A Study of Relation Between Robustness of Parameter Estimation and Model Order in System Identification

Mitsuru Matsubara and Sueo Sugimoto
Dept. of Electrical and Electronic Engineering, Ritsumeikan University
1-1-1 Noji-Higashi, Kusatsu City, Shiga 525-8577, Japan
E-mail: mbv707fs@hotmail.com

Abstract
In this paper, for the purpose of evaluating the degradation level of the performance in application with the parametric model in which the estimated parameters have the disturbance, we consider the relation between the robustness of parameter estimation and the model order. After the application performance cost function is derived, the gradient at the estimated parameter is focused to evaluate the disturbance level of the performance in the application. As a result, in this paper, it is shown that the robustness of parameter estimation is given as 2nd order function of the model order.

1 Introduction
In application using parametric model, to design the application which makes the performance the best, a number of the parametric models are generally estimated from two or more data sets. When the applications are respectively designed to those estimated parametric models, the disturbance of the performance of those applications arises due to the disturbance of those data sets (Fig. 1). In the case where the disturbance of the performance is small, then any parametric model can be employed to design the application. However, in the case where the disturbance is large, then the application performance by a parametric model may not be enough to various data sets. In this paper, the robustness of the estimated parameter of the parametric model shall be defined as the degradation level of the performance by such disturbance. Then the goal of this study is to clarify what is the factor that makes the disturbance large, and the purpose of this paper is to clarify that one of the factor is the model order.

Fig. 1: Disturbance of the application performance in the design process

2 Problem Settings
Consider the discrete-time causal linear time-invariant system:
\[ y(t) = G_0(q)u(t) + w(t) \]  \hspace{1cm} (1)
where \( y, u, w \) are the output, the input and the noise respectively, and \( q \) denotes the delay operator. In this paper, it is assumed that the objective system can be modeled as follows:
\[ y(t) = G(q, \theta)u(t) + w(t), \hspace{1cm} t = 1, 2, \cdots, N, \]  \hspace{1cm} (2)
where \( \theta \in \mathbb{R}^n (\theta \subset \Theta) \) denotes the model parameter, \( n \) denotes the model order and \( N \) denotes the sample number.

The application performance cost function to the estimated parametric model is defined as follows:

\[
V(\theta) = [F(\theta) - F(\theta_0)]^2, \tag{3}
\]

where \( \theta_0 \) is true parameter vector and \( F(\cdot) \) is the arbitrary scalar application cost function. Since \( V(\theta) \geq 0 \) with minimum \( V(\theta_0) = 0 \), \( \theta = \theta_0 \) is the optimal condition in the application.

Example:

Consider the state feedback \( \mathcal{H}_2 \) control problem to the following continuous-time plant model \( \mathcal{M} \):

\[
\mathcal{M}(\theta) : G(s, \theta) = C(sI - A)^{-1}B \tag{4}
\]

where \( A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times 1}, C \in \mathbb{R}^{1 \times n}, s \) denotes Laplace transform variable, and it is assumed that \((A, B)\) and \((A, C)\) are stabilizable and detectable, respectively. Let \( X = X^T \geq 0 \) be the stabilizing solution to the algebraic Riccati equation:

\[
A^T X + XA - XBB^T X + C^T C = 0. \tag{5}
\]

Since the solution becomes the function of \( \theta \), let \( X = X(\theta) \). Then the optimal state feedback gain \( K_{\text{opt}} \) is given as:

\[
K_{\text{opt}} = -B^T X(\theta), \tag{6}
\]

the application performance cost function is also defined as the best achievable control performance function:

\[
F(\theta) = H(C(\mathcal{M}(\theta)), \mathcal{M}(\theta)) = \text{tr} \{BX^T(\theta)B\}, \tag{7}
\]

where \( \text{tr}\{\cdot\} \) denotes trace of matrix \{\cdot\} and \( C(\mathcal{M}(\theta)) \) denotes the controller which is designed based on the plant model \( \mathcal{M}(\theta) \). In this example, since the controller which is designed to \( \mathcal{M}(\theta_0) \) can derive the best performance to the objective system, the parameter \( \theta \) which minimizes \( V(\theta) \) is \( \theta_0 \).

In (3), the more the norm of the gradient on \( V(\hat{\theta}_N) \) at a estimated parameter \( \hat{\theta}_N \) is large, the more the application performance cost with \( \hat{\theta}_N \) increases, and the more the disturbance of the application performance cost when the estimated parameter \( \hat{\theta}_N \) disturs becomes large. This means that when the norm of the gradient on \( V(\hat{\theta}_N) \) at a estimated parameter \( \hat{\theta}_N \) is large, the models with such disturbed estimates \( \hat{\theta}_N \) lack the robustness in the application. For this reason, in this paper, the robustness of the estimated parameter shall be defined by the norm of the gradient on \( V(\hat{\theta}_N) \) at a estimated parameter \( \theta_0 \).

### 3 Assumptions

Here, the following assumptions are introduced.

(i) The model with the true parameter \( \theta_0 \) sets the application cost to the best.

(ii) The application cost \( F(\theta) \) is monotonically increase by increasing of identification cost.

(iii) The identification cost is defined as log-likelihood.

From (iii), \( \hat{\theta}_N \) means maximum likelihood(ML) estimate. In this paper, from (ii) and (iii), we define \( F(\cdot) \) as \( \text{NE}_{G(z)}\{\log f(Z|\theta)\} \) where \( Z = \{z\} \) denotes unknown future data set. That is, \( F(\theta) \) is defined the average of the log-likelihood to the unknown future data set. Based on these assumptions, the degradation level of the application performance can be evaluated by the degradation level of the identification cost.

### 4 The relation between the robustness of the estimated parameter and the model order

Consider a second order Taylor series approximation to \( \text{E}_{G(z)}\{\log f(Z|\theta_N)\} \) around \( \hat{\theta}_N \):

\[
\text{E}_{G(z)}\{\log f(Z|\hat{\theta}_N)\} = \text{E}_{G(z)}\{\log f(Z|\theta_0)\} - \frac{1}{2}(\hat{\theta}_N - \theta_0)^T J(\theta_0)(\hat{\theta}_N - \theta_0) + f(\hat{\theta}_N), \tag{8}
\]

\[
J(\theta) = -\int g(z) \frac{\partial^2 \log f(z|\theta)}{\partial \theta \partial \theta^T} dz. \tag{9}
\]

Then, from (3) and (8), \( V \) with the ML estimate \( \hat{\theta}_N \) is approximately:

\[
V(\hat{\theta}_N) = \left[ \text{NE}_{G(z)}\{\log f(Z|\theta)\} - \text{NE}_{G(z)}\{\log f(Z|\theta_0)\} \right]^2 \geq 0
\]

\[
= \frac{N^2}{4} (\hat{\theta}_N - \theta_0)^T J(\theta_0)(\hat{\theta}_N - \theta_0)^T, \tag{10}
\]

where \( g(z) \) is true distribution which generates the measured data set \( W = \{w(t) = y(t) - G_0(q)u(t)\} \).

Our purpose is to evaluate the norm of the gradient of \( V(\theta) \) at \( \theta = \hat{\theta}_N \) to evaluate the robustness of the estimated parameter.

Let \( L(\theta) \) be tangent of \( V(\theta) \) at \( \theta_N \), then \( L(\theta) \) is given by:

\[
L(\theta) = \frac{dV(\theta)}{d\theta^T} \bigg|_{\theta = \hat{\theta}_N} (\theta - \hat{\theta}_N) + V(\hat{\theta}_N), \tag{11}
\]

\[
\frac{dV(\theta)}{d\theta^T} = N^2(\theta - \theta_0)^T J(\theta_0)(\theta - \theta_0)^T J(\theta_0)(\theta - \theta_0). \tag{12}
\]

Consider the intercept of \( L(\theta) \) at \( \theta_0 \), then from (10), (11) and (12), \( L(\theta_0) \) is given as follows:

\[
L(\theta_0) = \frac{dV(\theta)}{d\theta^T} \bigg|_{\theta = \hat{\theta}_N} (\theta_0 - \hat{\theta}_N) + V(\hat{\theta}_N)
\]
where 

In Fig. 2, it can be observed that when a point \( \hat{\theta}_N \) is fixed, the more \( L(\theta_0) \) is small, the more the gradient of \( V(\theta) \) at \( \hat{\theta}_N \) becomes large. This means that when a point \( \theta_N \) is fixed, the norm of the gradient of \( V(\theta) \) at \( \hat{\theta}_N \) can be evaluated by using \( V(\hat{\theta}_N) \). Therefore, the robustness of the estimated parameter \( \theta_N \) in the application shall be evaluated by \( V(\hat{\theta}_N) \).

![Fig. 2: Tangent \( L(\theta) \) and the intercept \( L(\theta_0) \)](image)

To evaluate the gradient by using \( V(\hat{\theta}_N) \), the estimated parameter \( \hat{\theta}_N \) has to be fixed because the estimated \( \hat{\theta}_N \) is the function of the data set which be observed by the present time from the past, i.e. \( \theta_N(w) \), and also the estimated \( \hat{\theta}_N(w) \) actually has the disturbance.

Thus, \( V(\hat{\theta}_N) \) shall be re-defined as follows:

\[
V_{ave}(\hat{\theta}_N) = \mathbb{E}_{G(w)} \left\{ V \left( \hat{\theta}_N(w) \right) \right\},
\]

where \( \mathbb{E}_{G(w)} \{ \cdot \} \) denotes the average to the measured data set \( W = \{w\} \).

In this paper, to derive the relation between \( V_{ave}(\hat{\theta}_N) \) and the model order, we attempt two approaches.

### 4.1 Approach 1

It is known that \( \sqrt{N} \Delta \theta \equiv \sqrt{N}(\hat{\theta}_N - \theta_0) \) has asymptotic distribution[5]:

\[
\sqrt{N} \Delta \theta \equiv \sqrt{N}(\hat{\theta}_N - \theta_0) \sim N_0(0, I(\theta_0)), \quad N \to \infty.
\]

where \( I(\theta_0) \) denotes Fisher’s information matrix:

\[
I(\theta) = \int g(w) \frac{\partial \log f(w|\theta)}{\partial \theta} \frac{\partial \log f(w|\theta)}{\partial \theta^T} dw.
\]

From (10) and (16), \( V_{ave}(\hat{\theta}_N) \) is also given by:

\[
V_{ave}(\hat{\theta}_N) = \frac{1}{4} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} J_{ij}(\theta_0) J_{kl}(\theta_0)
\]

\[
\times E_{G(w)} \left\{ \sqrt{N} \Delta \theta_i \sqrt{N} \Delta \theta_j \sqrt{N} \Delta \theta_k \sqrt{N} \Delta \theta_l \right\}
\]

From (15), since \( \sqrt{N} \Delta \theta_i \) for \( \forall i \) are asymptotically jointly normal with zero, 4th order moment[4] in (17) is given by:

\[
E_{G(w)} \left\{ \sqrt{N} \Delta \theta_i \sqrt{N} \Delta \theta_j \sqrt{N} \Delta \theta_k \sqrt{N} \Delta \theta_l \right\} =
\]

\[
E_{G(w)} \left\{ \sqrt{N} \Delta \theta_i \sqrt{N} \Delta \theta_j \right\} E_{G(w)} \left\{ \sqrt{N} \Delta \theta_k \sqrt{N} \Delta \theta_l \right\}
\]

Consequently, from (18), \( V_{ave}(\hat{\theta}_N) \) becomes as follows:

\[
V_{ave}(\hat{\theta}_N) = \frac{3}{4} \left( \text{tr} \left\{ I(\theta_0) J^{-1}(\theta_0) \right\} \right)^2 - h(I, J),
\]

where \( h \) denotes the function of \( I(\theta_0) \) and \( J(\theta_0) \). If \( \theta_0 \) which satisfies \( g(w) = f(w|\theta_0) \) exists in \( \Theta \), the relation \( I(\theta_0) = J(\theta_0) \) is obtained[3]. Then, \( V_{ave}(\hat{\theta}_N) \) becomes as follows:

\[
V_{ave}(\hat{\theta}_N) = \frac{3}{4} n^2 - h(I, J),
\]

where

\[
\begin{align*}
A_i & = \text{det} \left[ \left( \begin{array}{cc} j & k \\ i & l \end{array} \right) \right] \\
& \equiv A_{ij} a_{ik} a_{jk} a_{kl} \in \mathbb{R}^{2 \times 2}.
\end{align*}
\]

\[h(I, J) = 2 \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} \text{det} \left[ \Gamma \left( \begin{array}{cc} j & k \\ i & l \end{array} \right) \right] \cdot J(\theta_0) \left( \begin{array}{cc} j & k \\ i & l \end{array} \right)
\]
distribution with n degrees of freedom is n and 2n respectively. Therefore, from (10), the following equation is derived:

\[
V_{\text{ave}}(\hat{\theta}_N) = \frac{1}{4} \left( \text{Var}_{G(w)}(V(\hat{\theta}_N)) + (E_{G(w)}\{V(\hat{\theta}_N)\})^2 \right)
\]

\[
= \frac{1}{4} n(n + 2).
\]  

(25)

4.3 The relation between \(V_{\text{ave}}(\hat{\theta}_N)\) and the model order \(n\)

From (25), \(h(I, J)\) may be certainly expressed as follows:

\[
h(I, J) = \frac{1}{2} n(n - 1).
\]  

(26)

In (26), in the case where \(n\) is 1 and 2, then \(h(I, J)\) respectively becomes 0 and 1 as well as (21).

From (20) and (25), the fact that the more model order is large, the more the application performance cost function \(V_{\text{ave}}(\hat{\theta}_N)\) monotonically increase is derived. In other words, the more model order is large, the more the disturbance level of the application performance with the model becomes large. Even if the model order can minimize AIC, there is no change in this fact. Because as the similar form of \(V_{\text{ave}}(\hat{\theta}_N)\) in (14), the estimated bias \(B(= n)\) of the log-likelihood corrected by AIC is given as follows:

\[
B = 2E_{G(w)} \left\{ \sqrt{V(\hat{\theta}_N(w))} \right\},
\]  

(27)

where the penalty term of AIC is 2n[3].

5 The Similar Representations

As the relation between the robustness of the estimated parameter in the application and the model order, two similar formulas are already proposed at least. The one is given as follows[2]:

\[
\text{Var} \left\{ G(e^{j\omega}, \hat{\theta}_N) \right\} \approx n \frac{\Phi_u(e^{j\omega})}{N\Phi_u(e^{j\omega})},
\]  

(28)

where \(\Phi_u(e^{j\omega})\) and \(\Phi_u(e^{j\omega})\) are noise power spectrum and input power spectrum respectively. This means that the more the model order \(n\) is large, the more the disturbance level of the estimated model \(G(e^{j\omega}, \hat{\theta}_N)\) increases. Therefore, if the more \(G(e^{j\omega}, \hat{\theta}_N)\) is disturbed, the more the disturbance level of the application may be increased.

The other is given as follows:

\[
\frac{1}{2\pi} \int_{-\pi}^{\pi} N\Phi_u(e^{j\omega}) \text{Var} \left\{ G(e^{j\omega}, \hat{\theta}_N) \right\} d\omega = n \text{Var} \{ w(t) \}
\]  

(29)

There is a water-bed effect for the variance of \(G(e^{j\omega}, \hat{\theta}_N)\), i.e. if the variance is made small in some frequency regions it must be large in another region to correct the inequality. In the case where the input power spectrum is invariant, this means that the more model order \(n\) is large, the more the disturbance level of the estimated model \(G(e^{j\omega}, \hat{\theta}_N)\) is increased. Therefore, if the more \(G(e^{j\omega}, \hat{\theta}_N)\) is disturbed, the more the disturbance level of the application may be increased.

If the model order \(n\) is fixed in (29), the method minimizing the model variance \(\text{Var}[G(e^{j\omega}, \hat{\theta}_N)]\) in the intended frequency regions, which is related to the disturbance level of the application, is to only design the input \(\Phi_u(e^{j\omega})\) appropriately. From such a viewpoint, the method which designs the optimal input in system identification for control design is proposed in [6, 7].

6 Conclusions

In this paper, for the purpose of evaluating the degradation level of the performance in application with the parametric model in which the estimated parameters have the disturbance, we considered the relation between the robustness of the parameter estimation and the model order. After the application performance cost function was derived, the gradient at the estimated parameter was focused to evaluate the disturbance level of the performance in the application. As a result, in this paper, it was shown that the robustness of parameter estimation is given as 2nd order function of the model order. This result derived the fact that even if the model order can minimize AIC, the more the model order is large, the more the disturbance level of the application performance becomes large. In future work, we will verify the fact based on numerical analysis.

References


— 123 —

