = 2.8$.</td>
<td>20</td>
</tr>
<tr>
<td>2.8</td>
<td>Histogram of standard error of $\hat{\delta}$, where $se(\hat{\delta}) &lt; 30, T = 60, \beta = 2.8$.</td>
<td>21</td>
</tr>
<tr>
<td>2.9</td>
<td>Histogram of standard error of $\hat{\delta}_0$, where $se(\hat{\delta}_0) &lt; 30, T = 60, \beta = 2.8$.</td>
<td>21</td>
</tr>
<tr>
<td>2.10</td>
<td>Underage when datasets have 60 data points and $c_1 = c_2 = 30$.</td>
<td>28</td>
</tr>
<tr>
<td>2.11</td>
<td>Coverage when datasets have 60 data points and $c_1 = c_2 = 30$.</td>
<td>28</td>
</tr>
<tr>
<td>2.12</td>
<td>Overage when datasets have 60 data points and $c_1 = c_2 = 30$.</td>
<td>29</td>
</tr>
<tr>
<td>2.13</td>
<td>Underage when datasets have 120 data points and $c_1 = c_2 = 30$.</td>
<td>29</td>
</tr>
<tr>
<td>2.14</td>
<td>Coverage when datasets have 120 data points and $c_1 = c_2 = 30$.</td>
<td>30</td>
</tr>
<tr>
<td>2.15</td>
<td>Overage when datasets have 120 data points and $c_1 = c_2 = 30$.</td>
<td>30</td>
</tr>
<tr>
<td>2.16</td>
<td>Underage when datasets have 240 data points and $c_1 = c_2 = 30$.</td>
<td>31</td>
</tr>
<tr>
<td>2.17</td>
<td>Coverage when datasets have 240 data points and $c_1 = c_2 = 30$.</td>
<td>31</td>
</tr>
<tr>
<td>2.18</td>
<td>Overage when datasets have 240 data points and $c_1 = c_2 = 30$.</td>
<td>32</td>
</tr>
<tr>
<td>2.19</td>
<td>Underage when datasets have 480 data points and $c_1 = c_2 = 30$.</td>
<td>32</td>
</tr>
</tbody>
</table>
2.20 Coverage when datasets have 480 data points and $c_1 = c_2 = 30$ .......... 33
2.21 Overage when datasets have 480 data points and $c_1 = c_2 = 30$ .......... 33
2.22 Underage when datasets have 960 data points and $c_1 = c_2 = 30$ .......... 34
2.23 Coverage when datasets have 960 data points and $c_1 = c_2 = 30$ .......... 34
2.24 Overage when datasets have 960 data points and $c_1 = c_2 = 30$ .......... 35
2.25 Recursive computation of the log-likelihood ........................................ 36
2.26 Recursive computation of the score vectors ........................................ 39
2.27 Recursive computation of the observed information $j_t(\theta)$ .................. 41
2.28 Recursive computation of the Fisher information $i_t(\theta)$ ...................... 43

4.1 The surface plot of the externally computed log-likelihood function for the constrained likelihood of the 2-D tracking model......................... 72
4.2 Underage for $\rho$ when datasets have 60 observations ......................... 74
4.3 Coverage for $\rho$ when datasets have 60 observations............................ 74
4.4 Overage for $\rho$ when datasets have 60 observations............................. 75
4.5 Underage for $\rho$ when datasets have 120 observations.......................... 75
4.6 Coverage for $\rho$ when datasets have 120 observations........................... 76
4.7 Overage for $\rho$ when datasets have 120 observations............................ 76
4.8 Underage for $\rho$ when datasets have 240 observations........................... 77
4.9 Coverage for $\rho$ when datasets have 240 observations........................... 77
4.10 Overage for $\rho$ when datasets have 240 observations............................ 78
4.11 Underage for $\rho$ when datasets have 480 observations........................... 78
4.12 Coverage for $\rho$ when datasets have 480 observations........................... 79
4.13 Overage for $\rho$ when datasets have 480 observations............................ 79

B.1 A scatter plot of MLEs overlaid with the 95% predictive interval, $T = 120, \beta = 0.2$ 84
B.2 Histogram of standard error of $\hat{\beta}$, where $se(\hat{\beta}) < 30, T = 120, \beta = 0.2.$ .... 84
<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
</table>
| B.3  | Histogram of standard error of \( \hat{\delta} \), where \( se(\hat{\delta}) < 30, T = 120, \beta = 0.2 \) ........  
| B.4  | Histogram of standard error of \( \hat{\delta}_0 \), where \( se(\hat{\delta}_0) < 30, T = 120, \beta = 0.2 \) ........  
| B.5  | Histogram of standard error of \( \hat{\beta} \), where \( se(\hat{\beta}) < 30, T = 120, \beta = 2.8 \) ........  
| B.6  | Histogram of standard error of \( \hat{\delta} \), where \( se(\hat{\delta}) < 30, T = 120, \beta = 2.8 \) ........  
| B.7  | A scatter plot of MLEs overlaid with the 95% predictive interval, \( T = 120, \beta = 2.8 \)  
| B.8  | Histogram of standard error of \( \hat{\delta}_0 \), where \( se(\hat{\delta}_0) < 30, T = 120, \beta = 2.8 \) ........  
| B.9  | A scatter plot of MLEs overlaid with the 95% predictive interval, \( T = 240, \beta = 0.2 \)  
| B.10 | Histogram of standard error of \( \hat{\beta} \), where \( se(\hat{\beta}) < 30, T = 240, \beta = 0.2 \) ........  
| B.11 | Histogram of standard error of \( \hat{\delta} \), where \( se(\hat{\delta}) < 30, T = 240, \beta = 0.2 \) ........  
| B.12 | Histogram of standard error of \( \hat{\delta}_0 \), where \( se(\hat{\delta}_0) < 30, T = 240, \beta = 0.2 \) ........  
| B.13 | A scatter plot of MLEs overlaid with the 95% predictive interval, \( T = 240, \beta = 2.8 \)  
| B.14 | Histogram of standard error of \( \hat{\beta} \), where \( se(\hat{\beta}) < 30, T = 240, \beta = 2.8 \) ........  
| B.15 | Histogram of standard error of \( \hat{\delta} \), where \( se(\hat{\delta}) < 30, T = 240, \beta = 2.8 \) ........  
| B.16 | Histogram of standard error of \( \hat{\delta}_0 \), where \( se(\hat{\delta}_0) < 30, T = 240, \beta = 2.8 \) ........  
| B.17 | A scatter plot of MLEs overlaid with the 95% predictive interval, \( T = 480, \beta = 0.2 \)  
| B.18 | Histogram of standard error of \( \hat{\beta} \), where \( se(\hat{\beta}) < 30, T = 480, \beta = 0.2 \) ........  
| B.19 | Histogram of standard error of \( \hat{\delta} \), where \( se(\hat{\delta}) < 30, T = 480, \beta = 0.2 \) ........  
| B.20 | Histogram of standard error of \( \hat{\delta}_0 \), where \( se(\hat{\delta}_0) < 30, T = 480, \beta = 0.2 \) ........  
| B.21 | A scatter plot of MLEs overlaid with the 95% predictive interval, \( T = 480, \beta = 2.8 \)  
| B.22 | Histogram of standard error of \( \hat{\beta} \), where \( se(\hat{\beta}) < 30, T = 480, \beta = 2.8 \) ........  
| B.23 | Histogram of standard error of \( \hat{\delta} \), where \( se(\hat{\delta}) < 30, T = 480, \beta = 2.8 \) ........  
| B.24 | Histogram of standard error of \( \hat{\delta}_0 \), where \( se(\hat{\delta}_0) < 30, T = 480, \beta = 2.8 \) ........  
| B.25 | A scatter plot of MLEs overlaid with the 95% predictive interval, \( T = 960, \beta = 0.2 \)  
| B.26 | Histogram of standard error of \( \hat{\beta} \), where \( se(\hat{\beta}) < 30, T = 960, \beta = 0.2 \) ........  

xi
B.27 Histogram of standard error of $\hat{\delta}$, where $se(\hat{\delta}) < 30, T = 960, \beta = 0.2$. ............ 97
B.28 Histogram of standard error of $\hat{\delta}_0$, where $se(\hat{\delta}_0) < 30, T = 960, \beta = 0.2$. ............ 97
B.29 A scatter plot of MLEs overlaid with the 95% predictive interval, $T = 960, \beta = 2.8$ 98
B.30 Histogram of standard error of $\hat{\beta}$, where $se(\hat{\beta}) < 30, T = 960, \beta = 2.8$. ............ 98
B.31 Histogram of standard error of $\hat{\delta}$, where $se(\hat{\delta}) < 30, T = 960, \beta = 2.8$. ............ 99
B.32 Histogram of standard error of $\hat{\delta}_0$, where $se(\hat{\delta}_0) < 30, T = 960, \beta = 2.8$. ............ 99
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Underage, coverage, and overage simulation results for the local level model, ( T = {60, 120} )</td>
<td>23</td>
</tr>
<tr>
<td>2.2 Underage, coverage, and overage simulation results for the local level model, ( T = {240, 480} )</td>
<td>24</td>
</tr>
<tr>
<td>2.3 Underage, coverage, and overage simulation results for the local level model, ( T = 960 )</td>
<td>25</td>
</tr>
<tr>
<td>4.1 Underage, coverage, and overage simulation results for the two dimensional object tracking model</td>
<td>73</td>
</tr>
</tbody>
</table>
This thesis is dedicated to my parents Zi’an Du, Guilan Zhang, and my thesis advisor Dr. Ronald Butler, whose unwavering support empowers and inspires me.
Chapter 1
Introduction

1.1. Introduction and the Objectives

The objective of this thesis is to develop general procedures for constructing confidence intervals (CIs) of the error disturbance parameters (standard deviations) and transformations of the error disturbance parameters in time-invariant state space models (ssm). Wald intervals [Wald, 1947] are typically used for CIs of the error disturbance parameters in existing literature and programs, but there have been no evaluations of the performance of such Wald intervals. Moreover, there are no existing confidence interval procedures for error disturbance parameters in state space models by inverting the likelihood ratio test, the score test or using the higher-order asymptotics. We use these procedures to construct confidence intervals (CIs) for the error disturbance parameters and one-to-one transformations of these disturbance parameters in time-invariant state space models and evaluated their underage/coverage/overage performance via simulation. Our overall recommendation in making inference for disturbance parameters in time-invariant ssm is to use the likelihood ratio procedure because of its overall superior performance and efficiency in computation.

A state space model is a discrete-time stochastic model that has a state equation and an observation equation. The state equation is a Markov model that describes the transition of true states in time. The observation equation describes how observations of the true states with measurement error change in time. See [Durbin and Koopman, 2012], [Harvey, 1990] and [Meinhold and Singpurwalla, 1983] for a general presentation of the state space models and Kalman filter. Given a set of observations, estimating individual disturbance parameters in the presence of other unknown parameters in ssm is notoriously difficult, even for linear
time-invariant state space models [Auger-Méthé et al., 2016]. Two equivalent approaches can be used to estimate error disturbance parameters from the likelihood: 1) use the multivariate normal structure of the likelihood; 2) use Kalman recursions to compute the likelihood. For each of the two approaches, we attempted to construct CIs for error disturbance parameters using four different methods: 1) Wald method; 2) likelihood ratio method, i.e. inverting the likelihood ratio test; 3) score method; and 4) higher-order asymptotic method.

The Wald, likelihood ratio, and score methods are first-order methods [Cox and Hinkley, 1974]. The Wald method only requires estimation of the global maximum likelihood estimators (MLEs). The score method requires estimation of the constrained MLEs over a fine grid of hypothesized parameter values. The likelihood ratio method requires estimation of both global and constrained MLEs over the fine grid. For the likelihood ratio test, the signed root of the likelihood ratio test statistic denoted as \( \hat{w} \) follows an approximate standard normal distribution. Our proposed higher-order asymptotic method is based on an approximate continuous cumulative distribution function (CDF) of \( \hat{w} \) given ancillary \( a \). It was developed by [Barndorff-Nielsen, 1980, Barndorff-Nielsen, 1990] and is summarized in [Butler, 2007].

We first evaluated the performance of the four confidence intervals methods in a simple local level model [Durbin and Koopman, 2012]. It was shown that the higher-order asymptotic method and the likelihood ratio method have considerably better performance than the Wald and score methods. The former two methods have close to target underage, coverage and overage accuracy. Then, we used Kalman recursion to compute the three first-order methods and achieved the same results as those from the multivariate normal approach. We abandoned the Kalman recursion for the higher-order asymptotic method due to its inherent technical complexities.

Next, we considered the linear time-invariant state space models, with the aim of generalizing the methods for constructing confidence intervals (CIs) of the error disturbance parameters and transformations of error disturbance parameters. For the general procedure,
we focused on the theoretical development and simulation for an object tracking example in two-dimensional space. Our results show that the likelihood ratio method is the best approach for constructing confidence intervals (CIs) for the error disturbance parameters in the presence of nuisance parameters. In comparison, the Wald intervals have far inferior performance. We evaluated the performance in terms of accuracy in achieving the nominal underage, coverage, and overage. The score method and the higher-order asymptotic method were abandoned due to challenges in finding MLEs accurately.

1.2. Organization of the Thesis

In this thesis, we investigate procedures for constructing confidence intervals (CIs) of the error disturbance parameters and transformations of the error disturbance parameters in time-invariant state space models (ssm). The procedures are evaluated using a simple local level model as well as practical object tracking examples via simulation.

In Chapter 2, we introduce the local level model and developed four CI methods for the state error disturbance parameter. Performance of the CI methods was evaluated by underage, coverage, and overage via Monte Carlo simulation. Kalman recursions were employed to compute the quantities in the three first-order methods.

In Chapter 3, we introduce the general linear time-invariant state space model and the procedure for constructing CI for any disturbance parameters in the presence of other nuisance parameters.

In Chapter 4, we introduce two examples. Example one is about object tracking in one dimension. Example two is about object tracking in two dimensions. We are interested in constructing confidence intervals for the correlation of the state error disturbances in example two. Results and the various challenges arising in the process are discussed.
Chapter 2
Local Level Model

The local level model as presented in [Durbin and Koopman, 2012] is a simple state space model of the form

\[ x_t = \alpha x_{t-1} + \beta u_t, \quad (2.1) \]
\[ y_t = \gamma x_t + \delta \epsilon_t, \quad t = 1, \ldots, T, \quad (2.2) \]

where Equation 2.1 describes the state space \( \{x_t, t \geq 1\} \) and Equation 2.2 reflects the observations \( \{y_t, t \geq 1\} \). We shall focus on the case in which \( \alpha = 1, \gamma = 1, u_t \sim N(0,1), \epsilon_t \sim N(0,1) \), where \( N \) stands for a normal distribution, and \( \beta \) and \( \delta \) are unknown parameters. All four parameters \( \alpha, \beta, \gamma, \delta \) may be taken as unknown, but in the simple local level model, we will only take \( \beta \) and \( \delta \) as unknown parameters.

Assume the initial state \( x_0 \sim N(a_0, p_0) \), where \( x_0 \) is unobserved, and \( a_0, p_0 \) are known values for mean and variance of \( x_0 \). Also, assume \( \{u_t\} \) and \( \{\epsilon_t\} \) are all mutually independent and are independent of \( x_0 \). Since \( \beta u_t \sim N(0, \beta^2) \) and \( \delta \epsilon_t \sim N(0, \delta^2) \), the values \( \beta^2 \) and \( \delta^2 \) are variances and we are especially interested in the value of \( \beta \) (the interest parameter) and perhaps less interested in \( \delta \) (the nuisance parameter).

The local level model is a random walk model. In this model, we initially want to infer \( \beta \) from the knowledge of the observations. In particular, we will find a confidence interval for \( \beta \) in the presence of nuisance parameter \( \delta \).
Let $y^T = (y_1, \ldots, y_T)$, and rewrite the model in the form

$$y_t = \gamma(\alpha^t x_0 + \beta \sum_{j=1}^t u_j \alpha^{t-j}) + \delta \epsilon_t, \quad t = 1, \ldots, T. \quad (2.3)$$

When $\alpha = 1, \gamma = 1$, the model becomes

$$y_t = x_0 + \beta \sum_{j=1}^t u_j + \delta \epsilon_t, \quad t = 1, \ldots, T. \quad (2.4)$$

Using standard results from multivariate normal theory, we can derive that (see Appendix A for the derivation)

$$y \sim \mathcal{N}_T(1a_0, \Omega), \quad (2.5)$$

where $\mathcal{N}_T$ denotes a multivariate normal distribution of dimension $T$, $1 = (1, \ldots, 1)$ is $T \times 1$ and $\Omega = 11^T p_0 + \Sigma$. Here, $\Omega$ and $\Sigma$ are $T \times T$ symmetric matrices and $\Sigma = \{\sigma_{ij}\}$ has the form

$$\sigma_{ij} = \begin{cases} 
\beta^2 \min(i, j), & i \neq j \\
\beta^2 + \delta^2, & i = j \quad i, j = 1, \ldots, T.
\end{cases} \quad (2.6)$$

Without loss of generality, we assume $a_0 = 0$. This is possible because we can always subtract $1a_0$ from $y$ and center the distribution of $y$ around 0. Then $y \sim \mathcal{N}_T(0, \Omega)$. We want to work with the standard deviations $\beta, \delta$. With $\xi = (\beta, \delta)$, the log likelihood function is

$$l(\xi) = c - \frac{1}{2} \log |\Omega| - \frac{1}{2} y^T \Omega^{-1} y = c - \frac{1}{2} \log |\Omega| - \frac{1}{2} \text{tr}(\Omega^{-1} yy^T). \quad (2.7)$$

where $c$ is a constant. This takes the form of a curved exponential family with canonical parameter $\theta_\xi$ as the $T(T + 1)/2$ distinct elements of Equation 2.8 whose components are functions of $\xi = (\xi_1, \xi_2) = (\beta, \delta)$. 
This representation is awkward because

\[-\frac{1}{2} \text{tr}(\Omega^{-1}yy^T) = -\sum_{i=1}^{T} \frac{1}{2} \omega_{ii} y_i^2 - \sum_{i<j} \omega_{ij} y_i y_j\]

and \(\theta_\xi\) consists of \(\{-\omega_{ii}/2 : i = 1, \ldots, T\} \cup \{-\omega_{ij} : i < j\}\) with different weights on and off the diagonal. For computational simplicity, we will simply use all \(T^2\) elements of \(-\Omega^{-1}/2\) as the canonical parameter and allow the repeats as distinct parameters since this does not matter in either the computations or the theory. Thus \(\theta_\xi\) consists of all \(T^2\) elements of \(-\Omega^{-1}/2\). Accordingly, the canonical sufficient statistic vector \(x\) consists of all \(T^2\) elements of \(yy^T\). This results in a \((T^2, 2)\)-curved exponential family.

Taking

\[\theta_\xi = -\frac{1}{2} \text{vec}(\Omega^{-1})\]  
\[x = \text{vec}(yy^T)\]  
\[c(\theta_\xi) = \frac{1}{2} \log |\Omega|\]

then

\[l(\xi) = c - \frac{1}{2} \log |\Omega| - \frac{1}{2} \left\{ \text{vec}(\Omega^{-1}) \right\}^T \text{vec}(yy^T)\]

\[= c - c(\theta_\xi) + \theta_\xi^T x,\]

which is the form of a curved exponential family. The definition of the curved exponential families can be found in Section 2.1.4.1.
The $2 \times 2$ observed information matrix with respect to $\xi = (\xi_1, \xi_2) = (\beta, \delta)$ is

$$j(\xi) = -\frac{\partial^2 l(\xi)}{\partial \xi \partial \xi^T} = \{j_{kl} : k, l = 1, 2\} \quad (2.13)$$

and $j_{kl}$ is derived in Appendix A as

$$j_{kl} = -\frac{1}{2} \text{tr}(\Omega^{-1} \Omega_k \Omega^{-1} \Omega_l) + \frac{1}{2} \text{tr}\{\Omega^{-1} \Omega_{kl}(I - \Omega^{-1} y y^T)\} + \text{tr}(\Omega^{-1} \Omega_k \Omega^{-1} \Omega_l \Omega^{-1} y y^T) \quad (2.14)$$

where $\Omega_k = \partial \Omega / \partial \xi_k$ and $\Omega_{kl} = \partial^2 \Omega / \partial \xi_k \partial \xi_l$.

The expected value of $j_{kl}$ is

$$i_{kl} = E(j_{kl}) = \frac{1}{2} \text{tr}(\Omega^{-1} \Omega_k \Omega^{-1} \Omega_l) \quad (2.15)$$

and $i_{kl}$ is the $(k, l)$-element of the Fisher information matrix, which is

$$i(\xi) = E\{j(\xi)\} = \{i_{kl} : k, l = 1, 2\} = \left\{\frac{1}{2} \text{tr}(\Omega^{-1} \Omega_k \Omega^{-1} \Omega_l) : k, l = 1, 2\right\} \quad (2.16)$$

2.1. Constructing Confidence Intervals for the State Disturbance Standard Deviation

In this section, we review three standard statistical methods for confidence interval construction. A higher-order asymptotic confidence interval method is detailed in Section 2.1.4.
2.1.1. Constructing Confidence Intervals by Inverting the Likelihood Ratio Test

We can construct confidence intervals for the state disturbance standard deviation $\beta$ by inverting the likelihood ratio test (LRT). We test the hypothesis that

\[ H_0 : \beta = \beta_0 \]
\[ H_1 : \beta \neq \beta_0 \]  \hspace{1cm} (2.17)

using the signed likelihood ratio test. The likelihood ratio test statistic is

\[ \lambda = \frac{L(\beta_0, \hat{\delta}_0)}{L(\hat{\beta}, \hat{\delta})}, \]  \hspace{1cm} (2.18)

where $L(\cdot)$ is the likelihood of the local level model, $\beta_0$ is the hypothesized value for $\beta$ under the null hypothesis $H_0$, $(\hat{\beta}, \hat{\delta})$ are the global maximum likelihood estimators (MLEs), and $\hat{\delta}_0$ is the constrained MLE for $\delta$ when holding $\beta = \beta_0$ fixed.

Finding both of the maximum likelihood estimators can be efficiently performed using the maximum likelihood parameter estimation function (\texttt{estimate}) [MathWorks, 2023b] for state space models (ssm) [MathWorks, 2023a].

To test against the two-sided alternative hypothesis in Equation 2.17, we use the signed root of $\lambda$ given as

\[ \hat{w}(\beta) = sgn(\hat{\beta} - \beta)\sqrt{-2\log \lambda} = sgn(\hat{\beta} - \beta)\sqrt{-2\{l(\beta_0, \hat{\delta}_0) - l(\hat{\beta}, \hat{\delta})\}} \sim N(0, 1), \]  \hspace{1cm} (2.19)

where $sgn(\cdot)$ is the sign of the value, $l(\cdot)$ is the log-likelihood. Under the assumption that $\beta$ is the true parameter, $\hat{w}(\beta)$ has an approximate $N(0, 1)$ distribution based on standard likelihood ratio asymptotics applied to this model. A 95\% asymptotic confidence interval (CI) for $\beta$ is \{\(\beta : -1.96 < \hat{w}(\beta) < 1.96\}\}. 

8
To show that \(\hat{w}(\beta)\) may serve as a pivotal quantity for estimating such confidence intervals, we must show that \(g(\beta) = \Phi\{\hat{w}(\beta)\}\) is monotone in \(\beta\). This is true if \(\hat{w}(\beta)\) is monotone in \(\beta\). To show this, write

\[
\hat{w}^2(\beta) = -2\log \lambda = -2l\{\beta, \hat{\delta}(\beta)\} + \text{constant}
\]  

(2.20)

Taking derivatives with respect to \(\beta\) on both sides and applying the chain rule on the right side, we get

\[
2\hat{w}(\beta) \frac{\partial \hat{w}(\beta)}{\partial \beta} = -2 \left[ l'_\beta\{\beta, \hat{\delta}(\beta)\} + l'_\delta\{\beta, \hat{\delta}(\beta)\} \frac{\partial \hat{\delta}(\beta)}{\partial \beta} \right]
\]

(2.21)

where \(l'_\beta\{\beta, \hat{\delta}(\beta)\} = \partial l\{\beta, \hat{\delta}(\beta)\}/\partial \beta\) and \(l'_\delta\{\beta, \hat{\delta}(\beta)\} = \partial l\{\beta, \hat{\delta}(\beta)\}/\partial \delta\). Since \(l'_\delta\{\beta, \hat{\delta}(\beta)\} = 0\),

\[
\hat{w}(\beta) \frac{\partial \hat{w}(\beta)}{\partial \beta} = -l'_\beta\{\beta, \hat{\delta}(\beta)\}
\]

(2.22)

At \(\beta = \hat{\beta}\), both sides are 0.

When \(\beta < \hat{\beta}\), \(l'_\beta\{\beta, \hat{\delta}(\beta)\} > 0\) because \(l(\beta, \delta)\) is an increasing function on \(\beta < \hat{\beta}\), and the maximum must occur at \(\hat{\beta}\). Similarly, when \(\beta > \hat{\beta}\), \(l'_\beta\{\beta, \hat{\delta}(\beta)\} < 0\) because \(l(\beta, \delta)\) is an decreasing function on \(\beta > \hat{\beta}\). Hence,

\[
\frac{\partial \hat{w}(\beta)}{\partial \beta} = \frac{-l'_\beta\{\beta, \hat{\delta}(\beta)\}}{\hat{w}(\beta)} = \frac{-l'_\beta\{\beta, \hat{\delta}(\beta)\}}{\text{sgn}(\beta - \hat{\beta})\sqrt{-2\log \lambda}} = \begin{cases} 
0 & \text{if } \beta = \hat{\beta} \\
< 0 & \text{if } \beta \neq \hat{\beta}
\end{cases}
\]

(2.23)

This shows that \(\hat{w}(\beta)\) is monotone decreasing in \(\beta\).

Computationally, we can perform the hypothesis test for each value of \(\beta = \beta_0\) on a fine grid of \(\beta \in (0.001, u)\) for some chosen value of \(u\). If we use the significance level \(\alpha = 0.05\) for the hypothesis test, the 95% CI is \((\beta_L, \beta_U)\), where \(\beta_L\) solves \(\hat{w}(\beta_L) = -1.96\) and \(\beta_U\) solves \(\hat{w}(\beta_U) = 1.96\).
We simulated a dataset with 100 observations using the simple local level model below

\[
x_t = x_{t-1} + u_t, \quad (2.24)
\]
\[
y_t = x_t + \sqrt{2}\epsilon_t, \quad t = 1, \ldots, T, \quad (2.25)
\]

and constructed a 95% CI for \( \beta = 1 \) using the likelihood ratio test. The 95% CI for \( \beta \) based on the LRT is (0.940081, 1.636818). The confidence interval determination is depicted in Figure 2.1.

![Figure 2.1: Confidence interval determination using the likelihood ratio test](image)

2.1.2. Constructing Wald Confidence Intervals

In testing the hypothesis in Equation 2.17, the square root of the Wald statistic is

\[
\sqrt{W} = \frac{\hat{\beta} - \beta_0}{\hat{s}e(\hat{\beta})} \sim N(0, 1). \quad (2.26)
\]

where \( \hat{s}e(\hat{\beta}) = \sqrt{(1,1)} \)-element of \( i^{-1}(\hat{\xi}) \).
A 95% Wald confidence interval for $\beta$ is $\hat{\beta} \pm 1.96 \hat{se}(\hat{\beta})$. It is relatively easy to construct the Wald confidence intervals for the state disturbance standard deviation $\beta$ using the standard output for $\hat{\beta}$ and $\hat{se}(\hat{\beta})$ from the state space model estimates from MATLAB.

2.1.3. Constructing Score Confidence Intervals

We can construct the score confidence intervals for the state disturbance standard deviation $\beta$. We define the score vector as $s^T = (l'_\beta(\xi), l'_\delta(\xi))$. The score test statistic for testing $H_0 : \beta = \beta_0$ is

$$S(\beta_0) = s(\hat{\xi}_0)^T i(\hat{\xi}_0)^{-1} s(\hat{\xi}_0)$$

(2.27)

where $\hat{\xi}_0 = (\beta_0, \hat{\delta}_0)$ and $\hat{\delta}_0 = \hat{\delta}(\beta_0)$. The 95% score confidence interval for $\beta$ is $\{ \beta : \sqrt{S(\beta)} < 1.96 \}$.

2.1.4. Saddlepoint Methodology

2.1.4.1. Curved Exponential Families

Suppose $y$ is the observed data and let $x = x(y)$ be the $m \times 1$ canonical sufficient statistic for an exponential family with the form

$$f(y; \xi) = \exp \{ \theta_\xi^T x - c(\theta_\xi) - d(y) \}$$

(2.28)

where $\theta_\xi$ is the $m \times 1$ canonical parameter which depends on the lower dimensional vector $\xi \in \Xi \subset R^k$ with $k < m$ and $\Xi$ is an open subset in $R^k$. The parametric class $\{ f(\cdot; \xi) : \xi \in \Xi \}$ is called an $(m, k)$-curved exponential family.
The $k \times k$ Fisher information matrix is

$$i(\xi) = -E\left(\frac{\partial^2 l_\xi}{\partial \xi \partial \xi^T}\right) = \hat{\theta}_\xi^T \Sigma_\xi \hat{\theta}_\xi$$  \hspace{1cm} (2.29)

where $l_\xi = \ln f(y; \xi)$ and $\hat{\theta}_\xi = \partial \theta_\xi / \partial \xi^T$ is an $m \times k$ matrix.

Let $\mu_\xi = E(X; \xi)$ denote the mean parameterization for the canonical sufficient statistic $X$, then

$$\mu_\xi = E(X; \xi) = \frac{\partial c(\theta)}{\partial \theta} \bigg|_{\theta=\theta_\xi} = c'(\theta_\xi)$$  \hspace{1cm} (2.30)

is $m \times 1$ and $\dot{\mu}_\xi$ is $m \times k$ and given by

$$\mu_\xi = \frac{\partial \mu_\xi}{\partial \xi^T} = \frac{\partial \mu_\xi}{\partial \theta^T} \frac{\partial \theta^T}{\partial \xi^T} = c''(\theta_\xi) \hat{\theta}_\xi = \Sigma_\xi \hat{\theta}_\xi,$$  \hspace{1cm} (2.31)

where

$$\Sigma_\xi = c''(\theta_\xi) = \text{cov}(X; \xi).$$  \hspace{1cm} (2.32)

Thus, $i(\xi)$ can be written as $i(\xi) = \hat{\theta}_\xi^T \mu_\xi$. Moreover, the $k \times k$ observed information matrix can be expressed in terms of $i(\xi)$ as

$$j(\xi) = -\frac{\partial^2 l_\xi}{\partial \xi \partial \xi^T} = i(\xi) - D_\xi$$  \hspace{1cm} (2.33)

where

$$(D_\xi)_{ij} = \left\{ (x - \mu_\xi)^T \frac{\partial \theta_\xi}{\partial \xi_i \partial \xi_j} \right\}$$  \hspace{1cm} (2.34)

In the regular exponential family setting, $j(\xi) = i(\xi)$, whereas in the curved exponential family setting, $j(\xi) \neq i(\xi)$. 


2.1.4.2. Higher-order Asymptotics for the Signed Root of the Log-likelihood Ratio Test Statistic

We use $\xi^T = (\xi_1, \xi_2) = (\beta, \delta)$, where $\beta$ is the scalar interest parameter and $\delta$ is the scalar nuisance parameter. The signed root of the likelihood ratio test statistic is $\hat{w}(\beta)$ in Equation 2.19. The conditional CDF of $\hat{w}(\beta)|a; \xi$ given an approximate ancillary $a$ has strong dependence on only the $\beta$ component of $\xi$. Let $\hat{W}_\beta$ denote the random variable and $\hat{w}(\beta)$ as its observed value. Then, [Barndorff-Nielsen, 1990] derived an approximate continuous CDF approximation for the random variable $\hat{W}_\beta$ given ancillary $a$ as

$$
\hat{P}\{\hat{W}_\beta \leq \hat{w}(\beta)|a\} = \Phi\{\hat{w}(\beta)\} + \phi\{\hat{w}(\beta)\} \left\{\frac{1}{\hat{w}(\beta)} - \frac{1}{\hat{u}(\beta)}\right\}, \quad \hat{\beta} \neq \beta \tag{2.35}
$$

where $\Phi(\cdot)$ is the standard normal CDF, $\phi(\cdot)$ is the standard normal density, $\hat{w}(\beta)$ is defined in Equation 2.19, and $\hat{u}(\beta)$ is given in [Butler, 2007]. The term $\Phi\{\hat{w}(\beta)\}$ is the first-order asymptotic likelihood ratio pivot leading to first-order 95% confidence interval $\{\beta : 0.025 < \Phi\{\hat{w}(\beta)\} < 0.975\}$. Replacing $\Phi\{\hat{w}(\beta)\}$ with $G(\beta) = \hat{P}\{\hat{W}_\beta \leq \hat{w}(\beta)|a\}$ leads to the second-order 95% CI as $\{\beta : 0.025 < G(\beta) < 0.975\}$.

The value for $\hat{u}(\beta)$ is difficult to compute. So we use the [Skovgaard, 1996] approximation as given in [Butler, 2007] instead. The approximation for $\hat{u}(\beta)$ is

$$
\hat{u}(\beta) = \text{sgn}(\hat{\beta} - \beta) \sqrt{\frac{|i(\hat{\xi})|}{|j_{\delta\delta}(\hat{\xi}_0)|}} \left| \left( \frac{\partial \theta^T}{\partial \delta} \right|_{\xi = \hat{\xi}_0} \right) \frac{\partial \mu_\xi}{\partial \xi^T} \right|_{\xi = \hat{\xi}}. \tag{2.36}
$$
Here, $j_{\delta\delta}(\hat{\xi}_0)$ is the $(2,2)$-element of $j(\xi)$ evaluated at $\xi = \hat{\xi}_0$. Notation $||\cdot||$ denotes the absolute value of the determinant of a matrix, and $\hat{\theta}_0 = \hat{\theta}_{\hat{\xi}_0}$. We can derive that

$$
\left| \left| \left( \frac{\partial \theta^T}{\partial \delta} \right|_{\xi=\hat{\xi}_0} \right| \right| \left( \frac{\partial \mu}{\partial \xi^T} \right|_{\xi=\hat{\xi}} = \left| \left| \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix} \right| \right| \tag{2.37}
$$

where

$$
d_{11} = \frac{1}{2} \text{tr} \left\{ \Omega^{-1}(\hat{\xi}_0)\Omega_2(\hat{\xi}_0)\Omega^{-1}(\hat{\xi}_0)\Omega_1(\hat{\xi}) \right\} \tag{2.38}
d_{12} = \frac{1}{2} \text{tr} \left\{ \Omega^{-1}(\hat{\xi}_0)\Omega_2(\hat{\xi}_0)\Omega^{-1}(\hat{\xi}_0)\Omega_2(\hat{\xi}) \right\} \tag{2.39}
d_{21} = \frac{1}{2} \text{tr} \left\{ \{\Omega^{-1}(\hat{\xi}_0) - \Omega^{-1}(\hat{\xi})\}\Omega_1(\hat{\xi}) \right\} \tag{2.40}
d_{22} = \frac{1}{2} \text{tr} \left\{ \{\Omega^{-1}(\hat{\xi}_0) - \Omega^{-1}(\hat{\xi})\}\Omega_2(\hat{\xi}) \right\} \tag{2.41}
$$

and

$$
\Omega_k(\hat{\xi}_0) = \frac{\partial \Omega}{\partial \xi_k} \bigg|_{\xi=\hat{\xi}_0} \quad \Omega_k(\hat{\xi}) = \frac{\partial \Omega}{\partial \xi_k} \bigg|_{\xi=\hat{\xi}} \tag{2.42}
$$

We will show the derivation for Equation 2.37 below.

$$
\mu = E(X; \xi) = E \left\{ \text{vec}(yy^T) \right\} = \text{vec} \left\{ E(yy^T) \right\} = \text{vec} \left\{ \Omega(\xi) \right\} \tag{2.43}
$$

Denote

$$
\mu_{\hat{\xi}} = \frac{\partial \mu}{\partial \xi} \bigg|_{\xi=\hat{\xi}} = \frac{\partial \text{vec} \{ \Omega(\xi) \}}{\partial \xi} \bigg|_{\xi=\hat{\xi}} = \left[ \text{vec} \{ \Omega_1(\hat{\xi}) \}, \text{vec} \{ \Omega_2(\hat{\xi}) \} \right] \tag{2.44}
$$

where

$$
\text{vec} \{ \Omega_k(\hat{\xi}) \} = \frac{\partial \text{vec} \{ \Omega(\xi) \}}{\partial \xi_k} \bigg|_{\xi=\hat{\xi}}, \quad k \in \{1, 2\} \tag{2.45}
$$

with

$$
\Omega_k(\hat{\xi}) = \frac{\partial \Omega(\xi)}{\partial \xi_k} \bigg|_{\xi=\hat{\xi}}. \tag{2.46}
$$
Since \( \theta_{\xi} = -\frac{1}{2} \text{vec}(\Omega^{-1}) \),
\[
\dot{\theta}_{\xi} = \frac{\partial \theta}{\partial \xi^T} = \left[ \frac{\partial \theta}{\partial \xi_1}, \frac{\partial \theta}{\partial \xi_2} \right] \tag{2.47}
\]
where
\[
\frac{\partial \theta}{\partial \xi_k} = -\frac{1}{2} \text{vec} \left( \frac{\partial \Omega^{-1}}{\partial \xi_k} \right) = \frac{1}{2} \text{vec} \left( \Omega^{-1} \Omega_k \Omega^{-1} \right), \quad k \in \{1, 2\} \tag{2.48}
\]
Recall that \( \xi = (\xi_1, \xi_2) = (\beta, \delta) \). From Equation 2.48, it is easy to get
\[
\left. \frac{\partial \theta^T}{\partial \delta} \right|_{\xi = \hat{\xi}_0} = \left. \frac{\partial \theta^T}{\partial \xi_2} \right|_{\xi = \hat{\xi}_0} = \frac{1}{2} \left[ \text{vec} \{ \Omega^{-1}(\hat{\xi}_0)\Omega_2(\hat{\xi}_0)\Omega^{-1}(\hat{\xi}_0) \} \right]^T \tag{2.49}
\]
Thus,
\[
\left. \frac{\partial \theta^T}{\partial \delta} \right|_{\xi = \hat{\xi}_0} \left. \frac{\partial \mu_{\xi}}{\partial \xi^T} \right|_{\xi = \hat{\xi}} = \frac{1}{2} \left[ \text{tr} \{ \Omega^{-1}(\hat{\xi}_0)\Omega_2(\hat{\xi}_0)\Omega^{-1}(\hat{\xi}_0) \} \right] \left[ \text{tr} \{ \Omega^{-1}(\hat{\xi}_0)\Omega_1(\hat{\xi}) \} \right] \left[ \text{tr} \{ \Omega^{-1}(\hat{\xi}_0)\Omega_2(\hat{\xi}_0) \} \right] \tag{2.50}
\]
Moreover, \( \dot{\theta} = \dot{\theta}_{\xi} = -\frac{1}{2} \text{vec}(\Omega^{-1}(\hat{\xi})) \), thus \( \dot{\theta}^T = \dot{\theta}_{\xi}^T = -\frac{1}{2} \left[ \text{vec}(\Omega^{-1}(\hat{\xi})) \right]^T \).
\[
\dot{\theta}_{\xi}^T \mu_{\xi} = -\frac{1}{2} \left[ \text{vec}(\Omega^{-1}(\hat{\xi})) \right]^T \text{vec}(\Omega_1(\hat{\xi})) \left[ \text{vec}(\Omega^{-1}(\hat{\xi})) \right]^T \text{vec}(\Omega_2(\hat{\xi})) \tag{2.52}
\]
\[
= -\frac{1}{2} \left[ \text{tr} \{ \Omega^{-1}(\hat{\xi})\Omega_1(\hat{\xi}) \} \right] \left[ \text{tr} \{ \Omega^{-1}(\hat{\xi})\Omega_2(\hat{\xi}) \} \right] \tag{2.53}
\]
Thus,
\[
(\dot{\theta} - \dot{\theta}_0)^T \mu_{\xi} = \frac{1}{2} \left[ \text{tr} \{ \Omega^{-1}(\hat{\xi}_0) - \Omega^{-1}(\hat{\xi}) \} \Omega_1(\hat{\xi}) \right], \text{tr} \left[ \{ \Omega^{-1}(\hat{\xi}_0) - \Omega^{-1}(\hat{\xi}) \} \Omega_2(\hat{\xi}) \right] \tag{2.54}
\]
Combining results from Equation 2.51 and Equation 2.54, we get the expressions for Equation 2.37.
2.2. Simulation

2.2.1. Maximum Likelihood Estimators (MLEs) for the State Space Model

The parameter estimation of state-space models is performed using the `estimate` function in MATLAB, which maximizes the log-likelihood using numerical maximization algorithms [Durbin and Koopman, 2012]. The algorithms, however, do not guarantee that a global maximum of the likelihood is found. The likelihood surface of the state-space model may contain multiple local maxima, which leads to multiple roots of the score function. What can be said about any critical value found by MATLAB is that if the MLE $\hat{\xi}$ is chosen in a compact set which contains the true parameter $\xi$ in the interior, then $\hat{\xi}$ will be a consistent estimate of $\xi$. Furthermore, there will be a central limit theorem for such a critical value which allows the approximation $\hat{\xi} \sim N\{\xi, i(\xi)^{-1}\}$ for large $T$ [Lehmann and Casella, 1998].

Difficulties, however, arise for smaller values of $T$ such as $T = 60$ or $120$. In these “small sample” settings, the MLEs returned by MATLAB occurred on the boundary of the parameter space $[0, \infty)^2$ with either $\hat{\beta} = 0$ or $\hat{\delta} = 0$. Taking $\beta = 0.2$, $\delta = \sqrt{2}$ and $T = 60$, Figure 2.2 provides a scatter plot of 10000 pairs of $(\hat{\beta}, \hat{\delta})$ estimated from simulated datasets. The red asterisks denote $\hat{\beta}$ on the boundary 0 and those $\hat{\beta}$ also have large estimated standard error $\hat{se}(\hat{\beta}) \geq 30$. Overlaid on the plot is a 95% predictive ellipsoid for values of $(\hat{\beta}, \hat{\delta})$ based on the assumption of the Central Limit Theorem (CLT). The ellipse is

$$\{(\hat{\beta}, \hat{\delta}) : Q(\hat{\beta}, \hat{\delta}) = \chi^2_2(0.95)\} \quad (2.55)$$

where $\chi^2_2(0.95)$ is the 95th percentile of a $\chi^2_2$ distribution and

$$Q(\hat{\beta}, \hat{\delta}) = (\hat{\beta} - 0.2, \hat{\delta} - \sqrt{2})i(0.2, \sqrt{2})(\hat{\beta} - 0.2, \hat{\delta} - \sqrt{2})^T \quad (2.56)$$
Under the CLT, $Q(\hat{\beta}, \hat{\delta}) \sim \chi^2_2$ for large $T$. Clearly, for $T = 60$, the approximation of the $\chi^2_2$ distribution for $Q(\hat{\beta}, \hat{\delta})$ do not work well, given the overlap of the predictive interval outside the parameter space. Figure 2.3 shows the discrepancy between the large $\widehat{se}(\hat{\beta})$ values when $\hat{\beta} \approx 0$ and the other $\widehat{se}(\hat{\beta})$ values.

Settings in which $\hat{\beta} \approx 0$ and $\widehat{se}(\hat{\beta})$ is large reflect the simulation of datasets for which the data is uninformative about the parameter $\beta$, and central limit theory breaks down. We shall see that this phenomenon becomes less extreme for datasets with larger $T$. Thus, our construction of confidence intervals must account for cases in which $\hat{\beta} \approx 0$ and $\widehat{se}(\hat{\beta})$ is large. For such cases, we choose to assume that the confidence interval for $\beta$ is $(0, \infty)$, reflecting data that are uninformative about $\beta$. When computing coverage, we consider all four CI methods as covered.

The same problem occurs when simulating datasets with $\beta = 2.8$, $\delta = \sqrt{2}$ and $T = 60$. Figure 2.6 provides a scatter plot of the MLE with a 95% predictive ellipsoid overlaid. We see a similar situation in which $\hat{\delta} \approx 0$ and $\widehat{se}(\hat{\delta})$ is large. These anomalous datasets are uninformative about the value of $\delta$ in the chosen model. As in the setting with $\hat{\beta} \approx 0$, we choose $(0, \infty)$ as a confidence interval for $\beta$. The rationale for this choice is that the data do not support the chosen model and lacking such information, it makes no sense to attempt to limit the value of $\beta$ to a specific range. With no support for the chosen model, the designation of a finite range of any parameter seems unreasonable.

To our knowledge, this phenomenon of MLEs occurring on the boundary does not seem to have been discussed in the context of the state space model. As $T$ increases, datasets carry increasing information, and the frequencies of MLEs on the boundary diminish. However, our proposed confidence interval methods must propose an interval for $\beta$ even with anomalous datasets for which $\hat{\beta} \approx 0$ or $\hat{\delta} \approx 0$. So for all four confidence interval methods, we choose $(0, \infty)$ as a confidence interval for $\beta$ when $\hat{\beta} \approx 0$ or $\hat{\delta} \approx 0$. 

17
Next, Figure 2.3, Figure 2.4 and Figure 2.5 are the histograms of the estimated standard errors of MLEs when true $\beta = 0.2, T = 60$. Figure 2.7, Figure 2.8 and Figure 2.9 are the histograms of the estimated standard errors of MLEs when true $\beta = 2.8, T = 60$.

![A scatter plot of the MLEs of $\beta$ and $\delta$](image)

Figure 2.2: A scatter plot of MLEs overlaid with the 95% predictive interval, $T = 60, \beta = 0.2$

![Histogram of standard error of $\hat{\beta}$](image)

Figure 2.3: Histogram of standard error of $\hat{\beta}$, where $se(\hat{\beta}) < 30, T = 60, \beta = 0.2$.

Plots with $\beta = 0.2$ or 2.8 and $T = 120, 240, 480$ or 960 can be found in Appendix B.
Figure 2.4: Histogram of standard error of $\hat{\delta}$, where $se(\hat{\delta}) < 30, T = 60, \beta = 0.2$.

Figure 2.5: Histogram of standard error of $\hat{\delta}_0$, where $se(\hat{\delta}_0) < 30, T = 60, \beta = 0.2$. 
Figure 2.6: A scatter plot of MLEs overlaid with the 95% predictive interval, $T = 60, \beta = 2.8$.

Figure 2.7: Histogram of standard error of $\hat{\beta}$, where $se(\hat{\beta}) < 30$, $T = 60, \beta = 2.8$. 
Figure 2.8: Histogram of standard error of $\hat{\delta}$, where $se(\hat{\delta}) < 30$, $T = 60$, $\beta = 2.8$.

Figure 2.9: Histogram of standard error of $\hat{\delta}_0$, where $se(\hat{\delta}_0) < 30$, $T = 60$, $\beta = 2.8$
2.2.2. Simulation Procedure

We examine the performance of the four confidence interval (CI) methods in terms of underage, coverage, and overage. Underage (overage) occurs when the confidence interval misses the true $\beta$ and is to the left (right) of the true $\beta$ values. We consider simulations with sample size $T = \{60, 120, 240, 480, 960\}$. The number of simulated datasets is 10000 for $T = \{60, 120, 240, 480, 960\}$. The value of $\delta$ is fixed at $\sqrt{2}$. We take 10 evenly spaced $\beta$ values evenly spaced in the interval $[0.1, 3.0]$. We then plot the underage, coverage, and overage vs. true $\beta$ for all four CI methods, using significance level $\alpha = 0.05$. The general rule for underage, coverage, and overage is that if the CDF of the pivot is in the range $(\alpha/2, 1 - \alpha/2)$, then it is counted as coverage; if the CDF of the pivot is less than $\alpha/2$, then it is counted as underage; and if the CDF of the pivot is greater than $1 - \alpha/2$, then it is counted as overage. But in light of MLEs for $\hat{\beta}$ and $\hat{\delta}$ occurring on the boundary, this rule is modified so that if $\hat{\sigma}(\hat{\beta}) \geq c_1 = 30$ and $\hat{\sigma}(\hat{\delta}) \geq c_2 = 30$, then interval $(0, \infty)$ is used and $\beta$ is considered as covered for that dataset for all four confidence interval methods.

2.2.3. Evaluation of the Four Confidence Interval Methods

2.2.3.1. Simulation Results

For the local level model, we experimented with various cutoff values and found $c_1 = c_2 = 30$ is reasonably good at picking out noninformative datasets. Notice that these cutoffs assure good performance for the local level model and may not guarantee good performance for a general method.

Simulation results for the local level model are summarized in Table 2.1, Table 2.2, and Table 2.3.
Table 2.1: Underage, coverage, and overage simulation results for the local level model, $T = \{60, 120\}$

<table>
<thead>
<tr>
<th>$T$</th>
<th>$\beta$</th>
<th>Wald method</th>
<th>likelihood ratio method</th>
<th>score method</th>
<th>higher-order method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>underage</td>
<td>coverage</td>
<td>overage</td>
<td>underage</td>
</tr>
<tr>
<td>0.10</td>
<td>0.0077</td>
<td>0.9923</td>
<td>0.0000</td>
<td>0.0158</td>
<td>0.9842</td>
</tr>
<tr>
<td>0.42</td>
<td>0.0075</td>
<td>0.9116</td>
<td>0.0809</td>
<td>0.0193</td>
<td>0.9573</td>
</tr>
<tr>
<td>0.74</td>
<td>0.0088</td>
<td>0.9023</td>
<td>0.0889</td>
<td>0.0229</td>
<td>0.9370</td>
</tr>
<tr>
<td>1.07</td>
<td>0.0072</td>
<td>0.9165</td>
<td>0.0763</td>
<td>0.0183</td>
<td>0.9452</td>
</tr>
<tr>
<td>60</td>
<td>1.39</td>
<td>0.0075</td>
<td>0.9235</td>
<td>0.0690</td>
<td>0.0194</td>
</tr>
<tr>
<td></td>
<td>1.71</td>
<td>0.0043</td>
<td>0.9317</td>
<td>0.0640</td>
<td>0.0140</td>
</tr>
<tr>
<td></td>
<td>2.03</td>
<td>0.0026</td>
<td>0.9333</td>
<td>0.0641</td>
<td>0.0099</td>
</tr>
<tr>
<td></td>
<td>2.36</td>
<td>0.0010</td>
<td>0.9357</td>
<td>0.0633</td>
<td>0.0048</td>
</tr>
<tr>
<td></td>
<td>2.68</td>
<td>0.0011</td>
<td>0.9379</td>
<td>0.0610</td>
<td>0.0044</td>
</tr>
<tr>
<td>3.00</td>
<td>0.0004</td>
<td>0.9426</td>
<td>0.0570</td>
<td>0.0024</td>
<td>0.9644</td>
</tr>
<tr>
<td>120</td>
<td>1.39</td>
<td>0.0113</td>
<td>0.9306</td>
<td>0.0581</td>
<td>0.0217</td>
</tr>
<tr>
<td></td>
<td>1.71</td>
<td>0.0097</td>
<td>0.9397</td>
<td>0.0506</td>
<td>0.0194</td>
</tr>
<tr>
<td></td>
<td>2.03</td>
<td>0.0100</td>
<td>0.9359</td>
<td>0.0541</td>
<td>0.0215</td>
</tr>
<tr>
<td></td>
<td>2.36</td>
<td>0.0088</td>
<td>0.9439</td>
<td>0.0473</td>
<td>0.0172</td>
</tr>
<tr>
<td></td>
<td>2.68</td>
<td>0.0053</td>
<td>0.9488</td>
<td>0.0459</td>
<td>0.0124</td>
</tr>
<tr>
<td></td>
<td>3.00</td>
<td>0.0027</td>
<td>0.9517</td>
<td>0.0456</td>
<td>0.0072</td>
</tr>
</tbody>
</table>
Table 2.2: Underage, coverage, and overage simulation results for the local level model, $T = \{240, 480\}$

<table>
<thead>
<tr>
<th>$T$</th>
<th>$\beta$</th>
<th>Wald method</th>
<th>likelihood ratio method</th>
<th>score method</th>
<th>higher-order method</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>0.0068</td>
<td>0.9027</td>
<td>0.0905</td>
<td>0.0395</td>
<td>0.0464</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0188</td>
<td>0.9417</td>
<td>0.0319</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9481</td>
<td>0.0359</td>
</tr>
<tr>
<td></td>
<td>0.42</td>
<td>0.9293</td>
<td>0.0602</td>
<td>0.0228</td>
<td>0.9439</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9481</td>
<td>0.0359</td>
</tr>
<tr>
<td></td>
<td>0.74</td>
<td>0.9351</td>
<td>0.0531</td>
<td>0.0216</td>
<td>0.9488</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9481</td>
<td>0.0359</td>
</tr>
<tr>
<td></td>
<td>1.07</td>
<td>0.9392</td>
<td>0.0497</td>
<td>0.0216</td>
<td>0.9488</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9481</td>
<td>0.0359</td>
</tr>
<tr>
<td>240</td>
<td>1.39</td>
<td>0.9441</td>
<td>0.0419</td>
<td>0.0219</td>
<td>0.9501</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9469</td>
<td>0.0270</td>
</tr>
<tr>
<td></td>
<td>1.71</td>
<td>0.9426</td>
<td>0.0418</td>
<td>0.0261</td>
<td>0.9469</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9469</td>
<td>0.0270</td>
</tr>
<tr>
<td></td>
<td>2.03</td>
<td>0.9472</td>
<td>0.0405</td>
<td>0.0210</td>
<td>0.9509</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9469</td>
<td>0.0270</td>
</tr>
<tr>
<td></td>
<td>2.36</td>
<td>0.9448</td>
<td>0.0409</td>
<td>0.0236</td>
<td>0.9460</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9469</td>
<td>0.0270</td>
</tr>
<tr>
<td></td>
<td>2.68</td>
<td>0.9468</td>
<td>0.0397</td>
<td>0.0202</td>
<td>0.9529</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9469</td>
<td>0.0270</td>
</tr>
<tr>
<td></td>
<td>3.00</td>
<td>0.9517</td>
<td>0.0391</td>
<td>0.0168</td>
<td>0.9546</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9445</td>
<td>0.0353</td>
</tr>
<tr>
<td></td>
<td>0.10</td>
<td>0.0080</td>
<td>0.9124</td>
<td>0.0796</td>
<td>0.0202</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9445</td>
<td>0.0353</td>
</tr>
<tr>
<td></td>
<td>0.42</td>
<td>0.0133</td>
<td>0.9394</td>
<td>0.0493</td>
<td>0.0214</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9499</td>
<td>0.0287</td>
</tr>
<tr>
<td></td>
<td>0.74</td>
<td>0.0180</td>
<td>0.9409</td>
<td>0.0411</td>
<td>0.0257</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9476</td>
<td>0.0267</td>
</tr>
<tr>
<td></td>
<td>1.07</td>
<td>0.0155</td>
<td>0.9449</td>
<td>0.0396</td>
<td>0.0218</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9492</td>
<td>0.0290</td>
</tr>
<tr>
<td>480</td>
<td>1.39</td>
<td>0.9418</td>
<td>0.0410</td>
<td>0.0241</td>
<td>0.9460</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9460</td>
<td>0.0299</td>
</tr>
<tr>
<td></td>
<td>1.71</td>
<td>0.9425</td>
<td>0.0402</td>
<td>0.0248</td>
<td>0.9436</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9436</td>
<td>0.0316</td>
</tr>
<tr>
<td></td>
<td>2.03</td>
<td>0.9508</td>
<td>0.0335</td>
<td>0.0242</td>
<td>0.9505</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9505</td>
<td>0.0253</td>
</tr>
<tr>
<td></td>
<td>2.36</td>
<td>0.9436</td>
<td>0.0387</td>
<td>0.0244</td>
<td>0.9477</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9477</td>
<td>0.0279</td>
</tr>
<tr>
<td></td>
<td>2.68</td>
<td>0.9459</td>
<td>0.0359</td>
<td>0.0257</td>
<td>0.9465</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9465</td>
<td>0.0278</td>
</tr>
<tr>
<td></td>
<td>3.00</td>
<td>0.9449</td>
<td>0.0388</td>
<td>0.0232</td>
<td>0.9468</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.0200</td>
<td>0.9468</td>
<td>0.0300</td>
</tr>
<tr>
<td>$T$</td>
<td>$\beta$</td>
<td>Wald method</td>
<td>likelihood ratio method</td>
<td>score method</td>
<td>higher-order method</td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>-------------</td>
<td>-------------------------</td>
<td>-------------</td>
<td>-------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wald method</td>
<td>likelihood ratio method</td>
<td>score method</td>
<td>higher-order method</td>
</tr>
<tr>
<td>0.10</td>
<td>0.0116</td>
<td>0.9219</td>
<td>0.8355</td>
<td>0.0221</td>
<td>0.9520</td>
</tr>
<tr>
<td>0.42</td>
<td>0.0142</td>
<td>0.9474</td>
<td>0.0384</td>
<td>0.9820</td>
<td>0.9520</td>
</tr>
<tr>
<td>0.74</td>
<td>0.0154</td>
<td>0.9492</td>
<td>0.0344</td>
<td>0.9635</td>
<td>0.9520</td>
</tr>
<tr>
<td>1.07</td>
<td>0.0167</td>
<td>0.9452</td>
<td>0.0372</td>
<td>0.9672</td>
<td>0.9520</td>
</tr>
<tr>
<td>1.39</td>
<td>0.0170</td>
<td>0.9488</td>
<td>0.0384</td>
<td>0.9502</td>
<td>0.9520</td>
</tr>
<tr>
<td>1.71</td>
<td>0.0180</td>
<td>0.9486</td>
<td>0.0334</td>
<td>0.9513</td>
<td>0.9520</td>
</tr>
<tr>
<td>2.03</td>
<td>0.0188</td>
<td>0.9475</td>
<td>0.0337</td>
<td>0.9487</td>
<td>0.9520</td>
</tr>
<tr>
<td>2.36</td>
<td>0.0188</td>
<td>0.9509</td>
<td>0.0308</td>
<td>0.9328</td>
<td>0.9520</td>
</tr>
<tr>
<td>2.68</td>
<td>0.0188</td>
<td>0.9502</td>
<td>0.0308</td>
<td>0.9328</td>
<td>0.9520</td>
</tr>
<tr>
<td>3.00</td>
<td>0.0182</td>
<td>0.9515</td>
<td>0.0327</td>
<td>0.9523</td>
<td>0.9520</td>
</tr>
</tbody>
</table>
We plot the results for easy visualization and interpretation. Figure 2.10, Figure 2.11, and Figure 2.12 plot the underage, coverage, and overage versus true $\beta$ for the four CI methods when each of the 10000 simulated sample datasets has 60 data points. From Figure 2.10, we observe that for small $\beta \leq 0.2$, all four methods have low underage. When $0.7 < \beta < 3$, the second order method is right on target (0.025) for underage, whereas the likelihood ratio method has slightly higher underage; the underage of Wald method is too high (around 0.07) and the underage of the score method is too low (almost 0). Moreover, Figure 2.12 shows that for small $\beta < 0.4$, the overage from the second order method and the score method is above the target overage (0.025) and those from the likelihood ratio method and the Wald method are below the target. For a brief period when $0.5 \leq \beta < 1.5$, overage from the second order method and the likelihood ratio method is close to the target (0.025). When $\beta \geq 2.2$, all four methods have overage that is below the target, with the score method and the second-order method being slightly superior to the other two first-order methods.

Figure 2.11 shows that when $\beta < 1.7$, the second order method, the score method, and the likelihood ratio method have coverage that is close to the target (0.95), with the second order method having slightly superior performance. When $\beta \geq 2$, coverage from the second order method, the score method, and the likelihood ratio method start to drift above the target. In contrast, the Wald method has coverage that is too low in general.

Figure 2.13, Figure 2.14, and Figure 2.15 show the underage, coverage, and overage versus true $\beta$ for the four CI methods when each of the 10000 simulated sample datasets has 120 data points. We observe significant improvement in underage, coverage, and overage when the number of data points $T$ increases. From Figure 2.13, we observe that the second order method is right on target (0.025) for underage, whereas the likelihood ratio method has underage that is slightly higher than target. In contrast, the score method has underage that is too low and the Wald method has underage that is too high in general. Figure 2.14 shows that when $\beta < 2.4$, the second order method, the score method, and the likelihood ratio method have coverage close to the target (0.95). When $\beta \geq 2.4$, coverage from the
second order method, the score method, and the likelihood ratio method drifts above the
target. In contrast, the Wald method has coverage that is too low in general. Moreover,
Figure 2.15 shows that for small $\beta < 0.2$, the overage from the second order method and
the score method is above the target overage (0.025), and those from the likelihood ratio
method and the Wald method is below the target. When $0.2 < \beta < 2.0$, overage from the
second order method and the likelihood ratio method is close to the target (0.025). When
$\beta > 2.4$, all four methods have overage below the target.

Figure 2.16, Figure 2.17, and Figure 2.18 plot the underage, coverage, and overage versus
true $\beta$ for the four CI methods when each of the 10000 simulated sample datasets has 240 data
points. Figure 2.19, Figure 2.20, and Figure 2.21 plot the underage, coverage, and overage
versus true $\beta$ for the four CI methods when each of the 10000 simulated sample datasets
has 480 data points. Figure 2.22, Figure 2.23, and Figure 2.24 plot the underage, coverage,
and overage versus true $\beta$ for the four CI methods when each of the 10000 simulated sample
datasets has 960 data points. We observe significant improvement in underage, coverage,
and overage when the number of data points $T$ increases. From Figure 2.16, Figure 2.19
and Figure 2.22, we observe that the second order method is right on target (0.025) for
underage, whereas the likelihood ratio method has underage that is slightly higher than
target. In contrast, the score method has underage that is too low; and the Wald method
has underage that is too high in general. Figure 2.17, Figure 2.20 and Figure 2.23 show that
the second-order method, the score method, and the likelihood ratio method have coverage
close to the target (0.95). In contrast, the Wald method has coverage that is too low in
general. Moreover, Figure 2.18, Figure 2.21, and Figure 2.24 show that overage from the
second-order method and the likelihood ratio method is close to the target (0.025). The
overage from the score method is too high. In contrast, the Wald method has overage that
is too low in general. We also observe that as $T$ increases, coverage from all four methods
converges to the target (0.95).
Figure 2.10: Underage when datasets have 60 data points and $c_1 = c_2 = 30$

Figure 2.11: Coverage when datasets have 60 data points and $c_1 = c_2 = 30$
Figure 2.12: Overage when datasets have 60 data points and $c_1 = c_2 = 30$

Figure 2.13: Underage when datasets have 120 data points and $c_1 = c_2 = 30$
Figure 2.14: Coverage when datasets have 120 data points and $c_1 = c_2 = 30$

Figure 2.15: Overage when datasets have 120 data points and $c_1 = c_2 = 30$
Figure 2.16: Underage when datasets have 240 data points and $c_1 = c_2 = 30$

Figure 2.17: Coverage when datasets have 240 data points and $c_1 = c_2 = 30$
Figure 2.18: Overage when datasets have 240 data points and $c_1 = c_2 = 30$

Figure 2.19: Underage when datasets have 480 data points and $c_1 = c_2 = 30$
Figure 2.20: Coverage when datasets have 480 data points and $c_1 = c_2 = 30$

Figure 2.21: Overage when datasets have 480 data points and $c_1 = c_2 = 30$
Figure 2.22: Underage when datasets have 960 data points and $c_1 = c_2 = 30$

Figure 2.23: Coverage when datasets have 960 data points and $c_1 = c_2 = 30$
2.3. Kalman Recursions

Recall that the state space model is

\[ x_t = \alpha x_{t-1} + \beta u_t \]
\[ y_t = \gamma x_t + \delta \epsilon_t \quad t = 1, \ldots, T, \]

with \( u_t \sim N(0, 1) \), \( \epsilon_t \sim N(0, 1) \), and \( \{u_t\} \) and \( \{\epsilon_t\} \) are all mutually independent.

In this section, we provide the recursive computation for the log-likelihood, the score function, the observed information and the Fisher information.

2.3.1. Recursive Log-likelihood Computation

Log-likelihood computation for the state space model relies on the following well-known sequential characterization of the distribution of \( y_1, y_2, \ldots, y_t \). Initially, \( y_1 \sim N(m_1, g_1) \). For \( t = 2, \ldots, T \), the conditional distribution of \( y_t \) given \( \mathcal{F}_{t-1} = (y_1, y_2, \ldots, y_{t-1}) \) is \( N(m_t, g_t) \), where the conditional mean \( m_t = E(y_t|\mathcal{F}_{t-1}) \) depends on \( \mathcal{F}_{t-1} \) and the corresponding condi-
tional variance $g_t = \text{Var}(y_t|\mathcal{F}_{t-1})$ is independent of $\mathcal{F}_{t-1}$. With $l_0(\theta) = 0$, the log-likelihood $l_t(\theta)$ can be computed recursively as

$$l_t(\theta) = l_{t-1}(\theta) - \frac{1}{2} \left\{ \ln g_t + \frac{(y_t - m_t)^2}{g_t} \right\} \quad t = 1, 2, \ldots, T. \quad (2.58)$$

Here, $m_t$ and $g_t$ are computed recursively via

$$m_t = \gamma a_{t-1} \quad (2.59)$$

$$g_t = \gamma^2 r_t + \delta^2 \quad t = 1, 2, \ldots, T, \quad (2.60)$$

where $r_t$ and $a_t$ are computed recursively using Equation 2.68, Equation 2.69 and Equation 2.70.

The mechanism of the recursive computation of the log-likelihood is summarized in Figure 2.25.

![Figure 2.25: Recursive computation of the log-likelihood](image)

Assume the initial state $x_0 \sim N(a_0, p_0)$, where $x_0$ is unobserved, and $a_0, p_0$ are known values for mean and variance of $x_0$. Then, $x_1 \sim N(\alpha a_0, r_1)$ and Bayes update gives $x_1|y_1 \sim N(a_1, p_1)$. Next, $x_2|y_1 \sim N(\alpha a_1, r_2)$ and another Bayes update leads to $x_2|\mathcal{F}_2 = y_1, y_2 \sim N(a_2, p_2)$. Continuing the updates in similar fashion and we get $x_t|\mathcal{F}_{t-1} \sim N(\alpha a_{t-1}, r_t)$ and $x_t|\mathcal{F}_t \sim N(a_t, p_t)$. This is the recursion in the first line of Figure 2.25. From $a_{t-1}$, we compute $m_t$ via Equation 2.59. Using $r_t$, we get $g_t$ via Equation 2.60. The second line of Figure 2.25
shows this relationship. Together with \( y_t \), now we have all the ingredients to compute \( l_t(\theta) \) recursively, as shown in line 3 of Figure 2.25.

2.3.1.1. Details for the Recursion of the Log-likelihood

We now derive the recursion for computing \( l_t(\theta) \). With \( x_0 \) following the known distribution \( N(a_0, p_0) \), then \( x_1|x_0 \sim N(\alpha x_0, \beta^2) \), and \( x_1 \sim N(\alpha a_0, \alpha^2 p_0 + \beta^2) = N(\alpha a_0, r_1) \). A conjugate Bayes update gives \( x_1|y_1 \sim N(a_1, p_1) \), where we specify \( a_1 \) and \( p_1 \) as

\[
\begin{align*}
\alpha^2 p_0 + \beta^2 & \quad (2.61) \\
\alpha a_0 + \frac{\gamma r_1}{\delta^2 + \gamma^2 r_1}(y_1 - \alpha \gamma a_0) & \quad (2.62) \\
\frac{\delta^2 r_1}{\delta^2 + \gamma^2 r_1} & \quad (2.63)
\end{align*}
\]

Since \( x_1|y_1 \sim N(a_1, p_1) \), then

\[
x_2|y_1 \sim N(\alpha a_1, \alpha^2 p_1 + \beta^2) \sim N(\alpha a_1, r_2)
\]

where

\[
r_2 = \alpha^2 p_1 + \beta^2. \quad (2.64)
\]

Next,

\[
y_2|x_2 \sim N(\gamma x_2, \delta^2) \quad (2.65)
\]

Another conjugate Bayes update gives \( x_2|(y_1, y_2) \sim N(a_2, p_2) \), where we specify \( a_2 \) and \( p_2 \) as

\[
\begin{align*}
\alpha a_1 + \frac{\gamma r_2}{\delta^2 + \gamma^2 r_2}(y_2 - \alpha \gamma a_1) & \quad (2.66) \\
\frac{\delta^2 r_2}{\delta^2 + \gamma^2 r_2} & \quad (2.67)
\end{align*}
\]
More generally, $x_t | \mathcal{F}_t \sim N(a_t, p_t)$ for $t = 1, \ldots, T$, where $(a_t, p_t)$ are computed recursively as follows:

$$r_t = \alpha^2 p_{t-1} + \beta^2$$  \hspace{1cm} (2.68)
$$a_t = \alpha a_{t-1} + \frac{\gamma r_t}{\delta^2 + \gamma^2 r_t} (y_t - \alpha \gamma a_{t-1}) = \frac{\alpha \delta^2 a_{t-1} + \gamma r_t y_t}{\delta^2 + \gamma^2 r_t}$$  \hspace{1cm} (2.69)
$$p_t = \frac{\delta^2 r_t}{\delta^2 + \gamma^2 r_t}$$  for $t = 1, \ldots, T$.  \hspace{1cm} (2.70)

This recursion is depicted in the first line of Figure 2.25. Once the recursion for $\{r_t\}, \{a_t\}$ and $\{p_t\}$ are ready, we compute $\{m_t\}$ and $\{g_t\}$ via Equation 2.59 and Equation 2.60. Then, starting from $l_0(\theta) = 0$, we compute $l_t(\theta)$ recursively using Equation 2.58.

2.3.2. Recursive Computation of the Score Function

We now write the vector score function $\dot{l}_t(\theta) = \partial l_t(\theta)/\partial \theta$ as a recursion, restricting $\theta$ to the 2-dimensional parameters $\theta = (\beta, \delta)$. Define $\dot{l}_0(\theta) = 0_{2 \times 1}$. Using the product rule and chain rule for differentiation, the $(2 \times 1)$ vector recursion of $\dot{l}_t(\theta)$ is

$$\dot{l}_t(\theta) = \dot{l}_{t-1}(\theta) - \frac{1}{2 y_t} \dot{g}_t \left\{ 1 - \frac{(y_t - m_t)^2}{g_t} \right\} + \frac{(y_t - m_t)}{g_t} \dot{m}_t \quad t = 1, \ldots, T,$$  \hspace{1cm} (2.71)

where $\dot{g}_t = \partial g_t / \partial \theta$ and $\dot{m}_t = \partial m_t / \partial \theta$ are $(2 \times 1)$ vectors. Besides the recursion for $\{m_t\}$ and $\{g_t\}$, the expression of $\dot{l}_t(\theta)$ requires recursive computation of $\{\dot{g}_t\}$ and $\{\dot{m}_t\}$, which depend on the recursion for $\{\dot{r}_t\}$, $\{\dot{a}_t\}$ and $\{\dot{p}_t\}$, as shown in Equation 2.74, Equation 2.75 and Equation 2.77.

The flow chart for the recursive computation of the score vectors is summarized in Figure 2.26.
The first line in Figure 2.26 shows the recursion for \( \{r_t\}, \{a_t\} \) and \( \{p_t\} \), from which we can compute \( \{\dot{r}_t\}, \{\dot{a}_t\} \) and \( \{\dot{p}_t\} \) recursively, as shown in the second line in Figure 2.26. Furthermore, \( \dot{m}_t \) is a function of \( \dot{a}_{t-1} \), and \( \dot{g}_t \) is a function of \( \dot{r}_t \). Starting with \( \dot{l}_0(\theta) = 0_{2 \times 1} \), now we have all the ingredients to compute \( \dot{l}_t(\theta) \) recursively, as shown in line 4 of Figure 2.26.

### 2.3.2.1. Details for the Recursion of the Score Vectors

We now derive the recursion for computing \( \dot{l}_t(\theta) \). Since \( m_t = \gamma a_{t-1} \), then

\[
\dot{m}_t = \frac{\partial m_t}{\partial \theta} = \gamma \dot{a}_{t-1},
\]

and \( \dot{m}_t \) is a function of \( \dot{a}_{t-1} \).

Using \( g_t = \gamma^2 r_t + \delta^2 \), we get

\[
\dot{g}_t = \frac{\partial g_t}{\partial \theta} = \gamma^2 \dot{r}_t + 2\delta \xi_2,
\]

where \( \dot{r}_t = \partial r_t / \partial \theta \) and \( \xi_2 = (0, 1)^T \). This shows \( \dot{g}_t \) is a function of \( \dot{r}_t \).
We now update the Kalman recursion to include the $\dot{r}_t, \dot{a}_t$ and $\dot{p}_t$. Recall the recursive expressions for $r_t, a_t$ and $p_t$ given in Equation 2.68, Equation 2.69 and Equation 2.70. Because $a_0$ and $p_0$ are known constants, $\dot{a}_0 = 0_{2 \times 1} = \dot{p}_0$. Denote $\xi_1 = (1, 0)^T$. Differentiating $r_t$ gives

$$
\dot{r}_t = \alpha^2 \dot{p}_{t-1} + 2\beta \xi_1 \tag{2.74}
$$

This recursion is shown in the second row of Figure 2.26.

Differentiating $a_t$ gives

$$
\dot{a}_t = \frac{\partial a_t}{\partial \theta} = \frac{1}{\delta^2 + r_t \gamma^2} \{2\alpha \delta a_{t-1} \xi_2 + \alpha \delta^2 a_{t-1} + \gamma y_t \dot{r}_t \} - \frac{\alpha a_{t-1} \delta^2 + r_t \gamma y_t}{(\delta^2 + r_t \gamma^2)^2} \{2\delta \xi_2 + \gamma^2 \dot{r}_t \}
$$

$$
= d_t \{2\alpha \delta a_{t-1} \xi_2 + \alpha \delta^2 a_{t-1} + \gamma y_t \dot{r}_t \} - d_t^2 (\alpha a_{t-1} \delta^2 + r_t \gamma y_t) \{2\delta \xi_2 + \gamma^2 \dot{r}_t \} \tag{2.75}
$$

where

$$
d_t = \frac{1}{\delta^2 + r_t \gamma^2}. \tag{2.76}
$$

Thus, Equation 2.75 shows that $\dot{a}_t$ is a function of $r_t, \dot{r}_t, a_{t-1}$ and $\dot{a}_{t-1}$. This is seen in the update of $\dot{a}_t$ in the second row of Figure 2.26.

Differentiating $p_t$ gives

$$
\dot{p}_t = d_t \{\delta^2 \dot{r}_t + 2\delta r_t \xi_2 \} - r_t \delta^2 d_t^2 e_t \tag{2.77}
$$

where

$$
e_t = 2\delta \xi_2 + \gamma^2 \dot{r}_t \tag{2.78}
$$

Equation 2.77 shows that $\dot{p}_t$ is a function of $r_t$ and $\dot{r}_t$. From the recursion of $\{g_t\}, \{m_t\}, \{\dot{g}_t\}$ and $\{\dot{m}_t\}$, now we can compute $\dot{\theta}_t(\theta)$ using Equation 2.71.
2.3.3. Recursive Computation of the Observed Information $j_t(\theta)$

Taking the derivative of the $2 \times 1$ vector $\dot{l}_t(\theta)$ with respect to $\theta^T$, we get the $2 \times 2$ matrix $\ddot{l}_t(\theta) = -j_t(\theta)$. Using the product rule and the chain rule, and starting with $j_0(\theta) = 0_{2 \times 2}$, we get

\[
\begin{align*}
\dot{j}_t(\theta) &= j_{t-1}(\theta) + \frac{1}{2g_t} \left\{ \frac{(y_t - m_t)^2}{g_t} \right\} \dot{g}_t + \frac{1}{g_t} \dot{m}_t m_t^T \\
&\quad - \frac{1}{g_t^2} \dot{g}_t \dot{g}_t^T \left\{ \frac{1}{2} \frac{(y_t - m_t)^2}{g_t} - \frac{(y_t - m_t)}{g_t} \right\} \{ g_t \ddot{m}_t - \dot{g}_t m_t^T - \dot{m}_t \dot{g}_t^T \} \tag{2.79}
\end{align*}
\]

for $t = 1, \ldots, T$.

The recursions for $\ddot{m}_t$ and $\ddot{g}_t$ are in Equation 2.72 and Equation 2.73. We will derive the recursive computation of $\{\ddot{m}_t\}$ and $\{\ddot{g}_t\}$ as $2 \times 2$ matrices in Section 2.3.3.1.

The recursive computation of the observed information $j_t(\theta)$ is summarized in Figure 2.27.

![Figure 2.27: Recursive computation of the observed information $j_t(\theta)$](image)

With $\ddot{a}_0 = 0_{2 \times 1} = \ddot{p}_0$, we get $\ddot{a}_0 = 0_{2 \times 2} = \ddot{p}_0$, from which we compute $\ddot{r}_1$ via Equation 2.83. The recursion for computing $\ddot{p}_t$ from $r_t, \dot{r}_t$ and $\ddot{r}_1$ is given in Equation 2.84. The recursion of $\ddot{r}_t$ from $\ddot{p}_{t-1}$ is in Equation 2.83. The recursion of $\ddot{a}_t$ is given in Equation 2.81. It shows that $\ddot{a}_t$ is a function of $r_t, \dot{r}_t, \ddot{r}_t, a_{t-1}, \ddot{a}_{t-1}$ and $\ddot{a}_{t-1}$. These recursion are shown in the third row of
Figure 2.27. Using Equation 2.80 and Equation 2.82, we compute \( \{\ddot{m}_t\} \) and \( \{\ddot{g}_t\} \). Finally, starting with \( j_0(\theta) = 0_{2 \times 2} \), we are ready to compute \( j_t(\theta) \) recursively using Equation 2.79. This is shown on line four of Figure 2.27.

2.3.3.1. Details for the Recursion of the Observed Information

In this section, we provide details for the recursion of the observed information \( j_t(\theta) \). The recursive expression for \( j_t(\theta) \) is given in Equation 2.79. Since \( \dot{m}_t = \gamma \alpha \dot{a}_{t-1} \), we compute \( \partial \ddot{m}_t / \partial \theta^T \) as

\[
\ddot{m}_t = \gamma \alpha \ddot{a}_{t-1}
\]

(2.80)

where we need the recursion for \( \{\ddot{a}_t\} \). Differentiating the recursion for \( \{\dot{a}_t\} \) leads to \( \partial \dot{a}_t / \partial \theta^T \) as

\[
\begin{align*}
\dot{a}_t &= d_t \left\{ \alpha \delta \ddot{a}_{t-1} + 2 \alpha a_{t-1} \xi_2 \xi_2^T + 2 \alpha \delta (\dot{a}_{t-1} \xi_2^T + \xi_2 \dot{a}_{t-1}^T) + \gamma y_t \ddot{r}_t \right\} \\
&\quad - d_t^2 [(\alpha a_{t-1} \delta^2 + \gamma \dot{r}_t y_t)(2 \xi_2 \xi_2^T + \gamma^2 \ddot{r}_t) + \alpha \delta^2 \{\dot{a}_{t-1} e_t^T + e_t \dot{a}_{t-1}^T\} \\
&\quad + 2 \alpha \delta a_{t-1} \{\xi_2 e_t^T + e_t \xi_2^T\} + \gamma y_t \{\dot{r}_t e_t^T + e_t \dot{r}_t^T\}] + 2(\alpha a_{t-1} \delta^2 + \gamma r_t y_t) d_t^3 e_t e_t^T \\
&= 2 \alpha a_{t-1} \{\xi_2 e_t^T + e_t \xi_2^T\} + \gamma y_t \{\dot{r}_t e_t^T + e_t \dot{r}_t^T\}
\end{align*}
\]

(2.81)

Equation 2.81 shows that \( \dot{a}_t \) is a function of \( r_t, \dot{r}_t, \ddot{r}_t, a_{t-1}, \dot{a}_{t-1} \) and \( \ddot{a}_{t-1} \).

Using the recursion for \( \dot{g}_t \) in Equation 2.73, we compute \( \partial \dot{g}_t / \partial \theta^T \) as

\[
\dot{g}_t = \gamma^2 \ddot{r}_t + 2 \xi_2 \xi_2^T
\]

(2.82)

where

\[
\ddot{r}_t = \alpha^2 \ddot{p}_{t-1} + 2 \xi_1 \xi_1^T
\]

(2.83)

Equation 2.82 and 2.83 show that \( \dot{g}_t \) is a function of \( \ddot{r}_t \), and \( \dot{r}_t \) is a function of \( \ddot{p}_{t-1} \).
Next, we need the recursion for \( \{ \dot{p}_t \} \). From the recursion for \( \{ \ddot{p}_t \} \), we can derive that

\[
\ddot{p}_t = \frac{\partial \dot{p}_t}{\partial \theta} = d_t \left\{ \delta^2 \ddot{r}_t + 2 r_t \xi_2 \xi_2^T + 2 \delta (\dot{r}_t \xi_2^T + \xi_2 \dot{r}_t^T) \right\}
- d_t^2 \left[ \delta^2 \{ \dot{r}_t \xi_1^T + e_t \dot{r}_t^T \} + 2 \delta r_t \{ \xi_2 \xi_1^T + e_t \xi_2^T \} \right]
+ 2 \delta^2 r_t d_t^3 \xi_1 e_t^T - 2 \delta^2 r_t d_t^2 \xi_2 \xi_2^T - \delta^2 \gamma^2 r_t d_t^2 \ddot{r}_t
\]  

(2.84)

Equation 2.84 shows that \( \dot{p}_t \) is a function of \( r_t, \dot{r}_t \) and \( \ddot{r}_t \).

2.3.4. Recursive Computation of the Fisher Information

We take the expected value of the recursions for \( \{ j_t(\theta) \} \) to get recursions for the Fisher information matrix \( \{ i_t(\theta) \} \). That is, \( i_0(\theta) = 0_{2 \times 2} \) and for \( t = 1, \ldots, T \),

\[
i_t(\theta) = i_{t-1}(\theta) + \frac{\gamma^2 \alpha^2}{g_t} E(\dot{a}_{t-1} \dot{a}_{t-1}^T) + \frac{1}{2g_t^2 \dot{g}_t \dot{g}_t^T} \]  

(2.85)

where the details of the recursion for \( E(\dot{a}_t \dot{a}_t^T) \) is shown in Equation 2.93.

The recursive computation of the Fisher information \( i_t(\theta) \) is summarized in Figure 2.28.

![Diagram of recursive computation of the Fisher information](image)

Figure 2.28: Recursive computation of the Fisher information \( i_t(\theta) \)
The recursion of $E(\dot{a}_t \dot{a}_t^T)$ in the third row of Figure 2.28 depends on its previous value $E(\dot{a}_{t-1} \dot{a}_{t-1}^T)$, $\{r_t\}$, $\{\dot{r}_t\}$ and $\{g_t\}$. This recursion is shown in the second row of Figure 2.28. With $E(\dot{a}_{t-1} \dot{a}_{t-1}^T)$, $g_t$ and $\dot{g}_t$, we compute $i_t(\theta)$ recursively using Equation 2.85. This recursion is depicted in the fourth row of Figure 2.28.

2.3.4.1. Details for the Recursion of the Fisher Information $i_t(\theta)$

In this subsection, we provide details for the recursion of the Fisher information $i_t(\theta)$, as shown in Equation 2.85. First note that

$$i_t(\theta) = -E\left\{\ddot{i}_t(\theta)\right\} = -E^{F_{t-1}} E\left\{\ddot{i}_t(\theta) | F_{t-1}\right\} = E^{F_{t-1}} E\left\{j_t(\theta) | F_{t-1}\right\}.$$ (2.86)

Moreover, note that $\{g_t\}$, $\{r_t\}$ and $\{p_t\}$ do not depend on $F_{t-1}$, but $\{m_t\}$ and $\{\dot{m}_t\}$ depend on $F_{t-1}$, while $\{a_t\}$ depend on $F_{t-1}$ and $y_t - m_t$. Thus the conditional expectation

$$E\left\{\left(\frac{y_t - m_t}{g_t}\right)^2 | F_{t-1}\right\} = 1$$ (2.87)

since $E(y_t | F_{t-1}) = m_t$ and $Var(y_t | F_{t-1}) = g_t$.

Applying the conditional expectation to the recursion for $\{j_t(\theta)\}$ in Equation 2.79, the second the fifth terms have expected values zero and the the fourth term simplifies. This leads to

$$E\left\{j_t(\theta) | F_{t-1}\right\} = E\left\{j_{t-1}(\theta) | F_{t-1}\right\} + \frac{1}{g_t} E(\dot{m}_t \dot{m}_t^T | F_{t-1}) + \frac{1}{2g_t^2} \dot{g}_t \dot{g}_t^T$$ (2.88)

Now we need a recursions for $E(\dot{m}_t \dot{m}_t^T | F_{t-1})$. Recall that $\dot{m}_t \dot{m}_t^T = \gamma^2 \alpha^2 \dot{a}_{t-1} \dot{a}_{t-1}^T$. Then,

$$E(\dot{m}_t \dot{m}_t^T | F_{t-1}) = \gamma^2 \alpha^2 E(\dot{a}_{t-1} \dot{a}_{t-1}^T | F_{t-1})$$ (2.89)
where

\[ a_t = \alpha a_{t-1} + \frac{r_t \gamma}{\delta^2 + r_t \gamma^2} (y_t - \alpha \gamma a_{t-1}) \tag{2.90} \]

\[ \dot{a}_t = \frac{\alpha \delta^2}{\delta^2 + r_t \gamma^2} \dot{a}_{t-1} + \frac{\partial}{\partial \theta} \left( \frac{r_t \gamma}{\delta^2 + r_t \gamma^2} \right) (y_t - \alpha \gamma a_{t-1}) \]

\[ = \frac{\alpha \delta^2}{\delta^2 + r_t \gamma^2} \dot{a}_{t-1} + c_t (y_t - \alpha \gamma a_{t-1}) \tag{2.91} \]

with

\[ c_t = \frac{\partial}{\partial \theta} \left( \frac{r_t \gamma}{\delta^2 + r_t \gamma^2} \right) = d_t \gamma \dot{r}_t - \gamma r_t d_t^2 e_t \tag{2.92} \]

From the recursions of \( \{\dot{a}_t\} \), we compute the recursions for \( E(\dot{a}_t \dot{a}_t^T) \).

\[ E(\dot{a}_t \dot{a}_t^T) = \mathbb{E}^{F_{t-1}} E(\dot{a}_t \dot{a}_t^T | F_{t-1}) = \left( \frac{\alpha \delta^2}{\delta^2 + r_t \gamma^2} \right)^2 \mathbb{E}^{F_{t-1}} E(\dot{a}_{t-1} \dot{a}_{t-1}^T | F_{t-1}) \]

\[ + c_t c_t^T \mathbb{E}^{F_{t-1}} E \left\{ (y_t - \alpha \gamma a_{t-1})^2 | F_{t-1} \right\} + \text{cross terms} \]

\[ = \alpha^2 \delta^4 d_t^2 E(\dot{a}_{t-1} \dot{a}_{t-1}^T) + g_t c_t c_t^T \tag{2.93} \]

where

\[ \text{cross terms} = \frac{\alpha \delta^2}{\delta^2 + r_t \gamma^2} \left[ c_t^T E \left\{ \dot{a}_{t-1} (y_t - \alpha \gamma a_{t-1}) \right\} + c_t E \left\{ (y_t - \alpha \gamma a_{t-1}) \dot{a}_{t-1}^T \right\} \right] = 0 \]

Equation 2.93 shows that \( E(\dot{a}_t \dot{a}_t^T) \) is a function of its previous value and \( c_t \leftrightarrow (r_t, \dot{r}_t) \) and \( g_t \).

To see why the cross terms equal to 0, note that

\[ E \left\{ \dot{a}_{t-1} (y_t - \alpha \gamma a_{t-1}) | F_{t-1} \right\} = \dot{a}_{t-1} E \left\{ (y_t - \alpha \gamma a_{t-1}) | F_{t-1} \right\} = 0. \tag{2.94} \]

This is because \( \{\dot{a}_{t-1}\} \) are functions of \( F_{t-1} \), and \( E \left\{ (y_t - \alpha \gamma a_{t-1}) | F_{t-1} \right\} = 0. \)
2.3.5. A Summary of the Recursive Computations

Our results from the multivariate normal approach match exactly with those from the recursive computations for the log-likelihood, the score vectors, the observed information and the Fisher information. This not only confirms that computations from both approaches were performed correctly, but also offers an alternative way of performing simulation using the recursive approach, which is advantageous when the dataset is large and the MVN approach might be too slow or unstable in inverting large matrices.

We also attempted to derive and compute the quantities involved in the second-order method recursively. However, we eventually decided to abandon going that direction due to the complexities of the mathematical derivation.

We recommend using the likelihood ratio method with recursive computations when the dataset is very large. This is because the likelihood ratio method has comparable performance with the second-order method for large datasets, and it is much simpler than the second-order method.
The General Time-invariant State Space Model

The most general linear, multivariate, time-invariant Gaussian state space model uses the following system of equations [MathWorks, 2023c]

**state equation** \[ x_t = Ax_{t-1} + Bu_t, \]  
**observation equation** \[ y_t = Cx_t + D\epsilon_t, \quad t = 1, \ldots, T, \]

where \( x_t = (x_{t1}, \ldots, x_{tm})^T \) is an \( m \)-dimensional state vector at time \( t \). The initial state distribution \( x_0 \) is Gaussian with mean \( a_0 \) and covariance matrix \( P_0 \). That is, \( x_0 \sim N(a_0, P_0) \).

\( y_t = (y_{t1}, \ldots, y_{tn})^T \) is an \( n \)-dimensional observation vector at time \( t \). Matrix \( A \) is the \( m \times m \) state-transition matrix. Matrix \( B \) is the \( m \times k \) state-disturbance-loading matrix. Matrix \( C \) is the \( n \times m \) measurement-sensitivity matrix. Matrix \( D \) is the \( n \times h \) observation-innovation matrix. The matrices \( A, B, C, \) and \( D \) are called coefficient matrices. Error vector \( u_t = (u_{t1}, \ldots, u_{tk})^T \) is a \( k \)-dimensional Gaussian white-noise unit-variance vector of state disturbances at time \( t \). That is, \( u_t \sim N(0, I_{k \times k}) \). \( \epsilon_t = (\epsilon_{t1}, \ldots, \epsilon_{th})^T \) is an \( h \)-dimensional Gaussian white-noise unit-variance vector of observation innovations at time \( t \). That is, \( \epsilon_t \sim N(0, I_{h \times h}) \). The errors \( u_t \) and \( \epsilon_t \) are uncorrelated.

We can derive that

\[ y_t = Cx_t + D\epsilon_t \]
\[ = C \left( A^t x_0 + \Sigma^t_{j=1} A^{t-j} Bu_j \right) + D\epsilon_t, \quad t = 1, \ldots, T. \]
and \( y_t \) has an \( n \times 1 \) multivariate normal distribution with

\[
E(y_t) = CA_t^t a_0 \tag{3.4}
\]

\[
\text{cov}(y_t) = C \left\{ A_t^t P_0(A_t^t)^T + \sum_{k=1}^{t} A_t^{t-k} B B^T (A_t^{t-k})^T \right\} C^T + DD^T. \tag{3.5}
\]

Let \( y^T = (y_1^T, \ldots, y_T^T) \) be \( 1 \times nT \). Applying standard results from the multivariate normal theory, we can derive that

\[
y \sim N_{nT}(\mu, \Omega), \tag{3.6}
\]

where \( N_{nT} \) denotes a multivariate normal distribution of dimension \( nT \), and

\[
\mu = \text{vec}\{CAa_0, CA^2a_0, \ldots, CA^T a_0\} \tag{3.7}
\]

The \( nT \times nT \) block matrix \( \Omega = \{\Omega_{ij} = \text{cov}(y_i, y_j) : i, j = 1, \ldots, T.\} \) has \( T \times T \) subblocks of \( \Omega_{ij} \) with the form

\[
\Omega_{ij} = \begin{cases} 
C \left[ A_i^t P_0(A_i^t)^T + \sum_{k=1}^{\min(i,j)} A_{i}^{\min(i,j)-k} B B^T (A_{i}^{\min(i,j)-k})^T \right] C^T, & i \neq j \\
C \left\{ A_i^t P_0(A_i^t)^T + \sum_{k=1}^{i} A_i^{i-k} B B^T (A_i^{i-k})^T \right\} C^T + DD^T, & i = j
\end{cases} \tag{3.8}
\]

Without loss of generality, we assume \( a_0 = 0 \). This is possible because we can always substract \( \mu \) from \( y \) and center the distribution of \( y \) around \( 0 \). Then \( y \sim N_{nT}(0, \Omega) \).

Let \( \xi = (\xi_1, \cdots \xi_k) \). The log likelihood function is

\[
l(\xi) = c - \frac{1}{2} \log |\Omega| - \frac{1}{2} y^T \Omega^{-1} y = c - \frac{1}{2} \log |\Omega| - \frac{1}{2} \text{tr}(\Omega^{-1}yy^T). \tag{3.9}
\]
The $k^{th}$ element of the score vector is

$$s_k(\xi) = \frac{\partial l(\xi)}{\partial \xi_k} = -\frac{1}{2} \text{tr} \left( \Omega^{-1} \Omega_k \right) + \frac{1}{2} \text{tr} \left( \Omega^{-1} \Omega_k \Omega^{-1} yy^T \right)$$  \hspace{1cm} (3.10)$$

where $\Omega_k = \partial \Omega / \partial \xi_k$.

The $k \times k$ observed information matrix is

$$j(\xi) = -\frac{\partial^2 l(\xi)}{\partial \xi \partial \xi^T} = \{j_{kl}\}$$  \hspace{1cm} (3.11)$$

with

$$j_{kl} = \text{tr} \left\{ \Omega^{-1}_l \Omega^{-1}_k (\Omega^{-1} yy^T - \frac{1}{2} I_{nT}) \right\} + \frac{1}{2} \text{tr} \left\{ \Omega^{-1}_k \Omega^{-1}_{kl} (I_{nT} - \Omega^{-1} yy^T) \right\}$$  \hspace{1cm} (3.12)$$

where $\Omega_{kl} = \partial^2 \Omega / \partial \xi_k \partial \xi_l$.

The expected value of $j_{kl}$ is

$$i_{kl} = E(j_{kl}) = \frac{1}{2} \text{tr} \left( \Omega^{-1}_k \Omega^{-1}_l \right)$$  \hspace{1cm} (3.13)$$

and $i_{kl}$ is the $(k,l)$-element of the Fisher information matrix.
3.1. Derivation for the Covariance Matrix

From Equation 3.1 and Equation 3.2, we get

\[ y_t = Cx_t + D\epsilon_t \]
\[ = C(Ax_{t-1} + Bu_t) + D\epsilon_t \]
\[ = C\{A(Ax_{t-2} + Bu_{t-1}) + Bu_t\} + D\epsilon_t \]
\[ = C(A^2x_{t-2} + ABu_{t-1} + Bu_t) + D\epsilon_t \]
\[ = C(A^t x_0 + A^{t-1}Bu_1 + A^{t-2}Bu_2 + \cdots + ABu_{t-1} + Bu_t) + D\epsilon_t \]
\[ = C \left( A^t x_0 + \sum_{j=1}^{t} A^{t-j}Bu_j \right) + D\epsilon_t, \quad t = 1, \ldots, T. \tag{3.14} \]

Next, we will derive the \( T \times T \) subblocks in the \( nT \times nT \) block matrix \( \Omega = \{\Omega_{ij} = \text{cov}(y_i, y_j) : i, j = 1, \ldots, T.\} \).

The \( T \) subblocks on the diagonal of \( \Omega \) are \( \Omega_{ii} = \text{cov}(y_i, y_i) = \text{var}(y_i) : i = 1, \ldots, T. \} \).

\[ \text{var}(y_i) = \text{cov} \left\{ C \left( A^i x_0 + \sum_{k=1}^{i} A^{i-k}Bu_k \right) + D\epsilon_i, C \left( A^i x_0 + \sum_{k=1}^{i} A^{i-k}Bu_k \right) + D\epsilon_i \right\} \]
\[ = \text{cov} \left\{ C \left( A^i x_0 + \sum_{k=1}^{i} A^{i-k}Bu_k \right), C \left( A^i x_0 + \sum_{k=1}^{i} A^{i-k}Bu_k \right) \right\} + \text{cov}(D\epsilon_i, D\epsilon_i) \]
\[ = \text{cov}(CA^i x_0, CA^i x_0) + \text{cov} \left( \sum_{k=1}^{i} C A^{i-k}Bu_k, \sum_{k=1}^{i} C A^{i-k}Bu_k \right) + DD^T \]
\[ = CA^T P_0(A^i)^T C + C \left\{ \text{cov} \left( \sum_{k=1}^{i} A^{i-k}Bu_k, \sum_{k=1}^{i} A^{i-k}Bu_k \right) \right\} C^T + DD^T \]
\[ = C \left\{ A^T P_0(A^i)^T + \sum_{k=1}^{i} A^{i-k}BB^T(A^{i-k})^T \right\} C^T + DD^T \tag{3.15} \]
The off-diagonal subblocks of \( \Omega \) are \( \{ \Omega_{ij} = \text{cov}(y_i, y_j) : i, j = 1, \ldots, T \text{ and } i \neq j \} \). When \( i \neq j \), we have

\[
\text{cov}(y_i, y_j) = \text{cov} \left\{ C (A^i x_0 + \Sigma_{k=1}^i A^{i-k} B u_k) + D \epsilon_i, C (A^j x_0 + \Sigma_{m=1}^j A^{j-m} B u_m) + D \epsilon_j \right\}
\]

\[
= \text{cov} \left\{ C (A^i x_0 + \Sigma_{k=1}^i A^{i-k} B u_k), C (A^j x_0 + \Sigma_{m=1}^j A^{j-m} B u_m) \right\}
\]

\[
= C \left\{ \text{cov} \left( A^i x_0 + \Sigma_{k=1}^i A^{i-k} B u_k, A^j x_0 + \Sigma_{m=1}^j A^{j-m} B u_m \right) \right\} C^T
\]

\[
= C \left\{ A^i P_0 (A^j)^T + \Sigma_{\min(i,j)}^i \text{cov} \left( A^{i-k} B u_k, A^{j-k} B u_k \right) \right\} C^T
\]

\[
(3.16)
\]

If \( i < j \), then

\[
\text{cov}(y_i, y_j) = C \left\{ A^i P_0 (A^j)^T + \Sigma_{k=1}^i \text{cov} \left( \Sigma_{k=1}^i A^{i-k} B u_k, \Sigma_{m=1}^j A^{j-m} B u_m \right) \right\} C^T
\]

\[
= C \left\{ A^i P_0 (A^j)^T + \Sigma_{k=1}^i \text{cov} \left( A^{i-k} B u_k, A^{i-k} B u_k \right) \right\} C^T
\]

\[
= C \left\{ A^i P_0 (A^j)^T + \Sigma_{k=1}^i A^{i-k} B B^T (A^{i-k})^T \right\} C^T
\]

\[
(3.17)
\]

If \( i > j \), then

\[
\text{cov}(y_i, y_j) = C \left\{ A^i P_0 (A^j)^T + \Sigma_{k=1}^j \text{cov} \left( \Sigma_{k=1}^j A^{j-k} B u_k, \Sigma_{m=1}^j A^{j-m} B u_m \right) \right\} C^T
\]

\[
= C \left\{ A^i P_0 (A^j)^T + \Sigma_{k=1}^j \text{cov} \left( A^{j-k} B u_k, A^{j-k} B u_k \right) \right\} C^T
\]

\[
= C \left\{ A^i P_0 (A^j)^T + \Sigma_{k=1}^j A^{j-k} B B^T (A^{j-k})^T \right\} C^T
\]

\[
(3.18)
\]

Combining Equation 3.17 and Equation 3.18, we have that when \( i \neq j \),

\[
\Omega_{ij} = \text{cov}(y_i, y_j)
\]

\[
= C \left[ A^i P_0 (A^j)^T + \Sigma_{k=1}^{\min(i,j)} A^{\min(i,j)-k} B B^T \left\{ A^{\min(i,j)-k} \right\}^T \right] C^T, \quad i, j = 1, \ldots, T \quad (3.19)
\]
In summary, the \((i, j)\) subblock of size \(T \times T\) in the \(nT \times nT\) block matrix \(\Omega\) have the expressions in Equation 3.8.

### 3.2. Higher-order Asymptotics for the Signed Root of the Log Likelihood Ratio Test Statistic for the General Time-invariant State Space Model

We use \(\xi^T = (\xi_1, \xi_2, \ldots, \xi_k) = (\beta, \delta^T)\), where \(\beta\) is the scalar interest parameter and \(\delta^T = (\xi_2, \ldots, \xi_k)\) is the vector of nuisance parameters. The signed root of the likelihood ratio test statistic is \(\hat{w}(\beta)\). The conditional CDF of \(\hat{w}(\beta)|a; \xi\) given an approximate ancillary \(a\) has strong dependence on only the \(\beta\) component of \(\xi\). Let \(\hat{W}_\beta\) denote the random variable and \(\hat{w}(\beta)\) as its observed value. Then, the approximate continuous CDF approximation for the random variable \(\hat{W}_\beta\) given ancillary \(a\) has the same expression as in Equation 2.35.

For the general state-space model, the approximation for \(\hat{u}(\beta)\) is derived using Equation 2.36. We can derive that

\[
\begin{align*}
\left\| \left( \frac{\partial \theta^T}{\partial \xi} \right|_{\xi = \hat{\xi}_0} \frac{\partial \mu_{\xi}}{\partial \xi} \right)_{\xi = \hat{\xi}} &= \left\| \left( \frac{\partial \theta^T}{\partial \xi} \right|_{\xi = \hat{\xi}_0} \right) \left( \frac{\partial \mu_{\xi}}{\partial \beta} \right|_{\xi = \hat{\xi}_0}, \frac{\partial \mu_{\xi}}{\partial \delta^T} \right|_{\xi = \hat{\xi}} \\
&= \left\| \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} \right\| 
\end{align*}
\]

where \(D_{11}\) has dimension \((k - 1) \times 1\). Denote

\[
\Omega_k(\hat{\xi}_0) = \left. \frac{\partial \Omega}{\partial \xi_k} \right|_{\xi = \hat{\xi}_0}, \quad \Omega_k(\hat{\xi}) = \left. \frac{\partial \Omega}{\partial \xi_k} \right|_{\xi = \hat{\xi}} 
\]

Denote each element of \(D_{11}\) as \((D_{11})_j, j = 2, \ldots, k\). We can derive that

\[
(D_{11})_j = \frac{1}{2} \text{tr} \left\{ \Omega^{-1}(\hat{\xi}_0) \Omega_j(\hat{\xi}_0) \Omega^{-1}(\hat{\xi}_0) \Omega_1(\hat{\xi}) \right\} 
\]
Matrix $D_{12}$ has dimension $(k - 1) \times (k - 1)$. Denote each element of $D_{12}$ as $(D_{12})_{ij}$, $i, j = 2, \ldots, k$. We can derive that

$$
(D_{12})_{ij} = \frac{1}{2} \operatorname{tr} \left\{ \Omega^{-1}(\hat{\xi}_0)\Omega_i(\hat{\xi}_0)\Omega^{-1}(\hat{\xi}_0)\Omega_j(\hat{\xi}) \right\}.
$$

(3.23)

Matrix $D_{21}$ has dimension $1 \times 1$. We can derive that

$$
(D_{21})_{11} = \frac{1}{2} \operatorname{tr} \left[ \{\Omega^{-1}(\hat{\xi}_0) - \Omega^{-1}(\hat{\xi})\} \Omega_1(\hat{\xi}) \right].
$$

(3.24)

Matrix $D_{22}$ has dimension $1 \times (k - 1)$. Denote each element of $D_{22}$ as $(D_{22})_j$, $j = 2, \ldots, k$. We can derive that

$$
(D_{22})_j = \frac{1}{2} \operatorname{tr} \left[ \{\Omega^{-1}(\hat{\xi}_0) - \Omega^{-1}(\hat{\xi})\} \Omega_j(\hat{\xi}) \right].
$$

(3.25)
4.1. Example One: Using the Kalman Filter to Track a Moving Object in One Dimension

The Kalman filter is frequently used in tracking moving objects. We will consider an example where the real path of an object in one dimension follows the continuous time model [Sadli, 2020]:

\[ \text{position} = f(t) = t^2 - t, \quad \text{time } t \geq 0 \] (4.1)

In practical settings, position is measured in discrete time. Assuming the time interval between each measurement is \( \Delta t = 1 \) s (s stands for second) and applying the kinematic equations in discrete time, we obtain a state-space model that describes the path in Equation 4.1. (See Section 4.1.1 for the derivation.) We denote \( x_t^* = (s_t, v_t)^T \) as the position and velocity of the object at time \( t \). This sequence has a trend in time and the state-space model is written in terms of the detrended sequence \( x_t = x_t^* - E(x_t^*) \). The detrended observations are \( y_t = y_t^* - (1, 0)E(x_t^*) \), where \( y_t^* \) is the observed positions before detrending. The detrended state-space model is

\[ x_t = Ax_{t-1} + b\beta u_t, \quad t = 1, \ldots, T. \] (4.2)

\[ y_t = (1, 0)x_t + \delta\epsilon_t \] (4.3)
where \( u_t \overset{i.i.d.}{\sim} N(0, 1) \), \( \epsilon_t \overset{i.i.d.}{\sim} N(0, 1) \), and

\[
E(x^*_0) = (0, -1)^T
\]  
(4.4)

\[
A = \begin{pmatrix} 1 & \triangle t \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}
\]  
(4.5)

\[
b = \begin{pmatrix} \frac{1}{2} (\triangle t)^2 \\ \triangle t \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix}
\]  
(4.6)

\[
y^*_t = (1, 0)x^*_t + \delta \epsilon_t
\]  
(4.7)

In the above model, \( \beta \) and \( \delta \) are unknown parameters. We are interested in finding a confidence interval for the interest parameter \( \beta \) in the presence of the nuisance parameter \( \delta \). We may make inference about \( \beta \) with nuisance parameter \( \delta \) using the detrended model. This is because \( E(x^*_t) \) does not depend on \( (\beta, \delta) \), so detrending does not change the likelihood function in terms of \( (\beta, \delta) \).

### 4.1.1. Derivation for the State Space Model for One-dimensional Object Tracking

For an object following the continuous model in Equation 4.1, the velocity is

\[
f'(t) = 2t - 1, \quad \text{time } t \geq 0
\]  
(4.8)

and the acceleration is

\[
f''(t) = 2, \quad \text{time } t \geq 0.
\]  
(4.9)
The kinematic equations in discrete time are

\[
\begin{align*}
\text{position} & \quad s_t = s_{t-1} + v_{t-1} \Delta t + \frac{1}{2} a_{t-1} (\Delta t)^2 \\
\text{velocity} & \quad v_t = v_{t-1} + a_{t-1} \Delta t
\end{align*}
\] (4.10) (4.11)

where \( a_{t-1} \) is the acceleration at time \( t-1 \), \( \Delta t \) is the discretized time interval and we assume \( \Delta t = 1 \text{ s} \).

To get the state-space model, we assume the acceleration \( a_t \) is an i.i.d. normal sequence of random variables with mean \( \mu_a = 2 \text{ m/s}^2 \) (\( m \) stands for meter) and variance \( \beta^2 \). That is, \( a_t \overset{i.i.d.}{\sim} N(\mu_a = 2, \beta^2) \). Under these assumptions, the standard deviation of \( v_t \) given \( v_{t-1} \) is

\[
\begin{align*}
\text{sd}(v_t|v_{t-1}) &= \Delta t \text{ sd}(a_{t-1}) = (\Delta t)\beta \\
\text{sd}(s_t|s_{t-1}, v_{t-1}) &= \frac{1}{2}(\Delta t)^2 \text{ sd}(a_{t-1}) = \frac{1}{2}(\Delta t)^2 \beta
\end{align*}
\] (4.12) (4.13)

With \( x_t^* = (s_t, v_t)^T \) and

\[
A = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \quad b = \begin{pmatrix} \frac{1}{2}(\Delta t)^2 \\ \Delta t \end{pmatrix}
\] (4.14) (4.15)

we can write Equation 4.10 and Equation 4.11 in a vector form as

\[
x_t^* = Ax_{t-1}^* + ba_{t-1} = Ax_{t-1}^* + ba_t
\] (4.16)
The state recursions in Equation 4.16 reflect a dynamic change in which $E(x_t^*)$ depends on time $t$. To detrend the state equation, we will write it in terms of $x_t = x_t^* - E(x_t^*)$, which is the centered state value. Since

$$E(x_t^*) = A E(x_{t-1}^*) + b E(a_t) = A E(x_{t-1}^*) + b \mu_a,$$

(4.17)

substracting Equation 4.17 from Equation 4.16, we get

$$x_t^* - E(x_t^*) = A \{x_{t-1}^* - E(x_{t-1}^*)\} + b(a_t - \mu_a)$$

(4.18)

That is,

$$x_t = A x_{t-1} + b(a_t - \mu_a), \quad t = 1, \ldots, T.$$  

(4.19)

Equation 4.19 is the detrended state equation with

$$E(x_0^*) = E \begin{pmatrix} s_0 \\ v_0 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$$

(4.20)

Notice that Equation 4.19 centers the acceleration at mean 0 by substracting $\mu_a$ from $a_t$. Thus, the detrended state equation can be further simplified as

$$x_t = A x_{t-1} + b \beta u_t, \quad t = 1, \ldots, T.$$  

(4.21)

where $u_t \sim i.i.d. N(0, 1)$.

The observation equation can be expressed as

$$y_t^* = (1, 0)x_t^* + \delta \epsilon_t = (1, 0)\{x_t + E(x_t^*)\} + \delta \epsilon_t.$$
So that after centering,

\[ y_t = y_t^* - (1, 0) E(x_t^*) = (1, 0)x_t + \delta \epsilon_t, \quad t = 1, \ldots, T. \tag{4.22} \]

where \( \epsilon_t \sim N(0, 1) \).

Combining the state Equation 4.21 and the observation Equation 4.22, we obtain the state-space model that describe the path in Equation 4.1 for one-dimensional object tracking.

4.2. Example Two: Using the Kalman Filter to Track a Moving Object in Two Dimensions

Assume the real path of an object is two dimensional in Euclidean coordinates. Moreover, assume the real path of the object in both coordinates follow the same continuous time model in Equation 4.1. By applying the kinematic equations in discrete time, we obtain a detrended state-space model that describes the path of a moving object in two dimensions as (see Section 4.2.1 for the derivation)

\[ x_t = (A \otimes I_2)x_{t-1} + (b \otimes I_2)\Sigma u_t, \quad t = 1, \ldots, T. \tag{4.23} \]

\[ y_t = \{(1, 0) \otimes I_2\}x_t + D\epsilon_t. \tag{4.24} \]

Denote the raw position and velocity before detrending as \( x_t^* = (s_{1t}, s_{2t}, v_{1t}, v_{2t})^T \), where \( s_t = (s_{1t}, s_{2t})^T \), and \( v_t = (v_{1t}, v_{2t})^T \). The errors are \( u_t \overset{i.i.d}{\sim} N_2(0, I_2) \), \( \epsilon_t \overset{i.i.d}{\sim} N_2(0, I_2) \). We take \( \Delta t = 1 \). The detrended state space \( x_t \) and the detrended observations \( y_t \) are related to their
raw values according to

\[ x_t = x_t^* - E(x_t^*) \quad (4.25) \]

\[ E(x_0^*) = (0, 0, -1, -1)^T \quad (4.26) \]

\[ A = \begin{pmatrix} 1 & \triangle t \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad (4.27) \]

\[ b = \begin{pmatrix} \frac{1}{2} (\triangle t)^2 \\ \triangle t \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ 1 \end{pmatrix} \quad (4.28) \]

\[ \Sigma_a = \begin{pmatrix} \xi_2 & \xi_1 \\ \xi_1 & \xi_2 \end{pmatrix} \quad (4.29) \]

\[ y_t = y_t^* - E(y_t^*) \quad (4.30) \]

\[ E(y_t^*) = \{(1, 0) \otimes I_2\} E(x_t^*) \quad (4.31) \]

\[ D = \text{diag}\{\xi_3, \xi_3\}. \quad (4.32) \]

In the above model, \(\xi_1, \xi_2,\) and \(\xi_3\) are unknown parameters. The kinematic assumptions made in deriving this state-space model is that the acceleration process \(a_t = (a_{1t}, a_{2t})^T\) in the two-dimensional space is an i.i.d. sequence of \(N_2(u_a, \Sigma_a^2)\) responses representing stationary acceleration but with correlation in the two dimensions of acceleration. Notice that the likelihood function for the detrended model is the same as the likelihood function for the model before detrending since \(E(x_t^*)\) does not depend on the unknown parameters.
4.2.1. Derivation for the State Space Model for Two-dimensional Object Tracking

Denote $s_t = (s_{1t}, s_{2t})^T$, $v_t = (v_{1t}, v_{2t})^T$, and $x_t^* = (s_{1t}, s_{2t}, v_{1t}, v_{2t})^T$. Then building upon the derivation in Section 4.1.1 for one-dimensional object tracking, we can write the states for two dimensional object tracking as

$$
x_t^* = \begin{pmatrix}
    s_t \\
    v_t \\
    s_{t-1} + v_{t-1} \Delta t \\
    v_{t-1}
\end{pmatrix} + \begin{pmatrix}
    \frac{1}{2}(\Delta t)^2 a_t \\
    \Delta t a_t
\end{pmatrix}
$$

(4.33)

where $a_t = (a_{1t}, a_{2t})^T$ and $a_t \sim N_2(\mu_a, \Sigma_a^2)$ is a bivariate normal distribution with mean $\mu_a = (2, 2)^T$ and standard deviation matrix

$$
\Sigma_a = \begin{pmatrix}
    \xi_2 & \xi_1 \\
    \xi_1 & \xi_2
\end{pmatrix},
$$

(4.34)

defined such that $\Sigma_a^2$ is the covariance matrix.

Thus,

$$
x_t^* = \left( \begin{array}{cc}
I_2 & I_2 \Delta t \\
0 & I_2
\end{array} \right) \begin{pmatrix}
    s_{t-1} \\
    v_{t-1}
\end{pmatrix} + \begin{pmatrix}
    \frac{1}{2}(\Delta t)^2 I_2 \\
    \Delta t I_2
\end{pmatrix} a_t
$$

(4.35)

$$
x_t^* = (A \otimes I_2)x_{t-1}^* + (b \otimes I_2)a_t, \quad t = 1, \ldots, T.
$$

(4.36)
where

\[ A = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix} \]  
(4.37)

\[ b = \begin{pmatrix} \frac{1}{2}(\Delta t)^2 \\ \Delta t \end{pmatrix} \]  
(4.38)

We now detrend Equation 4.36 using

\[ E(x^*_t) = (A \otimes I_2)E(x^*_{t-1}) + (b \otimes I_2)\mu_a, \quad t = 1, \ldots, T. \]  
(4.39)

Subtracting Equation 4.39 from Equation 4.36, we get the state equation for two dimensional object tracking as

\[ x_t = x^*_t - E(x^*_t) = (A \otimes I_2)\{x^*_{t-1} - E(x^*_{t-1})\} + (b \otimes I_2)(a_t - \mu_a) \]  
(4.40)

\[ x_t = (A \otimes I_2)x_{t-1} + (b \otimes I_2)\Sigma a u_t, \quad t = 1, \ldots, T. \]  
(4.41)

where \( u_t \overset{i.i.d.}{\sim} N_2(0, I_2) \) and \( E(x^*_0) = (0, 0, -1, -1)^T \).

The raw observations for the two dimensional object tracking are

\[ y^*_t = \begin{pmatrix} s_{1t} \\ s_{2t} \end{pmatrix} + D \epsilon_t = \{(1, 0) \otimes I_2\}x^*_t + D\epsilon_t \]  
(4.42)

where \( \epsilon_t \overset{i.i.d.}{\sim} N_2(0, I_2), D = diag\{\xi_3, \xi_3\} \). We detrend using

\[ E(y^*_t) = \{(1, 0) \otimes I_2\}E(x^*_t). \]  
(4.43)
By subtracting Equation 4.43 from Equation 4.42, we obtain the observational equation for the two dimensional object tracking as

\[ y_t = y_t^* - E(y_t^*) = \{(1, 0) \otimes I_2\} x_t + D\epsilon_t \] (4.44)

### 4.3. Confidence Intervals for the Correlation Coefficient \( \rho \)

The two dimensional object tracking model is

\[ x_t = (A \otimes I_2)x_{t-1} + (b \otimes I_2)\Sigma_a u_t, \quad t = 1, \ldots, T. \] (4.45)

\[ y_t = Cx_t + D\epsilon_t \] (4.46)

where \( I_2 \) is the \( 2 \times 2 \) identity matrix, \( u_t \overset{i.i.d}{\sim} N_2(0, I_2) \), \( \epsilon_t \overset{i.i.d}{\sim} N_2(0, I_2) \), and

\[ A \otimes I_2 = \begin{pmatrix}
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{pmatrix} \] (4.47)

\[ b \otimes I_2 = \begin{pmatrix}
1/2 & 0 \\
0 & 1/2 \\
1 & 0 \\
0 & 1
\end{pmatrix} \] (4.48)
\[
\Sigma_a = \begin{pmatrix}
\xi_2 & \xi_1 \\
\xi_1 & \xi_2
\end{pmatrix}
\] (4.49)

\[
C = (1, 0) \otimes I_2 = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0
\end{pmatrix}
\] (4.50)

\[
D = \begin{pmatrix}
\xi_3 & 0 \\
0 & \xi_3
\end{pmatrix}
\] . (4.51)

The unknown parameter vector is \( \xi = (\xi_1, \xi_2, \xi_3)^T \). Assuming the initial state \( x_0 \) has distribution \( x_0 \sim N_4(\mathbf{0}_4, I_4) \), we want to construct the confidence intervals for the correlation coefficient

\[
\rho = \frac{2\xi_1\xi_2}{\xi_1^2 + \xi_2^2}
\] (4.52)

using four methods: 1) likelihood ratio; 2) Wald; 3) score; 4) higher-order asymptotic interval. In order to do that, we need to transform the original parameter vector \( \xi = (\xi_1, \xi_2, \xi_3)^T \) to

\[
\xi_\rho = \begin{pmatrix}
\rho = \frac{2\xi_1\xi_2}{\xi_1^2 + \xi_2^2} \\
\xi_2 \\
\xi_3
\end{pmatrix}
\] (4.53)

4.3.1. The Likelihood Ratio Interval

To construct the 95% likelihood ratio interval for \( \rho \), we need to find the \( \rho \) values such that

\[
\left\{ \rho : -2 \log \frac{L(\hat{\xi}_\rho)}{L(\xi_\rho)} < (\chi^2_1)^{-1}(0.95) \right\} .
\] (4.54)
Here $L$ stands for the likelihood, $\hat{\xi}_\rho$ is the global MLE of $\xi_\rho$, and $\hat{\xi}_{\rho_0}$ is the constrained MLE of $\xi_\rho$ when holding $\rho = \rho_0$ fixed.

Equivalently, denote

$$\lambda = \frac{L(\hat{\xi}_{\rho_0})}{L(\hat{\xi}_\rho)}$$  \hspace{1cm} (4.55)

we can use the signed root of $\lambda$ given as

$$\hat{w}(\rho) = \text{sgn}(\hat{\rho} - \rho)\sqrt{-2 \log \lambda} \sim N(0, 1),$$  \hspace{1cm} (4.56)

where $\text{sgn}(\cdot)$ is the sign of the value. Under the assumption that $\rho$ is the true parameter, $\hat{w}(\rho)$ has an approximate $N(0, 1)$ distribution based on standard likelihood ratio asymptotics applied to this model. A 95% asymptotic confidence interval (CI) for $\rho$ is $\{\rho : -1.96 < \hat{w}(\rho) < 1.96\}$.

4.3.2. The Wald Interval

The MLE of $\xi$ is $\hat{\xi}$ with approximate distribution $N(\xi, i^{-1}(\hat{\xi}))$. The MLE of $\xi_\rho$ is $\hat{\xi}_\rho$ with approximate distribution $N(\xi_\rho, i^{-1}_\rho(\hat{\xi}_\rho))$. We can show that $i_\rho = HiH^T$, where

$$H = \frac{\partial \xi^T}{\partial \xi_\rho} = \begin{pmatrix} \xi_2 \left( \frac{1}{\sqrt{1-\rho^2}} + \frac{\sqrt{1-\rho^2}}{\rho^2} - \frac{1}{\rho^2} \right) & 0 & 0 \\ \frac{1}{\rho} (1 - \sqrt{1-\rho^2}) & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$  \hspace{1cm} (4.57)

$$i = -E \left( \frac{\partial^2 l}{\partial \xi \partial \xi^T} \right)$$  \hspace{1cm} (4.58)

$$i_\rho = -E \left( \frac{\partial^2 l}{\partial \xi_\rho \partial \xi_\rho^T} \right)$$  \hspace{1cm} (4.59)

The 95% Wald interval for $\rho$ is $\hat{\rho} \pm 1.96 \sqrt{(1,1)}$-element of $i^{-1}_\rho(\hat{\xi}_\rho)$.  

64
4.3.3. The Score Interval

Holding \( \rho = \rho_0 \) fixed, from the log-likelihood function, we can compute \( \hat{\xi}_{\rho_0} \), from which we can compute \( \hat{\xi}_0 \), which is the constrained MLE of \( \xi \).

The score test statistic is invariant to the parameterization. We can show that the score test statistic in the \( \xi \) parameterization equals to the score test statistic in the \( \xi_\rho \) parameterization. That is,

\[
s_T^{\rho_0} i^{-1}_\rho s_\rho = s_T i^{-1}_0 s(\hat{\xi}_0) \tag{4.60}
\]

where \( s_\rho = \partial l / \partial \xi_\rho \), \( s = \partial l / \partial \xi \). See Appendix C for the proof.

When testing \( H_0 : \rho = \rho_0 \), the score test statistic is

\[
S(\rho_0) = s_T^{\rho_0} \hat{\xi}_{\rho_0} s_\rho(\hat{\xi}_{\rho_0}) = s_T(\hat{\xi}_0) i^{-1}_0 s(\hat{\xi}_0) \tag{4.61}
\]

The 95% score interval for \( \rho \) is \( \{ \rho : \sqrt{S(\rho)} < 1.96 \} \).

4.3.4. The Higher-order Asymptotic Interval

We use \( \xi_\rho^T = (\rho, \xi_2, \xi_3) = (\rho, \delta^T), \) where \( \rho \) is the scalar interest parameter and \( \delta^T = (\xi_2, \xi_3) \) is the vector of nuisance parameters. The signed root of the likelihood ratio test statistic is \( \hat{w}(\rho) \). The conditional CDF of \( \hat{w}(\rho)|a; \xi_\rho \) given an approximate ancillary \( a \) has strong dependence on only the \( \rho \) component of \( \xi_\rho \). Let \( \hat{W}_\rho \) denote the random variable and \( \hat{w}(\rho) \) as its observed value. Then, the Barndorff-Nielsen formula for an approximate continuous CDF approximation for the random variable \( \hat{W}_\rho \) given ancillary \( a \) is

\[
\hat{P} \left\{ \hat{W}_\rho \leq \hat{w}(\rho)|a \right\} = \Phi\{\hat{w}(\rho)\} + \phi\{\hat{w}(\rho)\} \left\{ \frac{1}{\hat{w}(\rho)} - \frac{1}{\hat{u}(\rho)} \right\}, \quad \hat{\rho} \neq \rho \tag{4.62}
\]

Let \( G(\rho) = \hat{P}\{\hat{W}_\rho \leq \hat{w}(\rho)|a \} \), then the second-order 95% CI is \( \{ \rho : 0.025 < G(\rho) < 0.975 \} \).
The value for \( \hat{u}(\rho) \) is difficult to compute and the approximation for \( \hat{u}(\rho) \) is

\[
\hat{u}(\rho) = \frac{\text{sgn}(\hat{\rho} - \rho)}{|i_\rho(\hat{\xi}_\rho)|} \sqrt{|j_\rho(\hat{\xi}_\rho)|} \left| \begin{pmatrix} \partial \theta_T / \partial \delta |_{\xi = \hat{\xi}_0} \\ (\hat{\theta} - \hat{\theta}_0)^T \end{pmatrix} \right| \frac{\partial \mu}{\partial \xi_T} |_{\hat{\xi}_\rho}. \tag{4.63}
\]

4.3.5. Derivation for \( \hat{u}_\rho(\rho) \)

Denote \( \xi = (\xi_1, \xi_2, \xi_3)^T \) and \( \xi_\rho = h(\xi) = (\rho, \xi_2, \xi_3)^T \), where the correlation coefficient \( \rho = 2\xi_1\xi_2/(\xi_1^2 + \xi_2^2) \). Denote the maximum likelihood estimator (MLE) for \( \xi \) as \( \hat{\xi} \), and the MLE for \( \xi_\rho \) as \( \hat{\xi}_\rho \). Holding \( \rho = \rho_0 \) fixed, denote the constrained maximum likelihood estimator (MLE) for \( \xi \) as \( \hat{\xi}_0 \) and the constrained MLE for \( \xi_\rho \) as \( \hat{\xi}_\rho_0 \). Furthermore, let \( \hat{u}_\rho(\rho) \) be \( \hat{u} \) in terms of the \( \xi_\rho \) parameterization, \( \hat{u}(\rho) \) is \( \hat{u} \) in terms of the \( \xi \) parameterization.

To compute \( \hat{u}_\rho(\rho) \), we need to compute \( |i_\rho(\hat{\xi}_\rho)| \), which is the determinant of the Fisher information in terms of the \( \xi_\rho \) parameterization evaluated at \( \hat{\xi}_\rho \). We also need \( |j_\rho(\hat{\xi}_\rho)| \), which is the determinant of the observed information in terms of the \( \xi_\rho \) parameterization evaluated at \( \hat{\xi}_\rho \). Moreover, we need \( |(j_\rho)_{22}(\hat{\xi}_\rho_0)| \), which is the determinant of the \( (\xi_2, \xi_3) \times (\xi_2, \xi_3) \) sub-block of \( j_\rho \) evaluated at \( \hat{\xi}_\rho_0 \). Furthermore, denoting \( \delta = (\xi_2, \xi_3)^T \), we need to compute

\[
\left| \begin{pmatrix} \partial \theta_T / \partial \delta |_{\xi = \hat{\xi}_0} \\ (\hat{\theta} - \hat{\theta}_0)^T \end{pmatrix} \right| \frac{\partial \mu}{\partial \xi_T} |_{\hat{\xi}_\rho}. \tag{4.64}
\]

We want to compute the above quantities in terms of the \( \xi \) parameterization.

Recall that we have shown \( i_\rho = HiH^T \), where \( H \) is a matrix with \( H_{ij} \) as its \( (i, j) \)-element and

\[
H = \frac{\partial \xi_T}{\partial \xi_\rho} = \left\{ H_{ij} = \frac{\partial \xi_j}{\partial (\xi_\rho)_i} \right\}. \tag{4.65}
\]
It is obvious that $|i_\rho| = |H|^2 |\hat{i}| = r^2 |\hat{i}|$, where

$$ r = |H| = \frac{\partial \xi_1}{\partial \rho} = \xi_2 \left( \frac{1}{\sqrt{1 - \rho^2}} + \frac{\sqrt{1 - \rho^2}}{\rho^2} - \frac{1}{\rho^2} \right) \quad (4.66) $$

We can show that $\hat{j}_\rho = \hat{H} \hat{j} \hat{H}^T$ at the global MLEs.

Proof:

$$ \frac{\partial l}{\partial (\xi_\rho)_i} = \sum_{k=1}^{3} \frac{\partial l}{\partial \xi_k} \frac{\partial \xi_k}{\partial (\xi_\rho)_i} \quad (4.67) $$

Applying the chain rule,

$$ (j_\rho)_{ij} = - \frac{\partial^2 l}{\partial (\xi_\rho)_i \partial (\xi_\rho)_j} = - \sum_{k=1}^{3} \sum_{l=1}^{3} \frac{\partial^2 l}{\partial \xi_k \partial \xi_l} \frac{\partial \xi_l}{\partial (\xi_\rho)_j} \frac{\partial \xi_k}{\partial (\xi_\rho)_i} - \sum_{k=1}^{3} \frac{\partial l}{\partial \xi_k} \frac{\partial^2 \xi_k}{\partial (\xi_\rho)_j} \quad (4.68) $$

In Equation 4.68, $\partial l / \partial \xi_k = 0$ at the global MLE $\hat{\xi}$. Hence,

$$ (j_\rho)_{ij} \bigg|_{\hat{\xi}} = \left( \hat{j}_\rho \right)_{ij} = - \sum_{k=1}^{3} \sum_{l=1}^{3} \frac{\partial^2 l}{\partial \xi_k \partial \xi_l} \frac{\partial \xi_l}{\partial (\xi_\rho)_j} \frac{\partial \xi_k}{\partial (\xi_\rho)_i} \bigg|_{\hat{\xi}} \quad (4.69) $$

$$ = \sum_{k=1}^{3} \sum_{l=1}^{3} \hat{j}_{kl} \hat{H}_{jl} \hat{H}_{ik} = (i, j) \text{ element of } \hat{H} \hat{j} \hat{H}^T \quad (4.70) $$

where

$$ \hat{j} = \frac{\partial^2 l}{\partial \xi \partial \xi^T} \bigg|_{\hat{\xi}} \quad (4.71) $$

$$ \hat{H} = \frac{\partial \xi^T}{\partial \xi} \bigg|_{\hat{\xi}} \quad (4.72) $$

Thus, $\hat{j}_\rho = \hat{H} \hat{j} \hat{H}^T$. It is easy to see that

$$ |\hat{j}_\rho| = \hat{r}^2 |\hat{j}| \quad (4.73) $$

where $\hat{r} = r \bigg|_{\hat{\xi}}$. 

67
Next, we will derive \(|(j_\rho)_{22}(\hat{\xi}_0)|\). Now, holding \(\rho = \rho_0\) fixed, we have

\[
\left. \frac{\partial l}{\partial (\xi_\rho)_2} \right|_{\hat{\xi}_0} = \left. \frac{\partial l}{\partial \xi_2} \right|_{\hat{\xi}_0} = \left. \frac{\partial l}{\partial (\xi_\rho)_3} \right|_{\hat{\xi}_0} = \left. \frac{\partial l}{\partial \xi_3} \right|_{\hat{\xi}_0} = 0
\]

(4.74)

Just like in Equation 4.68,

\[
(j_\rho)_{22}|_{\hat{\xi}_0} = -\sum_{k=1}^{3} \sum_{l=1}^{3} \left. \frac{\partial^2 l}{\partial \xi_k \partial \xi_l} \frac{\partial \xi_k}{\partial (\xi_\rho)_j} \frac{\partial \xi_l}{\partial (\xi_\rho)_i} \right|_{\hat{\xi}_0} - \sum_{k=1}^{3} \left. \frac{\partial l}{\partial \xi_k} \frac{\partial^2 \xi_k}{\partial (\xi_\rho)_i \partial (\xi_\rho)_j} \right|_{\hat{\xi}_0}, \quad i, j = 2, 3
\]

(4.75)

\[
\left. \frac{\partial^2 l}{\partial \xi_k \partial \xi_l} \frac{\partial \xi_k}{\partial (\xi_\rho)_i} \frac{\partial \xi_l}{\partial (\xi_\rho)_j} \right|_{\hat{\xi}_0} - 0
\]

(4.76)

\[
= (\hat{H}_{2,3;0}) j(\hat{\xi}_0)(\hat{H}^T_{2,3;0}),
\]

(4.77)

where \(\hat{H}_{2,3;0}\) is the second and third rows of \(H\) evaluated at \(\hat{\xi}_0\). \(\hat{H}^T_{2,3;0}\) is the transpose of \(\hat{H}_{2,3;0}\). \(j(\hat{\xi}_0)\) is \(j\) evaluated at \(\hat{\xi}_0\).

Notice that \(\hat{H}_{2,3;0}\) is a \(2 \times 3\) matrix. It has the following form

\[
\hat{H}_{2,3;0} = \begin{pmatrix}
\frac{1}{\rho} (1 - \sqrt{1 - \rho^2}) & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]

(4.78)
Next, with \( \delta = (\xi_2, \xi_3)^T \), we will compute

\[
\left| \left| \left( \frac{\partial \theta^T / \partial \delta}{(\hat{\theta} - \hat{\theta}_0)^T} \right) \frac{\partial \mu_{\xi^\rho} / \partial \xi}{\hat{\xi}^\rho} \right| \right| = \left| \left| \left( \frac{\partial \theta^T / \partial \delta}{(\hat{\theta} - \hat{\theta}_0)^T} \right) \frac{\partial \mu_{\xi^\rho} / \partial \xi}{\hat{\xi}^\rho} \right| \right| \hat{r} \quad (4.79)
\]

\[
\left| \left| \left( \frac{\partial \theta^T / \partial \delta}{(\hat{\theta} - \hat{\theta}_0)^T} \right) \frac{\partial \mu_{\xi^\rho} / \partial \xi}{\hat{\xi}^\rho} \right| \right| \hat{r} = \left| \left| (d_{ij} : i, j = 1, 2, 3) \right| \right| \hat{r} \quad (4.80)
\]

where the expressions for \( d_{ij} \) are

\[
d_{11} = \frac{1}{2} \text{tr} \left\{ \Omega^{-1}(\hat{\xi}_0) \Omega_2(\hat{\xi}_0) \Omega^{-1}(\hat{\xi}_0) \Omega_1(\hat{\xi}) \right\} \quad (4.82)
\]

\[
d_{12} = \frac{1}{2} \text{tr} \left\{ \Omega^{-1}(\hat{\xi}_0) \Omega_2(\hat{\xi}_0) \Omega^{-1}(\hat{\xi}_0) \Omega_2(\hat{\xi}) \right\} \quad (4.83)
\]

\[
d_{13} = \frac{1}{2} \text{tr} \left\{ \Omega^{-1}(\hat{\xi}_0) \Omega_3(\hat{\xi}_0) \Omega^{-1}(\hat{\xi}_0) \Omega_3(\hat{\xi}) \right\} \quad (4.84)
\]

\[
d_{21} = \frac{1}{2} \text{tr} \left\{ \Omega^{-1}(\hat{\xi}_0) \Omega_3(\hat{\xi}_0) \Omega^{-1}(\hat{\xi}_0) \Omega_1(\hat{\xi}) \right\} \quad (4.85)
\]

\[
d_{22} = \frac{1}{2} \text{tr} \left\{ \Omega^{-1}(\hat{\xi}_0) \Omega_3(\hat{\xi}_0) \Omega^{-1}(\hat{\xi}_0) \Omega_2(\hat{\xi}) \right\} \quad (4.86)
\]

\[
d_{23} = \frac{1}{2} \text{tr} \left\{ \Omega^{-1}(\hat{\xi}_0) \Omega_3(\hat{\xi}_0) \Omega^{-1}(\hat{\xi}_0) \Omega_3(\hat{\xi}) \right\} \quad (4.87)
\]

\[
d_{31} = \frac{1}{2} \text{tr} \left\{ \Omega^{-1}(\hat{\xi}_0) - \Omega^{-1}(\hat{\xi}) \right\} \Omega_1(\hat{\xi}) \quad (4.88)
\]

\[
d_{32} = \frac{1}{2} \text{tr} \left\{ \Omega^{-1}(\hat{\xi}_0) - \Omega^{-1}(\hat{\xi}) \right\} \Omega_2(\hat{\xi}) \quad (4.89)
\]

\[
d_{33} = \frac{1}{2} \text{tr} \left\{ \Omega^{-1}(\hat{\xi}_0) - \Omega^{-1}(\hat{\xi}) \right\} \Omega_3(\hat{\xi}) \quad (4.90)
\]

and

\[
\Omega_k(\hat{\xi}_0) = \frac{\partial \Omega}{\partial \xi_k} \bigg|_{\xi = \hat{\xi}_0} \quad \Omega_k(\hat{\xi}) = \frac{\partial \Omega}{\partial \xi_k} \bigg|_{\xi = \hat{\xi}} \quad k = 1, 2, 3 \quad (4.91)
\]
Putting things together, we can derive that the approximation for \( \hat{u}_\rho(\rho) \) as

\[
\hat{u}_\rho(\rho) = \frac{\text{sgn}(\hat{\rho} - \rho)}{|i(\hat{\xi})|} \sqrt{\frac{|j(\hat{\xi})|}{|((\hat{H}_{2,3,3}^T)(\hat{H}_{2,3,3}^-)|}} \cdot |(d_{ij} : i, j = 1, 2, 3)|.
\]

(4.92)

Note that all the ingredients in Equation 4.92 are expressed in terms of the original parameterization.

4.4. Simulation

4.4.1. Challenges With Constructing the Score Intervals and the Higher-order Asymptotic Intervals

Both the score interval and the higher-order asymptotic interval were computed using likelihood score and Hessian computations programmed outside the \texttt{ssm} in MATLAB. When performing calculations for the score interval using this external programming, we found that the second and third elements of the score vectors evaluated at the constrained MLE were not equal to 0, which in theory, should be 0. This led us to being suspicious about the accuracy of our computations outside of \texttt{ssm}. Moreover, when performing the higher-order asymptotic interval calculation, we found that the computation of the determinant of the denominator for the term under the square root of the \( \hat{u} \) expression was negative, which in theory, should have been positive. This led to an investigation of why our external programming was having difficulties.

By comparison, we found that the constrained MLEs computed from the \texttt{ssm} model were close to the true parameter values from which the data were simulated, but they differed from the constrained MLE computed from a grid search of the log-likelihood function computed outside of \texttt{ssm}. Therefore, we believe that our programming of score and log-likelihood computations outside of the MATLAB \texttt{ssm} program are subject to sufficient numerical issues.
that they cannot be used for this example to compute either the score interval or the higher-order asymptotic interval.

As an example, we simulated a dataset \( y \) with \( T = 100 \) observations using true parameter values \( \rho = 0.5 \) and \( \xi = (2.6795, 10, 5) \). Holding \( \rho = 0.5 \) fixed, the constrained MLEs computed from the \textit{ssm} model were \( \hat{\xi}_{\rho=0} = (2.8806, 10.7506, 5.2008) \). The score vector evaluated at this point using our external programming was \( (4.8598, -18.2894, -0.1934) \), and the last two components should be 0 in theory. The constrained MLEs computed from a grid search of the log-likelihood function were \( \hat{\xi}_{\rho=0} = (0.7906, 2.9506, 3.6008) \). The corresponding score vector from our external programming was \( (-0.5652, 0.0771, -0.0322) \). Figure 4.1 shows a surface plot of the log-likelihood function computed using external programming with \( \rho = 0.5 \) fixed for the 2-D tracking model. The blue dot marks the log-likelihood value at \( \hat{\xi}_{\rho=0} = (0.7906, 2.9506, 3.6008) \) computed outside of \textit{ssm} and the red dot marks the log-likelihood value at \( \hat{\xi}_{\rho=0} = (2.8806, 10.7506, 5.2008) \) computed using \textit{ssm}. Judging from this figure, we believe the log-likelihood computations performed using our external programming are subjected to sufficient numerical errors that they cannot be trusted for numerical computations for this particular example.

The discrepancy between our computations external to the \textit{ssm} and those from \textit{ssm} suggests that the computations involved are unstable enough that only results from the more sophisticated program \textit{ssm} can be trusted. Since \textit{ssm} does not provide the ingredients for computing the score vectors and Hessian of the log-likelihood, which are required for constructing the score interval and the higher-order asymptotic interval, we were not able to proceed with these two approaches. Our discussion and presentation therefore center around likelihood intervals and Wald intervals.
4.4.2. Simulation Procedure

We examine the performance of the confidence interval methods in terms of underage, coverage, and overage. We consider simulations with sample size $T = \{60, 120, 240, 480\}$. The number of simulated datasets is 10000 for each $T$ in $\{60, 120, 240, 480\}$. True $\rho$ takes 5 evenly spaced grid values on the closed interval $[0.1, 0.9]$. That is, true values of the interest parameter $\rho = \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The nuisance parameters are $\xi_2 = 10$ and $\xi_3 = 5$. Then we compute and plot the underage, coverage, and overage vs. true $\rho$ for the Wald method and the likelihood ratio method.

4.4.3. Evaluation of the Wald Method and the Likelihood Ratio Method

Table 4.1 summarizes the underage, coverage, and overage simulation results for the two dimensional object tracking model. The total number of simulation for each $T$ is 10000.
Table 4.1: Underage, coverage, and overage simulation results for the two dimensional object tracking model

<table>
<thead>
<tr>
<th>$T$</th>
<th>$\rho$</th>
<th>Wald method</th>
<th>Likelihood ratio method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>underage</td>
<td>coverage</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1378</td>
<td>0.7390</td>
<td>0.1232</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1487</td>
<td>0.7359</td>
<td>0.1154</td>
</tr>
<tr>
<td>60</td>
<td>0.5</td>
<td>0.1707</td>
<td>0.7314</td>
</tr>
<tr>
<td>0.7</td>
<td>0.1792</td>
<td>0.7352</td>
<td>0.0856</td>
</tr>
<tr>
<td>0.9</td>
<td>0.2103</td>
<td>0.7259</td>
<td>0.0638</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1371</td>
<td>0.7392</td>
<td>0.1237</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1491</td>
<td>0.7392</td>
<td>0.1117</td>
</tr>
<tr>
<td>120</td>
<td>0.5</td>
<td>0.1549</td>
<td>0.7379</td>
</tr>
<tr>
<td>0.7</td>
<td>0.1632</td>
<td>0.7346</td>
<td>0.1022</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1997</td>
<td>0.7148</td>
<td>0.0855</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1188</td>
<td>0.7574</td>
<td>0.1238</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1295</td>
<td>0.7543</td>
<td>0.1162</td>
</tr>
<tr>
<td>240</td>
<td>0.5</td>
<td>0.1436</td>
<td>0.7513</td>
</tr>
<tr>
<td>0.7</td>
<td>0.1531</td>
<td>0.7459</td>
<td>0.1010</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1748</td>
<td>0.7224</td>
<td>0.1028</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1238</td>
<td>0.7619</td>
<td>0.1143</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1283</td>
<td>0.7554</td>
<td>0.1163</td>
</tr>
<tr>
<td>480</td>
<td>0.5</td>
<td>0.1380</td>
<td>0.7442</td>
</tr>
<tr>
<td>0.7</td>
<td>0.1431</td>
<td>0.7521</td>
<td>0.1048</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1638</td>
<td>0.7254</td>
<td>0.1108</td>
</tr>
</tbody>
</table>

4.4.3.1. Interpretation of the Simulation Results

Figure 4.2, Figure 4.3, and Figure 4.4 plot the underage, coverage, and overage vs. true $\rho$ for the Wald and Likelihood ratio methods when each of the 10000 simulated datasets has 60 data points. Figure 4.5, Figure 4.6, and Figure 4.7 plot the underage, coverage, and overage vs. true $\rho$ for the two methods when each of the 10000 simulated datasets has 120 data points. Figure 4.8, Figure 4.9, and Figure 4.10 plot the underage, coverage, and overage vs. true $\rho$ for the two methods when each of the 10000 simulated datasets has 240 data points.
Figure 4.11, Figure 4.12, and Figure 4.13 plot the underage, coverage, and overage vs. true \( \rho \) for the two methods when each of the 10000 simulated datasets has 480 data points.

We see that across various values of \( T = \{60, 120, 240, 480\} \), the likelihood ratio method has underage, coverage, and overage that are right on target. For the Wald method, the general pattern is that it tends to have underage and overage that are higher than the target value 0.025, and hence the coverage is lower than the target value 0.95. As a result, for constructing confidence intervals for \( \rho \), we recommend using the likelihood ratio method.

Figure 4.2: Underage for \( \rho \) when datasets have 60 observations

Figure 4.3: Coverage for \( \rho \) when datasets have 60 observations
Figure 4.4: Overage for $\rho$ when datasets have 60 observations

Figure 4.5: Underage for $\rho$ when datasets have 120 observations
Figure 4.6: Coverage for $\rho$ when datasets have 120 observations

Figure 4.7: Overage for $\rho$ when datasets have 120 observations
Figure 4.8: Underage for $\rho$ when datasets have 240 observations

Figure 4.9: Coverage for $\rho$ when datasets have 240 observations
Figure 4.10: Overage for $\rho$ when datasets have 240 observations

Figure 4.11: Underage for $\rho$ when datasets have 480 observations
Figure 4.12: Coverage for $\rho$ when datasets have 480 observations

Figure 4.13: Overage for $\rho$ when datasets have 480 observations


Appendix A

Derivation for the Local Level Model

For the local level model,

\[ y_t = \gamma x_t + \delta \epsilon_t \]
\[ = \gamma (\alpha x_{t-1} + \beta u_t) + \delta \epsilon_t \]
\[ = \gamma (\alpha^2 x_{t-2} + \alpha \beta u_{t-1} + \beta u_t) + \delta \epsilon_t \]
\[ = \gamma (\alpha^t x_0 + \beta (\alpha^{t-1} u_1 + \alpha^{t-2} u_2 + \cdots + \alpha u_{t-1} + u_t) + \delta \epsilon_t \]
\[ = \gamma (\alpha^t x_0 + \beta \sum_{j=1}^{t} u_j \alpha^{t-j}) + \delta \epsilon_t, \quad t = 1, \ldots, T \tag{A.1} \]

When \( \alpha = 1, \gamma = 1 \), then
\[ y_t = x_0 + \beta \sum_{j=1}^{t} u_j + \delta \epsilon_t, \quad t = 1, \ldots, T \]

\[ E(y_t) = E(x_0 + \beta \sum_{j=1}^{t} u_j + \delta \epsilon_t) = E(x_0) = a_0, \quad \text{for all } t = 1, \ldots, T \tag{A.2} \]
\[
\text{cov}(y_i, y_j) = \text{cov}(x_0 + \beta \sum_{k=1}^{i} u_k + \delta \epsilon_i, x_0 + \beta \sum_{m=1}^{j} u_m + \delta \epsilon_j)
\]
\[
= \text{var}(x_0, x_0) + \text{cov}(\beta \sum_{k=1}^{i} u_k, \beta \sum_{m=1}^{j} u_m) + \delta^2 \text{cov}(\epsilon_i, \epsilon_j)
\]
\[
= p_0 + \beta^2 \text{cov}(\sum_{k=1}^{i} u_k, \sum_{m=1}^{j} u_m) + \delta^2 \text{cov}(\epsilon_i, \epsilon_j)
\]
\[
= \begin{cases} 
  p_0 + \beta^2 i, & i < j \\
  p_0 + \beta^2 i + \delta^2, & i = j, \quad i, j = 1, \ldots, T \\
  p_0 + \beta^2 j, & i > j 
\end{cases} \quad (A.3)
\]

We can derive the following expressions

\[
l_1 = \frac{\partial l(\xi)}{\partial \xi_1} = -\frac{1}{2} \text{tr} (\Omega^{-1} \Omega_1) + \frac{1}{2} \text{tr} (\Omega^{-1} \Omega_1 \Omega^{-1} yy^T) \quad (A.4)
\]

where \( \Omega_1 = \partial \Omega / \partial \xi_1 \). Also

\[
l_{12} = \frac{\partial^2 l(\xi)}{\partial \xi_1 \partial \xi_2} \\
= \frac{1}{2} \text{tr} (\Omega^{-1} \Omega_2 \Omega^{-1} \Omega_1 - \Omega^{-1} \Omega_{12}) - \frac{1}{2} \text{tr} \{ \Omega^{-1} (2 \Omega_2 \Omega^{-1} \Omega_1 - \Omega_{12}) \Omega^{-1} yy^T \} \\
= \frac{1}{2} \text{tr} \{ \Omega^{-1} (2 \Omega_2 \Omega^{-1} \Omega_1 - \Omega_{12}) \} - \frac{1}{2} \text{tr} \{ \Omega^{-1} (2 \Omega_1 \Omega^{-1} \Omega_2 - \Omega_{12}) \Omega^{-1} yy^T \} \\
= \frac{1}{2} \text{tr} (\Omega^{-1} \Omega_2 \Omega^{-1} \Omega_1) - \frac{1}{2} \text{tr} \{ \Omega^{-1} \Omega_{12} (I - \Omega^{-1} yy^T) \} - \text{tr} (\Omega^{-1} \Omega_1 \Omega^{-1} \Omega_2 \Omega^{-1} yy^T)
\]

The 2 \times 2 observed information matrix is

\[
\mathbf{j}(\xi) = -\frac{\partial^2 l(\xi)}{\partial \xi \partial \xi^T} = \{ j_{kl} : k, l = 1, 2 \} \quad (A.5)
\]
with
\[ j_{kl} = -\frac{1}{2} \text{tr} \left( \Omega^{-1} \Omega_k \Omega^{-1} \Omega_l \right) + \frac{1}{2} \text{tr} \left\{ \Omega^{-1} \Omega_{kl} \left( I - \Omega^{-1} yy^T \right) \right\} + \text{tr} \left( \Omega^{-1} \Omega_k \Omega^{-1} \Omega_l \Omega^{-1} yy^T \right) \]

The expected value of \( j_{kl} \) is
\[ i_{kl} = E(j_{kl}) = -\frac{1}{2} \text{tr} \left( \Omega^{-1} \Omega_k \Omega^{-1} \Omega_l \right) + \text{tr} \left\{ \Omega^{-1} \Omega_k \Omega^{-1} \Omega_l \Omega^{-1} E(yy^T) \right\} \]
\[ = \frac{1}{2} \text{tr} \left( \Omega^{-1} \Omega_k \Omega^{-1} \Omega_l \right) \]
and \( i_{kl} \) is the \((k,l)\)-element of the Fisher information matrix.

The Fisher information matrix is
\[ i(\xi) = E(j(\xi)) = \{i_{kl} : k, l = 1, 2\} = \left\{ \frac{1}{2} \text{tr}(\Omega^{-1} \Omega_k \Omega^{-1} \Omega_l) : k, l = 1, 2 \right\} \] (A.6)

For \( j(\xi) \), with
\[ j_{kl} = -\frac{1}{2} \text{tr} \left( \Omega^{-1} \Omega_k \Omega^{-1} \Omega_l \right) + \frac{1}{2} \text{tr} \left\{ \Omega^{-1} \Omega_{kl} \left( I - \Omega^{-1} yy^T \right) \right\} + \text{tr} \left( \Omega^{-1} \Omega_k \Omega^{-1} \Omega_l \Omega^{-1} yy^T \right) \]
\[ = \frac{1}{2} \text{tr} \left( \Omega^{-1} \Omega_k \Omega^{-1} \Omega_l \right) + \frac{1}{2} \text{tr} \left\{ \Omega^{-1} \Omega_{kl} \left( \Omega - yy^T \right) \right\} + \text{tr} \left\{ \Omega^{-1} \Omega_k \Omega^{-1} \Omega_l \Omega^{-1} (yy^T - \Omega) \right\} \]
\[ j_{kl} = i_{kl} - \frac{1}{2} \text{tr} \left\{ \Omega^{-1} \left( \Omega_{kl} - 2 \Omega_k \Omega^{-1} \Omega_l \right) \Omega^{-1} (yy^T - \Omega) \right\} \]
so that
\[ D_\xi = \{d_{kl}\} \quad d_{kl} = \frac{1}{2} \text{tr} \left\{ \Omega^{-1} \left( \Omega_{kl} - 2 \Omega_k \Omega^{-1} \Omega_l \right) \Omega^{-1} (yy^T - \Omega) \right\} \]
\[ j(\xi) = i(\xi) - D_\xi \]

This is expression 7.27 from [Butler, 2007]. To see that note
\[ -\frac{1}{2} \frac{\partial \Omega^{-1}}{\partial \xi_k} = \frac{1}{2} \Omega^{-1} \Omega_k \Omega^{-1} \]
so that

\[-\frac{1}{2} \frac{\partial^2 \Omega^{-1}}{\partial \xi_k \partial \xi_l} = \frac{1}{2} \left( \Omega^{-1} \Omega_{kl} \Omega^{-1} - \Omega^{-1} \Omega_k \Omega^{-1} \Omega_l \Omega^{-1} - \Omega^{-1} \Omega_l \Omega^{-1} \Omega_k \Omega^{-1} \right)\]

\[= \frac{1}{2} \Omega^{-1} \left( \Omega_{kl} - \Omega_k \Omega^{-1} \Omega_l - \Omega_l \Omega^{-1} \Omega_k \right) \Omega^{-1}\]

and

\[(D_{\xi})_{kl} = \sum_{ij} (y_i y_j - \omega_{ij}) \left(-\frac{1}{2}\right) \frac{\partial^2 \omega_{ij}}{\partial \xi_k \partial \xi_l}\]

\[= \frac{1}{2} \text{tr} \left\{ (yy^T - \Omega) \Omega^{-1} (\Omega_{kl} - \Omega_k \Omega^{-1} \Omega_l - \Omega_l \Omega^{-1} \Omega_k) \Omega^{-1} \right\}\]

\[= \frac{1}{2} \text{tr} \left\{ \Omega^{-1} (\Omega_{kl} - 2\Omega_k \Omega^{-1} \Omega_l) \Omega^{-1} (yy^T - \Omega) \right\}.\]
Appendix B
More Scatter Plots and Histograms for the Local Level Model

Figure B.1: A scatter plot of MLEs overlaid with the 95% predictive interval, $T = 120, \beta = 0.2$

Figure B.2: Histogram of standard error of $\hat{\beta}$, where $se(\hat{\beta}) < 30, T = 120, \beta = 0.2$. 
Figure B.3: Histogram of standard error of $\hat{\delta}$, where $se(\hat{\delta}) < 30, T = 120, \beta = 0.2$.

Figure B.4: Histogram of standard error of $\hat{\delta}_0$, where $se(\hat{\delta}_0) < 30, T = 120, \beta = 0.2$. 
Figure B.5: Histogram of standard error of $\hat{\beta}$, where $se(\hat{\beta}) < 30, T = 120, \beta = 2.8$.

Figure B.6: Histogram of standard error of $\hat{\delta}$, where $se(\hat{\delta}) < 30, T = 120, \beta = 2.8$. 
Figure B.7: A scatter plot of MLEs overlaid with the 95\% predictive interval, $T = 120, \beta = 2.8$.

Figure B.8: Histogram of standard error of $\hat{\delta}_0$, where $se(\hat{\delta}_0) < 30, T = 120, \beta = 2.8$.  

Figure B.9: A scatter plot of MLEs overlaid with the 95% predictive interval, $T = 240, \beta = 0.2$

Figure B.10: Histogram of standard error of $\hat{\beta}$, where $se(\hat{\beta}) < 30$, $T = 240, \beta = 0.2$. 
Figure B.11: Histogram of standard error of $\hat{\delta}$, where $se(\hat{\delta}) < 30, T = 240, \beta = 0.2$.

Figure B.12: Histogram of standard error of $\hat{\delta}_0$, where $se(\hat{\delta}_0) < 30, T = 240, \beta = 0.2$. 
Figure B.13: A scatter plot of MLEs overlaid with the 95\% predictive interval, \( T = 240, \beta = 2.8 \).

Figure B.14: Histogram of standard error of \( \hat{\beta} \), where \( se(\hat{\beta}) < 30 \), \( T = 240, \beta = 2.8 \).
Figure B.15: Histogram of standard error of $\hat{\delta}$, where $se(\hat{\delta}) < 30, T = 240, \beta = 2.8$.

Figure B.16: Histogram of standard error of $\hat{\delta}_0$, where $se(\hat{\delta}_0) < 30, T = 240, \beta = 2.8$. 
Figure B.17: A scatter plot of MLEs overlaid with the 95% predictive interval, $T = 480, \beta = 0.2$

Figure B.18: Histogram of standard error of $\hat{\beta}$, where $se(\hat{\beta}) < 30$, $T = 480, \beta = 0.2$. 
Figure B.19: Histogram of standard error of $\hat{\delta}$, where $se(\hat{\delta}) < 30$, $T = 480$, $\beta = 0.2$.

Figure B.20: Histogram of standard error of $\hat{\delta}_0$, where $se(\hat{\delta}_0) < 30$, $T = 480$, $\beta = 0.2$. 

93
Figure B.21: A scatter plot of MLEs overlaid with the 95% predictive interval, $T = 480, \beta = 2.8$

Figure B.22: Histogram of standard error of $\hat{\beta}$, where $se(\hat{\beta}) < 30$, $T = 480, \beta = 2.8$. 
Figure B.23: Histogram of standard error of \( \hat{\delta} \), where \( se(\hat{\delta}) < 30, T = 480, \beta = 2.8 \).

Figure B.24: Histogram of standard error of \( \hat{\delta}_0 \), where \( se(\hat{\delta}_0) < 30, T = 480, \beta = 2.8 \).
Figure B.25: A scatter plot of MLEs overlaid with the 95% predictive interval, \( T = 960, \beta = 0.2 \).

Figure B.26: Histogram of standard error of \( \hat{\beta} \), where \( se(\hat{\beta}) < 30 \), \( T = 960, \beta = 0.2 \).
Figure B.27: Histogram of standard error of $\hat{\delta}$, where $se(\hat{\delta}) < 30, T = 960, \beta = 0.2$.

Figure B.28: Histogram of standard error of $\hat{\delta}_0$, where $se(\hat{\delta}_0) < 30, T = 960, \beta = 0.2$. 
Figure B.29: A scatter plot of MLEs overlaid with the 95% predictive interval, $T = 960, \beta = 2.8$.

Figure B.30: Histogram of standard error of $\hat{\beta}$, where $se(\hat{\beta}) < 30, T = 960, \beta = 2.8$. 

98
Figure B.31: Histogram of standard error of $\hat{\delta}$, where $se(\hat{\delta}) < 30$, $T = 960$, $\beta = 2.8$.

Figure B.32: Histogram of standard error of $\hat{\delta}_0$, where $se(\hat{\delta}_0) < 30$, $T = 960$, $\beta = 2.8$. 
APPENDIX C
Score Test Statistic Is Invariant to the Parameterization

The score test statistic is invariant to the parameterization. We can show that the score in the $\xi$ parameterization equals to the score in the $\xi_\rho$ parameterization. That is,

$$s_T^{i\rho} s_\rho = s_T^{i\rho} s$$  \hspace{1cm} (C.1)

where $s_\rho = \partial l / \partial \xi_\rho$, $s = \partial l / \partial \xi$.

Proof:

$$\xi = \begin{pmatrix} \xi_1 = \frac{\xi_2}{\rho} (1 - \sqrt{1 - \rho^2}) \\ \xi_2 \\ \xi_3 \end{pmatrix}, \quad \xi_\rho = \begin{pmatrix} \rho = \frac{2\xi_1 \xi_2}{\xi_1 + \xi_2} \\ \xi_2 \\ \xi_3 \end{pmatrix}$$  \hspace{1cm} (C.2)

$$i = -E \left( \frac{\partial^2 l}{\partial \xi \partial \xi^T} \right)$$  \hspace{1cm} (C.3)

$$i_\rho = -E \left( \frac{\partial^2 l}{\partial \xi_\rho \partial \xi_\rho^T} \right)$$  \hspace{1cm} (C.4)

It is a fact that

$$i_\rho = \left( \frac{\partial \xi^T}{\partial \xi_\rho} \right) i \left( \frac{\partial \xi}{\partial \xi_\rho} \right) = H^T H$$  \hspace{1cm} (C.5)
where

\[
H = \frac{\partial \xi^T}{\partial \xi_\rho} = \begin{pmatrix}
\xi_2 \left( \frac{1}{\sqrt{1-\rho^2}} + \frac{\sqrt{1-\rho^2}}{\rho^2} - \frac{1}{\rho^2} \right) & 0 & 0 \\
\frac{1}{\rho} (1 - \sqrt{1-\rho^2}) & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]  \quad (C.6)

Using the chain rule,

\[
s_\rho = \frac{\partial l}{\partial \xi_\rho} = \frac{\partial \xi^T}{\partial \xi_\rho} \frac{\partial l}{\partial \xi} = H s
\]  \quad (C.7)

Now, we will prove that

\[
i_\rho = \left( \frac{\partial \xi^T}{\partial \xi_\rho} \right) i \left( \frac{\partial \xi}{\partial \xi_\rho} \right) = HiH^T
\]  \quad (C.8)

From \( s_\rho = Hs \), taking derivatives with respect to the \( i \)-th component of \( \xi_\rho \), we have

\[
\frac{\partial s_\rho}{\partial (\xi_\rho)_i} = H \frac{\partial^2 l}{\partial \xi \partial (\xi_\rho)_i} + \frac{\partial H}{\partial (\xi_\rho)_i} s
\]  \quad (C.9)

\[
- E \left\{ \frac{\partial s_\rho}{\partial (\xi_\rho)_i} \right\} = - E \left\{ H \frac{\partial^2 l}{\partial \xi \partial \xi^T} \frac{\partial \xi}{\partial (\xi_\rho)_i} \right\} + 0 = Hi \times (i\text{th column of } H^T)
\]  \quad (C.10)

Thus,

\[
i_\rho = - E \left\{ \frac{\partial s_\rho}{\partial \xi_\rho} \right\} = H i H^T
\]  \quad (C.11)

Next, it follows that

\[
s_\rho^T i_\rho^{-1} s_\rho = s^T H^T (HiH^T)^{-1} H s = s^T i^{-1} s
\]  \quad (C.12)

This holds for all values of \( \xi_\rho \), including the constrained MLE \( \hat{\xi}_{\rho_0} \).


MathWorks (2023c). What are state-space models? 47


