ECE 275A: Parameter Estimation I
Classical Estimation

Florian Meyer

Electrical and Computer Engineering Department
University of California San Diego

UC San Diego
JACOBS SCHOOL OF ENGINEERING

Statistical Family

- $\boldsymbol{\theta} \in \mathbb{R}^p$: unknown parameter of interest to be estimated (deterministic)
- $\mathbf{y} \in \mathbb{R}^m$: data which depends on the unknown parameter $\boldsymbol{\theta}$ (random)
- To mathematically model the data, which is inherently random, we consider a **statistical family** of parameterized distributions, i.e.,

$$\mathcal{P} = \left\{ p(\boldsymbol{y}; \boldsymbol{\theta}) \middle| \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p, \mathbf{y} \in \mathbb{Y} \subset \mathbb{R}^m \right\}$$

- Depending on the context, $p(\boldsymbol{y}; \boldsymbol{\theta})$ is either a probability density function (PDF) or a probability mass function (PMF)

Estimators

- An estimator $\hat{\boldsymbol{\theta}}(\cdot) : \mathbb{Y} \to \Theta$ is a function that does not depend on the unknown parameter $\boldsymbol{\theta}$; it can be considered as a rule that assigns a value to $\boldsymbol{\theta}$ for *each realization* of $\mathbf{y}$

- An estimate $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ is the value of $\boldsymbol{\theta}$ for a *fixed realization* of $\mathbf{y}$

- The primary goal of statistical parameter estimation is to find an estimator $\hat{\boldsymbol{\theta}}(\cdot)$ with the property of providing an estimate $\hat{\boldsymbol{\theta}}(\boldsymbol{y})$ that is *accurate* (close to the true unknown parameter $\boldsymbol{\theta}$) for most parameter values $\boldsymbol{\theta}$ and data realizations $\mathbf{y}$

- Another important property is that the estimator $\hat{\boldsymbol{\theta}}(\cdot)$ is *robust* to model mismatch (small changes in $p(\boldsymbol{y}; \boldsymbol{\theta})$ do not severely affect the performance of the estimator $\hat{\boldsymbol{\theta}}(\cdot)$)

The Mean Square Error

- A natural optimality criterion for estimators is the mean square error ("little mse"),
$$\text{mse}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = E(\|\tilde{\boldsymbol{\theta}}\|^2; \boldsymbol{\theta})$$
where $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ is the estimation error

- For future reference we also consider the matrix mean square error ("big MSE"),
$$\text{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = E(\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^{\mathsf{T}}; \boldsymbol{\theta})$$

- Note that $\text{mse}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = \text{tr}\big(\text{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})\big)$

- Unfortunately, this natural criterion leads to estimators that cannot be expressed only as a function of the data $\mathbf{y}$ and thus to unrealizable estimators

Example: Unrealizability of Mean Square Error Estimation

- Consider the data model $\mathrm{y} = A + \mathrm{n}$, with scalar parameter A to be estimated and additive Gaussian measurement noise $\mathrm{n} \sim \mathcal{N}(0, \sigma^2)$ with known σ^2
- Consider the estimator $\hat{\mathrm{A}} = a\mathrm{y}$ based on an arbitrary constant $a \in \mathbb{R}$, i.e., $\hat{\mathrm{A}} \sim \mathcal{N}(aA, a^2\sigma^2)$
- We aim to find the a which results in the minimum mean square error
- First, we rewrite the mean square error for a general unknown parameter $\boldsymbol{\theta}$ as

$$\begin{aligned}
\mathrm{mse}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\uptheta}}\right) &= E\left(\|\hat{\boldsymbol{\uptheta}} - \boldsymbol{\theta}\|^2\right) \\
&= E\left(\|(\hat{\boldsymbol{\uptheta}} - E(\hat{\boldsymbol{\uptheta}})) + (E(\hat{\boldsymbol{\uptheta}}) - \boldsymbol{\theta})\|^2\right) \\
&= E\left(\|\hat{\boldsymbol{\uptheta}} - E(\hat{\boldsymbol{\uptheta}})\|^2\right) + \|E(\hat{\boldsymbol{\uptheta}}) - \boldsymbol{\theta}\|^2 \\
&= \mathrm{tr}\left(\underbrace{\mathrm{Cov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\uptheta}})}_{\textit{variance}}\right) + \|\underbrace{E(\hat{\boldsymbol{\uptheta}}) - \boldsymbol{\theta}}_{\textit{bias}}\|^2
\end{aligned}$$

Example: Unrealizability of Mean Square Error Estimation

- For the scalar $\hat{\text{A}}$, the mean square error is obtained as

$$\begin{aligned} \text{mse}_A\big(\hat{\text{A}}\big) &= \text{var}\big(\hat{\text{A}}\big) + \big(E(\hat{\text{A}}) - A\big)^2 \\ &= a^2\sigma^2 + (a-1)^2 A^2 \end{aligned}$$

- Next, we differentiate the mean square error with respect to a and set it to zero, i.e.,

$$\frac{\partial \text{mse}_A\big(\hat{\text{A}}\big)}{\partial a} = 2a\sigma^2 + 2(a-1)A^2 \triangleq 0$$

- The optimal value for a is now obtained as

$$a_* = \frac{A^2}{\sigma^2 + A^2}$$

- a_* depends on the unknown parameter A and is thus unrealizable!

Uniformly Unbiased Estimators (UUBE)

- An alternative approach to obtain realizable estimators is to constrain the bias $E(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}) - \boldsymbol{\theta}$ to be zero:

Uniformly Unbiased Estimators (UUBE)

A UUBE $\hat{\boldsymbol{\theta}}$ is an estimator that satisfies

$$E(\hat{\boldsymbol{\theta}};\boldsymbol{\theta}) = \boldsymbol{\theta}, \forall \boldsymbol{\theta} \in \Theta$$

- Note that if $\hat{\boldsymbol{\theta}}$ is an UUBE, then $E(\tilde{\boldsymbol{\theta}};\boldsymbol{\theta}) = \mathbf{0}, \forall \boldsymbol{\theta} \in \Theta$

Uniformly Minimum Variance Unbiased Estimators (UMVUE)

- For all $\hat{\boldsymbol{\theta}}$ that are UUBEs, we have

$$\mathrm{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = \mathrm{Cov}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}) = \mathrm{Cov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})$$

- Thus, of particular interest, is the UUBE that minimizes the variance

Uniformly Minimum Variance Unbiased Estimator (UMVUE)

A UMVUE $\hat{\boldsymbol{\theta}}_*$ is an estimator that is defined as follows

$$\hat{\boldsymbol{\theta}}_* \text{ is an UUBE and } \mathrm{Cov}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}_*) \preccurlyeq \mathrm{Cov}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}),\ \forall \boldsymbol{\theta} \in \Theta, \forall \hat{\boldsymbol{\theta}} \text{ that are UUBEs}$$

How to construct a UMVUE?

- General Question: **How to construct a UMVUE?**
- One approach is to work with statistical families for which there exists a uniform lower bound of the error covariance matrix $\boldsymbol{B}_{\boldsymbol{\theta}}$, i.e.,

$$\boldsymbol{B}_{\boldsymbol{\theta}} \preccurlyeq \text{Cov}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}), \quad \forall \boldsymbol{\theta} \in \Theta, \hat{\boldsymbol{\theta}} \text{ that are UUBEs}$$

- If such a (matrix) lower bound exists and an UUBE $\hat{\boldsymbol{\theta}}'$ can be found such that

$$\text{Cov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}') = \boldsymbol{B}_{\boldsymbol{\theta}}, \forall \boldsymbol{\theta} \in \Theta$$

  then $\hat{\boldsymbol{\theta}}'$ is the UMVUE, i.e., $\hat{\boldsymbol{\theta}}' = \hat{\boldsymbol{\theta}}_*$

- For so-called **Regular Statistical Families (RSF)**, such a uniform lower bound exists and is referred to as the **Cramér-Rao Lower Bound (CRLB)**

Regular Statistical Families (RSF)

- Recall the definition to a statistical family

$$\mathcal{P} = \left\{ p(\boldsymbol{y};\boldsymbol{\theta}) \middle| \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p, \mathbf{y} \in \mathbb{Y} \subseteq \mathbb{R}^m \right\}$$

Regular Statistical Families (RSF)

An RSF is a statistical family $\mathcal{P}$ that satisfies the three conditions:

R1 The support of $p(\boldsymbol{y};\boldsymbol{\theta})$ independent of the parameter vector $\boldsymbol{\theta}$

R2 $p(\boldsymbol{y};\boldsymbol{\theta})$ is differentiable (i.e. $\nabla_{\boldsymbol{\theta}} p(\boldsymbol{y};\boldsymbol{\theta})$ exists)

R3 $p(\boldsymbol{y};\boldsymbol{\theta})$ is doubly-differentiable (i.e. $\nabla^2_{\boldsymbol{\theta}} p(\boldsymbol{y};\boldsymbol{\theta})$ exists)

- Let us define the score of $\mathcal{P}$ as $\boldsymbol{s}_{\boldsymbol{\theta}}(\mathbf{y}) = \nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{y};\boldsymbol{\theta})$
- If $\mathcal{P}$ is an RSF, then $E\left(\boldsymbol{s}_{\boldsymbol{\theta}}(\mathbf{y})\right) = \mathbf{0}$

Cramér-Rao Lower Bound (CRLB)

- If $\mathcal{P}$ is an RSF and $\hat{\boldsymbol{\theta}}$ is an UUBE, we also have

$$E\left(\boldsymbol{s}_{\boldsymbol{\theta}}(\mathbf{y})\tilde{\boldsymbol{\theta}}^{\mathsf{T}}\right) = \boldsymbol{I}$$

- The **Fisher Information Matrix (FIM)** $\boldsymbol{J}_{\boldsymbol{\theta}}$ of the RSF $\mathcal{P}$ is defined as the covariance matrix of the score

$$\boldsymbol{J}_{\boldsymbol{\theta}} = \mathrm{Cov}_{\boldsymbol{\theta}}\big(\boldsymbol{s}_{\boldsymbol{\theta}}(\mathbf{y})\big) = E\left(\boldsymbol{s}_{\boldsymbol{\theta}}(\mathbf{y})\,\boldsymbol{s}_{\boldsymbol{\theta}}^{\mathsf{T}}(\mathbf{y})\right) = -E\left(\nabla_{\boldsymbol{\theta}}^{2} \ln p(\boldsymbol{y};\boldsymbol{\theta})\right)$$

- If the FIM is positive definite, $\boldsymbol{J}_{\boldsymbol{\theta}}^{-1}$ exists and is equal to the CRLB, i.e.,

$$\mathrm{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = \mathrm{Cov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) \succcurlyeq \mathrm{CRLB} = \boldsymbol{J}_{\boldsymbol{\theta}}^{-1}, \quad \forall \boldsymbol{\theta} \in \Theta, \forall \hat{\boldsymbol{\theta}} \text{ that are UUBEs}$$

with equality iff $\tilde{\boldsymbol{\theta}} = \boldsymbol{J}_{\boldsymbol{\theta}}^{-1}\boldsymbol{s}_{\boldsymbol{\theta}}(\mathbf{y})$

Example: Linear-Gaussian Model

- Let $\mathbf{y} = \boldsymbol{A}\boldsymbol{\theta} + \mathbf{n}$ where the additive noise $\mathbf{n}$ is Gaussian distributed, i.e., $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C})$, and both $\boldsymbol{A} \in \mathbb{R}^{m \times p}$ and $\boldsymbol{C} \in \mathbb{R}^{m \times m}$ are known
- Note that $\mathbf{y} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{C})$
- **Goal:** Estimate the unknown parameter $\boldsymbol{\theta}$ from observed data $\mathbf{y}$
- A common choice is the Maximum Likelihood (ML) estimator $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{y}; \boldsymbol{\theta})$
- Assuming that $\boldsymbol{A}$ is a tall matrix and both $\boldsymbol{C}$ and $\boldsymbol{A}$ are full rank, the ML estimator is given by

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \ln p(\boldsymbol{y}; \boldsymbol{\theta}) = (\boldsymbol{A}^{\mathsf{T}}\boldsymbol{C}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{C}^{-1}\mathbf{y}$$

- It is easy to verify that $E(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}; \boldsymbol{\theta}) = \boldsymbol{\theta} \Rightarrow \hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is an UUBE

Example: Linear-Gaussian Model

- Next, we calculate the score and FIM as follows

$$\boldsymbol{s}_{\boldsymbol{\theta}}(\mathbf{y}) = \boldsymbol{A}^{\mathsf{T}} \boldsymbol{C}^{-1} (\mathbf{y} - \boldsymbol{A}\boldsymbol{\theta})$$
$$\boldsymbol{J}_{\boldsymbol{\theta}} = \boldsymbol{A}^{\mathsf{T}} \boldsymbol{C}^{-1} \boldsymbol{A}$$

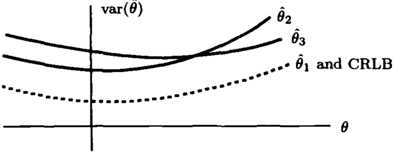
- Since the ML estimation error can be expressed as $\tilde{\boldsymbol{\theta}}_{\mathrm{ML}} = \boldsymbol{J}_{\boldsymbol{\theta}}^{-1} \boldsymbol{s}_{\boldsymbol{\theta}}(\mathbf{y})$, the ML estimator $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ can attain the CRLB

- **The MLE for linear-Gaussian models is also the UMVUE!**

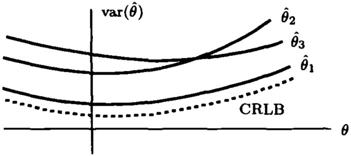
Efficient Estimators

- An estimator which is unbiased and attains the CRLB is said to be *efficient* in that it efficiently uses the data

- An UMVUE may or may not be efficient [*Kay, 1993*] :



(a) $\hat{\theta}_1$ **efficient and MVU**



(b) $\hat{\theta}_1$ **MVU but not efficient**

- In (a), the UMVU $\hat{\boldsymbol{\theta}}_1$ is efficient in that it attains the CRLB

- In (b), $\hat{\boldsymbol{\theta}}_1$ is the UMVU but does not attain the CRLB, and hence it is not efficient

Best Linear Unbiased Estimator (BLUE)

- Again we consider the linear model $\mathbf{y} = \boldsymbol{A\theta} + \mathbf{n}$, where the additive noise $\mathbf{n}$ has zero-mean and covariance $\boldsymbol{C}$ but does not have to be Gaussian
- Furthermore, the statistical family $\mathcal{P}$ induced by $p(\boldsymbol{y}; \boldsymbol{\theta})$ does not have to be an RSF
- Note that $E(\mathbf{y}; \boldsymbol{\theta}) = \boldsymbol{A\theta}$ and $\mathrm{Cov}_{\boldsymbol{\theta}}(\mathbf{y}) = \boldsymbol{C}$
- **Linear Unbiased Estimator (LUE)**: We aim to develop an UUBE $\hat{\boldsymbol{\theta}}$ that is a linear function of the data (i.e., $\hat{\boldsymbol{\theta}} = \boldsymbol{K}\mathbf{y}$)

Best Linear Unbiased Estimator (BLUE)

- For $\hat{\boldsymbol{\theta}}$ unbiased and linear

$$\Rightarrow E(\hat{\boldsymbol{\theta}}) = E(\boldsymbol{K}\mathbf{y}) = E(\boldsymbol{K}(\boldsymbol{A}\boldsymbol{\theta} + \mathbf{n})) = \boldsymbol{K}\boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{\theta}$$

$$\Rightarrow \boldsymbol{K}\boldsymbol{A} = \boldsymbol{I}$$

$$\Rightarrow \tilde{\boldsymbol{\theta}} = \boldsymbol{K}\mathbf{n}$$

- Thus, the MSE can be obtained as

$$\mathrm{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = \mathrm{Cov}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = \mathrm{Cov}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}) = E(\boldsymbol{K}\mathbf{n}\mathbf{n}^{\mathsf{T}}\boldsymbol{K}^{\mathsf{T}}) = \boldsymbol{K}\boldsymbol{C}\boldsymbol{K}^{\mathsf{T}}$$

- **Gauss-Markov Theorem (GMT):** The **Best Linear Unbiased Estimator (BLUE)** is given by

$$\hat{\boldsymbol{\theta}}_o = \boldsymbol{K}_o\mathbf{y} \qquad \boldsymbol{K}_o = (\boldsymbol{A}^T\boldsymbol{C}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{C}^{-1} \tag{1}$$

Proof of the Gauss-Markov Theorem

Proof:

- Based on Eq. (1) the $\mathrm{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_o)$ can be obtained as

$$\begin{aligned}\mathrm{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_o) &= \boldsymbol{K}_o \boldsymbol{C} \boldsymbol{K}_o^{\mathsf{T}} \\ &= (\boldsymbol{A}^T \boldsymbol{C}^{-1} \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{C}^{-1} \boldsymbol{C} \boldsymbol{K}_o^{\mathsf{T}} \\ &= (\boldsymbol{A}^{\mathsf{T}} \boldsymbol{C}^{-1} \boldsymbol{A})^{-1}\end{aligned}$$

where we used $\boldsymbol{K}_o \boldsymbol{A} = \boldsymbol{I}$

- Furthermore, note that

$$\begin{aligned}E(\tilde{\boldsymbol{\theta}}_o \tilde{\boldsymbol{\theta}}^{\mathsf{T}}) &= E(\boldsymbol{K}_o \mathbf{n}\mathbf{n}^{\mathsf{T}} \boldsymbol{K}^{\mathsf{T}}) \\ &= \boldsymbol{K}_o \boldsymbol{C} \boldsymbol{K}^{\mathsf{T}} \\ &= (\boldsymbol{A}^{\mathsf{T}} \boldsymbol{C}^{-1} \boldsymbol{A})^{-1} \boldsymbol{A}^{\mathsf{T}} \boldsymbol{C}^{-1} \boldsymbol{C} \boldsymbol{K}^{\mathsf{T}} \\ &= (\boldsymbol{A}^{\mathsf{T}} \boldsymbol{C}^{-1} \boldsymbol{A})^{-1}\end{aligned}$$

where we used $\boldsymbol{K}\boldsymbol{A} = \boldsymbol{I}$

Proof of the Gauss-Markov Theorem

- We can now develop the following positive-semidefinite expression

$$
\begin{aligned}
E\left((\tilde{\boldsymbol{\theta}}_o - \tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_o - \tilde{\boldsymbol{\theta}})^{\mathsf{T}}\right) &= E(\tilde{\boldsymbol{\theta}}_o\tilde{\boldsymbol{\theta}}_o^{\mathsf{T}} - \tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}_o^{\mathsf{T}} - \tilde{\boldsymbol{\theta}}_o\tilde{\boldsymbol{\theta}}^{\mathsf{T}} + \tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^{\mathsf{T}}) \\
&= \underbrace{E(\tilde{\boldsymbol{\theta}}_o\tilde{\boldsymbol{\theta}}_o^{\mathsf{T}})}_{\mathrm{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_o)} - E(\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}_o^{\mathsf{T}}) - E(\tilde{\boldsymbol{\theta}}_o\tilde{\boldsymbol{\theta}}^{\mathsf{T}}) + \underbrace{E(\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}^{\mathsf{T}})}_{\mathrm{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})} \\
&= (\boldsymbol{A}^{\mathsf{T}}\boldsymbol{C}^{-1}\boldsymbol{A})^{-1} - 2(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{C}^{-1}\boldsymbol{A})^{-1} + \boldsymbol{K}\boldsymbol{C}\boldsymbol{K}^{\mathsf{T}} \\
&= -(\boldsymbol{A}^{\mathsf{T}}\boldsymbol{C}^{-1}\boldsymbol{A})^{-1} + \boldsymbol{K}\boldsymbol{C}\boldsymbol{K}^{\mathsf{T}} \\
&= -\mathrm{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_o) + \mathrm{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) \\
&\succcurlyeq \mathbf{0}
\end{aligned}
$$

- Finally, we get $\mathrm{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_o) \preccurlyeq \mathrm{MSE}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}})$ with equality iff $\hat{\boldsymbol{\theta}}_o = \hat{\boldsymbol{\theta}}$.

Asymptotic Properties of ML Estimation

- Recall that $\mathbf{y} \in \mathbb{R}^m$; for a finite number of data records m, in general, the ML estimator is not the UMVUE

- However, for $p(\mathbf{y}, \theta)$ being from a RSF and $m \to \infty$, it can be shown that the ML estimate is *asymptotically unbiased, asymptotically Gaussian, and asymptotically efficient/UMVUE*, i.e.,

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{J}_{\boldsymbol{\theta}}^{-1}) \quad \text{for } m \to \infty$$

  where $\boldsymbol{J}_{\boldsymbol{\theta}}$ is the Fisher information matrix

- Since (apart from pathological situations) $\boldsymbol{J}_{\boldsymbol{\theta}}^{-1} \to 0$ for $m \to \infty$, it follows that $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\boldsymbol{y})$ is consistent, i.e.,

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \to \boldsymbol{\theta} \quad \text{for } m \to \infty$$

- Proof: See Kay, Appendix 7B

Parameter Transformation

- Let $\boldsymbol{\alpha} = \boldsymbol{g}(\boldsymbol{\theta})$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ may have different dimensions
- We recall that $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\boldsymbol{y}) \triangleq \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{y}; \boldsymbol{\theta})$
- The ML estimate of $\boldsymbol{\alpha}$ can be defined as

$$\hat{\boldsymbol{\alpha}}_{\mathrm{ML}}(\boldsymbol{y}) \triangleq \arg\max_{\boldsymbol{\alpha}} \tilde{p}(\boldsymbol{y}; \boldsymbol{\alpha})$$

  where $\tilde{p}(\boldsymbol{y}; \boldsymbol{\alpha})$ is given as follows

  1. if $\boldsymbol{g}(\boldsymbol{\theta})$ is invertible, then $\tilde{p}(\boldsymbol{y}; \boldsymbol{\alpha}) = p\big(\boldsymbol{y}; \boldsymbol{g}^{-1}(\boldsymbol{\alpha})\big)$
  2. if $\boldsymbol{g}(\boldsymbol{\theta})$ is not invertible, i.e., to a given $\boldsymbol{\alpha}$ belonging to the range of $\boldsymbol{g}(\cdot)$, there exist several $\boldsymbol{\theta}_i$ such that $\boldsymbol{\alpha} = \boldsymbol{g}(\boldsymbol{\theta}_i)$, then $\tilde{p}(\boldsymbol{y}; \boldsymbol{\alpha}) = \max_{i:\boldsymbol{g}(\boldsymbol{\theta}_i)=\boldsymbol{\alpha}} p\big(\boldsymbol{y}; \boldsymbol{\theta}_i\big)$

- In either case, we have $\hat{\boldsymbol{\alpha}}_{\mathrm{ML}}(\boldsymbol{y}) = \boldsymbol{g}\big(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\boldsymbol{y})\big)$