Notes on Estimation Theory

Gyubeom Edward Im*

February 11, 2024

Contents

*blog: alida.tistory.com, email: gyurse@gmail.com

1 Introduction

추정 이론(estimation theory)는 관측된 데이터를 바탕으로 모델의 파라미터나 상태를 예측하는 다양한 방법을 정리한 이론이다. 이는 데이터 분석, 신호처리, 기계학습, 금융, 로봇공학 등 다양한 분야에서 널리 쓰이고 있으며 주로 불확실성을 다루는 과정에서 정확한 결정을 내리기 위한 필수적인 도구로 사용된다. 추정 이론의 응용 분야는 매우 넓은데 통신에서는 신호의 품질을 추정하거나 기계학습에서는 데이터로부터 알고리즘의 파라미터를 결정하는데 사용된다. 또한 금융 분야에서는 시장의 미래 동향을 예측하기 위한 변수를 추정하는데 필수적으로 사용되고 있다.

1.1 The Mathematical Estimation Problem

좋은 추정값(estimator)을 얻기 위해서는 우선 수학적으로 데이터를 잘 모델링해야 한다. 데이터는 랜덤성을 띄기 때문에 확률 밀도 함수(probability density function, pdf) $p(x[0], x[1], \cdots, x[N-1]; \theta)$를 사용하여 데이터를 표현한다. 이 때 $x[n]$은 N개의 데이터를 의미하며 θ는 미지의 모델 파라미터를 의미한다. 만약 $N = 1$인 경우 pdf는 아래와 같이 나타낼 수 있다.

$$\boxed{p(x[0];\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[0]-\theta)^2\right]} \tag{1}$$

- $p(x; A)$: 확률 분포가 파라미터 A에 의해 정의됨. (pdf of x paramterized by A)

위 식에서 보다시피 파라미터 θ는 $x[0]$의 확률에 직접적인 영향을 주기 때문에, 역으로 $x[0]$의 값을 보고 θ를 추정하는 것이 가능하다.

예를 들면 아래와 같은 그림이 주어졌다고 했을 때 $x[0]$ 값이 음수가 관측되면 우리는 $\theta = \theta_2$보다는 $\theta = \theta_1$이라고 일반적으로 유추할 수 있다. 하지만 실제 문제에서는 위와 같은 pdf는 주어지지 않기 때문에 데이터 x와 파라미터 θ의 관계를 정의해야 한다. 랜덤한 노이즈를 $w[n]$라고 했을 때 둘 사이의 관계는 아래와 같이 정의할 수 있다.

$$x[n] = A + Bn + w[n] \qquad n = 0, 1, \cdots, N-1 \tag{2}$$

일반적으로 $w[n]$는 평균이 0인 white gaussian noise(WGN)으로 설정한다. 이 때, $\theta = [A, B]^\intercal$는 미지의 모델 파라미터를 말한다.

$$w[n] \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

- $a \sim \mathcal{N}(\mu, \sigma^2)$: 확률변수 a가 평균이 μ이고 분산이 σ^2인 가우시안 분포를 따른다.

N개의 데이터를 벡터화하여 $\mathbf{x} = [x[0], x[1], \cdots, x[N-1]]^\intercal$라고 표현하면 $\mathbf{x}$에 대한 pdf는 다음과 같다.

$$\boxed{p(\mathbf{x};\theta) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}(x[n]-A-Bn)^2\right]} \tag{4}$$

pdf에 기반한 (4)과 같은 추정은 고전적인 추정 방법으로써 파라미터 θ를 우리가 모르지만 고정된 상수로 보는 빈도주의(frequentist) 관점으로 해석될 수 있다. 이와 달리 현대적인 추정 방법은 파라미터 θ 또한 별도의 확률변수로 해석하는 방법을 베이지안(bayesian) 관점을 주로 사용한다.

$$\begin{aligned} &\text{Frequentist:} \quad \underbrace{x[n]}_{\text{r.v.}} = \underbrace{\theta}_{\text{deterministic}} + w[n] \\ &\text{Bayesian:} \quad \underbrace{x[n]}_{\text{r.v.}} = \underbrace{\theta}_{\text{r.v.}} + w[n] \end{aligned} \tag{5}$$

베이지안 관점에서는 데이터 $\mathbf{x}$와 파라미터 θ가 둘 다 확률변수이므로 둘 사이의 결합 확률 분포(joint pdf)을 사용하여 확률을 표현한다.

$$p(\mathbf{x}, \theta) = p(\mathbf{x}|\theta)p(\theta) \tag{6}$$

- $p(\mathbf{x}, \theta)$: $jointpdf$
$-$p(x | θ) : conditional pdf : θ를 알고 있는 상태에서 데이터 $\mathbf{x}$에 대한 우리의 지식
- $p(\theta)$: prior pdf: 어떤 데이터 $\mathbf{x}$가 관측되기 전 θ에 대한 우리의 경험, 지식

1.2 Assessing Estimator Performance

위와 같이 100일간 측정한 노이즈를 포함한 몸무게 데이터가 주어졌다고 가정하자. 위 데이터는 아래와 같이 모델링할 수 있다.

$$x[n] = A + w[n] \tag{7}$$

- $w[n]$: 평균이 0인 노이즈
- $x[n]$: 측정된 데이터
- A : 추정하고자 하는 파라미터

일반적으로 A를 다음과 같이 데이터들의 평균으로 추정하는 것이 합리적일 것이다.

$$\boxed{\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]} \tag{8}$$

- $\hat{A}$: A의 추정값 1

여기에 다음과 같은 질문을 할 수 있다.

- $\hat{A}$는 실제 A와 얼마나 가까울까?
- 평균말고 더 좋은 추정 방법은 없을까?

다음과 같이 다른 추정 방법을 사용하여 A를 추정할 수 있다.

$$\boxed{\breve{A} = x[0]} \tag{9}$$

- $\breve{A}$: A의 추정값 2

직관적으로 우리는 모든 데이터(또는 무한개의 데이터)를 관측하는게 아닌 이상 좋은 추정값을 얻는 것이 어렵다는 것을 알 수 있다. 실제로 $\breve{A} = 69.8$이고 $\hat{A} = 71.1$이어서 $\breve{A}$가 $A = 70$에 더 가깝다. 이런 경우 $\breve{A}$가 더 좋은 추정값이라고 볼 수 있을까? 당연히 아니다.

추정값(estimator) $\hat{A}$는 확률변수 $x[n]$에 대한 함수이므로 $\hat{A}$ 역시 확률변수가 된다. 따라서 추정값 역시 노이즈로 인해 다양한 결과물들을 도출할 수 있다. $\breve{A}$가 A에 더 가깝다는 사실은 주어진 $x[n]$의 예제에 대해서만을 의미한다. 따라서 추정값의 성능을 평가하기 위해서는 반드시 통계적으로 접근해야 한다. 예를 들어 여러번의 실험을 통해 데이터를 수집하고 이를 반복적으로 추정하는 방법이 존재한다.

데이터를 여러번 수집한 후 추정값 $\hat{A}, \breve{A}$의 기대값(expectation)을 계산하면 다음과 같다.

$$\begin{aligned}
\mathbb{E}(\hat{A}) &= \mathbb{E}\left(\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right) \\
&= \frac{1}{N}\sum_{n=0}^{N-1} \mathbb{E}(x[n]) \\
&= \frac{1}{N}\sum_{n=0}^{N-1} A \\
&= A
\end{aligned} \tag{10}$$

$$\begin{aligned}
\mathbb{E}(\breve{A}) &= \mathbb{E}(x[0]) \\
&= A
\end{aligned} \tag{11}$$

따라서 두 추정값의 성능은 동일한 것일까? $\hat{A}$가 $\breve{A}$보다 더 좋은 추정값임을 증명하기 위해서는 추정의 분산이 더 작음을 입증해야 한다.

$$
\begin{aligned}
\mathrm{var}(\hat{A}) &= \mathrm{var}\left(\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right) \\
&= \frac{1}{N^2}\sum_{n=0}^{N-1} \mathrm{var}(x[n]) \\
&= \frac{1}{N^2} N\sigma^2 \\
&= \frac{\sigma^2}{N}
\end{aligned} \tag{12}
$$

$$
\begin{aligned}
\mathrm{var}(\breve{A}) &= \mathrm{var}(x[0]) \\
&= \sigma^2 \\
&> \mathrm{var}(\hat{A})
\end{aligned} \tag{13}
$$

따라서 이를 통해 얻을 수 있는 결론은 다음과 같다.

추정값(estimator)는 확률변수(random variable)이다. 따라서 추정값의 성능은 반드시 통계적 방법이나 pdf를 통해 판단되어야 한다. 컴퓨터 시뮬레이션을 사용하여 추정값을 평가하는 방법은 파라미터에 대한 통찰을 얻기엔 충분히 좋지만 이를 절대적인 값으로 해석하면 안된다. 운이 좋은 경우 추정값은 소수점 오차를 가진 정확도로 구할 수 있지만 운이 나쁜 경우에는(데이터가 부족하거나 에러 값이 들어 있거나) 잘못된 추정값을 얻을 수 있다.

2 Minimum Variance Unbiased Estimation

본 섹션에서 나오는 추정값들은 고전적인 빈도주의 관점에서 파라미터 θ가 고정된 값으로 주어졌다고 가정한다.

2.1 Unbiased Estimators

추정값이 불편(unbiased)되었다라는 의미는 추정값의 평균이 파라미터의 참 값과 동일하다는 의미와 동치이다. 일반적으로 파리미터는 특정 범위 $a < \theta < b$ 안에 존재하므로 불편추정값(unbiased estimator, 또는 불편추정량)이란 다음과 같이 수학적으로 정의할 수 있다.

$$
\mathbb{E}(\hat{\theta}) = \theta \tag{14}
$$

만약 추정값이 편향(biased)되어 있다면 $\mathbb{E}(\hat{\theta}) \neq \theta$이고 편향 $b(\theta)$은 다음과 같이 계산할 수 있다.

$$
\boxed{b(\theta) = \mathbb{E}(\hat{\theta}) - \theta} \tag{15}
$$

2.1.1 Example 2.1 and Example 2.2

다음과 같은 데이터가 주어졌다고 하자.

$$
x[n] = A + w[n] \quad n = 0, 1 \cdots, N-1 \tag{16}
$$

- A: 추정하고자 하는 파라미터, $-\infty < A < \infty$
- $w[n]$: WGN

이에 대한 일반적인 추정값 $\hat{A}$는 다음과 같이 데이터의 평균으로 예측할 수 있다.

$$
\hat{A} = \frac{1}{N}\sum_{n=0}^{N-1} x[n] \tag{17}
$$

선형성에 의해 기대값(expectation)은 다음과 같이 정의된다.

$$
\begin{aligned}
\mathbb{E}(\hat{A}) &= \mathbb{E}\left(\frac{1}{N}\sum_{n=0}^{N-1} x[n]\right) \\
&= \frac{1}{N}\sum_{n=0}^{N-1} \mathbb{E}(x[n]) \\
&= \frac{1}{N}\sum_{n=0}^{N-1} A \\
&= A
\end{aligned} \tag{18}
$$

따라서 평균 추정값 $\hat{A}$는 불편추정값(unbiased estimator)이다. 만약 다음과 같은 추정값 $\breve{A}$가 있다고 가정해보자.

$$\breve{A} = \frac{1}{2N}\sum_{n=0}^{N-1} x[n] \tag{19}$$

$\breve{A}$의 기대값은 다음과 같다.

$$\begin{aligned}\mathbb{E}(\breve{A}) &= \frac{1}{2}A \\ &= A \text{ if } A = 0 \\ &\neq A \text{ if } A \neq 0\end{aligned} \tag{20}$$

따라서 $\breve{A}$는 편의추정값(biased estimator)이다.

불편추정값이 반드시 좋은 추정값을 의미할까? 추정값이 불편성을 지닌다고 해서 반드시 좋은 추정값이라는 의미는 아니다. 불편성의 의미는 오직 추정값의 기대값(expectation)이 실제 값과 동일하다는 의미만 지닌다. 이와 반대로 편의추정값은 시스템의 노이즈를 포함하여 모델링한 값일 수 있다. 하지만 영구적인 편향성은 항상 안 좋은 추정값을 가진다. 예를 들면 다음과 같이 동일 파라미터 θ에 대한 여러 추정값 $\{\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_n\}$이 주어졌을 때 가장 합리적인 방법은 이들의 기대값을 구하는 것이다.

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i \tag{21}$$

만약 모든 추정값들이 불편성을 지니고 서로 독립이라면 다음 공식이 성립한다.

$$\mathbb{E}(\hat{\theta}) = \theta \tag{22}$$

$$\begin{aligned}\text{var}(\hat{\theta}) &= \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(\hat{\theta}_i) \\ &= \frac{\text{var}(\hat{\theta}_1)}{n}\end{aligned} \tag{23}$$

따라서 위 식에서 보다시피 많은 수($n \uparrow$)의 추정값을 사용할 수록 분산은 작아진다. 만약 $n \to \infty$이면 분산은 0이 되고 $\hat{\theta} \to \theta$가 된다. 하지만 편의추정량의 경우 $\mathbb{E}(\hat{\theta}_i) = \theta + b(\theta)$이므로 다음과 같은 기대값을 가진다.

$$\begin{aligned}\mathbb{E}(\hat{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\hat{\theta}_i) \\ &= \theta + b(\theta)\end{aligned} \tag{24}$$

따라서 n이 충분히 많다고 하더라도 편향 $b(\theta)$ 값은 제거되지 않으므로 실제 추정값으로 수렴하지 않는다.

2.2 Minimum Variance Criterion

최적의 추정값을 찾기 위해서는 최적의 criterion을 사용해야 한다. 널리 사용되는 criterion 중 하나가 mean square error(MSE)이다.

$$\boxed{\text{mse}(\hat{\theta}) = \mathbb{E}\Big[(\hat{\theta} - \theta)^2\Big]} \tag{25}$$

- θ : 추정해야 할 파라미터. 빈도주의 관점에 의해 θ는 고정된 상수 값을 의미한다. 즉, 확률변수가 아니다.

MSE는 실제 값 θ과 추정값 $\hat{\theta}$의 평균 제곱 편차를 측정한다. MSE는 널리 사용되는 criterion 중 하나이지만 아쉽게도 편향에 의한 영향을 받는다. 위 식에 $\pm\mathbb{E}(\hat{\theta})$를 추가한 후 식을 전개하면 다음과 같다.

$$\begin{aligned}\text{mse}(\hat{\theta}) &= \mathbb{E}\Big\{\Big[\Big(\hat{\theta} - \mathbb{E}(\hat{\theta})\Big) + \Big(\mathbb{E}(\hat{\theta}) - \theta\Big)\Big]^2\Big\} \\ &= \text{var}(\hat{\theta}) + \Big[\mathbb{E}(\hat{\theta}) - \theta\Big] \\ &= \text{var}(\hat{\theta}) + b^2(\theta)\end{aligned} \tag{26}$$

따라서 최적의 추정값을 찾기 위한 criterion을 고려할 때 MSE를 최소화해주는 minimum MSE(MMSE) 추정값을 고려하면 안된다. MMSE에 대한 대안으로 편향이 0이고 분산을 최소화하는 추정값을 사용해야 하는데 이를 minimum variance unbiased estimator(MVUE)라고 한다. 불편추정값의 분산을 최소화하는 과정은 pdf 관점에서 $p(\hat{\theta} - \theta)$를 0 주변에 집중시키는 효과가 있다. 따라서 추정 오차가 커질 가능성이 작아진다.

2.3 Existence of the Minimum Variance Unbiased Estimator

독자는 MVUE가 실제로 존재하는지 여부에 대해 궁금증이 생길 수 있다. 즉, 모든 파라미터 θ에 대해 최소의 분산을 가지며 불편향된 추정값이 존재하는지 궁금할 수 있다. 결론만 말하자면 MVUE는 항상 존재하는 것은 아니다.

2.4 Finding the Minimum Variance Unbiased Estimator

만약 MVUE가 존재한다고해도 이를 찾는 것이 불가능할 수 있다. MVUE를 찾을 수 있는 절대적인 방법이란 아직 알려지지 않았다. 하지만 이를 가능하게 해주는 몇몇 기준들은 존재한다.

- Cramer-Rao lower bound(CRLB)를 결정하고 추정값들이 이를 만족하는지 확인한다.
- Rao-Blackwell-Lehmann-Scheffe(RBLS) 이론을 적용한다.
- 추정값이 불편성(unbiased) 뿐만아니라 선형(linear) 특성이 있다는 제약조건 하에 분산을 최소화하는 MVUE를 구한다.

1,2 방법을 사용하면 MVUE를 구할 수 있고 3 방법은 MVUE가 데이터에 대하여 선형인 경우에만 적용된다.

- 1,2) CRLB는 임의의 불편추정값에 대하여 분산의 하한선(lower bound)를 결정하게 해준다. 만약 어떤 추정값이 CRLB와 동일한 분산 값을 가진다면 이는 반드시 MVUE가 된다. CRLB와 동일한 분산 값을 가지지 않는다고 하더라도 MVUE가 존재할 수 있다. 이럴 때는 RBLS를 적용한다. RBLS는 충분통계량(sufficient statistics)을 먼저 구한 후 충분통계량에 대한 추정을 수행하는데 이 때 추정값이 θ에 대한 불편추정값이 된다.
- 3) 이는 추정값이 선형이어야 하는 제약조건을 가진다. 이는 때때로 강력한 제약조건이지만 최적의 선형 추정값을 구할 수 있다.

3 Cramer-Rao Lower Bound

어떠한 불편추정값(unbiased estimator)의 분산의 하한선(lower bound)를 결정할 수 있다는 것은 실제 추정 문제에서 매우 유용하게 사용된다. 최선의 경우 특정 추정값이 MVUE임을 바로 구할수도 있다. 그렇지 않은 경우라 하더라도 여러 불편추정량에 대한 벤치마크 용도로 활용될 수도 있다. 이는 신호 추정 분야에서 매우 유용하게 활용된다. Cramer-Rao lower bound(CRLB) 이외에도 [McAulay and Hofstetter 1971, Kendall and Stuart 1979, Seidman 1970, Ziv and Zakai 1969]와 같이 분산의 한계(bound)를 결정하는 알고리즘들이 존재하지만 CRLB가 이들 중 가장 쉽게 한계를 구할 수 있다.

3.1 Estimator Accuracy Considerations

추정에 사용되는 정보는 일반적으로 관측된 데이터로부터 얻을 수 있고 관측 데이터는 노이즈를 포함하기 때문에 일반적으로 pdf로 표현될 수 있다. 따라서 추정의 정확도는 당연히 pdf와 직접적인 관련이 있다. 만약 파라미터가 pdf에 영향을 거의 주지 않는 최악의 경우에는 좋은 추정값을 얻는 것이 어려울 것이다. 따라서 파라미터가 pdf에 영향을 많이 줄수록 추정의 정확도는 올라간다.

3.1.1 Example 3.1 - PDF Dependence on Unknown Parameter

만약 하나의 데이터가 샘플링되었다고 가정해보자

$$x[0] = A + w[0] \tag{27}$$

이 때, $w[n]$는 $\mathcal{N}(0, \sigma^2)$을 따르는 white gaussian noise(WGN)이다. 일반적으로 좋은 추정값(estimator)이란 σ^2가 작은 추정값임을 알 수 있다. 그리고 좋은 불편추정값은 $\hat{A} = x[0]$임을 알 수 있다.

분산 σ^2 값이 작을 수록 좋은 추정값임을 설명하기 위해 아래와 같은 예제를 들 수 있다. 만약 서로 다른 두 분산 $\sigma_1 = 1/3$와 $\sigma_2 = 1$이 주어졌고 $x[0] = 3$인 경우를 가정해보자. 이에 대한 pdf는 다음과 같다.

$$p_i(x[0]; A) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2\sigma_i^2}(x[0] - A)^2 \right] \quad i = 1, 2 \tag{28}$$

$\sigma_1^2 < \sigma_2^2$이므로 우리는 $p_1(\cdot)$이 A를 더 잘 추정하고 있다고 판단할 수 있다. 예를 들어 $A = 4.5$일 확률은 p_1보다 p_2가 더 높으므로 우리는 p_1이 더 좁은 범위에 대한 확실한 확률분포를 가짐을 알 수 있다.

Likelihood Function v.s. Probability Distribution Function 만약 pdf $p(x; \theta)$가 x는 고정된 값이면서 동시에 파라미터 θ에 대한 함수라면 이를 일반적으로 가능도함수(likelihood function)라고 부른다. 이와 반대로, $p(x; \theta)$가 θ는 고정된 값이면서 동시에 x에 대한 함수라면 이는 일반적인 확률분포함수(probability distribution function, pdf)라고 부른다. 두 개념을 비교한 그림은 아래와 같다.

그림에서 보는 것과 같이 pdf와 가능도함수는 모양만 같을 뿐 이를 해석하는 관점이 서로 다르다. pdf의 경우 파라미터가 주어졌을 때 특정 구간 내에서 확률을 구하는 것에 관심이 있다면 가능도함수는 데이터가 주어졌을 때 이를 가장 잘 설명하는 파라미터는 무엇인가? 에 관심이 있다.

지금까지 설명한 pdf는 전부 데이터 x가 주어졌을 때 파라미터 A를 찾는 형태이기 때문에 이는 가능도함수(likelihood function) 관점에서 해석할 수 있다. 가능도함수의 뾰족한 정도(sharpness)는 추정값이 얼마나 정확한지에 대한 정확도를 판단하는데 사용된다. 수학적으로 곡선의 뾰족한 정도는 함수의 2차 미분을 수행하여 곡률(curvature)를 구함으로써 얻을 수 있다. 이 때, pdf 특성 상 exponential 항이 존재하여 계산이 어렵기 때문에 일반적으로 log를 취한 값을 사용하는데 이를 로그 가능도함수(log likelihood function)이라고 한다.

$$\ln p(x[0];A) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x[0]-A)^2 \tag{29}$$

파라미터 A에 대한 1차 미분을 수행하면 다음과 같다.

$$\frac{\partial \ln p(x[0];A)}{\partial A} = \frac{1}{\sigma^2}(x[0]-A) \tag{30}$$

다시 한번 2차 미분을 취한 후 양변에 음수를 곱하면 다음과 같다.

$$-\frac{\partial^2 \ln p(x[0];A)}{\partial A^2} = \frac{1}{\sigma^2} \tag{31}$$

위 식의 의미는 σ^2가 감소할수록 원래 함수의 곡률(curvature)을 의미하는 $-\frac{\partial^2 \ln p(x[0];A)}{\partial A^2}$는 증가하는 것을 의미한다. 앞서 말한 추정값 $\hat{A} = x[0]$의 분산은 다음과 같다.

$$\mathrm{var}(\hat{A}) = \sigma^2 \tag{32}$$

여기에 (31)을 대입하면 다음과 같다.

$$\mathrm{var}(\hat{A}) = \frac{1}{-\frac{\partial^2 \ln p(x[0];A)}{\partial A^2}} \tag{33}$$

위와 같은 간단한 예제에서는 로그 가능도함수의 2차 미분값이 데이터 $x[0]$에 독립이지만 일반적으로는 2차 미분값은 데이터 $\mathbf{x} = [x[0], \cdots, x[n]]$에 종속적이다. 곡률의 정확한 표현은 다음과 같이 나타낼 수 있다.

$$\boxed{-\mathbb{E}\left[\frac{\partial^2 \ln p(x[0];A)}{\partial A^2}\right]} \tag{34}$$

위 식을 통해 다양한 관측 데이터가 주어졌을 때 로그 가능도함수의 평균적인 곡률을 측정할 수 있다.

3.2 Cramer-Rao Lower Bound

3.2.1. Theorem 3.1 (Cramer-Rao Lower Bound - Scalar Parameter) pdf $p(\mathbf{x};\theta)$가 다음과 같은 정규 조건(regularity condition)을 만족하면

$$\mathbb{E}\left[\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta}\right] = 0 \quad \text{for all } \theta \tag{35}$$

임의의 불편추정값(unbiased estimator) $\hat{\theta}$의 분산은 다음 조건을 반드시 만족한다.

$$\mathrm{var}(\hat{\theta}) \geq \frac{1}{-\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right]} \tag{36}$$

이 때 편미분은 실제 파라미터의 참 값 θ에 대하여 수행되었다. 추가적으로 불편추정값의 분산이 하한선(lower bound)에 도달하려면 다음 조건을 반드시 만족해야 한다. (필요충분조건)

$$\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta} = I(\theta)(g(\mathbf{x}) - \theta) \tag{37}$$

위 식은 로그 가능도함수의 1차 미분이 위와 같은 임의의 함수 I, g의 곱셈 형태로 표현되어야 한다는 뜻이다. 이 때 불편추정값은 $\hat{\theta} = g(\mathbf{x})$가 되며 $\hat{\theta}$는 MVUE를 만족한다. 이 때의 최소 분산 값은 $1/I(\theta)$이 되며 이를 Fisher information이라고 한다. Fisher information은 일반적으로 다음과 같이 정의한다.

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right] \tag{38}$$

따라서 MVUE의 분산은 다음과 같이 표현하기도 한다.

$$\text{var}(\hat{\theta}) = \frac{1}{I(\theta)} \quad \cdots \text{ for MVUE} \tag{39}$$

(36)에서 2차 미분값은 $\mathbf{x}$에 종속적이기 때문에 기대값은 정의에 따라 다음과 같이 전개된다.

$$\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2}\right] = \int \frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta^2} p(\mathbf{x};\theta) d\mathbf{x} \tag{40}$$

3.2.1 Example 3.3 - DC Level in White Gaussian Noise

다음과 같은 여러 관측 데이터들이 주어졌다고 하자.

$$x[n] = A + w[n] \quad n = 0, 1, \cdots, N-1 \tag{41}$$

- $w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN

파라미터 A에 대한 CRLB를 유도해보면 다음과 같다. 우선 모든 관측값 $\mathbf{x}$에 대한 가능도함수를 구한다.

$$\begin{aligned} p(\mathbf{x}; A) &= \Pi_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x[n]-A)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] \end{aligned} \tag{42}$$

로그 가능도 함수는 다음과 같다.

$$\ln p(\mathbf{x}; A) = -\ln[(2\pi\sigma^2)^{\frac{N}{2}}] - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2 \tag{43}$$

1차 미분을 수행하면 다음과 같다.

$$\begin{aligned} \frac{\partial \ln p(\mathbf{x}; A)}{\partial A} &= \frac{\partial}{\partial A}\left[-\ln[(2\pi\sigma^2)^{\frac{N}{2}}] - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] \\ &= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A) \\ &= \frac{N}{\sigma^2}(\bar{\mathbf{x}} - A) \end{aligned} \tag{44}$$

- $\bar{\mathbf{x}}$: $\mathbf{x}$의 평균

2차 미분을 수행하면 다음과 같다.

$$\frac{\partial^2 \ln p(\mathbf{x}; A)}{\partial A^2} = -\frac{N}{\sigma^2} \tag{45}$$

위 식은 파라미터가 없는 상수임에 유의한다. (36) 식에 이를 대입하면 다음과 같다.

$$\text{var}(\hat{A}) \geq \frac{\sigma^2}{N} \tag{46}$$

위 식이 정규 조건을 만족하는가? (35)를 보면 적용해보면 만족하는 것을 알 수 있다.

$$\mathbb{E}\left[\frac{\partial \ln p(\mathbf{x};\theta)}{\partial\theta}\right] = \mathbb{E}\left[\frac{N}{\sigma^2}(\bar{\mathbf{x}} - A)\right] = 0 \tag{47}$$

- $\mathbb{E}(\bar{\mathbf{x}}) = A$

따라서 CRLB 정의에 따라 임의의 불편추정값 $\breve{A}$의 분산이 $\text{var}(\breve{A}) = \frac{\sigma^2}{N}$을 만족하면 이는 반드시 MVUE가 된다. 위 예제에서는 $\breve{A} = \bar{\mathbf{x}}$가 MVUE가 된다. 그리고 이러한 추정값을 efficient하다라고 한다. 관측 데이터를 효율적(efficient)으로 사용하여 추정하였다는 의미이다.

3.3 Transformation of Parameters

실제 추정 문제에서는 우리가 추정하고자 하는 파라미터가 단순한 θ가 아닌 θ의 함수 형태를 추정해야 하는 일이 자주 발생한다. 이전 예제에서도 단순히 A를 추정하는 것이 아닌 A^2를 추정하고 싶을 수도 있다. 만약 A의 CRLB를 알고 있는 경우에는 A^2의 CRLB도 쉽게 구할 수 있고 A와 관련된 어떤 함수라도 구할 수 있다.

추정하고자 하는 파라미터가 $\alpha = g(\theta)$와 같이 θ에 대한 함수일 경우 CRLB는 다음과 같다.

$$\text{var}(\hat{\alpha}) \geq \frac{\left(\frac{\partial g}{\partial \theta}\right)^2}{-\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \theta^2}\right]} \tag{48}$$

만약 $\alpha = g(A) = A^2$인 경우 CRLB는 다음과 같다.

$$\text{var}(\hat{\alpha}) \geq \frac{(2A)^2}{N/\sigma^2} = \frac{4A^2\sigma^2}{N} \tag{49}$$

이전 Example 3.3 예제에서 $\hat{A} = \bar{\mathbf{x}}$는 A에 대하여 efficient함을 보였다. 따라서 $\bar{\mathbf{x}}^2$ 역시 A^2에 대하여 efiicient하다고 예상할 수 있으나 이는 사실이 아니다. 그 이전에 $\hat{A}^2 = \bar{\mathbf{x}}^2$는 A^2에 대한 불편추정값 조차 아니다. 즉, 편향(bias)이 존재한다.

$$\begin{aligned} \mathbb{E}(\bar{\mathbf{x}}^2) &= \mathbb{E}^2(\bar{\mathbf{x}}) + \text{var}(\bar{\mathbf{x}}) \\ &= A^2 + \frac{\sigma^2}{N} \\ &\neq A^2 \end{aligned} \tag{50}$$

따라서 우리는 위 예제를 통해 추정값의 efficiency는 비선형 변환에 의해 보존되지 않는 것을 알 수 있다. 하지만 linear 또는 affine 변환에 대하여는 efficiency가 보존됨을 쉽게 보일 수 있다.

만약 $\hat{\theta}$가 θ에 대하여 efficient하고 $\alpha = g(\theta) = a\theta + b$와 같은 affine 변환이 주어진 경우를 확인해보자.

$$\hat{\alpha} = g(\hat{\theta}) = a\hat{\theta} + b \tag{51}$$

$$\begin{aligned} \mathbb{E}(a\hat{\theta} + b) &= a\mathbb{E}(\hat{\theta}) + b \\ &= a\theta + b \\ &= g(\theta) \end{aligned} \tag{52}$$

$$\begin{aligned} \text{var}(\hat{\alpha}) &= \text{var}(a\hat{\theta} + b) \\ &= a^2\text{var}(\hat{\theta}) \end{aligned} \tag{53}$$

$g(\theta)$의 CRLB를 보면 다음과 같다.

$$\begin{aligned} \text{var}(\hat{\alpha}) &\geq \frac{\left(\frac{\partial g}{\partial \theta}\right)^2}{I(\theta)} \\ &= \left(\frac{\partial g}{\partial \theta}\right)^2 \text{var}(\hat{\theta}) \\ &= a^2\text{var}(\hat{\theta}) \end{aligned} \tag{54}$$

위 식에서 (53), (54)의 분산이 동일하기 때문에 $\hat{\alpha}$ 역시 MVUE이면서 동시에 efficient함을 알 수 있다.

앞서 보았듯이 efficiency는 linear 또는 affine 변환에서만 보존되는 것을 확인하였다. 하지만 데이터가 충분히 큰 경우에는 비선형 변환도 근사적으로(approximately) efficiency가 보존된다. 다시 이전 예제 $\alpha = g(A) = A^2$로 돌아가서 데이터 $x[n]$의 평균을 $\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=0}^{N-1} x[n]$이라고 할 때 $\bar{\mathbf{x}}^2$는 N이 충분히 클 경우 근사적으로 편향이 제거된다.

$$\begin{aligned} \text{var}(\bar{\mathbf{x}}^2) &= \mathbb{E}(\bar{\mathbf{x}}^4) - \mathbb{E}^2(\bar{\mathbf{x}}^2) \\ &= \frac{4A^2\sigma^2}{N} + \frac{2\sigma^2}{N^2} \\ &\approx \frac{4A^2\sigma^2}{N} \quad \cdots \text{ for } N \to \infty \end{aligned} \tag{55}$$

$g(A) = A^2$의 CRLB는 다음과 같다.

$$\begin{aligned}\mathrm{var}(\hat{\alpha}) &\geq \frac{(2A)^2}{N/\sigma^2} \\ &= \frac{4A^2\sigma^2}{N}\end{aligned} \tag{56}$$

(55), (55)가 N이 충분히 큰 경우에 대하여 서로 동일하기 때문에 데이터가 많은 경우에는 비선형 변환에 대한 추정값도 MVUE가 되며 동시에 efficient함을 알 수 있다.

다른 방법을 사용하여 비선형 변환이 근사적으로 efficient함을 보일 수 있다. 확률분포 관점에서 봤을 때 N이 커질수록 $\bar{\mathbf{x}}$는 A 주변으로 뾰족해지는 경향이 있다. 이에 따라 $\pm 3\sigma$ 사이의 간격은 좁아지게 되고 좁은 영역에 대해여 비선형 변환을 수행하면 근사적으로 선형 변환을 한 것과 유사한 효과를 얻는다. 이를 $\bar{\mathbf{x}} = A$ 지점에서 테일러 1차 근사를 통해 표현하면 다음과 같다.

$$g(\bar{\mathbf{x}}) \approx g(A) + \frac{dg(A)}{dA}(\bar{\mathbf{x}} - A) \tag{57}$$

이러한 경우를 점근적으로(asymptotically) efficient하다고 한다. 이 때 기대값은 다음과 같다.

$$\mathbb{E}[g(\bar{\mathbf{x}})] = g(A) = A^2 \tag{58}$$

분산은 다음과 같다.

$$\begin{aligned}\mathrm{var}[g(\bar{\mathbf{x}})] &= \left[\frac{dg(A)}{dA}\right]^2 \mathrm{var}(\bar{\mathbf{x}}) \\ &= \frac{(2A)^2\sigma^2}{N} \\ &= \frac{4A^2\sigma^2}{N}\end{aligned} \tag{59}$$

즉 추정값은 CRLB에 점근적으로(asymptotically) 도달하는 것을 알 수 있다. 따라서 비선형 변환은 점근적으로 efficient하다.

3.4 Extension to a Vector Parameter

지금까지는 추정하려는 파라미터 θ가 스칼라 값이었다. 해당 섹션에서는 파라미터가 여러개인 벡터 파리미터 $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_p]^\intercal$ 케이스에 대해 다룬다. $\hat{\boldsymbol{\theta}}$는 $\boldsymbol{\theta}$에 대한 불편추정값(unbiased estimator)이라고 가정한다. 벡터 파라미터의 각 원소의 분산은 다음과 같이 나타낼 수 있다.

$$\mathrm{var}(\hat{\theta}_i) \geq [\mathbf{I}^{-1}(\boldsymbol{\theta})]_{ii} \tag{60}$$

- $\mathbf{I}(\boldsymbol{\theta}) \in \mathbb{R}^{p\times p}$: Fisher information 행렬

이는 (39)의 벡터 버전임을 알 수 있다. 일반적으로 스칼라 버전에서 행렬은 벡터 버전에서 역행렬로 표현된다. Fisher information 행렬을 자세히 나타내면 다음과 같다.

$$\big[\mathbf{I}(\boldsymbol{\theta})\big]_{ij} = -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial\theta_i \partial\theta_j}\right] \tag{61}$$

- $i = 1, 2, \cdots, p$
- $j = 1, 2, \cdots, p$

$p = 1$인 스칼라 케이스는 $\mathbf{I}(\boldsymbol{\theta}) = I(\theta)$가 된다. 스칼라 버전과 동일하게 불편추정값의 분산이 하한선(lower bound)에 도달하려면 다음 조건을 반드시 만족해야 한다. (필요충분조건)

$$\frac{\partial \ln p(\mathbf{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta}) \tag{62}$$

위 식은 로그 가능도함수의 1차 미분이 위와 같은 임의의 함수 $\mathbf{I}, \mathbf{g}$의 곱셈 형태로 표현되어야 한다는 뜻이다. 이 때 불편추정값은 $\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x})$가 되며 $\hat{\boldsymbol{\theta}}$는 MVUE를 만족한다. 이 때의 최소 분산 값은 $1/\mathbf{I}(\boldsymbol{\theta})$이 된다.

$$\mathrm{var}(\hat{\boldsymbol{\theta}}) = 1/\mathbf{I}(\boldsymbol{\theta}) \quad \cdots \text{ for MVUE} \tag{63}$$

3.4.1 Example 3.6 - DC Level in White Gaussian Noise (Revisited)

예제 3.3과 같이 관측 데이터들이 주어졌다고 하자.

$$x[n] = A + w[n] \quad n = 0, 1, \cdots, N-1 \tag{64}$$

$w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN
이 때, 추정하고자 하는 파라미터가 $\theta = [A, \sigma^2]^\intercal$인 경우를 생각해보자. 즉 $p = 2$이다. Fisher information 행렬은 다음과 같이 나타낼 수 있다.

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial A^2}\right] & -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial A \partial \sigma^2}\right] \\ -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial \sigma^2 \partial A}\right] & -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial {\sigma^2}^2}\right] \end{bmatrix} \in \mathbb{R}^{2\times 2} \tag{65}$$

Fisher information 행렬은 대칭이며 동시에 positive definite한 특징을 가지고 있다. 예제 3.3의 로그 가능도함수는 다음과 같다.

$$\ln p(\mathbf{x};\boldsymbol{\theta}) = -\frac{N}{2}\ln 2\pi - \frac{N}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2 \tag{66}$$

로그 가능도함수의 1,2차 편미분은 다음과 같이 쉽게 구할 수 있다.

$$\begin{aligned} \frac{\ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A} &= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A) \\ \frac{\ln p(\mathbf{x};\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=0}^{N-1}(x[n]-A)^2 \\ \frac{\ln^2 p(\mathbf{x};\boldsymbol{\theta})}{\partial A^2} &= -\frac{N}{\sigma^2} \\ \frac{\ln^2 p(\mathbf{x};\boldsymbol{\theta})}{\partial A \partial \sigma^2} &= -\frac{1}{\sigma^4}\sum_{n=0}^{N-1}(x[n]-A) \\ \frac{\ln^2 p(\mathbf{x};\boldsymbol{\theta})}{\partial {\sigma^2}^2} &= \frac{N}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{n=0}^{N-1}(x[n]-A)^2 \end{aligned} \tag{67}$$

이를 Fisher information 행렬에 대입하면 다음과 같다.

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix} \tag{68}$$

흔한 경우는 아니지만 예제의 케이스는 역행렬을 매우 쉽게 구할 수 있다. 단순히 역수를 취해줌으로써 역행렬을 구하면 (60)와 같이 CRLB를 구할 수 있다.

$$\begin{aligned} \mathrm{var}(\hat{A}) &\geq \frac{\sigma^2}{N} \\ \mathrm{var}(\hat{\sigma^2}) &\geq \frac{2\sigma^4}{N} \end{aligned} \tag{69}$$

이 중 $\mathrm{var}(\hat{A})$는 스칼라 케이스에서 σ^2 값이 이미 주어진 경우와 동일한 CRLB를 가지는 것을 알 수 있다. 다시 말하자면 이러한 예제는 일반적인 상황에서는 참이 아니지만 예제의 경우 참이다.

3.4.2 Example 3.7 - Line Fitting

다음과 같은 line fitting 문제가 주어졌다고 가정해보자

$$x[n] = A + Bn + w[n] \quad n = 0, 1, \cdots, N-1 \tag{70}$$

$w[n] \sim \mathcal{N}(0, \sigma^2)$: WGN
이 때, y절편의 값 A와 기울기 B의 값을 찾아야 한다. 추정하고자 하는 파라미터는 $\theta = [A,\ B]^\intercal$이다. $p = 2$ 케이스이므로 Fisher information 행렬은 다음과 같이 나타낼 수 있다.

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial A^2}\right] & -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial A \partial B^2}\right] \\ -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial B \partial A}\right] & -\mathbb{E}\left[\frac{\partial^2 \ln p(\mathbf{x};\theta)}{\partial B^2}\right] \end{bmatrix} \in \mathbb{R}^{2\times 2} \tag{71}$$

가능도함수는 다음과 같다.

$$p(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A - Bn)^2 \right] \tag{72}$$

로그 가능도함수의 1,2차 미분은 다음과 같다.

$$\begin{aligned} \frac{\ln p(\mathbf{x};\boldsymbol{\theta})}{\partial A} &= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A - Bn) \\ \frac{\ln p(\mathbf{x};\boldsymbol{\theta})}{\partial B} &= \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A - Bn)n \\ \frac{\ln^2 p(\mathbf{x};\boldsymbol{\theta})}{\partial A^2} &= -\frac{N}{\sigma^2} \\ \frac{\ln^2 p(\mathbf{x};\boldsymbol{\theta})}{\partial A \partial B} &= -\frac{1}{\sigma^2} \sum_{n=0}^{N-1} n \\ \frac{\ln^2 p(\mathbf{x};\boldsymbol{\theta})}{\partial \sigma^{2\,2}} &= -\frac{1}{\sigma^2} \sum_{n=0}^{N-1} n^2 \end{aligned} \tag{73}$$

이를 Fisher information 행렬에 대입하면 다음과 같다.

$$\begin{aligned} \mathbf{I}(\boldsymbol{\theta}) &= \frac{1}{\sigma^2} \begin{bmatrix} N & \sum_{n=0}^{N-1} n \\ \sum_{n=0}^{N-1} n & \sum_{n=0}^{N-1} n^2 \end{bmatrix} \\ &= \frac{1}{\sigma^2} \begin{bmatrix} N & \frac{N(N-1)}{2} \\ \frac{N(N-1)}{2} & \frac{N(N-1)(2N-1)}{6} \end{bmatrix} \end{aligned} \tag{74}$$

$\mathbf{I}(\boldsymbol{\theta})$의 역행렬을 구해보면 다음과 같다.

$$\mathbf{I}(\boldsymbol{\theta})^{-1} = \sigma^2 \begin{bmatrix} \frac{2(2N-1)}{N(N+1)} & -\frac{6}{N(N+1)} \\ -\frac{6}{N(N+1)} & \frac{12}{N(N^2-1)} \end{bmatrix} \tag{75}$$

(60)와 같이 CRLB를 구해보면 다음과 같다.

$$\begin{aligned} \mathrm{var}(\hat{A}) &\geq \frac{2(2N-1)\sigma^2}{N(N+1)} \\ \mathrm{var}(\hat{B}) &\geq \frac{12\sigma^2}{N(N^2-1)} \end{aligned} \tag{76}$$

위 예제에서 보다시피 스칼라 파라미터 $\theta = A$만 추정했을 때와는 달리 $\theta = [A,\ B]^\intercal$처럼 벡터 파라미터를 추정하면 CRLB는 커지는 것을 알 수 있다. 스칼라 파라미터만 추정했을 때 CRLB는 다음과 같다.

$$\mathrm{var}(\hat{A}) \geq \frac{\sigma^2}{N} \tag{77}$$

따라서 $N \geq 2$인 경우에는 벡터 파라미터의 CRLB가 항상 스칼라 파라미터의 CRLB보다 크다.

$$\frac{2(2N-1)\sigma^2}{N(N+1)} > \frac{\sigma^2}{N} \quad \cdots \text{ for } N \geq 2 \tag{78}$$

또한 특정 파라미터는 다른 파라미터보다 데이터 개수 N에 민감하게 반응할 수 있다.

$$\frac{\mathrm{CRLB}(\hat{A})}{\mathrm{CRLB}(\hat{B})} = \frac{(2N-1)(N-1)}{6} > 1 \quad \cdots \text{ for } N \geq 3 \tag{79}$$

$\mathrm{CRLB}(\hat{B})$는 데이터 증가에 $1/N^3$으로 감소하는 반면, $\mathrm{CRLB}(\hat{A})$는 데이터 증가에 $1/N$ 비율로 감소한다. 이러한 차이로 인해 $x[n]$이 B를 변경하는 것에 A를 변경하는 것보다 더 민감하게 반응한다는 것을 알 수 있다.

3.5 Vector Parameter CRLB for Transformations

4 Linear Models

5 General Minimum Variance Unbiased Estimation

6 Best Linear Unbiased Estimation

7 Maximum Likelihood Estimation

8 Least Squares

9 The Bayesian Philosophy

10 General Bayesian Estimators

11 Linear Bayesian Estimators

12 Kalman Filters

13 References

[1] Kay, Steven M. Fundamentals of statistical signal processing: estimation theory. Prentice-Hall, Inc., 1993.

[2] Simon, Dan. Optimal state estimation: Kalman, H infinity, and nonlinear approaches. John Wiley Sons, 2006.

14 Revision log

- 1st: 2024-02-11